****
****

Proceedings of

National Journal of Parallel and Soft Computing

Volume 03, Issue 01

Organized by

University of Computer Studies, Yangon

Ministry of Science and Technology, Myanmar

December, 2022

EDITORIAL BOARD

REVIEWER BOARD

National Journal of Parallel and Soft Computing Volume 03, Issue 01

December,2022

CONTENTS

****

Big Data and Cloud Computing

Embedded System

Data Mining, Web Mining and Machine Learning

Big Data and Cloud Computing

Multi-Objective Task Scheduling using K-means Algorithm in Cloud Computing

Chue Theingi Kyaw, Yu Mon Zaw
University of Computer Studies, Yangon
chuetheingikyaw@ucsy.edu.mm, yumonzaw@ucsy.edu.mm

Abstract

Nowadays, cloud computing has become a popular technology among users, and the volume of data is increasing day by day. Cloud computing provides virtualized resources to customers on a demand and pay-per-use basis via the internet. The scheduling of tasks is a very important aspect of cloud computing in order to reduce task completion time. This paper proposes a multi-objective task scheduling algorithm using the K-means clustering technique for better outcomes. Here, the tasks are clustered based on task length (cloudlet) and deadline. Multi-objective task scheduling using the K-means algorithm is compared with the existing First Come First Serve (FCFS) algorithm on CloudSim. Results reveal that the proposed scheduling algorithm has excelled over the existing traditional algorithm (FCFS), and comparative analysis shows better results in terms of a reduction in makespan and execution time. In this research, the CloudSim simulator version 3.0.3 is used to test this experiment.

KEYWORDS: Task Scheduling, Multi-objective, K-means, Execution time, Makespan

1. Introduction

Today, Cloud computing led to the popular technology and most of the IT environment start using cloud computing. Cloud computing has brought utility-oriented IT services as a new trend in the enterprise business for users, worldwide because of its economic advantage. Cloud computing offers several of services and resources delivered to the user over the internet. It shared resources, software, and information, are delivered to computers and other devices on demand because of that is Internet-based computing. Many existing cloud service platforms, such as Amazon EC2, Google App Engine and Microsoft Azure, have proven their success and opened to the public in a pay-as-you-go manner. Different cloud have different requirements and business objectives, and thus provide on demand different kinds of services to the client, such as Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS). And then, cloud environments can be simply categorized into private cloud, public clouds, hybrid cloud and community cloud. In Fig.1 show the different service of cloud computing and their usage.

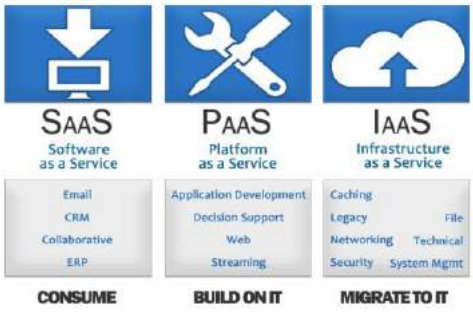


Figure 1. Cloud Computing Services

Development of cloud computing slower down the efficiency, performance and utilization of resources for which cloud computing need to be evolved. Optimal management of the available resources is proving to be the big challenge in cloud computing and task scheduling plays a pivotal role in improving the utilization of computational devices. Basically, scheduling is a process of assigning incoming tasks to available resources in cloud datacenters, where a resource exists in terms of a virtual machine (VM). As cloud computing is increasing day by day, the optimal resource utilization is becoming more and more convoluted. Fig.2 defines the flow of task scheduling.

The main objective of task scheduling is to maximize resource utilization, to minimize the makespan and response time, and to have a balanced load on all the machines. A good scheduling algorithm yields good system performance. An effective task scheduling

method requires not only meeting the user's needs but also improve the performance of the whole system. This system proposed multi-objective task scheduling algorithm using two criteria "Tasklength and Deadline". This algorithm integrated with K-means clustering algorithm for assigning tasks to VMs.

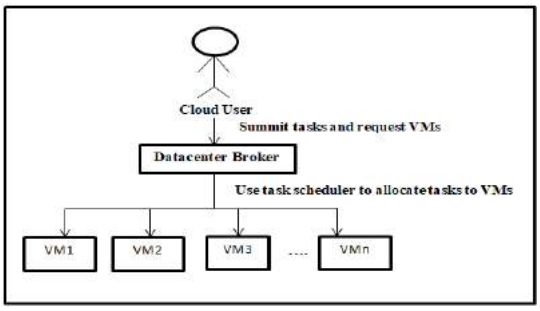


Figure.2 Task Scheduling flow

The rest of the paper is arranged as follows: In Section 2 describes related works of task scheduling. Section 3 depicts the implementation details of the proposed system. Experimental evaluation is illustrated in Section 4. Finally, in Section 5, discusses the conclusion and future work of the system.

2. Related Work

Cloud Computing is the trend advancements in the developing stage. In task scheduling main problem is to map the task received by broker to particular VM in the way to minimize the execution time and makespan [6].

Multi-objective task scheduling algorithms are defined by many researchers. Grouped Task Scheduling (GTS) algorithm used in cloud computing uses both TS algorithm and Min_Min algorithm to achieve both minimum latency and execution time of the tasks [1]. GTS algorithm has five categories (Urgent User& Task, Urgent Task, Urgent User, Long Task and Normal Task). Other used non-dominated sorting [2] to improve the throughput of the datacenter and reduce the cost. In this paper, task's priority is assigned according to the QoS of tasks and the aim of this paper is to minimize the execution time of a task. In [3] author proposed dynamic cloud task scheduling (DCTS) algorithm proposed in classifies the tasks and creates the necessary VMs in advance, based on the scheduling history. By this way, it reduces the time to create the VMs and hence reduces the completion time of the tasks. This algorithm makes use of the Bayes classifier to classify the tasks. In [4] author proposed to schedule a task on two parameters that are task length and its deadline. The comparative analysis shows a reduction in makespan and average waiting time. The author [5] developed a selective algorithm that uses the Max-Min and Min-Min algorithms. It determines to select one of these two algorithms, depending on the standard deviation of the expected completion times of the tasks on each of the resources. A comparison of this algorithm with the FCFS algorithm shows that this algorithm is more efficient in minimizing the makespan.

In cloud computing, although many contributions have been made by the research community towards providing a solution for the scheduling problem, still there is a need for new solutions to reduce the makespan further and also to improve utilization rate. Hence, a new scheduling technique namely "multi-objective task scheduling using K-means algorithm in cloud computing" is proposed to minimize the makespan, execution time in workload among the VMs.

3. System Implementation

There are several classes in CloudSim that provide the simulation environment for cloud computing. It is necessary to understand existing allocation policies and the classes that enable these allocation methods in order to implement task scheduling policies. Since clustering is the new concept in the CloudSim, additional classes have been added to compute our working policies.

3.1. Mathematical Model

A mathematical model for the research problem is proposed with the following assumptions:

- There are n independent heterogeneous tasks (cloudlets) to be scheduled and executed in m heterogeneous virtual machines (VMs).
- Each task has length L in terms of number of instructions (Million Instructions - MI).
- Each task has Deadline DL (Second-s).

- Each machine has a processing capacity in terms of million instructions per second (MIPS).

In proposed system K-mean clustering algorithm is used to create the clusters for tasks. In which for k clusters centroids are calculated based on multi-objectives Tasklengh and Deadline using equation (1), (2) and then Centroid is calculated using Euclidean Distance.

L_i= Number of Instructions (MI) (1)

$DL_i = L_i/VM_{mips}$ (2)

where L_i= Cloudlet length

DL_i=Deadline

VM_{mips}= MIPS of Average VM

3.2. Multi-objective Task Scheduling using K-means Algorithm

Both Data Mining techniques and Cloud Computing helps the business organizations to achieve maximized profit and reduce costs in different possible ways. Mainly Clustering is the method which includes the grouping of similar type objects into one cluster and a cluster which includes the objects of data set is chosen in order to minimize some measure of dissimilarity. For task scheduling, K-Means clustering algorithm is used. Fig.2 depicts the system overall flow chat of multi-objective task scheduling using K-means algorithm:

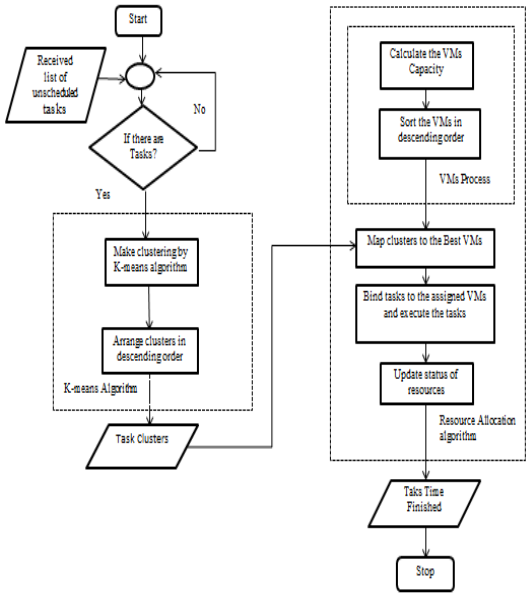


Figure.3 System overall flow

Algorithm1: Multi-objective Task Scheduling using K-means Algorithm

Input: List of unscheduled Tasks (Cloudlets) Set
Output: Task Clusters
Start
Step1: Select k points according to number of VMs
Step2: Form K clusters of cloudlets by assigning each point to its closest centroid.
Step3: Calculate the centroid using Euclidean distance of all the task clusters

dist ((x, y), (a, b)) = $\sqrt{(x - a)^2 + (y - b)^2}$

Where x= cloudlet length

y=deadline

Step4: For each task cluster, find the mean as

$$\mu_k = \frac{1}{N_k}\sum_{q=1}^{N_k} x_q$$

where N_k = the number of instances belonging to cluster k

μ_k = the mean of the cluster k.

Step5: Recomputed the centroid for each cluster.
Step6: Repeat Step2 to Step5 Until centroid do not change.
Step7: Arrange clusters in descending order
Step8: Map clusters to the Best VM using Resource Allocation algorithm.
End

Algorithm2: Resource Allocation Algorithm

Input : Task (Cloudlet) Clusters

: List of Virtual Machines (VMs)

Output: Allocated tasks to VMs
Start
Step1: Set i=0
Step2: **While** VM[i] is not empty **do**
Step3: Calculate the capacities of VM[i] using equation (1),

$$C_i = Pro_{ni} * P_{mips} + VM_{bwi} \text{ (1)}$$

where Pro_{ni} is the number of processors in VM_i

P_{mips} is millions of instructions per second of all processors in VM_i

VM_{bwi} is the communication bandwidth ability of VM_i

Step4: Set i=i+1
Step5: **End While**

Step6: Sort the VMs in **descending order** based on their maximum capacity
Step7: Set j =0
Step8: **While** cluster [j] is not empty **do**
Step9: Assign cluster[j] to VM and execute it
Step10: Update status of resources.
Step11: Set j=j+1
Step12: **End While**
End

3.3 Implementation of Proposed Algorithm in CloudSim

CloudSim is a java-based simulation tool, so it can be used either with the eclipse IDE or NetBeans IDE. CloudSim have different classes to support the simulation environment for the cloud computing. In order to apply new scheduling algorithm, it is essential to have knowledge about existing allocation policies and the classes that support these allocation strategies. Since clustering is the new concept in the CloudSim, so some new classes are also created in CloudSim. Eclipse IDE and Eclipse oxygen are used in the proposed system.

4. Experimental Results

The simulation environment is a 64-bit windows 10 operating system with core i3 and 8 GB RAM. Google cluster workload is used in conducting the experiments. Analyzing the Google cluster workload traces, a realistic Google-like workload is generated using Monte-Carlo simulation [9]. The task size ranges from 15,000MI to 900,000MI. In this system, both the tasks and the resources exhibit heterogeneous characteristics.

4.1. Performance Measure

The performance metric is a standard definition of a measurable quantity that indicates some aspect of performance. Most of scheduling problem needs to minimize the makespan and execution time of the tasks. Therefore, the proposed method was calculated using these metrics.

Makespan: Makespan is defined as the maximum time taken by a VM to complete the tasks in the task queue. It is denoted as the maximum of the completion time of all the tasks which is given by equation (3).

MCT=max $\{CT_{ij} | i \in 1, 2, 3, .., m, j \in 1, 2, 3... n\}$ **(3)**

where CT_{ij} ,represents the time taken by VM j to complete task i

Execution Time: It is the amount of time taken by a VM to run or execute a task. Here, average execution time defined in equations (4) and (5) is used to measure the performance of the datacenter.

$$ET_{ij} = \frac{T_{length(i,j)}}{VM_{comp(i,j)}} \quad (4)$$

$$\overline{ET} = \frac{\sum_{j=1}^{m}\sum_{i=1}^{n} ET_{ij}}{n} \quad (5)$$

where ET_{ij}, represents the execution time of task i on virtual machine j

$T_{length(i,j),}$ represents tasks instruction length

$VM_{comp(i,j),}$represents the calculation ability of the j virtual machine

$VM_{pesnum(i,j)}$, indicates the number of CPU virtual machine j

n represents the total number of tasks

4.2. Performance Evaluation

The proposed method is demonstrated using CloudSim 3.0.3, which is one of the open source frameworks that provide modeling, simulation, experimentation, and management of infrastructure services on cloud computing infrastructure. [10]. Table 1 presents the simulation parameters for the demonstration.

Table 1 Simulation Parameters

Entity	Quantity
Datacenter	1
Physical Machine (Host)	2-4
Processing capacity of each PE in the hosts	100000MIPS
Memory (RAM)	2048MB,3096MB
PEs in each VM	1-2
Processing capacity of each PE in the VMs	500-3500MIPS
Memory capacity of VMs	512-4196 MB
Task Length	15000-900000 MI
Number of tasks	500-1000
Number of VMs	5
Number of task clusters(k)	5

4.3. Experimental results for proposed system

In the proposed system, multi-objective task scheduling is performed with the aim of minimizing the makespan and execution time. The first comparison is performed between traditional algorithm (FCFS) and proposed algorithm based on calculated average execution time and the second comparison is performed between traditional algorithm (FCFS) and proposed algorithm based on calculated makespan. Traditional scheduling algorithm (FCFS) and proposed algorithm have been implemented on the same dataset. At first, average execution time and makespan are measured and the resultant values are given in Table 2 and Table 3 for 5 VMs. When compared to FCFS, the proposed algorithm reduces average execution time by 4%, 9%, 14%, and 24%, respectively.

Table 2 Average Execution Time Comparison for 5VMs with difference Task size

Execution Time (milli sec)		
Number of Tasks	FCFS	Proposed System
500	187.7291795	180.0403048
600	197.1855512	181.2537222
800	216.116459	189.12425
1000	248.973795	201

Table 3 Makespan Comparison for 5 VMs with difference Task size

Makespan (milli sec)		
Number of Tasks	FCFS	Proposed System
500	1187.5	1161.401099
600	1283.332	1185.43956
800	1500	1257.85714
1000	1649.9994	1353.571429

In this system, after task clustering, tasks with large clusters are scheduled to the VMs which are of high capacity and those with a smaller cluster to the VMs of less capacity which increases the resource utilization. And then, this reduces the average execution time which in turn reduces the makespan. When compared to FCFS, the proposed algorithm reduces the makespan by 2%, 8%, 19%, and 22%, respectively. The average execution time and makespan measured for these algorithms have been registered in Fig.4 and Fig.5.

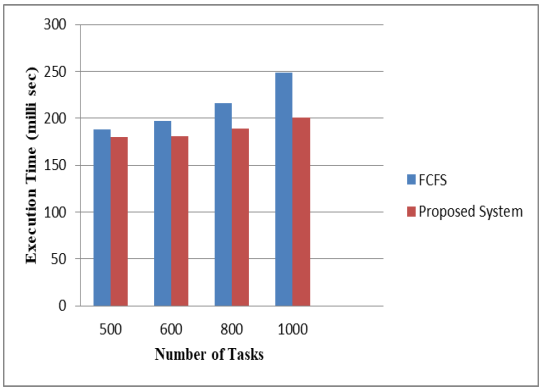


Figure.4 Average Execution Time

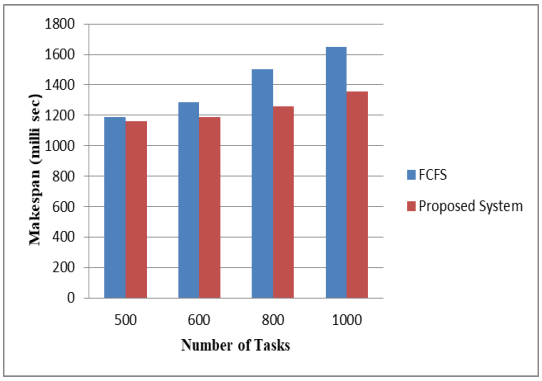


Figure.5 Makespan

The proposed algorithm provides an optimal scheduling method. Most of the algorithms schedule tasks based on single criteria (i.e. execution time). The proposed system uses two criteria to give minimum execution time and makespan. Above experiments conclude that multi-objective tasks scheduling algorithm is proposed which is integrated with K-means clustering algorithm for assigning tasks to VMs that produces better results in terms of execution time and makespan than the traditional scheduling algorithm (FCFS).

5. Conclusion and Future Works

The proposed scheduling algorithm generates task clusters based on attributes of the task by making use of K-Means clustering which is a simple and efficient clustering algorithm. Tasks in the clusters are scheduled to suitable VMs

based on the VM capacity. This system can give an optimal task scheduling algorithm which provides the minimum execution time and Makespan. As the future works, the system will test and compare with other clustering algorithms and techniques for cloud task scheduling purpose. This system can also add and modify more QoS parameters for task clustering.

ACKNOWLEDGMENT

To complete this paper, many things are needed like my hand work as well as the supporting of many people. I am extremely grateful to my supervisor **Dr. Yu Mon Zaw**, Lecturer, Faculty of Computer Science (FCS), the University of Computer Studies, Yangon for their valuable guidance, supervision, patience, encouragement and editing during the period of study towards completion of this piece of research work.

References

[1] H. G. E. D. H. Ali, I. A. Saroit, and A. M. Kotb, Grouped Tasks Scheduling Algorithm Based on QoS in Cloud Computing Network,Egyptian Informatics Journal, Vol. 18, No. 1, pp. 11-19, March, 2017.

[2] Atul Vikas Lakraa, Dharmendra Kumar Yadav, “Multi-Objective Tasks Scheduling Algorithm for Cloud Computing Throughput Optimization” Elsevier International Conference on Intelligent Computing, Communication & Convergence, Vol 48, pp 107-113, 2015.

[3] P. Y. Zhang and M. C. Zhou, Dynamic cloud task scheduling based on a two-stage strategy, IEEE Transactions on Automation Science and Engineering, Vol. 15, No. 2, pp. 772-783, April, 2018.

[4] K. Etminani, M. A. Naghibzadeh, A Min-Min Max-Min selective algorithm for grid task scheduling, 3rd IEEE/IFIP International.

[5] Arnav Wadhonkar, Deepti Theng, “A Task Scheduling Algorithm Based on Task Length and Deadline in Cloud Computing”, International Journal of Scientific & Engineering Research, Volume 7, Issue 4, April-2016.

[6] S. Mittal, A. Katal, An optimized task scheduling algorithm in cloud computing, IEEE 6th International Conference on Advanced Computing, Bhimavaram, India, 2016, pp.197-202.

[7] A. Hussain, M. Aleem, GoCJ: Google Cloud Jobs Dataset, Mendeley Data, v1, 2018.

[8] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, A. Y.Wu, An efficient k-means clustering algorithm: analysis and implementation, IEEE Transcations on Pattern Analysis and Machine.

[9] Tahani Aladwani Mecca, Saudi Arabia, Types of Task Scheduling Algorithms in Cloud Computing Environment, 2020. http://creativecommons.org/

[10] Fang, Y., Wang, F., & Ge, J. (2010). A Task Scheduling Algorithm Based on Load Balancing In Cloud Computing. In Web Information Systems and Mining (pp. 271-277). Springer Berlin Heidelberg.

Breast Cancer Predictive Analysis Using Big Data Environment

Yee Mon Ei, Thida Aung
University of Computer Studies, Yangon
yeemon.ei@ucsy.edu.mm, tdathida@ucsy.edu.mm

Abstract

Nowadays, big data is very popular in healthcare applications. Breast cancer is the most occurred cancer disease in the world that commonly occurs in a woman. If this disease is detected in early stages, there will be a better chance in treatment. In this system, a scalable and fault tolerant pipeline model is proposed for analyzing breast cancer data and predicting the breast cancer. This system is implemented on apache spark infrastructure applying random forest algorithm and the input data source is UCI breast cancer dataset. This system implementation using random forest algorithm is compared with naïve bayes in terms of accuracy, precision, recall and f-measure. The analysis of evaluation results describes the achievement of the proposed system with the accuracy of 98.2% in the big data analytics environment.

Keywords: Apache spark; Random Forest; Naïve Bayes

1. Introduction

Nowadays, the data rate increment based on the technology development occurs volume of data with many characteristics becomes big data. The big data exists by many formats of data. Although data storage occurs no expensive, this increasing data in many resources of data by many formats of data happens new problems relating to data processing [4].

Breast cancer is a cancer disease which occurs from the breast tissue and is the most common cancer type distinguished in western countries. The survival rates improvement and the deaths decrement regarding with this disease can be achieved by early breast cancer detection. Two breast cancer cells are: (a) cancer cells, malignant and (b) non-cancer cells, Benign. The malignant specification by breast cells are involved in the breast cancer prediction. Various methods are utilized by pathologists for breast cancer forecasting. Various machine learning and statistical methods are used by researchers for breast cancer forecasting.

Big data analytics has become more interested in the huge amount of raw data analysis and management. In order to perform the parallelization and operation distribution of various tasks in various clusters, the Map Reduce framework was initially developed by google [1]. This framework has been applied for the implementation of the parallelization and operation distribution system with the operation of batch data in various data nodes. For this implementation, apache hadoop has been utilized as the best framework for implementation [11]. On the other hand, apache spark provides the achievement with more attractiveness and high performance with the expansion of hadoop's capabilities and permitting the processing on the real time stream data [13].

In this paper, the breast cancer prediction model is developed with apache spark based random forest. This system is implemented with the infrastructure of big data analytics (Apache Spark) for the analyzing high amount of velocity and large amount of breast cancer data. UCI breast cancer data set is used as input data source for forecasting. This infrastructure has the ability of fault tolerance and scalability memory operating engine for huge amount of data. It can perform the implementation of machine learning approaches efficiently.

There are four main modules in this proposed system: collection of data; preprocessing of data, labeling for data, and classification module. The development of these modules is implemented on the four layers: layer of data ingestion, layer for storage, layer for processing and layer for data analytics. In the data ingestion layer, data collection of breast cancer dataset is performed and then this collection of ingested data is stored

at layer of storage, hadoop distributed file system (HDFS). Apache spark is used as a processing layer for big data analysis and breast cancer data prediction analysis in a distributed computing environment with reliability and fault-tolerance. To construct the prediction models, data cleaning and preprocessing, and model generation are performed at off-line training in layer of data analytics.

In this paper, the section 2 describes about the related works of the proposed system and then section 3 presents about the background theory of the proposed system. In section 4, the proposed system is presented and the experimental results of the proposed system are discussed in section 5. Finally, the presentation of the system is concluded in section 6.

2. Related Works

The authors presented the breast cancer prediction model with grid computing and support vector machine [2]. They initially predicted breast cancer by support vector machine without grid computing. Then, they predicted breast cancer with support vector machine model by grid computing. Finally, they compared the analysis result and then the construction of new model is performed. The developed new model was performed with grid computing on data before fitting it for prediction, that achieves the prediction results improvement.

The authors analyzed the hybrid approach by combining the fuzzy and traditional decision tree in the breast cancer classification which can be utilized by experts in the SEER dataset [3]. Three performance evaluation parameters were used to perform the comparison of two algorithms: sensitivity, accuracy, and specificity. The performance evaluation showed that the performance of fuzzy decision tree is significantly higher than the traditional J8 decision tree.

The hybrid approach for breast cancer prediction was proposed [7]. This paper contains two parts. They initially performed the dimensionality reduction by fuzzy-artificial immune system (FAIS). Then, the breast cancer classification was done by using k-nearest neighbor according to the selected features. The experimental results showed that the presented system achieves the prediction accuracy increment and the execution time decrement.

The authors demonstrated the prediction approach for the diseases using the medical data features [5]. Later, this paper applied rough approach by back propagation neural network. Then, the prediction was done utilizing UCI breast cancer, statlog heart disease, and hepatitis datasets. The system evaluation presented that breast cancer, heart disease, and hepatitis achieved the accuracy of 90.4 %, 98.6%, and 97.3 %. The authors proposed the unsupervised and supervised approaches with weka on BCDR-F01 malignancy data prediction with Naive Bayes algorithm [8]. This paper computes the misclassification cost by Naive Bays and compared with the misclassification cost by applying other algorithms.

In this paper, the authors proposed a prediction algorithm for the breast cancer recurrence in SEER (surveillance, epidemiology, and end results) dataset of program of the national cancer institute (NCI) with Map Reduce framework [9]. Preprocessing was initially done in this paper. After data cleansing and preparation, the final dataset building was done. Finally, the analysis of this dataset was performed for breast cancer recurrences occurred in the initial 5 years after breast cancer treatment.

3. Background Theory

Big data analytics combines the traditional analytics with mining methods along for huge amount of data to fundamental infrastructure creation for analysis, model construction and forecasting the markets, behaviour, products, and services by establishing the needs of the enterprise for the market and customer portion. Three types of big data analytics are:

- **Descriptive Analytics**, apply data mining and aggregation approaches to support understanding the last period and result: “What has happened?”
- **Prescriptive Analytics**, apply simulation and scalarization methods to give the advice for possible results and outcomes: “What should we do to occur in future?”
- **Predictive Analytics**, apply statistical and forecasting methods the future and result: “What could happen in future?”

Predictive big data analytics contains approaches for forecasting future results according to history and ongoing data [6]. Statistics is good analytical approach for big data analysis as big data is the consideration of data in statistics.

3.1. Random Forest

Random forest is a supervised ensemble learning algorithm. It makes decision trees aggregation randomly. The principal of this is a binary tree that's built by recursive partitioning (RPART). This is generally developed that the iterative binary dividing of the tree to same nodes. The dividing of data from the parent tree into its son nodes is done with a good binary dividing such that the unity enhancement in the son nodes is done by the parent. This is constructed with many trees, in where the construction of every tree is done with a bootstrap instance. It distinct with CART as non-deterministic construction with two steps of randomization function. It considers only a subset of features. Random forest possesses same parameters like decision tree. Advantages are that it performs very well on large scale dataset. It can handle classification and regression. It becomes a better model with the model randomness addition.

The random forest is built with the followings:

- Let J is the number of variables and I is the testing events in the classification.
- Let j is the number of input variables for the decision at a given node; j<J.
- Select a training set and apply the rest of the testing events.
- At each node of the tree, the random selection of j variables is done on the decision. Calculate the best partition of the training set from the j variables.

For forecasting, a new instance is pushed down the tree. After, it is assigned the label of the terminal node where it ends. This iteration of process is performed by all tree's aggregation, and the label that gets the most incidents is provided as the forecasting.

3.2. Apache Spark

This is an open-sourced cluster infrastructure constructed above hadoop distributed file system (HDFS). Data processing and data assignment to worker nodes are performed for processing. Worker nodes are controlled by the master node with scheduling and dispatching of distributed tasks. Therefore, this infrastructure requires a cluster manager and distributed storage system. It possesses API that is familiar with developer and faster in-memory data engine.

Spark Core is the fundamental element and operation engine in spark. Spark core supports scheduling, I/O functionalities and remitting by distributed jobs. These operations are taken by an application programming interface (API) in Scala, Python, R, and Java. This API can support various languages with interfaces. It supports faster speed for in-memory calculating abilities. The high-level programming approach is provided by a driver program for the parallel operations calling on a resilient distributed dataset (RDD) in function passing in spark. The scheduling of parallel operation in clusters is done with this core. Resilient distributed dataset (RDD) was the fundamental API until spark version 2.1.x. RDD has the ability of read-only and fault-tolerance dataset collection assigned on the cluster for processing.

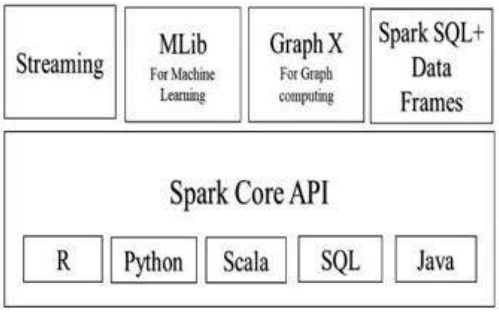


Figure 1. Apache Spark Ecosystem [12]

Spark streaming, graphX, spark sql, and machine learning library are the four elements of spark core. Spark sql has the association with structured data by utilizing data frames. This sql possess an interface for query data processing. The addition of real-time data processing to batch processing in spark is performed by spark streaming.

Various machine learning approaches: classification, collaborative filtering, clustering, dimensionality reduction, and regression are implemented by a distributed manner. Feature selection, extraction, and transformation are taken by these approaches on structured dataset. Moreover, they support tools for constructing and tuning machine learning pipelines and

implementing with loading and keeping models, algorithms, and pipelines [10].

4. Proposed System

The main target of this system is to develop high performance and scalable breast cancer prediction system on big data analytics platform, Apache Spark. This model is developed on apache spark infrastructure using machine learning and the used data source is UCI breast cancer dataset that distinguishes breast tumor into malignant or benign according to 10 features for prediction [14]. Random Forest is applied for breast cancer prediction. This cancer prediction system will provide for medical professionals for better decisions according to the clinical data of patients with desired accuracy. Figure 2 shows the proposed system flow diagram.

This system develops an apache spark-based model for handling huge amount of data in order to be accurate and fast. UCI breast cancer dataset is used as input dataset for prediction. Spark has the ability of fault tolerance and scalability memory operation engine for big data. It can perform the implementation of machine learning approaches by efficient manner. The substituting of missing values from the input dataset is performed for data preprocessing. After data preprocessing, the data storage is done HDFS with the splitting into training and testing set. For Spark-based Random Forest training, the training set is loaded into the Hadoop File System (HDFS). Then, the training is performed on hadoop with the loaded training set. After model training completion, the generated trained model is saved in HDFS and then the copy is sent to local file system for performance evaluation.

After the training completion, performance evaluation is performed. For performance implementation, the testing set from random split of the dataset is done. After the testing data loading, the target class of testing data is predicted using random forest model. The testing of the proposed system calculates the output values by using the trained model values. Accuracy, precision, recall, and f-measure are the key metrics of performance evaluation of the proposed system. The detail of performance evaluation is described in the next section.

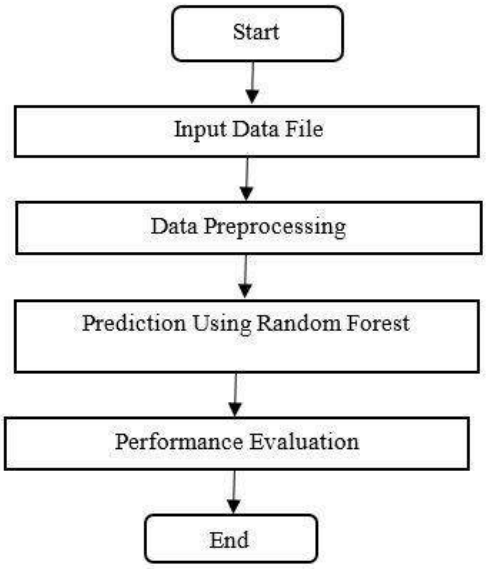


Figure 2. System Flow Diagram

Each machine learning pipeline includes:

- Data Frame: The principal element for solving data frames. This is applied in the maintenance of feature vectors, predictions, labels, and data.
- Transformer: The machine learning algorithm for the data frame features conversion to predictions.
- Estimator: The machine learning algorithm for the data frames training to the model generation.
- Pipeline: The task for a workflow creation with binding many transformers and estimators.
- Parameters: The specification set for the sharing with estimators and transformers.

The pipeline process contains:

- Ingestion of data: Loading of input data into spark and conversion to data frames is done.
- Preparation of data: Preparation of the suitable input data with the missing values elimination is performed. This output data is utilized in model construction and training.
- Training of model and construction: Features addition to data frames and conversion into prediction is done with the model training by data frames.
- Prediction: Implementation of the model prediction with training and testing data.

For the system implementation, breast cancer diagnostic data at UCI machine learning repository is used as the input data source. The input data source contains tumor features from digital image of breast fine needle aspirates (FNA). The 10 attributes for each patient contain: texture, radius, area, perimeter, compactness, smoothness, concavity, symmetry, fractal portion, and concave points. These attributes are the specifications of the cell nuclei at the fine needle aspirate of a breast mass. There are two types of breast cancer: Malignant and Benign.

The characterization of the dataset is follows: (a) cell shape/size equality: the size and shape of cancer cell are not same. In order to define whether the cell is cancerous or not, this attribute is utilized. (b) size of epithelial cell: The enlargement of epithelial cells is performed whether the impact factor of cancer. (c) bland chromatin: This belongs to the sample cancer cell surface. Although normal cell surface is smooth, the cancer cell surface is not smooth. (d) thickness of clump: accumulation of cancer cells is at layers in that normal cells grouping at monophonic layers are performed. (e) marginal adhesion: the penetration is done by normal cells although the penetration is not performed by cancer cells. (f) bare nuclei: This factor is only available for normal cells. (g) normal nucleoli: The small visible nucleus structures is known as Nucleoli. Though the Nucleoli is significant in cancer cells, this is invisible and small at normal cell.

In Data preprocessing, the conversion of unstructured data into structured data is performed. Data preprocessing includes

(i) dataset loading
(ii) checking of missing values existence
(iii) dividing data to independent and dependent data
(iv) label encoding
(v) one hot encoding
(vi) splitting data to training and testing.

After data preprocessing, the breast cancer forecasting is performed with our proposed model apache spark-based random forest. The prediction steps are

- Loading preprocessed dataset by RDD format
- Transforming RDD into data frame
- Reading labels and features from data frame
- Encoding the non-numeric features
- Indexing string with each encoded feature
- Aggregation of vector by numeric and one-hot-encoded features
- Converting the aggregated vector to a pipeline
- Converting the pipeline to a readable spark form
- Training the model by random forest-based features with the training data
- Testing on the whole data to achieve forecasting prediction label value

The step of random forest is as follows:

i. Reading the breast cancer data that has “t” features.
ii. Selecting features subsets and naming randomly by “r” from “t” features
iii. Calculating the node “t” on “r” features regarding to finest fit
iv. Splitting the node to child nodes according to the best split
v. Iterating from (ii) to (iv) until taken “j” nodes
vi. Iterating from (ii) to (v) by “m” times to get “m” number of trees for forest construction
vii. Testing features and generated trees are utilized for forecasting and then the result is kept as target
viii. Computing the voting value for every forecasted target
ix. Considering the high voting value as final output

5. Performance Evaluation

In this system, breast cancer dataset from UCI is used. This system is implemented with dataset size 600,000 breast cancer data. This system is tested with data size ratio: (Training – 80%, Testing – 20%). Firstly, randomly selected 80% records as training data and 20 % records for testing from this dataset. The popular performance measures (accuracy, recall, f-measure, and precision) are evaluated for this proposed system analysis. The experimental setup is illustrated in Table 1.

Table 1. System Specification

Operating System	Ubuntu 20.04 LTS
Host Specification	Intel ® Core i7-8550U CPU @ 3.7GHz, 8GB Memory, 1TB Hard Disk
VMs Specification	6 GB RAM, 200 GB Hard Disk
Software Component	- Hadoop 3.2.2 - Flume 1.9.0 - Spark 3.1.2 -Scala 2.12.3

The comparative results for the performance of proposed random forest-based classifier and naïve bayes classification are illustrated in Figure 3.

Machine Learning Algorithm	Benign			Malignant		
	Precision	Recall	F-measure	Precision	Recall	F-measure
Random Forest	0.99	0.98	0.98	0.96	0.99	0.97
Naïve Bayes	0.91	0.94	0.92	0.89	0.84	0.86

Figure 3. Performance comparisons of Random Forest and Naïve Bayes

The comparison of accuracy between Random Forest and Naïve Bayes on breast cancer dataset is illustrated in Figure 4.

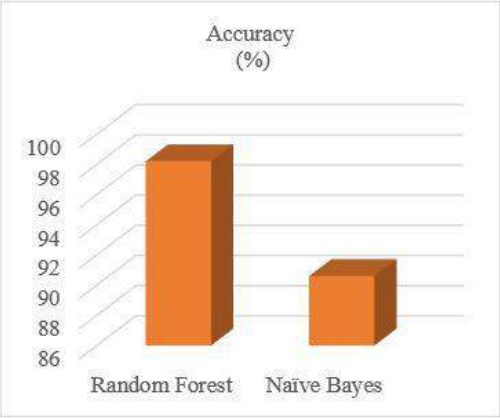


Figure 4. Accuracy comparisons of Random Forest and Naïve Bayes

According to the evaluation results, Random Forest classifier with proposed breast cancer classification model achieves the best optimal accuracy than Naïve Bayes classifier.

6. Conclusion

In this system, this proposed prediction model is adaptable to classical machine learning algorithms with data generally and forecasting methods is possible for achieving a result to the phenomenon of big data. In order to improve the survivability rate among the patients of breast cancer, the system is implemented for breast cancer prediction method. In this system, spark based random forest approach for breast cancer prediction. The proposed system approach is evaluated and compared using wisconsin breast cancer dataset. The advantage of the proposed system is that the computational complexity and response time are reduced than traditional machine learning approaches as the implementation on big data analytics platform, Apache Spark. The experimental outcomes also demonstration the superiority of the random forest classifier in terms of accuracy, recall, precision, and f-measure.

References

[1] J. Dean, and S. Ghemawat, "Mapreduce: Simplified Data Processing on Large Clusters," *in Proceedings of the 6th Conference on Symposium on Operating Systems Design & Implementation*, San Francisco, CA, USA, 6–8 December, 2004, pp. 137-150.

[2] V. Deshwal, and M. Sharma, "Breast Cancer Detection using SVM Classifier with Grid Search Technique", *International Journal of Computer Applications (0975 – 8887)*, Volume 178 – No. 31, Foundation of Computer Science, USA, July 2019, pp. 18-23.

[3] M. J. Domingo, B. D. Gerardo, and R. P. Medina, "Fuzzy Decision Tree for Breast Cancer Prediction*",* *In Proceedings of 2019 International Conference on Advanced Information Science and System (AISS'19)*, ACM, Singapore, November 15 – 17, 2019, pp. 1-6.

[4] M. Lněnička, R. Máchová, J. Komárková, and I. Čermáková, "Components of Big Data Analytics for Strategic Management of Enterprise Architecture", *2nd International Conference on Strategic Management*, VSB-Technical University of Ostrava, Ostrava, Czech Republic, 25-26 May, 2017, pp.398-406.

[5] K. Nahato, K. Harichandran, and K. Arputharaj, "Knowledge Mining from Clinical Datasets Using Rough Sets and Backpropagation Neural Network*",* *Computational and Mathematical Methods in Medicine*, Hindawi Limited, United Kingdom, 04 March, 2015, pp.1-13.

[6] E. Ricciarddelli, A. Cersosimo, D. Cimini, and F. D Paola, "Analysis of Heavy Rainfall Events Occurred in

Italy By Using Numerical Weather Prediction, Microwave and Infrared Technique", *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, IEEE, Valencia, Spain, 22-27 July, 2018, pp.931-934.

[7] S. Sahana, K. Polat, H. Kodaz, and S. Günes, "A new hybrid method based on fuzzy-artificial immune system and knn algorithm for breast cancer diagnosis", *Computers in Biology and Medicine*, Elsevier, USA, Volume 37, Issue 3, March 2007, pp. 415-423.

[8] T. A. Shaikh, and R. Ali, "A CAD Tool for Breast Cancer Prediction using Naive Bayes Classifier", *2020 International Conference on Emerging Smart Computing and Informatics (ESCI)*, IEEE, Pune, India, 12-14 March, 2020, pp. 351-356.

[9] H. Thottathy, K. K. Pavan, and R. P. Panchadula, "Microarray Breast Cancer Data Clustering Using Map Reduce Based K-Means Algorithm", *Revue d'Intelligence Artificielle (RIA)*, International Information and Engineering Technology Association (IIETA), Vol. 34, No. 6, 31 December, 2020, pp. 763-769.

[10] MLlib: Main guide Spark 2.3.0. Documentation https://spark.apache.org/docs/2.3.0/ml-guide.html", 2018.

[11] "Welcome to Apache Hadoop!", Available at "http://hadoop.apache.org/." [Online; accessed 28-December- 2017].

[12] What is Apache Spark? https://databricks.com/ spark/ about", 2018.

[13] "Spark-3.1.2", Available at https://spark.apache .org/ docs/3.1.2/ml-guide.html." [Online; accessed - May-2021].

[14]https://archive.ics.uci.edu/ml/datasets/breast+ cancer

Weather Prediction Analytics Using MapReduce-Based Logistic Regression

Su Hlaing Mon Than, Hmway Hmway Tar
Information Technology Supporting and Maintenance Department,
University of Computer Studies, Hinthada,
Faculty of Computer Science, University of Information Technology, Yangon
suhlaingmon.than@ucsy.edu.mm, hmwaytar34@gmail.com

Abstract

Nowadays, the prediction of weather has become a challenging task. A strong weather prediction system is very useful and important for our agricultural. Weather prediction is important for investigating of many businesses and decrease crop damage. Agricultural is the vital role for our country's business and most of people are depend on farming activities. Regression is one of the main methods used in rainfall data prediction. The proposed system will predict the weather condition of Hinthada Region, by Multinomial Logistic Regression and in order to achieve desire prediction accuracy. The proposed system will apply weather data prediction such as temperature, humidity and wind speed. The proposed system will applie Multinomial Logistic Regression and MapReduce platform. The proposed will store the various famous of weather data in Hadoop Distributed File System (HDFS) such that maximum and minimum of temperature, humidity and wind speed.

Keywords: Big Data, Multiple Linear Regression, Hadoop, MapReduce

1. Introduction

Weather prediction is one of the most important roles in agricultural countries like Myanmar. Weather prediction is the main challenging problem for our country. Weather prediction techniques are available for the prediction of weather forecasting. Traditional data analysis is suitable for only structured data and is not appropriate for large amount of data. Today, analysis of a large amount of data is difficult and conventional methods are not get the estimate accuracy. Therefore, we need the advanced method for better prediction of weather. Weather prediction becomes by collecting quantitative data about the maximum and minimum of temperature, humidity and wind speed and so on. Weather prediction is important analysis flood production and storm forecasting. Regression analysis is the measurement of relationship between outcome variable and one or more predictors. Regression more than two variables among which one is independent variable and all others are independent variables known as multinomial logistic regression. In the propose system we use Multinomial logistic Regression for the system's model, store the data Hadoop Distributed File System and process with MapReduce Algorithm. Multinomial Logistic Regression model plays the process of extract the previously unknown and useful information from large quantities of incomplete data for weather prediction. The paper proposes an approach Apache Hadoop for processing huge number of meteorological datasets. This paper improves the forecasting model of Hadoop framework system use proficient and unique able. It is mainly part for business, agriculture, farmers and related organizations to understand the natural disaster. The proposed system will help the agriculturist; they will know which crop to grow at different rainfall condition. The main purpose of this analysis is to know the prediction of the weather data. The system uses multinomial logistic regression method for data forecasting and map reduce programming model for data significant. The meteorological data extremely growled and valued which is big data. Big Data identifies specific kinds of data sets, such as unstructured data, which populate the data layer of scientific computing applications. Big data analytics involves the voluminous amount of data in which meaningful and useful values that are hidden to be extracted in systematics analysis. Big data

analytics is the process to check the large data sets containing a variety of data types. The meteorological data will use from Wikipedia Hinthada. The proposed system uses weather data more than 6 years of data online with 2 gigabytes collect. At the proposed system, uses meteorological data to get the prediction by using multinomial logistic regression and to reduce the datasets by using map reduce model. The goal of analytics is to improve the weather forecasting by gaining information. The accuracy indicates the proposed system improves the performance in terms of efficiency.

2. Related Works

The calculation of weather prediction is a challenging matter. Prediction modelling involves artificial neural network, data mining, K-means clustering, methodology and so on. MapReduce is becoming a useful programming model for solving scattered file systems and processing large datasets. MapReduce is a Java platform and an efficient distributed scheduling model.

Weather Prediction based Big Data using Hadoop Map Reduce Technique was [13] proposed the application of the weather report using previous studies with the concept of Big Data Hadoop. It analyzes each day's climate record and predicts the same day's climate using datasets. This model is not suitable for calculation desire accuracy.

A time series model for rainfall forecasting was [7] proposed the performance of accuracy by using data mining techniques. This system did not provide more accurate prediction and is not suitable for large amount of data.

A Storm analysis model from Rainfall dataset using Artificial Neural Network and Min Max Algorithms was [1] proposed methodology to develop a Storm analysis. This system did not provide more accurate prediction and was not suitable for large datasets. Moreover, Artificial Neural Network must require much processing time.

Meteorological Data Analysis Using MapReduce for the weather forecasting system was [13] used MK-Means Clustering Algorithm. MK-Means Clustering Algorithm similar the K-means algorithm. This algorithm did not allow development of an optimal set of clusters and for effective results, Moreover K-means clustering result in constant as it give varying results on different runs of an algorithm. A random choice of cluster patterns yields different clustering results resulting in inconsistency.

A non-linear rainfall-runoff model using radial basis function network was [6] presented the supervised learning algorithm. In this system, non-linear regression was not easy to analyze and neural system was complex.

Rainfall prediction using modified linear regression was [11] calculated the average value of result but did not estimate values. The weakness of this system was that did not provide an estimate of rainfall prediction.

Rainfall Prediction using Data Mining Techniques [7] was used data mining techniques and artificial neural network. In this system, data preprocessing steps were waste the time and artificial neural network did not easy to analyze by using conventional method.

3. Background Theory

Big Data, which is datasets, that are enormous, additionally high in velocity and variety. Big data analytics becomes with the solution since it turns out to be simple and nearly more reliable to store huge amount of data. The key word of Big Data is currently useful which is in our regular life and it is a recently condition and going to control the world in future. Big Data has many dimensions. Among them, volume, velocity, variety.

Volume: The amount of data known as the volume of data that contains the structured and unstructured.

Velocity: Capturing the growing data production rates. More and more data produced and must be collected in shorter times.

Variety: The multiplication of data such as structured, semi-structured, unstructured data with various formats which represents as text, audio, and hybrid data.

3.1. Hadoop

Hadoop is an open-source framework for processing a huge amount of data that supports the processing of large chunks of data with using high-level languages. Two main components of Hadoop are Hadoop Distributed File System (HDFS) and MapReduce. Hadoop clusters have

many parallel paradigms that store and process large amount of data sets. These paradigms applied easy to use languages and a set of consumer product machines in one location. Data storage and all processes occur these machines. Hadoop is able to improve the processing time for many clusters.

3.1.1 Hadoop Distributed File System

Hadoop Distributed File System (HDFS) is a distributed file system to manage for large datasets. The main properties of HDFS can store the large amount of data cluster which is a group of machines networked together in a single location. In HDFS large data separates into small task and then stores in multi locations called Data nodes. These data nodes are connected to a Name node. HDFS processes parallel processing and the storage of data is in informal fashion also avoiding error. The node works as the server and controls the data nodes using Hadoop paradigm. Using the HDFS that this core is possible to develop and execute distributed application. All other activities make it simple and increase the capabilities of the paradigm.

3.1.2 Map Reduce

Map Reduce is a programming paradigm that was develop to handle very large dataset and distribute the files across thousands of nodes. The Map Reduce software framework is a programming model, which now adopted by Apache Hadoop. Map Reduce is not only a simple programming model but also an efficient distributed scheduling model. MapReduce is a parallel programming model and a task scheduling model. Large amount of data is cut into unrelated blocks by Map program and task to lots of computers to process, receiving distributed computing. The proposed system will use map reduce algorithm to predict based on weather conditions on daily in agricultural big data environment to gather huge amount of data. MapReduce process the huge number of datasets in parallel using lots of computer running in a cluster. Map Reduce will split the large amount of dataset into small parts of a task and assign them to different system called nodes. Reduce function processes all the years of rainfall data to extract key and value pairs. MapReduce performs shuffle, sort and reduce. The Figure 1 shows the step of this system.

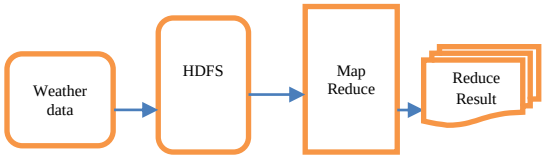


Figure 1. System Flow Diagram

4. Proposed Architecture

The data collected from Department of Meteorology and Hydrology, Hinthada for big data analytic. The data use for in our experiment's average maximum and minimum of temperature, average maximum and minimum of rainfall, humidity and wind speed. For the large amount of weather datasets, traditional methods do not give accurate results. Due to the voluminous of weather data sets, no simple process for determining of the proposed system. This data uses to formulate the equation for predicting the rainfall with Multinomial Logistic Regression. The function of Regression is to summarize the data as simple, useful and structural. The proposed system uses Multinomial Logistic Regression methods for more complicated data analysis and less for error percentage. The Multinomial Logistic Regression applied on the dataset and the coefficients will use to predict the weather data. To get the most accuracy result, we use the latest coefficients to forecast. To get the prediction of the weather system, we use Multinomial Logistic Regression model. Using Multinomial Logistic Regression approach, it can predict weather in anyone of the future year by using climate factors. Multinomial Logistic Regression will predict future weather information efficiency and computational time for the process. The data files store in Hadoop Distributed File System. Then, these files l split and go to different mappers. The output of each mapper processes a set of pairs (key, value). In the proposed system, key is station name, date and value are parameters: Temperature, Wind speed and Humidity. In the proposed system. For the mapper function, the output of mappers merge and sort by key. Finally, all results receive to the reducers. Each reducer calculates the accuracy result and store the final result in Hadoop Distributed File System. Figure 2 shows the using

weather dataset and Figure 3 shows the system's processing steps.

Figure 2. The Weather datasets

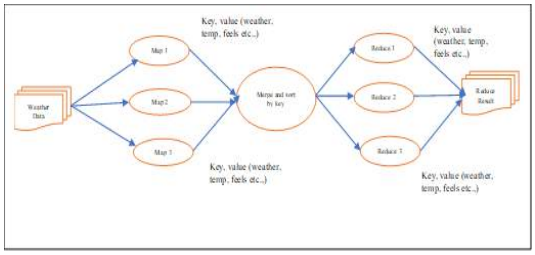


Figure 3. Proposed MapReduce Platform

5. Performance Evolution

The proposed system use dataset of Hinthada Wikipedia contains the following parameters Weather, Temperature, Feels, Gust, Rain, Humidity, Cloud and Pressure. The data files are stored in HDFS. Then, weather files split and goes to different mappers. The output of each mapper is a set of pairs (key, value), where key consists of Weather, Temperature, Feels, Gust, Rain, Humidity, Cloud, Pressure, vis is the contain parameters. Then the output of mapper is merge and sort by key. Finally, all results sent to the reducer. Reducer store the result in HDFS. Reducer calculate the annual weather condition. To get the final weather prediction, multinomial logistic regression uses on reduce result. The analysis is executed in Hadoop standalone mode.

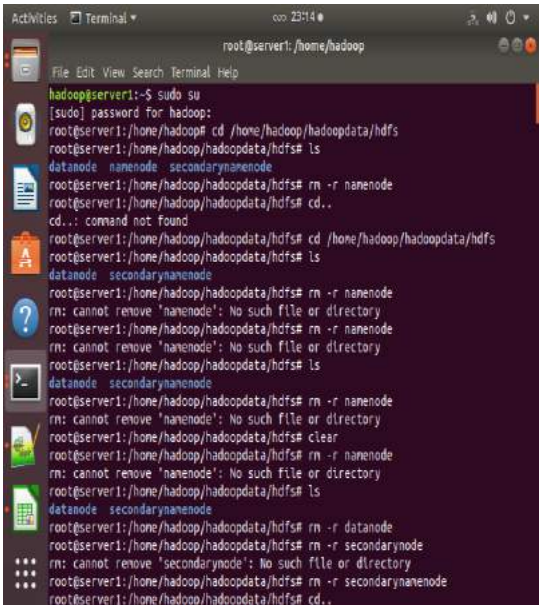


Figure 4. Hadoop cluster working

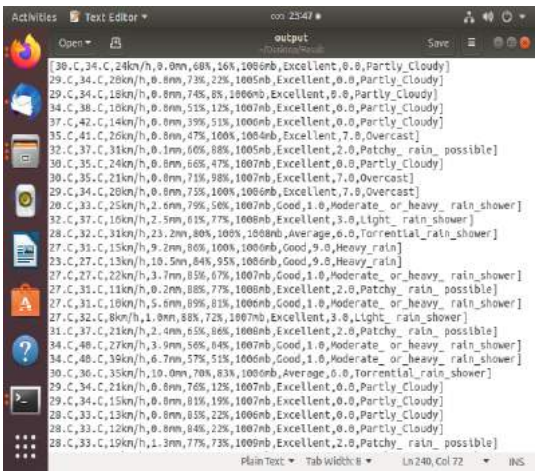


Figure 5. The final result of weather prediction

6. Conclusion and Future Work

Climate data is unique and clamorous in nature and has a huge volume of data. Precise weather prediction is essential for agriculture dependent countries like Myanmar. Using MapReduce can process a huge amount of data and can analyze effectively.

References

[1] Basvanth Reddy', Prof. B.A Patil "Weather Prediction Based on Big Data using Hadoop Map Reduce Technique" (2018) International Journal vol. 5, Issue 6, June 2016

[2] Anjana Joseph Joseph and M. Lakshmi "Storm Analysis with Raw Rainfall Dataset by using

Artificial Neural Network and Min-Max Algorithms", Indian Journal of Science and Technology Vol 9(10), DOI March (2016)

[3] Joko azhari Suyatno, Fhira Nhita and Aniq Atiqi Rohmaeati Rainfall forecasting in Bandaung Regency using C4.5 Algorithm", May (2018)

[4] E.Ricciarddelli, A.Cersosimo, D. Cimini, and F. D Paola, "Analysis Of Heavy Rainfall Events Occurred in Italy By Using Numerical Weather Prediction, Microwave and Infrared Technique"

[5] B. Anurag, M. Prakash, V. Kanna, and P. Choudhary, "Weather Forecasting using MapReduce", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 5, No. 9, pp.1-8, 2017.

[6] P. ChandrashakerReddy and A. Sureshbabu, "Survey on Weather Prediction using Big Data Analystics", In: Proc. of International Conf. On Electrical, Computer and Communication Technologies, pp.1-6, 2017.

[7] M. Senthilkumar, N. Manikandan, U. Senthilkuma and R. Samy, "Weather Data Analysis Using Hadoop", International Journal of Pharmacy and Technology, Vol.8, No.4, pp.21827-21834, 2016.

[8] V. Dagade, L. Mahesh, A. Supriya. and K. Priya, "Big Data Weather Analytics Using Hadoop", International Journal Technology in Computer Science & Electronics, Vol.14, No.2, pp.194-199,2015.

[9] M. Joshi, S. Shaikh, and P. Waghmode, "Farmer Buddy-Weather Prediction and Crop Suggestion using Artificial Neural Network on Map-Reduce Framework", International Journal of Computer Applications, Vol. 159, No. 7, pp. 1-3,2017.

[10] C.P. Shabariram, K.E. Kannammal, and T. Manojpraphakar, "Rainfall Analysis and Rainstorm Prediction using MapReduce Framework", In: Proc. of International Conference on Computer Communication and Informatics, pp.1-6, 2016.

[11] Q. Xiaoyun, K. Xiaoning, Z. Chao, J. Shuai and M. Xiuda, "Short-Term Prediction of Wind Power Based on Deep Long Short-Term Memory", In: Proc. of International Conference on Asia-Pacific Power and Energy, pp.1148-1152, 2016.

[12] R. Basvanth and B.A. Patil, "Weather Prediction Based on Big Data Using Hadoop Map Reduce Technique", International Journal of Advanced Research in Computer & Communication Engineering, Vol.5, No.6, pp.1-6, 2016.

[13] K.A. Ismail and M. Abdulmajid, "Big Data Prediction Framework for Weather Temperature Based on MapReduce Algorithm", In: Proc. of International Conference on Open Systems, pp. 1-6,2016.

[14] Doreswamy and G. Ibrahim, "Big Data Techniques: Hadoop And Map Reduce for Weather Forecasting", International Journal of Latest Trends in Engineering and Technology, Special Issue, pp.194-199, 2016.

[15] S. Selvaragini, and E. Venkatesan, "Big Data Techniques For Weather Forecasting", International Journal of Pure and Applied Mathematics, Vol.116, No.18, pp.195-201, 2017. [16] Gwo-Fong Lin, Lu-Hsein Chen, "A non-linear rainfall-runoff model using radial basis function.

[17] Wei- Fang, Xuezhi, Wen, Victor Sheng, Wubin Pan Meteorological Data Analysis Using MapReduce, The Scientific World Journal, February 2014.

[18] S.Prabakaran, P.Naveen Kumar and P.Sai Mani Tarun "Rainfall Prediction Using Modified Linear Regression".

Embedded System

Implementation of Voice-Controlled Wheelchair for Physically Disabled Persons Based on Pulse-Width Modulation (PWM)

Moe Moe Aye[1], Htet Thazin Tike Thein[2]
[1]*University of Computer Studies (Myitkyina)*
[2]*University of Computer Studies, Yangon*
moe804aye@gmail.com, htztthein@gmail.com

Abstract

The physically disabled persons cannot move from one place to another on their own and they continuously require someone to help them in getting the wheelchair moving. Voice commands can activate the wheelchair to make it easy for people who cannot control their movements by hands. In this paper, voice recognition module v 3.1 is applied on user's recognized voice commands which depend on the specified directions. Moreover, Arduino Uno Microcontroller Circuit and DC gear motors are utilized to control all the movements of wheelchair, and the motor driver (L298N) is applied to manage the speed and the direction of DC gear motors. In addition, the buzzer is utilized for giving alarm and Ultrasonic Sensors (HC-SR 04) is also utilized to detect the obstacles in between wheelchair and the way of direction. In this system, voice-controlled wheelchair is implemented based on Pulse-Width Modulation (PWM) technique which performs by changing the average voltage sent to the motor. The ultimate purpose of the system is to provide independent and easy access to physically disabled person who cannot control their movements for wheelchair especially by hands.

Keyword: *Voice Recognition Module; Ultrasonic Sensor; Arduino; DC Gear Motors; Ultrasonic Sensors; Pulse-Width Modulation (PWM)*

1. Introduction

Voice recognition is one of the approaches which commonly used to control the electrical and electronic utilizations because of easily being reproduced by human. It is a key technology which can provide human interaction with machines for controlling wheelchairs. In general, a wheelchair is the most widely used mechanical device by physically disabled patients and aged people to move. According to the statistical records for health and diseases, many people with physical disability usually depend on other people for moving from one place to another. Manually operated ordinary wheelchairs do not require any electrical system.

Moreover, the user needs an extra supporting person or self-assistance by hands to move on. However, people with arms and hands weaken find it difficult to use an ordinary wheelchair as their hands are not capable to operate to move it to any direction. A voice can typically activate the wheelchair to make it easy for physically disabled persons who cannot control their movements by hands. Some patients such as quadriplegic, cerebral palsy and multiple sclerosis usually depend on other people to move from one place to another because they don't have the freedom of mobility. Therefore, voice controlled based wheelchair is needed to develop to overcome these problems. Typically, the wheelchair can be operated using the voice commands input through the user. These voice commands to the wheelchair are given by the user using unilateral microphone as per user effort. The voice recognition will be done by voice recognition module and the output from this module is then received by Arduino. The goal of the system is to control access to voice commands by specific user to implement a wheelchair using small words recognition system.

2. Related Works

M. Senthil Sivakumar, Jaykishan Murji et al. [6] have implemented a voice-controlled wheelchair robot by detecting by voice capture module compared with predefined voices loaded in the system by voice recognition module. Khyati Meena, Shubham Gupta et al. [2] have developed a motorized wheelchair which works on user's gesture control. Their wheelchair

targets on old people and youngsters needing temporary rehabilitation. Mritha Ramalingam and Elanchezhian [4] have developed a wheelchair which can be operated using simple voice commands given by the user. This wheelchair is intended for people suffering physical disability providing easy access and more safety from obstacles with automatic protection service. The implementation of a voice-controlled wheelchair for disabled people applying Arduino and Bluetooth module is presented by Pramila Kupkar, Prajakta Pandit et al. [3]. They are targeted directly for physically disabled persons in the community with enhanced lifestyle.

3. The Proposed Voice-Controlled Pulse-Width Modulation (PWM) Based Wheelchair

The main purpose of the proposed voice-controlled Pulse-Width Modulation (PWM) based wheelchair is to develop a wheelchair which will move as per the user's voice commands. It performs on voice commands given by the wheelchair users. It consists of Voice recognition module, Arduino Uno, Ultrasonic sensor, DC gear motors, motor driver and Buzzer to complete the whole system. As the proposed system is to control all the movements of the wheelchair as per user's voice commands, the five basic movements of the wheelchair are described as follows:

- Moving Forward
- Moving Backward
- Turning Left
- Turning Right
- Stop Condition

The forward command moves the wheelchair forward until the obstacle is detected. Similarly, the backward command for the backward direction, to run the opposite movement of wheel rotation. The left command will make right wheel move backward and left wheel moves forward. The right command makes left wheel move backward and right wheel forward. For the stop command, the rotation of both motors will stop. Moreover, in this system, the wheelchair can be controlled by some angle where the user wishes to rotate its wheelchair by like 30˚, 45˚, 60˚ etc. The proposed system uses five major devices:

- Voice Recognition Module
- HC–SR04 Ultrasonic Sensor
- ATMEL Mega 328P Arduino
- L298 N Motor Driver
- Buzzer

3.1. Voice Recognition Module

To control the wheelchair, in this system, the voice recognition module v 3.1 is applied which is a speaker-dependent device enables to store 15 pieces of voice instruction for each 1500ms. Before using this module, the user needs to record his/her voices by using Arduino IDE software so that the device can recognize the voice of the use user. This voice module compares the received command to the pre-recorded commands by the user as training voice commands. If the command doesn't match with the pre-recorded commands, the wheelchair will not be activated to operate. In this system, the voice commands for the wheelchair are trained by using Arduino IDE software. There are totally seven commands such as Forward, Back, Right, Left, Stop, Low, and Help have been recorded in our system to achieve the target. The recording of speaker dependent voice commands can be received in any language that is comfortable for the user. The quality of microphone also plays the important role for the performance of the wheelchair. For training voice commands with Arduino, an USB cable is used to connect the Arduino with the laptop or desktop computer. After connecting the Arduino Uno with the laptop or desktop computer, the desired program can be opened to upload the code into the Arduino by clicking on the upload button and wait for a while. After uploading is done, the next step is to open the Serial Monitor which follows the commands and start the training process.

3.2. HC–SR04 Ultrasonic Sensor

HC–SR04 Ultrasonic Distance Sensor is used to determine the distance from the target object applying high frequency ultrasonic sound and it consists of 4 pins, Vcc, Trigger (input), Echo (output) and Ground. HC–SR04 can measure transparent and dark object insensitively to interference like smoke, different lighting condition and water vapour. The sensor used in the system is capable of detecting objects 80 cm (2feets and 8 inches) from the object's position and above 3inches(7.6cm) from the ground.

3.3. ATMEL Mega 328P Arduino

The ATMEL Mega 328P Arduino uno has 14 digital input/output pins (six capable of PWM output), 6 analog inputs, a 16 MHz quartz crystal, a USB connection, a power jack, an ICSP header and a reset button. Arduino receives the input given to the voice recognition module and converts into the format accepted by the motors.

3.4. L298 N Motor Driver

The Motor Driver L298N is a dual H-Bridge (a special circuit) type motor driver for Arduino which allows speed and direction control of two DC motors or stepper motors at the same time. The module can drive DC motors that have voltages between 5 and 35 V, with a peak current up to 2A for each. L298N applies pulse width modulation technique (PWM) to control the speed of the DC motors.

3.5. Buzzer

Buzzer module makes the simplest sound. Just change the frequency, the user can hear different sound (like alarm). In this system, buzzer can be alerted the guardians of the paralyzed person when buzzer makes a sound and take necessary care. If the paralyzed person needs any help, the wheelchair user speaks help command to turn on the buzzer.

3.6. Pulse-Width Modulation (PWM)

To achieve the speed control, the proposed wheelchair is implemented on Pulse-Width Modulation (PWM) technique which generates High and Low pulses to vary the speed in the motor. PWM is a powerful technique for controlling analog circuits with a microcontroller's digital outputs. It is a commonly used technique for generally controlling DC power to an electrical device, made practical by modern electronic power switches. PWM is used in many applications, ranging from communications to power control and conversion. For example, the PWM is commonly used to control the speed of electric motors, the brightness of lights, in ultrasonic cleaning applications, and many more. PWM is basically a digital unipolar square wave signal where the duration of the ON time can be adjusted (or modulated) as desired. This way the power delivered to the load can be controlled from a microcontroller. In this system, the proposed wheelchair is applied the output logic 1 level of the microcontroller which is +3.3 V. The frequency depends on the application. Therefore, the system generates a PWM signal with a frequency of 40 kHz. Moreover, the amplitude of the motor voltage remains constant, and the motor is always at full strength. The result is that the motor can be rotated much more slowly without it stalling.

4. Workflow of the Proposed System

The detail workflow of the proposed system is shown in figure 3.2. All the modules are needed to mount onboard as to ease the wheelchair movement. Firstly, a microphone which is located nearest position to the user to make it handy and easy to use. Generally, the input voice level of the user affects the recognition accuracy of the given voice commands. Then, the system is triggered by the voice commands produced by the user through the microphone. In fact, the voice of the user for the system is already trained and stored in the module.

When the user gives the commands, the module matches them with the existing commands and offers the suitable output if the voice and the command match. The system starts working by applying the supply voltage to the speech recognition circuit. All these commands are fed into the voice recognition kit via a PC/Laptop. The wheelchair will go back to the standby condition or end the whole system by turning off the power supply of the speech recognition board.

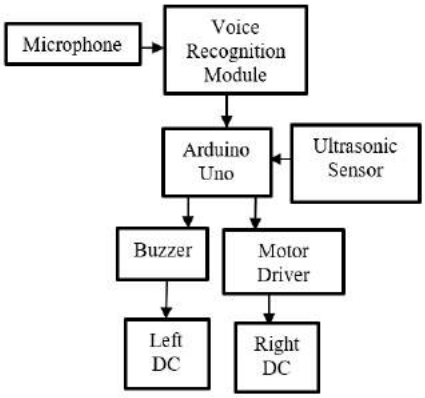


Figure 3.1 The architecture of the proposed system

The architecture for the implementation of the wheelchair in this system is shown in figure 3.1. Moreover, the prototype of the wheelchair is described in figure 3.3.

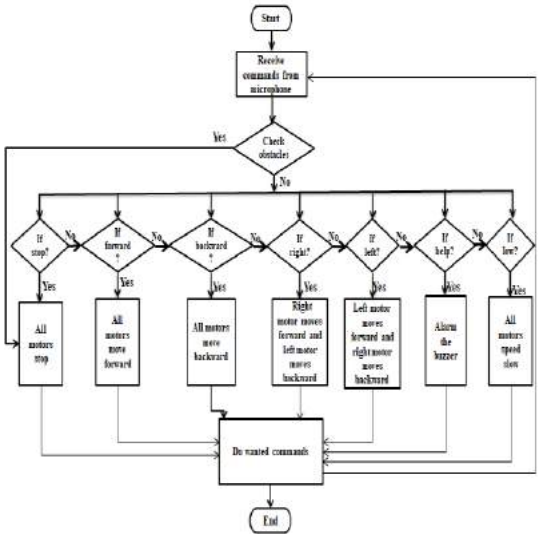


Figure 3.2 The workflow of the proposed system

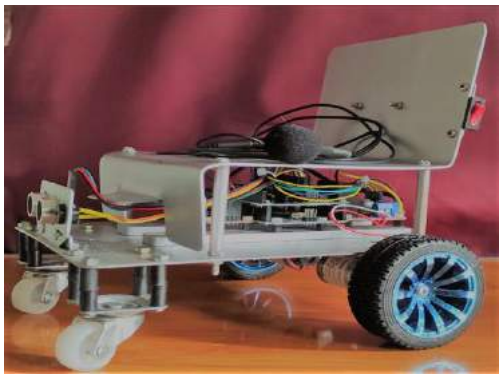


Figure 3.3 The design of wheelchair for the system

5. Experimentation of the Proposed System

To evaluate the performance of the voice controlled PWM based wheelchair, the recorded samples with voice commands are divided into two sets: training set and testing set. The experimental outcomes resulted from the proposed system present the performance of the system. By comparing the output results i.e., the environmental noise with the silent area with the purpose of calculating the accuracy of the proposed system. It is the measurement of the difference between the environmental noise area and the silent area results. In fact, there are totally 50 samples for each command word spoken by the wheelchair user. The percentage of accuracy for each command in environmental noise area and silent area in table 1. Table 1 shows the results of the wheelchair movements tested in the environmental noise and silent area. In total there are 50 samples for each command word spoken by the wheelchair user. The number of errors depending upon the commands of the user that cannot catch correctly by the voice module because of the quality of microphone and Arduino uno has a single hardware serial. The goal of the experiment is to find the accuracy and the correctness of the wheelchair in response to the voice commands in different conditions. The experimental results of wheelchair movement experimented after replacing new jumper wires are shown in Table 2. These results are recorded after replacing new jumper wires weekly and monthly tested results of 70 samples depending upon the motions of the wheelchair.

Table 1. Experimental results based on voice commands in environmental noise area and silent area

Voice commands	Result in environmental noise area		Result in silent area	
	No of correct commands	No of errors	No of correct commands	No of errors
Stop	41	9	45	5
Forward	43	7	46	4
Backward	37	13	40	10
Right	40	10	45	5
Left	35	15	40	10
Help	36	14	40	10
Low	40	10	45	5

Table 2. Results of the wheelchair movement experimented after replacing new jumper wires

Commands	One Week (70 Trials)	One Month (70 Trials)
Stop	10	9
Forward	9	9
Backward	9	8
Right	10	10
Left	9	8

Help	10	10
Low	10	9
Accuracy (%)	94.3	90

Moreover, the percentage of the accuracy of the wheelchair in environmental area is 78% (Accuracy = 350-78/350 x100%= 77.7%) and in silent area is 86 % (350-49/350x100%= 86%). The accuracy depends on the pronunciation spoken by the user. Based on the overall observation from the results above, there are four command words that have 80% of accuracy. They are "Forward"," low", "Right", and "Stop". This means that all of them are easy to be pronounced by the user. The hardest work to recognize by the system is "Left" which scored about 70% of accuracy in environmental area. This system is using US English and the command language is also used in the local language.

5. Conclusion

In this paper, the implementation of a voice controlled PWM based wheelchair using Arduino and voice recognition module intended controlling the motion of a wheelchair is developed for disabled especially paralyzed people. The direction of the wheelchair can be selected using the specified voice commands by the user. The system is intended to offer fully independent for users because they do not need another person to move the wheelchair. The design of the wheelchair not only reduce the manufacture cost compared with current market but also give great competitive with other types of electrical wheelchairs. The only thing needed to ride the wheelchair is the synthetic voice commands of the user. The voice-controlled wheelchair can also enhance safety for users who use ordinary joystick-controlled wheelchair, by preventing collision with walls, fixed objects, furniture, and other people. Therefore, all the drawbacks of the joystick-controlled wheelchairs are overcome by using the proposed voice-controlled wheelchair.

References

[1] N. Aktar1, J. Israt, L. Bijoya, "Voice Recognition based Intelligent Wheelchair and GPS Tracking System", *International Conference on Electrical, Computer and Communication Engineering (ECCE)*, February, 2019.

[2] M. Khyati, G. Shubham, K. Vijay, "Voice Controlled Wheelchair", *International Journal of Electronics, Electrical and Computational System (IJEECS)*, vol. 6, issue- 4, April, 2017.

[3] K. Pramila, P. Prajakta, D. Nikita, "Android controlled wheelchair", *Imperial Journal of Interdisciplinary Research (IJIR)*, vol. 2, issue-6, 2016.

[4] R. Puviarasi, R. Mritha, Elanchezhian ,"Self-Assistive Technology for Disabled People – Voice Controlled Wheel Chair and Home Automation System", *IAES International Journal of Robotics and Automation*, September, 2014.

[5] R.C. Simpson, "Smart wheelchairs: A literature review", *Journal of rehabilitation research and development*, July, 2005.

[6] M.S. Sivakumar, J. Murji, "Speech controlled automatic wheelchair", *International Conference on Information Science, Computing and Telecommunications (PACT)*, July, 2013.

[7] B.D. Vijendra, P. Subramanian, R Karthikeyan, "Voice Controlled Wheelchair", *International Journal of Pure and Applied Mathematics*, vol. 119, 2018.

Embedded based Smart Car Parking System Using Graph Theory

Nan Chaw Su Kyi, Khant Kyawt Kyawt Theint
University of Computer Studies, Yangon
nanchawsukyi@gmail.com, khantkyawtkyawttheint@ucsy.edu.mm

Abstract

An effective car parking of a condo is conducted in this system to reduce the time for finding the free parking slots. The condo has a two-way road for each entrance and exit. Smart car parking enables the user to find the nearest parking areas. Push Buttons are used for validation when the user arrives the entrance. If the user is pressed to Push Button, the system will check for the nearest empty slot and show to the user in LCD and the gate is opened by using servo motor. Graph Theory is applied to find the shortest path. As soon as a car has left a slot, this system redefines the state of the slot as free. This system is implemented by using Arduino mega 2560 microcontroller, Push Button, LCD, Ultrasonic sensor and Servo Motor. This system is implemented using Matlab and C++ programming language with Arduino software IDE.

Keywords: Arduino Mega 2560, LCD, Push Button, Servo motor.

1. Introduction

Nowadays, electronic control technology is advancing rapidly around the world. New electronic technologies and manufacturing systems are also rapidly changing.
In recent years, many control modes used in automobiles were developed by microcontrollers (minicomputers).

In densely populated cities, parking systems are an important part of life. The problem of parking leads to air population, traffic congestion and driver frustration in public places like shopping malls, cinemas, hotels and condos. The problem can be solved by making a smart car parking lot management system using open source hardware and programmable sensors. It is mainly on embedded systems. An embedded system is an electronic/electromechanical to perform a specific function, a combination of firmware and hardware [8]. It is safe and efficient for the parking system. Essentially, electronics are integrated circuits that perform computations for real-time performance.

The paper is organized as follows. In the next section, the related work for the smart parking lot management system is described. In Section 3, discusses the background theory of graph algorithm for the development of smart car parking system. The proposed system design and hardware implementation is presented in Section 4 and the experimental results and analysis are discussed in Section 5 and Section 6. The conclusion of this paper is summarized in the last section

2. Related Work

Many researchers have been worked for classification. Sifat Hassan [1] proposed for Automated Parking System using Graph Algorithm. The system reduces the hassle of encountering vehicles parked in automatic parking spaces. For most cases, the algorithm for automatic parking is O (n3) times. This algorithm is quite efficient compared to it.

Mohammad [2] developed Smart Indoor Parking System Based on Dijkstra's Algorithm. Parking space is determined by the size of the car to be used. Dijkstra's algorithm is the shortest path algorithm for graphs with non-negative dots.

JAAFAR [3] used Modifies Dijkstra's Algorithm for Intelligent Guidance Parking System. This system has been proved to print a mini-map to guide drivers and also equipped with GUI system. The Dijkstra's algorithm decides the shortest path from one single node to another single node to find the shortest path. Define the node that does not yet go directly to infinite, find the minimum weight from it and get the shortest path from that node.

3. Background Theory

In graph theory, the shortest path problem is the problem that a path between two verticals (or nodes) in a graph reduces the sum of the weights of its structural edges.

3.1 Graph Theory

Graph are data structures used to represent "connections" between elements. These elements are called nodes. They are objects in real life; Represents people and objects. Connections between nodes are called edges. Graph theory is the mathematical theory of the properties and

applications of graphs. Graph theory is ultimately the study of relationships. A graph in this content is made up of lines connected by dots. A graph G = (V, E) is a set of vertices V and E where each edge (u,v) is a connection between vertices u,v ∈ V.

In Graph Theory, a path is a sequence of distinct vertices and edges connection two nodes. There came be a plethora of paths that lead from 1 source node to a destination node. The number of paths can grow exponentially in terms of the input, but only a subset of those paths minimizes the sum of edges weights, and those are called shortest paths [9].

In this system, Graph theory uses direct graph and calculates using it procedures. The edges of the floor are calculated with different weights. For edges on the floor, weight values 3, 2 and 1 are used. The weights for these edges are set to find the nearest parking lot to the exit. In this selection, the system calculated the weights specified for the edges and direct graph theory to find the vacant parking lot nearest to the exit.

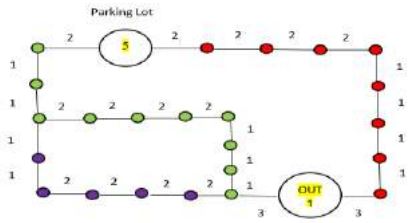


Figure 13: Three paths with different weight values of a node

3.1.1. Directed Graph and Weight Graph

Does not indicate the shortest route problem can be specified for graphs, whether directed or mixed. The definition for directional graphs is that the direction of the path needs to be connected in series vertices with the appropriate indication points. The indicated graph or graph has a direction at the edges. This system uses a directed graph because it has inputs and outputs.

Weight graphs: Many graph can have edge that contain a certain weight to represent an arbitrary value such as cost, distance, quantity, etc. A weighted graph is a graph in which each branch is given a numerical weight. They are used to analyze electrical circuits. To draw up project schedules, find the shortest routes they can be used to analyze social relationships and to design models for the analysis and solution of many other problems [10].

4. Proposed System

In the proposed system, an effective car parking of a condo is conducted with the aim of reducing the time for finding the free parking slots.

4.1 Hardware Requirements

In Hardware implementation, Arduino Mega2560, Push Buttons, LCDs, Servo Motors, and Ultrasonic Sensors have been used to realize the system.

4.1.1 Arduino Mega 2560

Arduino has greater academic applications. Working on a computer for compatible Arduino projects using hardware and software. The Arduino Mega 2560 is compatible with most other Arduino boards. The Arduino mega 2560 has 54 digital I / O pins and 16 analog inputs, as well as a 16 MHz crystal oscillator, USB connection, Power outlet in addition to the ICSP header and reset button [6].

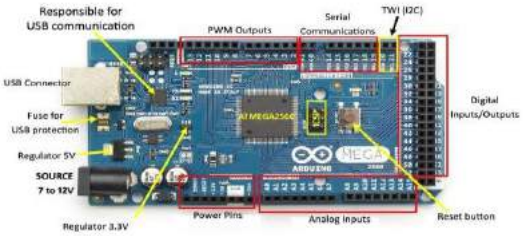


Figure 1. Arduino Mega 2560

4.1.2 Liquid Crystal Display (LCD)

LCD is used as a user interface. LCD will display the shortest path number by using Mega 2560. LCD circuit diagram is shown in Figure 2.

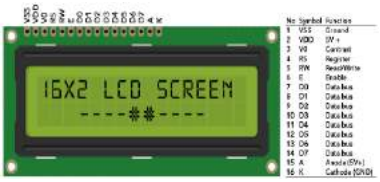


Figure 2. Liquid Crystal Display (LCD)

The LCD used is a 16 x 2 display with 32 alphanumeric characters [11]. The supply voltage is Vcc = 5V.

4.1.3 Servo Motor

The servo motor has an angular position that is not present in a normal motor. A rotary actuator or motor that allows precise control of acceleration and velocity [5].

Figure 3. Servo Motor

4.1.4 Ultrasonic Sensor

Uses ultrasonic sensor and determines the distance to a target. The module automatically sends eight 40 KHz to see if the pulse signal returns. Ultrasonic sensor (HC-SR04) provides 2cm-400cm or 1″ to 13 feet distance measurement function, the accuracy can reach to 3mm [7].

Figure 4. Ultrasonic Sensor

4.1.5 Push Button

A button is a simple push button to control certain parts of a machine or process. The surface is usually flat or shaped and made to fit a person's finger or hand [12].

Figure 5. Push Button

4.2 System Design

The flow chart of the smart car parking lot management system is illustrated as in Figure 6. As soon as a car enters, the initialization process begins. The gate will not be opened if the parking is not free. With only five parking spaces left, entrance1 (IN 2) will be closed. If there are more than five parking spaces, both entrances will be reopened. When a car enters, the user selects the exit port via a push button. Receiving data from empty parking lots via sensor, it calculates the closest exit to display on LCD. If the car comes to the exit without parking, the gate will open. If the car does not enter the parking lot, this system will be started. After leaving the car, the system will be given a record of leaving the car park.

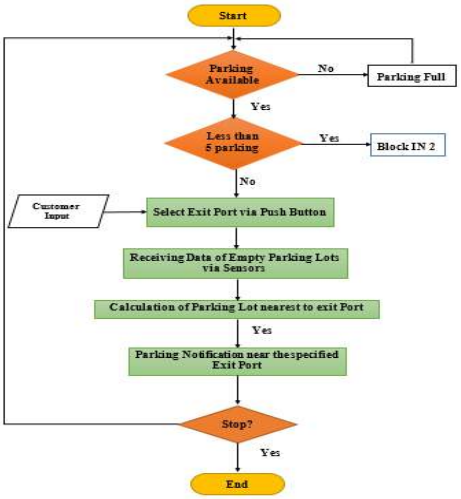


Figure 6. Flow chart of Smart Car Parking System

4.3 System Block Diagram

The diagram shows the input and output components connected to the Arduino Mega 2560. The input component is comprised of ultrasonic sensors, push buttons, an LCD, and a servo motor. The output components comprise an exit gate, ultrasonic sensors, and a servo. The system's operation starts when the car is placed in front of the gate. Then, when the driver presses the button, the counter circuit in the system will calculate whether there is an available vacant parking lot in the system and the shortest path between the two exit gates. If the vacant parking lot is available, the LCD will display the information about the parking lot available and the parking placement to the driver, and the entrance gate will open to allow the car to enter the parking. All of these can be summarized as shown in Figure 7.

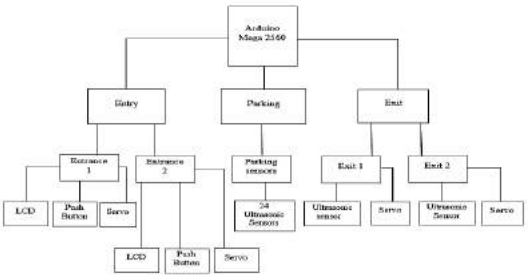


Figure 7. System Block Diagram of Smart Car Parking

5. System Experiments

This model would consider a car park in a condominium. Our analysis is to find a vacant spot closest to the exit for easy parking in the building. The four locations define two entrances and two exits. Therefore, the parking lot area is set from 5. The architecture has two entrances (IN1 and IN2)

and two exits (OUT1 and OUT2), and 24 parking spaces (lot5 to lot28).

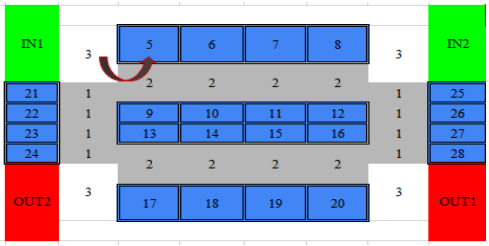


Figure 8. Architectural model of the parking system

Firstly, many methods from the shortest path problem use Graph theory. The shortest path for the direct graph is calculated with the weights specified in the Figure 8. The calculated with different weights for the floor of car parking.

ArwDs	IN1	IN2	OUT1	OUT2
IN1	-	-	-	-
IN2	-	-	-	-
OUT1	-	-	-	-
OUT2	-	-	-	-
5	-	-	15	9
6	-	-	13	11
7	-	-	11	13
8	-	-	9	15
9	-	-	15	9
10	-	-	13	11
11	-	-	11	13
12	-	-	9	15
13	-	-	11	5
14	-	-	9	7
15	-	-	7	9
16	-	-	5	11
17	-	-	11	5
18	-	-	9	7
19	-	-	7	9
20	-	-	5	11
21	-	-	15	7
22	-	-	14	6
23	-	-	13	5
24	-	-	12	4
25	-	-	7	15
26	-	-	6	14
27	-	-	5	13
28	-	-	4	12
	IN1	IN2	OUT1	OUT2

Figure 9. Shortest path number needed is taken for OUT1 (Exit3) and OUT2 (Exit4)

This data is taken from the calculated weights and directions. The theory is that the graph uses the graph to check for the nearest parking lot with exits OUT1 (Exit3) or OUT2 (Exit4). If space is available, indicate the destination. Theoretically, the weights and directions for the shortest path are calculated using Matlab. Then, it's written in the Arduino code with an if-then rule, depending on its weight.

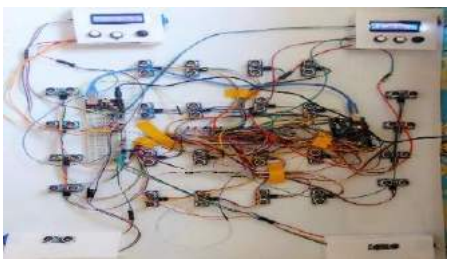


Figure 10. Construct Smart Car Parking Lot Management System

The input will be selected from the push buttons. The entrance numbers are IN1 (Entrance1) and IN2 (Entrance2), and the exit numbers are OUT1 (Exit3) and OUT2 (Exit4). Each entrance will have an LCD display and three push buttons (number 3, number 4 and number A). The system will indicate the parking space nearest to the exit when the car enters. Selecting push button number OUT1 (Exit3) will find the shortest path closest to OUT1 (Exit3) or select push button number OUT2 (Exit4) to find the shortest path closest to OUT2 (Exit4). Alternatively, if the user selects the A button, the best shortest path will be auto-selected as shown in Figure 11.

Figure 11. Display the first steps

The second part that works is the input part. There are two entrances to the entrance, but if there are fewer than five cars, IN2 (Entrance2) will be automatically closed and the system will only be able to enter from IN1 (Entrance1), as shown in Figure 12. If there are more than five parking spaces, both entrances will be reopened.

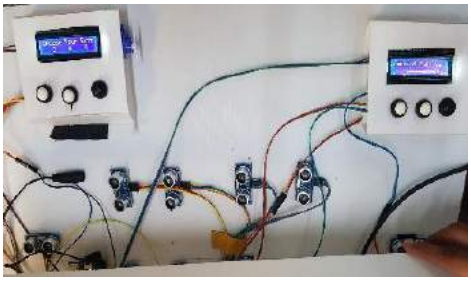


Figure 12. Display for IN2 (Entrance2) will be automatically closed

When all the parking spaces are full, both LCDs will show parking full, as shown in Figure 13.

Figure 13. LCDs will show parking full

If the user selects OUT1 (Exit3), the parking lot closest to the exit will be displayed as shown in Figure 14.

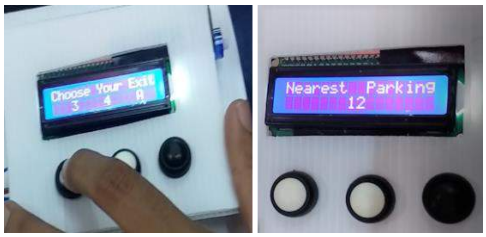


Figure 14. LCDs will show Nearest Parking

6. Analysis for Car Parking Searching

In this system, each lot has two lanes leading to the exit. The value of their paths is analyzed based on weight. Figures 15 and 16 show the analysis of the shortest path closest to the exit, depending on the distance of each of these routes.

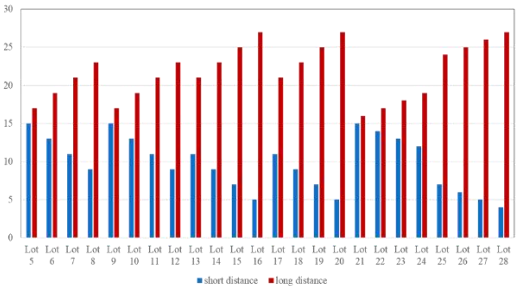


Figure 15. Analysis for Car Parking Searching in OUT1 (Exit3)

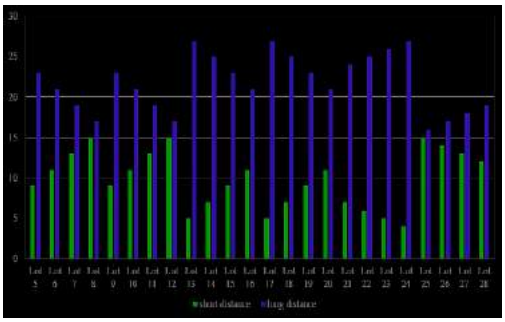


Figure 16. Analysis for Car Parking Searching in OUT2 (Exit4)

7. Limitation and Future Work

The system has two entrances and two exits, so it is not suitable for a condo in the middle of the street. It can only be used for a condo on the corner. When entering the parking lot, it will only be directed to the nearest parking lot with the exit, so it will not be able to point to the nearest entrance to the condo.

This work can be extended to autonomous car parks where the display will be used to detect the availability of parking lots as well as accepting different payment methods. In this system, it would be better for drivers to add sensors and lights and arrows along the lane at the entrance to the parking lot.

8. Conclusion

In this system, the proposed system requires minimal manpower and there is a minimum chance for human errors. The advantages of this will allow drivers to check for available parking spaces nearby. This system gives drivers a friendly car parking experience. The drivers can easily find a place for parking and save time in doing so. The system was helpful in reducing the wasted time of searching for parking lots and also improving the parking lot estimation.

REFERENCES

[1] Md. Sifat Hassan, Nushrat Naima Islam, Asim Md. Ahsan Fahim, Tanvir Hasan Turja, Shahrin Chowdhury: Automated Parking System using Graph Algorithm, Janaury 2020.

[2] K.I. Mohammad Ata*, A. Che Soh*, A. J. Ishak, H. Jaafar, N. A. Khairuddin: Smart Indoor Parking System Based on Dijkstra's Algorithm, April 2019.

[3] H. JAAFAR*, M. H. ZABIDI, A. C. SOH, T. P. HOONG, S. SHAFIE, S. A. AHMAD: Intelligent Guidance Parking System Using Modifies Dijkstra's Algorithm, October 2014.

[4] Mr. kush Shah, Ms. Priya Chaudhari, Arduino Based Smart Parking System International Research Journal of Engineering and Technology, January 2017.

[5] https://www.google.com/search?channel=nrow5 & client=firefox-b-d&q=what+is+servo+motor

[6] https://www.arduino.cc / en /Main/arduinoBoard Mega2560

[7] https://www.fierceelectronics.com/sensors/what-ultrasonic-sensor

[8] https://www.heavy.ai/technicalglossary/embedded-systems

[9] https://www.tutorialspoint.com/basic-concepts-of-graphs

[10] https://www.researchgate.net/figure/Weighted-directed-graph-i-i-iR-i_fig9_281632076

[11] https://www.google.com/search?channel=nrow5 &client=firefox-b-d&q=what+is+lcd+display

[12] https://www.google.com/search?channel=nrow 5&client=firefox-b-d&q=what+is+push+button

An Obstacle Avoidance Person Tracking Robot using Bubble Rebound Algorithm

Shwe Yi Paing, Htar Htar Lwin
University of Computer Studies, Yangon
shweyipaing8@gmail.com, htarhtarlwin@ucsy.edu.mm

Abstract

The current ultrasonic impediment evasion robot just purposes an ultrasonic sensor during the time spent obstruction aversion, which must be tried not to accord to the decent snag evasion course. Hindrance evasion can't follow extra data. An individual following portable robot is a creative versatile robot, which can perform individual following and impediment evasion undertakings all the while. Ultrasonic Position based approach is used in this framework for recognizing and finding the objective individual. This system also used the obstacle avoidance bubble rebound algorithm which can avoid the obstacles on the way of tracked person. The objective of these system is to investigate the feasibility of developing a person-tracking robot system using ultrasonic positioning for person tracking and ultrasonic sensor for obstacle avoidance. This system is implemented with C Language on Arduino IDE by using ultrasonic position sensors and Arduino MEGA board.

Keywords: obstacle avoidance, person-tracking, ultrasonic positioning, mobile robot, bubble rebound algorithm

1. Introduction

With the development of artificial intelligence technology, mobile robots are broadly utilized in canny processing plants, present day planned operations, security, accuracy horticulture and different angles. Wheeled portable robots have been generally utilized away and transportation fields. The focal point of this framework is to keep away from deterrents between the following individual the robot and follow an individual. The main thing to understand the independent movement control of a versatile robot is to get the data of the general climate and move it to the principal regulator to change over it into control order, to guarantee that the robot can securely and steadily keep away from all snags while moving to the objective, which can be accomplished when the portable robot has major areas of strength for a framework. Various sorts of sensors are expected for various data. The detecting advances of portable robots incorporate aloof detecting in light of different cameras, sound system vision and infrared cameras and dynamic detecting utilizing different sensors to identify dynamic or fixed snags continuously. Laser running is utilized to investigate the wheel slip of the four-wheel sliding guiding portable robot. A few different investigations have proposed target following of wheeled portable robots in light of visual strategies [3].

For an obscure climate, sensors are normally utilized for insightful snag aversion and way arranging. The early strategy for deterrent evasion and way arranging is to identify the stickers on the ground by infrared beam for route. This strategy must be utilized in a known climate.

As of now, the exploration on a snag aversion robot is generally about the engine driving rule, engine speed guideline conspires and going standard, and the examination on hindrance evasion is likewise about impediment evasion. An individual following versatile robot is a robot that follows an individual while at the same time carrying out impediment evasion. Individual following is a procedure utilized by robot and independent vehicles to follow a human inside a particular reach. The robot follows the objective individual and evasion the obstruction between the objective and robot.

Not many individuals concentrate on versatile robots when they experience pits during programmed travel.

2. Related Work

Many researchers have developed person tracking mobile robot and obstacle avoidance mobile robot using various method and various components.

In related work [1] proposed the tracking vehicle is programmed to maintain a fixed distance from the moving object and follow it. This system makes to follow a moving object based on automatic tracking system with three class: distance stabling process, line tracking system and main control system. The distance stabling process is the main process to control the motor's PWM (Pulse-Width- Modulation) speed as the feedback of ultrasonic sensor.

In related work [2] proposed automated obstacle avoidance system for mobile robots is designed and implemented in embedded system. The purpose of the system is to find the obstacle by using five Ultrasonic sensors and to avoid this obstacle by controlling the speed and direction of two DC motors using PIC. This work mainly deals with the automated obstacle avoidance (AOA) system of this paper by using Peripheral Interface Controller (PIC) 16F877A microcontroller that is implemented to design for AOA system and it is simply evaluated on PIC.

3. Background Theory

There are several approaches to the person-tracking robot and they are briefly explained as follow:

3.1. Vision-Based Approach

This is methodology utilizing a camera to catch the picture of the objective individual. The picture must be refreshed continuously. These strategy expects that the objective individual recognition is fruitful, albeit this may continuously be a test. Subsequent to distinguishing the objective individual in a picture, the control data, including headings and distances, will be processed from the varieties of the objective position and size in the picture.

The robot ought to then have the option to push toward the objective individual in light of this data. Various explores [2] have taken on and adjusted this way to deal with foster the individual following portable robots. In any case, a few vulnerabilities can in any case be sufficiently critical to impact the productivity of target identification. One element that influences the identification is light condition. Deciding the objective individual in the picture can be somewhat more troublesome when the variety or brilliance of the objective isn't sufficiently remarkable to make it not quite the same as that of the foundation or different deterrents. Another element that influences recognition is the concurrent movement of both individual and robot.

The vision sensor can undoubtedly lose the objective individual when the objective individual moves excessively quick. A few scientists utilized dynamic cameras. This diminished the issue of losing the objective individual, however expanded the trouble in the calculation plan. This approach isn't appropriate for the robot to perform impediment aversion. It is hard for the robot to differentiate between the objective individual and different snags. The circumstance may be more regrettable when there are a few people moving around in a similar climate. It is conceivable and reasonable for the robot to lose the objective individual assuming that the climate is unstructured.

3.2. Non-vision Based Approach

A non-vision-based approach utilizes a few sorts of rangefinders, like sonar sensors, infrared sensors, and others. Every rangefinder on the robot can decide the distance between the closest article and the actual rangefinder. Since the robot can't recognize item and target individual, this approach must be taken on to carry out either deterrent aversion while viewing every one of the articles as hindrances or individual following when the objective individual is generally the closest item to the robot with no in the middle between. Utilizing a Nomad 200 versatile robot furnished with 16 sonar rangefinders, the distance of the item can be processed by the closest sonar unit, and the rough heading likewise not entirely settled from the general area of the sonar unit, which identifies the closest distance. The robot can be productively customized to carry out hindrance evasion. Be that as it may, to furthermore execute individual following undertaking, even in a climate with a decent condition, is as yet troublesome and not viable.

3.3. Transmitter-and-receiver Based Approach

By using a transmitter-and-receiver approach, the transmitters situated on the objective individual send signals, like ultrasonic waves or squinting LED. The collectors situated on the robot get those signs. Subsequent to figuring the distance and the point of the objective individual from those signals, the robot knows where to push to turn itself toward the objective individual and abatement the in the middle between. In [1],

two transmitter-and-beneficiary based approaches have been talked about.

a. Person Tracking Using Blinking LED Devices: This approach requires outfitting the objective individual with two infrared LED gadgets with fixed distance among them and utilizing a camera on the robot to identify the two gadgets. This is like the vision-based approach. The principal contrast is that the signs from infrared LED gadgets ought to be firmer and not impacted by the aggravation in that frame of mind, for however long they are not hindered by any hindrance. The camera can pivot to keep the objective individual in the picture. By registering the distance between two LED lights and the deviation of the two lights from the focal vertical hub in the picture, the reach and the direction of the objective individual can be acquired separately by the robot.

b. Person Tracking Using an Ultrasonic Positioning System: This approach is to prepare the ultrasonic transmitters on the objective individual and the beneficiaries on the robot. By processing the time stretch among sending and getting the ultrasonic sign, the distance between the objective individual and the robot not entirely set in stone. The point can likewise be figured from the time postpone between a few recipients. These methodologies are straight-forward for individual following; however, they are not reasonable when there are obstructions between the objective individual and the robot. The discovery of obstructions will be an issue utilizing these methodologies. With practically no extra instrument, the robot can't execute deterrent aversion.

3.4. Intelligent Space Approach

The intelligent space approach uses a few sensors, for example, visual or non-visual sensors that are situated in the climate to recognize both the robot and the objective individual. In this way, the position data of the robot and target individual will be in the worldwide not entirely set in stone by the sensors in the astute space. From the general places of the robot and the objective individual, the robot movement will be arranged by this smart space and controlled through the organization. Notwithstanding, the ideal methodology in this exploration is to plan an independent robot that carries out assignments in unstructured conditions. This approach then, at that point, becomes unsatisfactory despite the fact that it could be all around worked.

4. System flow and Algorithm

This system aims to prevent the robot from collision when tracking and which could damage the robot system. When the bumper sensor sets an input signal as 1, the process should go directly to “stop.” If the bumper sensor signal value is not set, person tracking can be executed. When the sensor ranges are smaller than the 20 cm threshold, the action for obstacle avoidance should be made. In the event that the former situation no longer exists, the robot is carried out in tracking the target. The detail processing steps are explained as following sections.

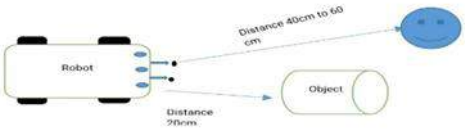


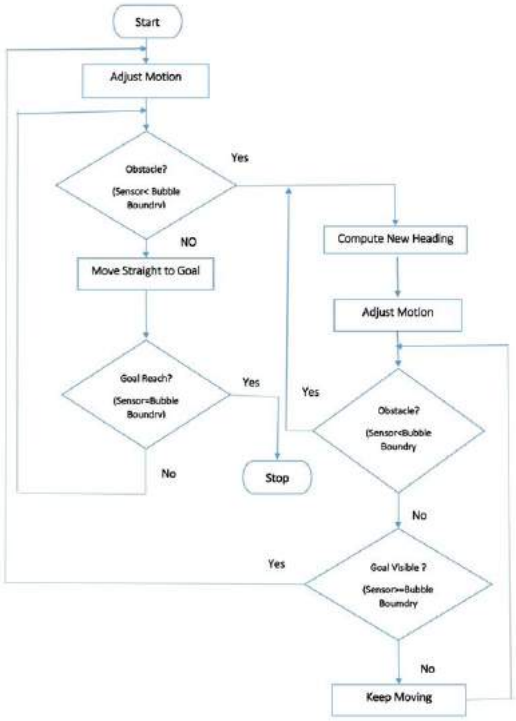
Figure 1. The Overview of System Flow

4.1. Bubble Rebound Algorithm for Person Tracking and Obstacle Avoidance

According the algorithm, the robot reacts to obstacles detected within an area called “sensitivity bubble”, whose shape and size are dynamically adjusted, depending on the chassis of the robot. Upon detection of an obstacle, the robot “rebounds” in a direction and then continues its motion in this direction until the goal becomes visible. (Example: Firstly, the robot moves straight towards the goal. If an obstacle is detected within the sensitivity bubble, the robots “rebounds” in a direction found as new direction and continues its motion in this new direction until the goal becomes visible.) If the person is moving condition, the tracking system maintains the distance between the Person and Robot at 60cm. If the person is detected the front view of the robot and the distance of the robot and person

is greater than 40cm, it well keeps going on. Although the person is detected the front view of the robot, but the person is not moving; the robot go straight to track person until the distance is greater than 30 cm.

5. Flow Chart of the Algorithm



Algorithm for Obstacle Avoidance

```
Input   RightSensor ← distance;
        FrontSensor ← distance;
        LeftSensor  ← distance;
BEGIN
        Adjust heading to goal;
        Detect the obstacle ( );

        if (obstacle found == 1)
        {
           Step 1:  Compute new heading;
                    Check orientation ( );
           Step 2:  Adjust motion;
        }
        else
        {
           Moving straight to goal;
        }
        end if
END
```

```
Check orientation ( )
  {
    if  (LeftSensor  <  bubble_boundary  &&
        FrontSensor  <  bubble_boundary  &&
        RightSensor < bubble_boundary )
    {
     Stop moving;
    }
    end if

    if  (LeftSensor  >  bubble_boundary  &&
         FrontSensor  >  bubble_boundary  &&
         RightSensor > bubble_boundary)
    {
     Moving straight to goal;
     }
     end if
     if  (LeftSensor  <  bubble_boundary  &&
          FrontSensor  >  bubble_boundary  &&
          RightSensor > bubble_boundary)
     {
      Change orientation to right;
      }
      end if

      if  (LeftSensor  >  bubble_boundary  &&
           FrontSensor  >  bubble_boundary  &&
           RightSensor < bubble_boundary)
      {
       Change orientation to left;
      }
      end if
      if  (LeftSensor  <  bubble_boundary  &&
           FrontSensor  <  bubble_boundary  &&
           RightSensor > bubble_boundary)
      {
        Stop moving;
        Change orientation to right;
       }
       end if
      if  (LeftSensor  >  bubble_boundary  &&
            FrontSensor  <  bubble_boundary &&
           RightSensor < bubble_boundary)
      {
       Stop moving;
       Change orientation to Left;
       }
        end if
```

Algorithm for Person Tracking

```
Input     RightSensor2 ← distance;
          LeftSensor2  ← distance;
BEGIN
          Adjust heading to goal;
          Detect the person ();
          if (person found == 1)
          {
          Step 1:   Adjust motion;
          Step 2:   Tracking;
          }
END
```

```
Detect the person ();
          if (LeftSensor2 < bubble_boundary &&
              RightSensor > bubble_boundary)
          {
           Change orientation to left for track;
           }
            end if

          if (LeftSensor > bubble_boundary &&
              RightSensor < bubble_boundary)
          {
           Change orientation to right for track;
          }
          end if
```

4.2 Components List and Explanation

- **GM25-370-24140 DC Gear Motor (4pcs)**

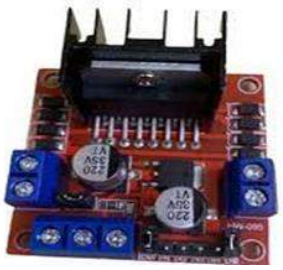
It consists of a electric DC motor and a gearbox or gearhead. These gearheads are used to reduce the DC motor speed, while increase the DC motor torque.

- **25mm Motor Bracket (4pcs)**

- **DC Motor Wheel (4pcs)**

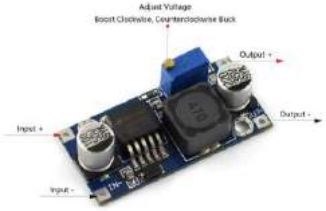
- **Motor Driver HW-095 (1pcs)**

The function of motor drivers is to take a low-current control signal to turn it into a higher-current signal that can drive a motor.



- **LM2596 Adjustable Power Supply (1pcs)**

The DC-DC Step-down Adjustable Power Supply Module with 3digit LED Display is based on monolithic integrated circuit.



- **UNO Prototpye Shield (1pcs)**

Arduino UNO Prototype Shield (Robot DYN) is specifically developed for easy connection between Arduino Mega and other devices.

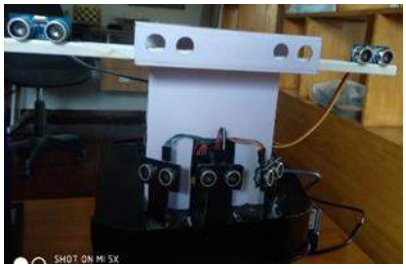
- **Arduino Mega Board (1pcs)**

Arduino Mega depends on the Atmega 2560 microcontroller. It includes digital input/output pins-54, where 16 pins are analog inputs, 14 are used PWM outputs hardware serial ports, an ICSP header, a power jack, a USB connection and reset button. The operating voltage of this board is 5volts, but the input Voltage will range from 7volts to 12volts.

- **3 Cells Lipo Battery 1500mAh (1pcs)**

Premium 103450 3.7V, 1500mAh rechargeable Li-Po battery. High Quality and durable with large capacity of 1500 mAh, less worry about battery's going dead with a short time.

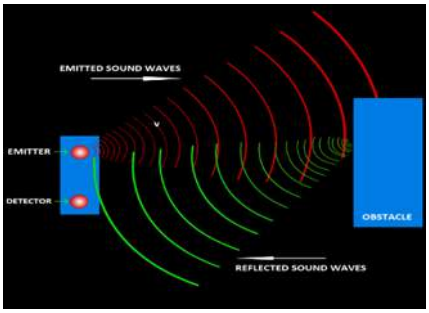
- **DIY Robot Kit (1pcs)**



- **Ultrasonic HC-SR04 Distance Measuring Sensor (3pcs)**

The HC-SR04 ultrasonic sensor uses sonar to determine distance to an object like bats do. There are only four pins that are VCC (Power), Trig (Trigger), Echo (Receive), and GND (Ground). It works by sending sound waves from the transmitter, which then bounce off of an object and then return to the receiver. By calculating the travel time and the speed of sound, the distance can be calculated.

- **Ultrasonic Sensor Work**



6. System Implementation

This system aimed to detect the person to track and to avoid the obstacle on the running route. Firstly, this system is implemented for obstacle avoiding process by using Arduino Uno-Board with motor drivers on four forward/reverse motors and combination of three ultrasonic sensors. It was a good robot in avoiding obstacle.

Second, this system adds the remaining process for the tracking, the Arduino Uno-Board is not compact for tracking and obstacle avoiding process. So, this system development changes the Arduino Mega board and add the new compactable shield. In the implementation phase for person tracing and obstacle avoiding, this system used five ultrasonic sensors (same sensors) are used. But it is not work correctly, because it is wrongly detect the person as obstacle and avoid instead of tracking. (Mixing and conflict tracking and obstacle avoiding)

So, this system added two new ultrasonic sensors (Ultrasonic Position Sensors) for tracking. After changing the sensors, it can work as perfectly. The newly changed sensors can detect more range and distance than avoidance sensors (three ultrasonic sensors).

7. Experimental Result

(I) Experiment result for Tracking

Criteria	Obstacle Avoidance Sensor (Left) Unit-cm	Obstacle Avoidance Sensor (Front) Unit-cm	Obstacle Avoidance Sensor (Right) Unit-cm	Tracking Sensor (Left) Unit-cm	Tracking Sensor (Ring) Unit-cm	Route	Velocity
Detect Range 1	>30	>30	>30	>60	>60	Continue Current route	10cm/s
Detect Range 2	>30	>30	>30	<50	>60	Change Orientate to left and continue	0cm/s
Detect Range 3	>30	>30	>30	>60	<50	Change Orientate to right and continue	0cm/s
Detect Range 4	>30	>30	>30	<30	<30	stop	0cm/s

(II) Experiment result for Obstacle Avoidance

Criteria	Obstacle Avoidance Sensor (Left) Unit-cm	Obstacle Avoidance Sensor (Front) Unit-cm	Obstacle Avoidance Sensor (Right) Unit-cm	Route	Velocity
Detect Range 1	>30	>30	>30	Continue Current route	10cm/s
Detect Range 2	<30	>30	>30	Change Orientate to right and continue	0cm/s
Detect Range 3	>30	>30	<30	Change Orientate to left and continue	0cm/s
Detect Range 4	>30	<30	>30	stop	0cm/s

(III) Experimental Result (Time consuming Point of View)

Criteria	Setting Delay Time	Result for Respond	Result of Serial Monitor
Time to Complete one whole circuit	10ms	Enough to detect and respond in time	Can't catch by human vision
Time to Complete one whole circuit	100ms	Not enough to detect and respond the rules of the proposed system	Enough time to catch and trace by human vision

8. Conclusion

The primary target of this framework is to explore the possibility of fostering an individual following robot framework utilizing ultrasonic situating framework. In this framework utilized C Programming to foster individual following portable robot. The information created by the ultrasonic situating framework can be used by the robot. Utilizing the ultrasonic situating framework has been proficiently settled by sequencing the execution request of those two controls.

This framework was to examine the practicality of fostering an individual following robot framework utilizing a ultrasonic situating framework. To achieve this goal, the accompanying objectives have been accomplished in this framework. 1. Made the connection point between the ultrasonic situating framework and the robot framework in the working system. 2. Fostered the plan of the calculation that can at the same time keep away from hindrances and track the assigned individual in an unstructured climate. 3. Finished individual following analysis when there is no impediment between the robot and the objective individual. 4. Finished individual following examination when there is an impediment between the robot and the objective individual. 5. Shown individual following when the objective individual makes a turn at a corner. 6. Shown individual following in an unstructured climate.

References

[1] Hung-Kwan Chan. "*The Implementation of a Person Tracking Mobile Robot*." Master's Thesis, The Chinese University of Hong Kong, Hong Kong, China, July 2004.

[2] Nader Samir "*Collision Avoidance and Simple Path Planning for Autonomous Robotic Exploration*", Lulea University of Technology, Sep 2014.

[3] *"A study on Techniques of person Following Robo*t", Mohd Fahmi Mohamad Amran, University of Malaysia, Sep 2015.

[4] *"Obstacle Avoidance Path Planning Design for Autonomous Driving Vehicles Based on an Improved Artificial Potential Field Algorithm*" Shandong University of Technology, China, June 2019.

Fire Detection and Alarm System Using Fuzzy Logic Control

Yee Mon Thaw, Htar Htar Lwin
University of Computer Studies, Yangon
yeemonthaw0099@gmail.com, htarhtarlwin@ucsy.edu.mm

Abstract

Ordinary fire discovery and caution frameworks utilize a solitary smoke alarm which is associated with a fire the board framework to give early caution before the fire spreads to harm level. Nonetheless, found just fire identification frameworks in light of smoke alarms are not compelling and brilliant on the grounds that they generate false alarm in case someone smokes. It is fundamental that a keen fire observing framework in light of multiple sensors that uses various parameters, such as the presence of temperature, humidity, smoke and flames. To get such a savvy arrangement, a multi-sensor arrangement is required that can keenly utilize sensor information and create genuine caution for additional fire control and the board. This framework proposed a savvy fire identification arrangement and caution framework in light of fluffy rationale to recognize the genuine presence of fire. The advancement of home fire ready framework is assembled in light of Arduino board. The fire is identified at a beginning phase and the framework produces an alert with various crisis level. This system can help users to improve their safety standards with immediate response by preventing accidents. This will eventually allow both the lives and the properties from the disaster.

Keywords: Temperature, Humidity, Smoke, Flame, Arduino, Alarm, Fire detection, Fuzzy logic system

1. Introduction

Fire and smoke are among the significant reasons of the incidental losses [1]. Fire recognition is significant as the fire makes serious harms both human existence and non-living resources. The majority of houses miss the mark on alarm framework that causes the occupant a serious gamble ablaze breakout in homes [2]. The fire breakout can likewise happen without the occupants. A large portion of alarm frameworks accessible in market are in wired association mode. These kinds of framework need constraints. It likewise doesn't meet the new programmed savvy home's necessities [3]. Consequently, an insightful remote alarm framework is required to have been created with lower upkeep alongside more secure and simpler.

Today there is countless financially open sensors used for fire acknowledgment. These sensors can be requested similarly as their ability to distinguish temperature levels, moistness and smoke. These sensors have shown fruitful for explicit applications can work as autonomous component or as free substances can be integrated into existing alert structures.

Due to movements in sensor development, numerous mechanical assemblies have been made to make life safer. The foremost point of convergence of this structure is an early fire area and watchfulness system using various sensors. This intends to make a proliferation of the fire peril finder using soft reasoning. The commitments from the structure are fire, smoke, and gas, finally, as needs be, the sign will be started and conveys sound of moving tones depending upon the level of the gamble. The level of potential gamble suggests that the chime isn't unnecessarily speedy while it ends up being speedier when the level is risky. In addition, when the level of fire achieved conceivable gamble and dangerous, the water sensor will establish as well as if the level of gas achieved expected risk and hazardous the vapor fan sensor will dynamic.

The recreation of fire danger identifier utilizes fluffy rationale and the Arduino microcontroller with comparative data sources, yet in this examination, the Arduino microcontroller is restricted exclusively to actuating the ringer.

2. Related Work

In the paper [1] presents the design and implementation of cost effective and reliable automated GSM based fire alarm system by using DHT-11 sensor and MQ-2 Sensor and buzzer. The system with the alarm algorithm detected fires that were alarmed by smoke sensors and temperature sensor. A fire detection system is developed based on the simultaneous measurements of temperature and smoke. This system does not use flame sensor and not detect humidity. The paper [2] proposed IoT based fire alarm prototype system involves a temperature sensor and MQ-2 gas sensor. The system detects the temperature of threshold value (104˚F For more) and shows an alert notification on the LCD display and sending SMS message. Fire detection and avoidance of fire accidents is one of the necessary and important application of home automation. In paper [3], the System Safe from Fire (SFF) by using multi-sensor, actuators, and operated by micro-controller unit (MCU) is an intelligent self-controlled smart fire extinguisher system. If any sensor is high, the system ring alarm. Sensors placed in different areas for monitoring purpose and input signals are taken from that sensors and combine integrated fuzzy logic to detect fire breakout location and severity and discard false fire situation, such as cigarette, smoke, welding etc. SFF notifies fire services and other by text messages and telephone calls when the fire is detected.

3. Problem Statements

Fire is very useful, it fills a ton of need for us for however long it is our taken care of, yet when it leaves our control, it very well may be the reason for lives and property harms. For a really long time, man has relied upon fire to cook and warming to a significant wellspring of lighting, subsequently huge number of individuals pass on every year and it harm property too each year very nearly billion-dollar property misfortune is assessed. Fire can be tracked down in two spots, indoor and open air. Outside will be fierce blaze and indoor will construct flames and house fires. Within the sight of fire, it produces heat, smoke, fire and in outcome suffocation increment, hard to take breath and air is straightforwardly impacted by it. Fire needs three things to begin and to proceed is oxygen, heat, fuel. To stop fire, we should take out something like one of them to extinguish a fire.

There are many reasons for a fire at home; not many of them are portrayed beneath:

- Environmental changes
- Carelessness
- Unattended stove and gas
- Re-ignite cigarette not properly extinguished
- Faulty wiring and reckless use of the electrical appliance
- Flammable liquids
- Lighting

Therefore, to overcome fire disasters it is necessary to take some action as soon as possible to control fire.

4. Background Theory

The Arduino IDE runs on various stages like Macintosh, windows and Linux. Straightforward and clear writing computer programs is conceivable in the event of Arduino IDE. The Arduino libraries assume a significant part in making the programming more straightforward by giving more extensive scope of libraries. There are many implicit libraries accessible in the Arduino IDE and it permits to add extra libraries that are accessible in the open hotspot for download. Adding of new sheets to Arduino IDE is conceivable. Since, Arduino C is gotten from C and C++ programming and is a lot simpler when looked at another regulator programming.

4.1. Research Methodology

The research methodology described the implementation of the Arduino fire alarm system by utilizing the standards of past fluffy rationale. To frame a fluffy set, this exploration utilizes triangle and trapezoidal capabilities connection to characterize the participation capability. The result got will be in the fluffy set. Every one of the factors utilized to make the On the Off chance that Principles.

The qualities will then, at that point, go through the fuzzification interaction which is the method involved with changing the fresh worth into a fluffy set. Subsequent to completing the fuzzification interaction, the worth will then check the situation with the fire happened in view

of the consequences of fluffy standards. Some rules are as follow:

(1) If the fire is "extremely huge" and smoke is "very little" and gas is "exceptionally low" then the alert is "not hazardous".

(2) If the fire is "exceptionally huge" and smoke is "very little" and gas is "high" then, at that point, the caution is "possible risk".

(3) If the fire is "extremely enormous" and smoke is "exceptionally minimized" and gas is "exceptionally high" then, at that point, the alert is "risky".

(4) If the fire is "enormous" and smoke is "conservative" and gas is "medium" then, at that point, the caution is "likely risk".

(5) If the fire is "huge" and smoke is "minimal" and gas is "extremely high" then the caution is "risky".

(6) If the fire is "typical" and smoke is "very little" and gas is "exceptionally low" then, at that point, the caution is "not perilous".

(7) If the fire is "little" and smoke is "medium" and gas is "exceptionally low" then the alert is "possible risk".

(8) If the fire is "little" and smoke is "medium" and gas is "high" then, at that point, the caution is "risky".

(9) If the fire is "tiny" and smoke is "exceptionally minimal" and gas is "high" then, at that point, the caution is "perilous".

(10) If the fire is "tiny" and smoke is "nearly nothing" and gas is "extremely low" then, at that point, the caution is "likely risk".

4.2. Fuzzy Logic Control

Fuzzy Logic Control (FLC) is one of the most famous methodologies read up widely for fire the executives' purposes. FLC is a numerical device used to deal with rationale of imprecision, practically equivalent to the capacities of the human mind. As displayed in figure 1, the FLC cycle model, which has five fundamental sub-processes, in particular:

(i) fuzzification
(ii) application of fuzzy operations,
(iii) implication
(iv) aggregation and
(v) defuzzification

In FLC, the inputs are initially fuzzified or modelled into membership functions then eventually applied with logical operators whether to intersect (AND) or disjoint (OR).

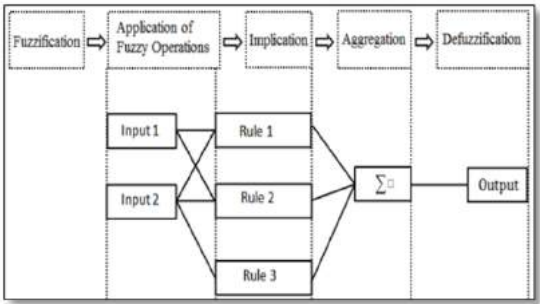


Figure 1: Fuzzy Login Control Process

Afterwards, the subsequent fluffy sets are exposed to suggestion technique where a Standard based arrangement of "on the off chance that calculations" are made. These fluffy sets are then additionally collected and deffuzified remove the result for explicit application. While FLC is viewed as one of the top decisions in planning and creating fire-related frameworks because of its effortlessness, adaptability, and strength in demonstrating defective signs, the distinguished fire-related gives actually stay unanswered, subsequently, the headings of examination ought to concentrate more in tracking down mediations and arrangements on these issues.

This framework presents the usage of FLC in planning a structure toward the improvement of a fire location and caution framework. It plans to investigate an expected strategy for further developing adaptability and responsiveness of sensors, and to guarantee dependability of fire location and caution framework.

5. Implementation of the System

There is a dire need to plan and foster savvy and smart fire identification and alert framework utilizing Fluffy Rationale which assumes a significant part in accomplishing wellbeing conditions and structures. To accomplish the previously mentioned objectives, a locally situated fire discovery and caution framework presents the Prudent and Reasonable Caution Framework utilizing Arduino Uno and simple to purchase sensor set. The proposed framework is powerful in distinguishing a fire within the sight of smoke and blazes at a level with the specific increment at room temperature. The circuit graph of the framework and the block chart of the

proposed framework is displayed in figure 2 and 3.

The proposed system uses a fuzzy logic-based system to identify the true fire incidents that may lead to a critical and dangerous situation and alerts the fire alarm. The proposed system is effective in detecting false fires and reporting only real fire incidents. Here, the false fire incidents are incidents that are not real fires, but a fire detection system gives an alarm for a real fire.

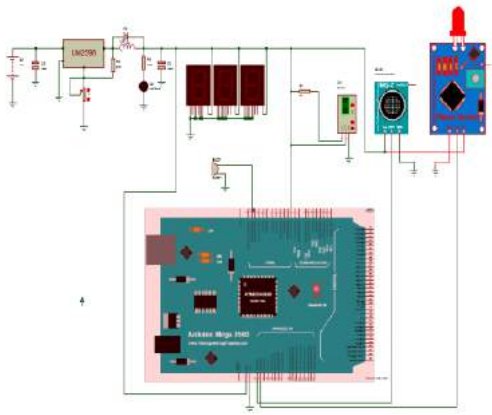


Figure 2: Circuit Diagram

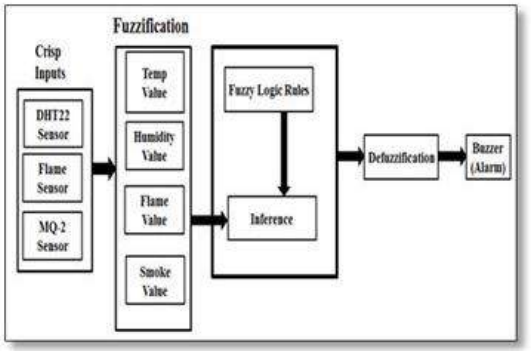


Figure 3: The Block Diagram of System

5.1. Membership Functions of the System

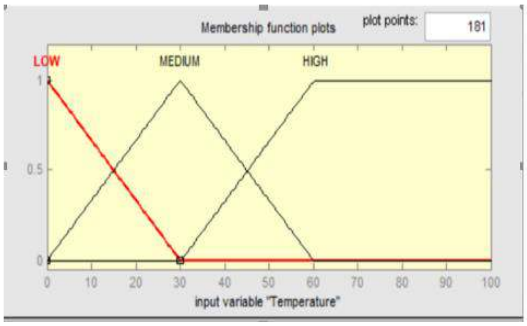


Figure 4: Memberships function for Temperature (˚C)

- $\mu(x) = \frac{30-x}{30}$, $0 \le x \le 30$ (for Low)
- $\mu(x) = \frac{x}{30}$, $0 \le x \le 30$ (for Medium)
- $\mu(x) = \frac{60-x}{30}$, $30 \le x \le 60$ (for Medium)
- $\mu(x) = \frac{x-30}{30}$, $30 \le x \le 60$ (for High)
- $\mu(x) = 1$, $x \ge 60$ (for High)

Example: Input x = **38.70** ℃

$$\mu(x) = \frac{0.7}{M} \qquad \mu(x) = \frac{0.3}{H}$$

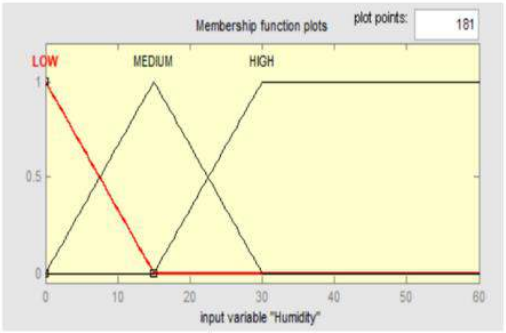


Figure 5: Memberships function for Humidity (%)

- $\mu(x) = \frac{15-x}{15}$, $0 \le x \le 15$ (for Low)
- $\mu(x) = \frac{x}{15}$, $0 \le x \le 15$ (for Medium)
- $\mu(x) = \frac{30-x}{15}$, $15 \le x \le 30$ (for Medium)
- $\mu(x) = \frac{x-15}{15}$, $15 \le x \le 30$ (for High)
- $\mu(x) = 1$, $x \ge 30$ (for High)

Example: Input x = 22.30%

$$\mu(x) = \frac{0.5}{M} \qquad \mu(x) = \frac{0.5}{H}$$

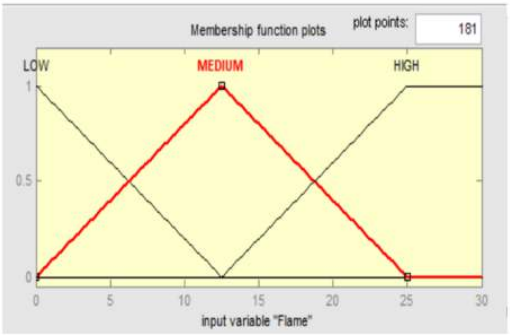


Figure 6: Memberships function for Smoke (ppm)

$$\mu_{trap}(x;a,b,c,d) = \begin{cases} 0, & x \le a \\ \frac{x-a}{b-a}, & a \le x \le b \\ 1, & b \le x \le c \\ \frac{d-x}{d-c}, & c \le x \le d \\ 0, & d \le x \end{cases}$$

$$= max\ (min\ (\frac{x-a}{b-a}, \frac{d-x}{d-c}), 0)$$

input x = 435.00 ppm

$$\mu(\boldsymbol{smoke}) = \{\frac{1}{\mathbf{H}}\}$$

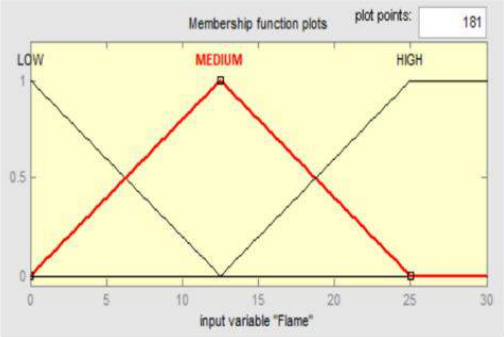


Figure 7: Memberships function for Flame (nm)

- $\mu(x) = \frac{12.5-x}{12.5}$, $0 \le x \le 12.5$ (for Low)
- $\mu(x) = \frac{x}{12.5}$, $0 \le x \le 12.5$ (for Medium)
- $\mu(x) = \frac{25-x}{12.5}$, $0 \le x \le 25$ (for Medium)
- $\mu(x) = \frac{x-12.5}{12.5}$, $12.5 \le x \le 25$ (for High)
- $\mu(x) = 1$, $x \ge 25$ (for High)

Example: Input x = 1021 nm

$$\mu(x) = \frac{1}{H}$$

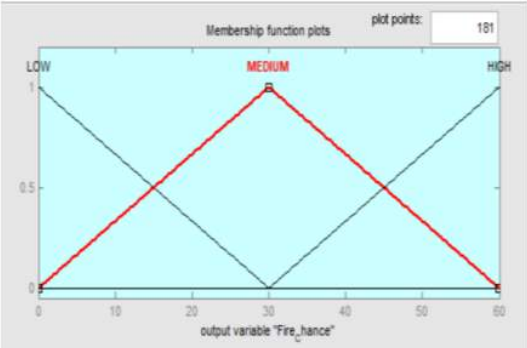


Figure 8: Memberships function for Fire Chance

5.2 Defuzzification

Defuzzification is a process that converts a fuzzified output into crisp value to fuzzy set. There are many well-known defuzzification method, Weight Average Method have been used to perform the defuzzification in this system. It provides less computationally intensive.

Weight Average Method is

$$x^* = \frac{\sum \mu_c(\bar{x}).\bar{x}}{\sum \mu_c(\bar{x})}$$

$$\boldsymbol{Flexibility} = \frac{(0.3 \times 30)+(0.7 \times 60)}{0.3+0.7} \approx 50$$

where,

$\sum$ = algebraic summation

x = element with maximum membership function

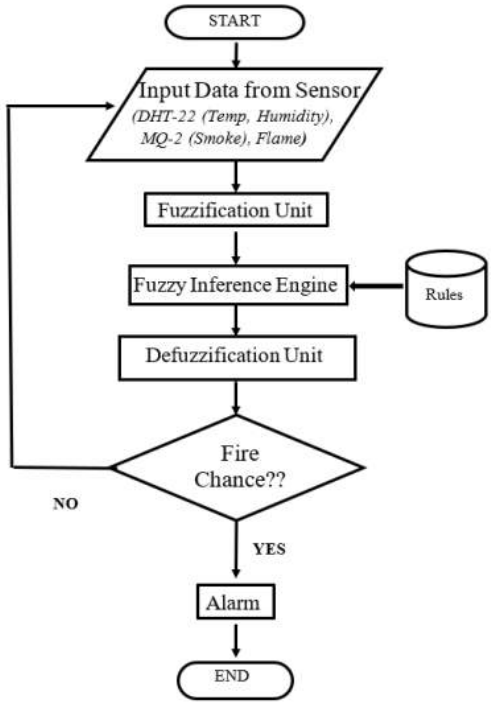


Figure 9: Flow chart of the system

5.3. Components List and Explanation

Figure 10: Flame Sensor

Flame Sensor: This sensor can recognize a fire by detecting light frequency between 760 - 1100 nanometers. The test distance relies upon the fire size and responsiveness settings. The identification point is 60 degrees, so the fire doesn't need to be directly before the sensor.

Figure 11: DHT22

DHT22: The DHT22 was utilized as one of the sensors that gives two estimations, like temperature and dampness, and these two qualities assume a significant part in the execution of fire observing and alert framework. DHT22 estimates temperature between −40◦C to 80◦C and 0 to 100 percent dampness. The explanation of involving DHT22 in this approach is a result of estimating two data sources utilizing single sensor. DHT22 is joined with microcontroller with three pins, which are Vcc, Information and Ground pins.

Table 1: DHT22 Properties Table

Model	AM2302
Power supply	3.3-6 V DC
Output Signal	Digital Signal via 1-wire bus
Sensing element	Polymer capacitor
Operating range	Humidity 0-100%RH, temperature –40~80 Celsius
Accuracy	Humidity ±2% RH (Max ±5% RH), Temperature < ±0.5 Celsius
Resolution of sensitivity	Humidity 0.1% RH; Temperature 0.1 Celsius
Repeatability	Humidity ±1% RH; Temperature ±0.2 Celsius
Humidity hysteresis	±0.3% RH
Long-term stability	±0.5% RH/year
Sensing period	Average: 2 s
Interchange ability	Fully interchangeable
Dimensions	Small size 14,185.5 mm, Big size 22,285 mm

Figure 12: MQ-2

The MQ-2 smoke sensor is delicate to smoke and gases (e.g CO2). It can distinguish gases in the convergence of reach 200 to 10000ppm. The smoke sensor has an implicit potentiometer that permits you to change the sensor responsiveness as indicated by how exact you need to distinguish gas.

Figure 13: Buzzer

Suitable for button prompt, prompt alarm and so on.

- Size: 26 * 21mm
- Voltage: 5V
- Port: Digital level
- Platform: Arduino, microcontroller

Arduino Mega Board (1pcs): Arduino Mega relies upon the Atmega 2560 microcontroller. It incorporates computerized input/yield pins-54, where 16 pins are simple sources of info, 14 are utilized like PWM yields equipment sequential ports, an ICSP header, a power jack, a USB association, as well as a RST button. The power supply of this board should be possible by interfacing it to a PC utilizing a USB link, or battery or an air conditioner DC connector. The working voltage of this microcontroller is 5volts, however the info Voltage will go from 7volts to 2volts.

Figure 14: Arduino Mega Board (2560)

Input Voltage	: DC 7-12V
Operating Voltage	: DC 5V
Digital I/O Pins	: 54 (D0 - D53)
PWM Digital I/O Pins	: 15 (D2-D13; D44 -D46)
Analog Input Pins	: 16 (A0 - A15)
DC Current per I/O Pin	: 20mA
DC Current for 3.3V Pin	: 50mA
Flash Memory	: 256KB
SRAM	: 8KB
EEPROM	: 4KB
Clock Speed	: 16MHz
Dimensions	: 108*53. 5mm, H-15mm

6. Test Results

The data are gathered using different sensors as shown in table 2.

Table 2: Results of Test

No	Temp (°C)	Humidity (%)	Smoke (ppm)	Flame (nm)	Fire Chance	TestCase (Low = 0, Med = 1, High =2)
1	52	28.70%	199	1021	50	2
2	54.7	25.60%	435	1021	50	2
3	43.9	25.50%	456	1021	45	2
4	38.7	22.30%	366	1021	50	2
5	60	19.80%	329	1021	49	2
6	60	20.20%	345	1021	49	2
7	60	16.80%	340	25	50	2
8	56	15.30%	480	25	52	2
9	58	22.30%	140	1008	51	2
10	30.1	74.70%	125	1016	17	0
11	30.1	73.80%	172	1009	19	1
12	29.6	73.80%	210	1008	20	0
13	36.3	67.40%	159	1013	36	1
14	55.7	18.60%	335	1021	46	2
15	49.9	30.30%	155	1009	34	2
16	29.6	73.80%	210	1008	35	2

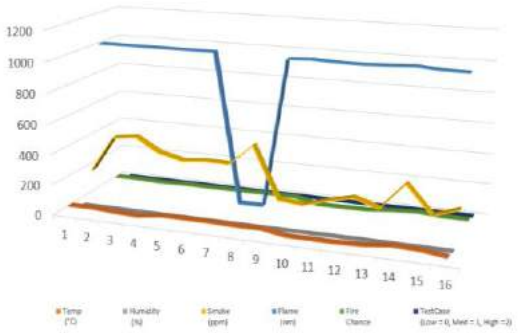


Figure 15: Testing results of the system

7. Performance Evaluation

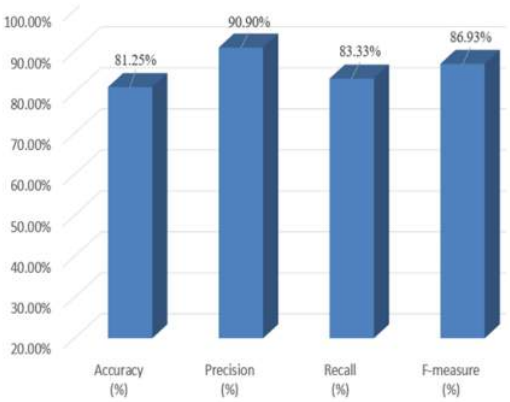


Figure 16: Evaluation of the system

This paper emphasizes on the Fire Detection and Alarm system because fire chance levels is becoming one of the most important tasks. The following figure 16 shows the performance analysis of the system with different data size. Results are evaluated by four evaluation methods (accuracy, precision, recall and F-measure) called confusion matrix. F-measure is the mean of recall and precision, that is F = 2PR/ (P+R).

8. Discussion and Conclusion

This system is to identify fire at beginning phase and diminish deception rate by involving various sensors to save human lives and properties. Ever, there was huge number of fire fiascos because of need access of fire system. The identification of fire become a difficult issue for human security and because of this reason a fire regulator system was planned and carried out in light of fluffy rationale in this exploration work.

This fluffy rationale-based Fire Recognition and Alert System is introduced to save lives and property harms. This system utilized fluffy rationale which upheld execution prerequisites without any problem. Numerous sensors are utilized to come by the exact outcomes to decrease misleading problem rate. Sensors perusing information are utilized as info, for example, the change pace of temperature, the change pace of moistness, and presence of fire and smoke. The system will alarm individuals on the off chance that any undesirable circumstance happens anyplace.

In this exploration, fire, temperature and dampness sensors was utilized however in future work more sensors can be used for example smoke, light sensor for fire regulator system to obtain more precise outcomes. The system was intended for identifying fire at explicit spot, in future by utilizing numerous system coordinated with one another can be planned. A work area and portable based application can be created to screen fire online when you are nowhere near home. The proposed work can be taken on in future work by applying control components to framework which are depicted in this examination for example Water Shower.

References

[1] W. H. Dong, L. Wang, G. Z. Yu, and Z. Bin Mei, “Design of Wireless Automatic Fire Alarm System,” *Procedia Eng.*, vol. 135, pp. 413–417, 2016.

[2] A. Imteaj, T. Rahman, M. K. Hossain, M. S. Alam, and S. A. Rahat, “An IoT based Fire Alarming and

Authentication System for Workhouse using Raspberry Pi 3," *ECCE 2017 - Int. Conf. Electr. Comput. Commun. Eng.*, no. February 2010, pp. 899–904, 2017.

[3] A. Mahgoub, N. Tarrad, R. Elsherif, A. Al-Ali, and L. Ismail, "IoT-Based Fire Alarm System," in *2019 Third World Conference on Smart Trends in Systems Security and Sustainablity (WorldS4)*, 2019, pp. 162–166.

[4] N. N. Mahzan, N. I. M. Enzai, N. M. Zin, and K. S. S. K. M. Noh, "Design of an Arduino-based home fire alarm system with GSM module," *J. Phys. Conf. Ser.*, vol. 1019, no. 1, 2018.

[5] W. L. Hsu, J. Y. Jhuang, C. S. Huang, C. K. Liang, and Y. C. Shiau, "Application of Internet of Things in a kitchen fire prevention system," *Appl. Sci.*, vol. 9, no. 17, 2019.

Indoor Air Pollution Detection and Monitoring System Using Fuzzy Logic

Kyaw Win Thu, Htar Htar Lwin
University of Computer Studies, Yangon
kyawwinthu@ucsy.edu.mm, htarhtarlwin@ucsy.edu.mm

Abstract

The most important ecological problems is air pollution that effect on human health and economic balance. It can cause the environmental effects that are acid rain, smog and global warming. Therefore, monitoring of air pollution is important of air pollutants condition. To solve this problem, we implemented Arduino-Based Indoor Air Pollution Detection and Monitoring System using Fuzzy Logic to do monitoring and detection of air quality for carbon dioxide (CO_2), carbon monoxide (CO) and Liquefied Petroleum Gas (LPG) gases. The proposed system can maintain gases for safety level based on the results using fuzzy logic rules. The proposed system uses three gas sensors as an input and it detects concentration of gas and displays air pollution status on LEDs based on air quality index (AQI). Fuzzy logic system-controlled ventilation fan can maintain the safety level of indoor air pollution.

Keywords: Indoor Air Pollution Detection, CO_2, CO, LPG, Fuzzy Logic System

1. Introduction

Air pollution is the biggest environmental issue of the world. World Health Organization (WHO) surveys that 3 million people annually get the health problem because of air pollution [1]. Over the ten years ago, air pollution become a life-dangerous factor because of human activities, industries and urbanization. Air pollution can cause both indoors and outdoors, People used about 90% of their activities indoor. Carbon dioxide can occur in indoor environment because of breathing and human activities inside the room. For Carbon monoxide is the most dangerous gas for indoor room and it causes the serious health problem and sometime people can die [5]. CO, LPG and other gases in the atmosphere can also be harmful to human health. CO 9 ppm is the maximum indoor safe level over 8 hours (ASHRA, WHO) [3]. The current OSHA (Occupational Safety and Health Administration) standard for LPG is 1000 ppm of air averaged over an eight-hour work shift [7].

The proposed system uses Fuzzy logic on an Arduino Uno microcontroller board and uses PM Gas sensors to monitor and detect gases. It also includes an alarm system with a buzzer and LEDs to detect air pollution conditions.

The proposed system can detect and monitor the concentration of CO_2, CO, and LPG. It makes display Good, Unhealthy and Harmful to the user when the concentration level of gases changes with LEDs.

Good = Green
Unhealthy = Yellow
Harmful = Red

It makes power usage of ventilation fan efficient when the concentration gases changes can maintain safety level of indoor air.

2. Related Work

There are many studies and research air quality of air pollution detection and monitoring. Mukesh and Sakula proposed indoor air quality monitoring using COZIR-A sensor, MQ7 sensor, and raspberry pi 3. This system detects CO_2 and CO indoor, the output of system is DC fan, it only has ON/OFF state. This system monitors the indoor air, but the fan has to be operated manually from the smart phone [4].

Fadli Pradityo and Nico Surantha proposed indoor air quality monitoring system with fuzzy logic control based on thing of internet (IOT). This system detects CO_2 and particulate matter (PM10) then control automatically ventilation fan but it has two inputs and output is interval of the

fan. This system used MQ 135 gas sensor, Sharp GP2Y1010AUOF sensor, Arduino UNO and raspberry pi 3. This system fuzzy logic operated in raspberry pi and the sensor's data import using raspberry pi 3 to cloud server. Arduino UNO convert the analog sensor value into digital value as an interface [3].

D.M.G. Preethicandra's proposed system for measuring the density of PM10 and CO2 concentrations in indoor air The Raspberry Pi 3 pushes sensor data to a cloud server and displays the data on a dashboard. This technology uses a ventilation fan to lower the concentration of those pollutants inside the room and bring it to a safe level for people [5].

3. Background Theory

Fuzzy is a synonym for uncertain, ambiguous, vague, or unclear. A computational method based on the degree of truth is fuzzy logic. A fuzzy logic system generates a specific output by using linguistic variables and the degree of truth of the input. The nature of the output depends on the condition of this input.

Unlike Boolean logic, which only employs two categories, this method uses multiple categories (true or false). To present the objects in the logic of Boolean, the digits 0 and 1 are utilized. As an example, the water temperature in the glass can be High and Low, and High represents 1 and Low represents 0. The water can be representing more types in fuzzy logic; however, they fell only into such two digits (0 or 1) representation categories. In such case, the condition of the water can be very cold, very warm, or warm. The answered value in Boolean Logic will only yes or no. The answered value in fuzzy logic will be in these two categories. On the other hand, some of the possible answers may be possibly yes, possibly no, or certainly no.

Instead of using precise category in previous mentioned examples, the fuzzy logic system utilizes the degrees of possibilities. To produce an explicit output, these can be used.

It addresses the issue of uncertainty in the engineering field. When accurate reasoning is unavailable, it gives an accurate level of reasoning. The structure of Fuzzy logic is simple and easy to understand. It supports to solve various industrial issues (especially for making decisions). It needs few data to execute.

4. Hardware Components used in Proposed System

Indoor air pollution monitoring sensors have various types of sensors such as satellite-based sensors and low-cost commercially sensors that are used in bicycle, cars, trucks and drones. The proposed system uses three input sensors for detection and monitoring of air pollution. These sensors are CO_2 sensor, CO sensor and LPG sensor.

In the proposed system, the MQ-135 gas sensor is used to detect the carbon dioxide gas in the room. The MQ-135 gas sensor is used to monitor harmful gases such as ammonia (NH3), sulfur (S), Benzene (C6H6), and CO_2.

As shown in figure 1, this sensor module has digital and analog output pins. The digital pin reaches a high level when the level of gas reaches over the threshold of air. This threshold value can be set by using the on-board potentiometer. The analog output pin outputs an analog voltage which can be used to approximate the level of these gases in the atmosphere.

The MQ135 sensor module requires 5V and 150 mA in order to operate. Accurate sensor values can be obtained by preheating the module before use. All MQ sensors have to be powered up for a pre-heat duration for the sensor to warm up before it can start working. When the module powers on, the power LED will turn on. Leave the module in this state till the pre-heat duration is completed.

Figure 1. MQ135 sensor Module

The CO gas in the space is found using the MQ7 module sensor as shown in figure 2. The MQ-7 can detect CO2-gas concentrations. This sensor responds quickly and with excellent sensitivity. The output of the sensor is an analog resistance.

Sensitive material of MQ-7 gas sensor is SnO_2, which with lower conductivity in clean

air. It makes detection by method of cycle high and low temperature, and detect CO when low temperature (heated by 1.5V). The sensor's conductivity is higher along with the gas concentration rising.

Figure 2. MQ7 sensor Module

The LPG gas can measure by the MQ2 module sensor as shown in figure 3 and it is useful device for indoor room. It can detect many other gases such as H_2, LPG, CH_4, CO, Alcohol, Smoke or Propane. The sensor values can take quickly because this sensor response time is very fast and it is high sensitivity. This device requires a protection resistor and an adjustable resistor on board. The MQ-2 gas sensor is sensitive to LPG, i-butane, propane, methane, alcohol, Hydrogen and smoke.

In home and office building, the MQ2 sensor module could be used as a gas leakage detecting equipment. The resistance of the sensitive component changes as the concentration of the target gas changes.

Figure 3. MQ2 sensor Module

The proposed system uses the Arduino Uno board for implementation of detection and monitoring of indoor air pollution. Arduino UNO is a microcontroller board based on the ATmega328P. It has 14 digital input/output pins (of which 6 can be used as PWM outputs), 6 analog inputs, a 16 MHz ceramic resonator, a USB connection, a power jack, an ICSP header and a reset button.

Arduino UNO is a not expensive, portable and it has many open-source codes for microcontroller programming. It can be integrated into many electronic projects. This board can be interfaced with other Arduino boards, Arduino shields, and Raspberry Pi boards and can control relays, LEDs, servos, and motors as an output. Arduino Uno board is as shown in figure 4.

Figure 4. Arduino Uno Board

Indoor air pollution monitoring and detection system uses the three LED for presenting the type of air quality. The red LED is used to present that air quality in the room is hazard condition. The yellow LED is used to present that air quality in the room is unhealthy condition. The green LED is used to present the air quality in the room is good condition. Figure 5 shows LEDs.

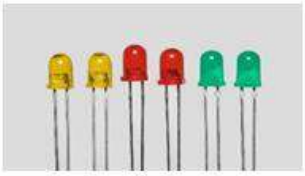


Figure 5. Light Emitting Diode (LED)

The proposed indoor air pollution monitoring and detection system used buzzer in order to give the alarm of air pollution when the air quality is unhealthy and hazard. The buzzer is a sounding device that can convert audio signals into sound signals. It is usually powered by DC voltage. Figure 6 shows Buzzer.

Figure 6. Buzzer

Figure 7 is the relay module that is a convenient board. The motor, valves, lamps DC fans require the relay module as a switch that can control the high voltage and current load into the

fixed value as the voltage of a battery. Typical voltage values of DC fans are 5V, 12V, 24V and 48V. In the proposed system, the 12V fun is used to absorb the bas air in the room. DC fan is shown in figure 8.

Figure 7. The Single Channel Relay Module

Figure 8. 12V DC Fun

5. Indoor Air Pollution Monitoring and Detecting System

The proposed system uses fuzzy logic system for an indoor air pollution monitoring and detection. The fuzzy logic technique can be implemented in various systems (hardware and software). The proposed system monitors the air quality in the room for CO_2, CO and LPG. The System can maintain CO_2, CO and LPG for safety level based on the results using fuzzy rules.

The hardware conceptual overview of the system is described in figure 9. The main purpose of the proposed system is developing the fuzzy rules to detect and monitor the interval of air pollution level and to maintain the air quality of room. In this system, there are five parts: Input data, Fuzzification of system, Rule inference, defuzzification and output of air quality.

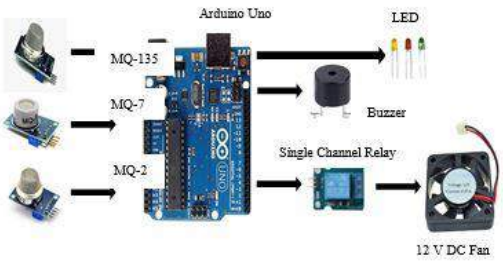


Figure 9. System Design of Proposed System

The flowchart of the proposed system is shown in figure 10. After power on, input crisp values from sensors are fuzzified using input membership functions. These fuzzy input values and Rules database are used in inference process. Out of the inference engine are converted into crisp values with Weight Average defuzzification methods. The output of this process is AQI value. If the value is less than or equal to 15, turn off ventilation fan and turn on green color LED. If the value is greater than 15 and less than 25, turn on ventilation fan, yellow LED and make alarm with short beep. If the value is greater than or equal to 25, turn on ventilation fan, Red LED and make alarm with continuous beep.

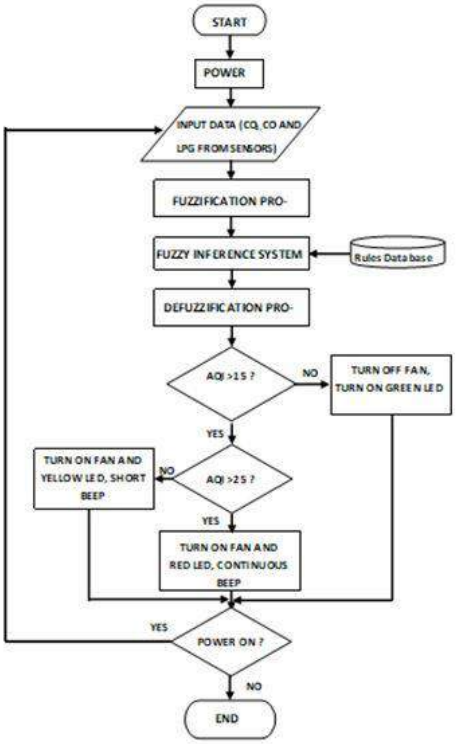


Figure 10. Flowchart of Proposed System

5.1 Fuzzification

The system first defines the linguistic terms variables of input sensors. The input data crisp value converted into fuzzy values using trapezoidal membership function and triangular membership function based on nature of input values. Three input sensors can be categorized in the following. The CO_2 sensor input is categorized into three main categories: Normal, Concentrate and Dangerous. The CO sensor input is categorized into two main categories: Safe and Dangerous. The LPG sensor input is categorized

into three main categories: Normal, Moderate and Hazard. The output is categorized into three categories: Good, Unhealthy and Harmful. Membership functions of each input and output are shown in figure 11 (a), (b), (c) and (d).

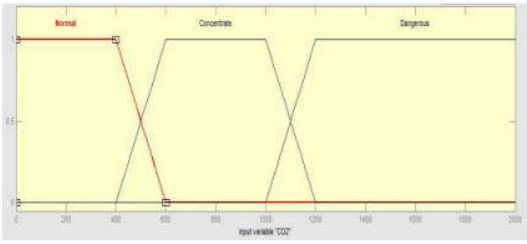


Figure 11(a). Membership Function of CO_2

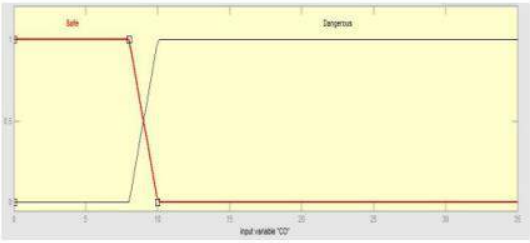


Figure 11(b). Membership Function of CO

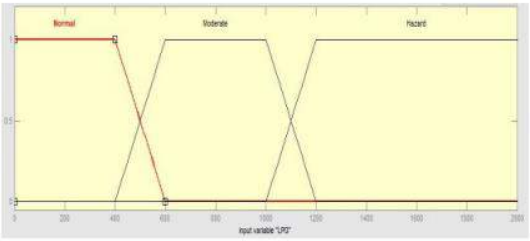


Figure 11(c). Membership Function of LPG

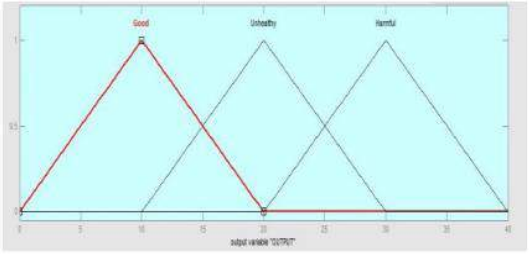


Figure 10(d). Membership of OUTPUT

5.2 Fuzzy Inference Engine

In the proposed system eighteen ruses are constructs the IF-THEN logic is applied to set effective rules. The construct rules are in the following:

1. If (CO_2 is Low) and (CO is Safe) and (LPG is Normal) then (output is Good)
2. If (CO_2 is Low) and (CO is Safe) and (LPG is Moderate) then (output is Unhealthy)
3. If (CO_2 is Low) and (CO is Safe) and (LPG is Hazard) then (output is Unhealthy)
4. If (CO_2 is Low) and (CO is Dangerous) and (LPG is Normal) then (output is Harmful)
5. If (CO_2 is Low) and (CO is Dangerous) and (LPG is Moderate) then (output is Harmful)
6. If (CO_2 is Low) and (CO is Dangerous) and (LPG is Hazard) then (output is Harmful)
7. If (CO_2 is Concentrate) and (CO is Safe) and (LPG is Normal) then (output is Unhealthy)
8. If (CO_2 is Concentrate) and (CO is Safe) and (LPG is Moderate) then (output is Unhealthy)
9. If (CO_2 is Concentrate) and (CO is Safe) and (LPG is Hazard) then (output is Harmful)
10. If (CO_2 is Concentrate) and (CO is Dangerous) and (LPG is Normal) then (output is Harmful)
11. If (CO_2 is Concentrate) and (CO is Dangerous) and (LPG is Moderate) then (output is Harmful)
12. If (CO_2 is Concentrate) and (CO is Dangerous) and (LPG is Hazard) then (output is Harmful)
13. If (CO_2 is Dangerous) and (CO is Safe) and (LPG is Normal) then (output is Unhealthy)
14. If (CO_2 is Dangerous) and (CO is Safe) and (LPG is Moderate) then (output is Unhealthy)
15. If (CO_2 is Dangerous) and (CO is Safe) and (LPG is Hazard) then (output is Harmful)
16. If (CO_2 is Dangerous) and (CO is Dangerous) and (LPG is Normal) then (output is Harmful)
17. If (CO_2 is Dangerous) and (CO is Dangerous) and (LPG is Moderate) then (output is Harmful)
18. If (CO_2 is Dangerous) and (CO is Dangerous) and (LPG is Hazard) then (output is Harmful)

5.3 Defuzzification

Defuzzification is the last algorithm step. It is the process of producing a quantifiable result in fuzzy logic, given fuzzy sets and corresponding membership degrees. It is typically needed in fuzzy control systems. These will have a number of rules that transform a number of variables into a fuzzy result, that is, the result is described in terms of membership in fuzzy sets.

Defuzzification is interpreting the membership degree of fuzzy sets into a specific decision or real values. In this step, the

defuzzifier uses the membership function to establish the output of air quality in the room. After the Fuzzification step, to forecast the production, the weight-average method is applied in defuzzification.

Weight Average Method is

$$x^* = \frac{\sum \mu_c(\bar{x}).\bar{x}}{\sum \mu_c(\bar{x})} \quad \text{Eq. 1}$$

Here $\sum$ denotes the algebraic summation and x is the element with maximum membership function.

6. Experiment Results

The main purpose of the proposed system is to keep indoor air pollution in safe level based on ASHRAE standard. If AQI level in the room exceed than safe level, system will turn on exhaust fan and if AQI level back to normal, exhaust fan will stop. The output levels are shown in table 1.

Table 1. Indoor Air Pollution Categories

AQI Levels	Indoor Air Pollution Categories
AQI <= 15	Good
15 < AQI < 25	Unhealthy
AQI >= 25	Harmful

If the output is less than or equal to 15, it is Good condition and the system will turn off exhaust fan and display green LED. If the output is greater than 15, the fan will be turn on and display red LED. Otherwise, yellow LED will be displayed. If it is in normal condition, the fan will be turn off. The test result is shown in table 2.

Table 2. Test Result

No.	CO_2 (ppm)	CO (ppm)	LPG (ppm)	Output of device	Calculation of Weight Average Method
1	978	10	1195	29	30
2	411	3	415	11	10.75
3	1405	3	441	22	22.05
4	451	9	441	16	17.55
5	502	13	446	23	23.19
6	696	5	1450	30	30
7	402	5	1346	21	20.1
8	399	5	1349	20	20
9	596	7	848	20	20
10	549	7	98	16	17.45
11	802	9	847	25	25
12	703	9	176	20	20
13	444	3	399	12	12.2
14	401	3	637	20	20
15	1233	10	996	30	30

7. Performance Evaluation

This paper emphasizes on the detecting and monitoring the indoor air pollution because air pollutant levels is becoming one of the most important tasks. In order to measure the performance of the system, we test with 15 values for each sensor. The performance of this system is evaluated in terms of precision, recall and F-measure. Precision (P) means the percentage of the correct detecting air quality suggested by the system which is divided by total number of errors detected by the system. Recall (R) means the percentage of correct detecting suggested by the system which is divided by total number of testing values. F-score is the mean of recall and precision, that is F = 2PR/ (P+R). The indoor air pollution levels are divided into three classes: (1) Good Condition, (2) Unhealthy Condition (3) Harmful Condition. Figure 12 shows how the system correctly detected the air quality of room.

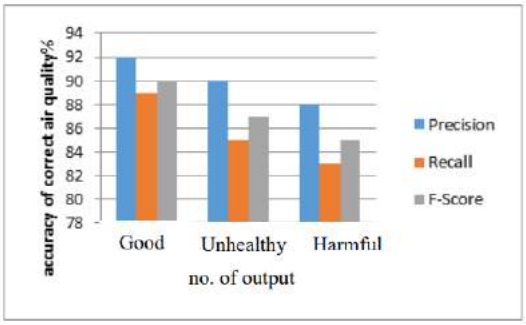


Figure 12. Precision, Recall and F-score on test sentences

8. Conclusion

The regulation of air pollutant levels is becoming one of the most important tasks. People need to know that it is important to solve the indoor air pollution around them. The indoor air pollution detection and monitoring system is implemented for a single room which can detect three types of gases. Fuzzy Logic control system is applied for monitoring the gas level using 18 rules. Embedded system and fuzzy logic rules can be efficiently used to detect the quality of air and the proposed system will help to detect and alarm pollution levels. It can be extended to monitor the other dangerous gases which are harmful the citizens.

References

[1] "Air pollution", World Health Organization, [Online]. Available: who.int [Accessed 17 April 2018]
[2] ASHRAE/ANSI Standard 62.1- 2013 Ventilation for Acceptable Indoor Air Quality [Online]. Available: http://www.myiaire.com/product-docs /ultraDRY/ASHRAE62.1.pdf
[3] Mukesh, D. M., & Akula, S. K. Automated Indoor Air Quality Monitor and Control. In: International Journal of Computer Applications, 159 (6) (2017) 0975 - 8887.
[4] Preethicandra, D. M. G. Design of a smart indoor air quality monitoring wireless sensor network for assisted living. In: IEEE International Instrumentation and Measurement Technology Conference (I2MTC), 2013.
[5] https://gaslab.com/blogs/articles/carbon-monoxide - levels-chart
[6] https://www.cdc.gov/niosh/docs/81-123/pdfs/0372 .pdf

Data Mining, Web Mining and Machine Learning

Weather Forecasting System Using Gaussian Naïve Bayes

May Theingi Kyaw, Ah Nge Htwe
University of Computer Studies, Yangon
maytheingikyaw@ucsy.edu.mm, anhtwe@ucsy.edu.mm

Abstract

Forecasting is the course of assessment in obscure circumstances from the authentic information. Weather conditions determining is one of the utilizations of information mining innovation to foresee the condition of environment for a future time frame and a given area as respects heat, darkness, dryness, wind, downpour and so on. It is imperative to distinguish the weighty precipitation and to give the data of admonitions in regards to the normal catastrophes. The proposed system intends to help the weather pattern for the future hourly estimating in light of authentic climate information for areas (Yangon) and assist for use in stations of weather. The system can be utilized in numerous ways: condition progressing development projects, transportation exercises, agrarian undertakings, flight tasks, flood circumstance and so on. The proposed system uses Gaussian Naïve Bayes for classification purposes in PHP programming language and MySQL database server to save the historical weather data. The dataset 52608 samples of Yangon from 2015 January to 2020 December have been gathered from OpenWeather API. Relative humidity, wind speed, temperature, clouds, pressure, visibility and dew_point is used as the required weather parameters to determine the class labels such as clouds, thunderstorm, rain, drizzle, clear, mist, haze, smoke and fog. Several capabilities of Gaussian Naïve Bayes have achieved good enough accuracy due to given above 72% certain results.

Keywords: Gaussian Naïve Bayes, weather forecasting, datamining

1. Introduction

Since ages, weather prediction remained the most thrilling and testing task for the specialists and meteorologists. Estimating climate unequivocally and precisely assists humankind with being ready for any regular catastrophes before time. Today, foreseeing weather conditions needs aptitude in different fields and furthermore includes processing complex numerical computations. The fast development in innovation has empowered researchers to precisely anticipate climate more. Precise forecast of climate boundaries is a troublesome errand because of the powerful idea of environment.

Weather forecasting is the most common way of gathering information on air conditions, which records the temperature, dampness, precipitation, wind speed and its heading utilizing fast PCs, wired and remote sensors, satellites and climate radars [12]. With the headway in data innovation complex numerical and factual models has empowered us to get better weather conditions estimating. Various weather conditions gauging models are proposed by scientists to further develop exactness of the models [2] [7] [10]. Various weather research and forecasting models (WRF), numerical weather prediction models (NWP), and automatic weather stations (AWS) are created and carried out to foresee climate all the more precisely [3]. Gigantic measure of climate informational index accessible in metrological focuses contains enormous volume of climate information which is utilized for the climate expectation. From the beyond couple of many years, information mining strategies is generally broadly utilized for climate expectation and have shown a wonderful degree of precision and pertinence in expectation.

Data mining is the investigation of how to decide hidden designs in the information to assist with pursuing ideal choices on PCs when the data set included is voluminous, difficult to portray precisely, and continually evolving. It sends strategies in light of AI, close by the ordinary techniques. These strategies can deliver choice or forecast models, in view of the huge volumes of real verifiable information. Along these lines, they address genuine proof based choice help [6]. Different order AI calculations are executed to anticipate the atmospheric conditions. Decision

Trees, Artificial Neural Networks (ANN), Naive Bayes Networks, Support Vector Machines (SVM), Fuzzy Logic, Genetic Algorithms are some of the commonly used data mining techniques that are predominantly implemented for weather prediction. In this system, the weather values are talked about temperature, humidity, wind speed, cloud, pressure, visibility and dew_point using Gaussian Naïve Bayes algorithm with acceptable results in estimation of weather.

2. Related Work

R. Siddhant and D. Shaba depicted "Weather conditions Forecast Using Data Mining". Expectation model is utilized for precipitation forecast. Among of the 15 elements, 5 highlights are applied for estimating. Climate information gathered from NOAA (National Oceanic Atmospheric Administration). KNN and Gaussian Naïve Bayes are utilized to gauge the precipitation. Mathematical traits are utilized and result for downpour or not. Gaussian Naïve Bayes are more precise than KNN [6].

Nikam, B. Valmik and B. B. Meshram introduced "Displaying Rainfall Prediction Using Data Mining Method - A Bayesian Approach". Information is acquired from Indian Meteorological Department. Among of the 36 qualities, just 7 ascribes which are generally applicable to precipitation forecast are thought of, and to decide downpour or not. Guileless Bayesian is utilized for precipitation expectation. Mathematical quality is utilized. This framework works with effective precision [14].

In [11], Santhanam et al, proposed Neural organization-based model for weather conditions determining. Analysts and researchers are involving Neural Network for climate expectation because of his effortlessness, strength and adequacy. In this work the creators have carried out and analyzed the exhibitions of engendering brain organization (BPN) and revolutionary premise worked brain organization (RBF). For this exploration work, the total climate information of a decade is gathered from meteorological division, Kanyakumari, Tamil Nadu, India. The outcomes showed that extreme premise worked brain organization (RBF) have better exactness, quicker and more dependable for climate expectation. The prescient precision of RBF was 88.49% which makes it more helpful for quick continuous weather conditions gauging.

In [4], KavithaRani et al, proposed a clever precipitation expectation model utilizing half breed classifier executing fake honey bee state calculation is coordinated effort with the hereditary calculation for preparing the feed forward brain organization. For the work the creators gathered the genuine climate informational index from Rayalaseema, Aandhra and Telangana areas of India. The precipitation expectation is finished utilizing the crossover classifier. From the exploratory outcomes it was found the proposed crossover classifier has preferable execution and prescient capacity over Artificial Bee Colony with Neural Network.

In [13], Sharma et al, fostered a precipitation forecast model in view of Bayesian organization. For this examination, the creators gathered the month to month climate information of a long time from 1981 to 2000 of 21 stations in Assam, India. K2 calculation is carried out on the climate informational index and contingent likelihood is tracked down utilizing most extreme probability approximations. Five distinct climatic boundaries viz. Temperature, Cloud cover, Relative moistness, Wind speed and Southern Oscillation Index (SOI) are utilized. Trial results showed that temperature is generally proficient and wind speed least one. Southern Oscillation Index is additionally tracked down significant in working on the outcomes. Some station got effectiveness above 95% though other station came by palatable outcomes.

3. Background Theory .

This system implemented classification technique which is a machine learning data mining technique in which the data from the indicators or autonomous factors is utilized to sort the information tests into nine classes. In light of different air factors, have carried out. Gaussian Naïve Bayes approach to build weather prediction model which will predict with appreciable accuracy.

3.1 Machine Learning for Weather Forecasting

From the past few periods, machine learning techniques are most widely used for weather estimation and have shown a remarkable level of

certain and applicability in prediction. Data mining is the learning of how to find out underlying patterns in the data to help make optimal determination on computers when the database involved is voluminous, hard to characterize certainly, and constantly changing. It deploys techniques based on machine learning, alongside the conventional methods. These techniques can produce decision or prediction models, based on the large volumes of actual historical data. Thus, they stand for true evidence-based decision support [6]. Forecasting weather exactly and certainly helps mankind to be prepared for any natural calamities before time. Today, predicting weather needs expertise in multiple fields and also regard as computing complex mathematical calculations.

Machine learning is a technique which creates a model from dataset training. When the weights are learned as accurately as possible, a model can estimate the certain output or value target given the test data record. The level of accurately arranged examples by the classifier (model) known as order precision gives us the presentation proportion of the classifier (model). This framework will construct a classifier (model) that will foresee the weather pattern for future occasion will considerable precision.

3.2 Gaussian Naïve Bayes Classifier

The Gaussian probability density function can be utilized to make expectations by subbing the boundaries with the new info worth of the variable and accordingly, the Gaussian capacity will give a gauge for the new information worth's likelihood.

The Naïve Bayes classifier expects that the worth of one element is autonomous of the worth of some other component. Innocent Bayes classifiers need preparing information to assess the boundaries expected for characterization. Because of straightforward plan and application, Naïve Bayes classifiers can be appropriate in some genuine situations.

The Gaussian Naïve Bayes classifier is a fast and straightforward classifier method that functions admirably without an excess of exertion and a decent degree of precision.

Gaussian Naïve Bayes is not difficult to fabricate and especially helpful for enormous datasets. It is particularly utilized for mathematical information (persistent and discretize information). Gaussian Naïve Bayes reckon as following Equation 1, 2, 3 and 4.

Mean formula

$$\mu = (\sum x_j)/N \quad (1)$$

where, $\sum x_j$ = Sum of All Data Points

N = Number of Data Points

μ = *Mean*

Variance formula

$$\sigma^2 = \sum(x_i - \mu)^2 \ /n-1 \quad (2)$$

where, σ^2 = Variance

x_i = Data Points

μ = *Mean*

n = Number of Data Points

Likelihood (3)

$$P(x|c) = (1 / (\text{sqrt}(2 * 3.14 * \sigma^2)) * \exp((-(x-\mu)^2)/(2*\sigma^2)))$$

P(x|c) is the likelihood which is the probability of predictor (attribute) given class

$$\text{Posterior} = \frac{prior * likelihood}{evidence} \quad (4)$$

3.3 Weather Dataset

Weather data is collected for this work from "https://docs.openweather.co.uk/history-bulk" website (OpenWeather API) to foster an expectation model that predicts weather patterns in view of verifiable climate information. The acquired climate information comprises of eight estimated which are temperature, dew point, humidity, sea level pressure, visibility, wind speed, cloud and events as shown in table 1.

Setting up the informational collection to suit an information mining task is the most tedious piece of the cycle. Seldom information are accessible in the structure expected by the information mining calculations [5]. This framework executed mean and change techniques in light of existing information. Adding complete dataset to the classifiers to be prepared on. The got climate informational index contains mathematical trait values, but model requires clear cut values.

Table 1: Description of Weather Attributes

Attribute	Type	Description
Temperature	Numerical	Temp is in ℃
Humidity	Numerical	Humidity in %
Wind Speed	Numerical	WindSpeed in meter/sec
Cloud	Numerical	Cloud in %
Pressure	Numerical	Pressure in hpa
Visibility	Numerical	Visibility in metres
Dew Point	Numerical	Dew Point in ℃
Events	String	Clear, Clouds, Rain, Drizzle, Thunderstorm, Mist, Fog, Haze and Smoke

4. Weather Forecasting System

This is the theoretical portrayal of the information and its connections in a given informational collection. Information mining models can be characterized into the accompanying classifications: grouping, relapse, affiliation examination, bunching, and exception or peculiarity recognition. Every class has a couple dozen unique calculations; each adopts a marginally unique strategy to take care of the issue. Arrangement and relapse undertakings are prescient strategies since they foresee a result variable in light of at least one information factors. Prescient calculations need a realized earlier informational index to "learn" the model. The level of accurately arranged examples by the classifier (model) known as order precision gives us the presentation proportion of the classifier (model). This framework will have constructed a classifier (model) that will have foreseen the weather pattern for future occasion will considerable precision.

This system will have used a Gaussian Naive-Bayes classification strategy to show the informational index and sum up the connection between the chose credits and the objective characteristic. This paper employs 6 years (2015-2020) hourly data from month January to December as training dataset that contains 9 attributes (temperature, humidity, wind speed, cloud, pressure, visibility and dew-point). And events are label of class Clear, Clouds, Drizzle, Rain, Thunderstorm, Mist, Fog, Haze and smoke. This framework will utilize the pre-arranged climate informational index to make the forecast model utilizing Gaussian Naive-Bayes calculation which will be then used to anticipate weather patterns as displayed in figure 1. This likewise ascertains the exactness of the created precipitation forecast model by utilizing different precision measures.

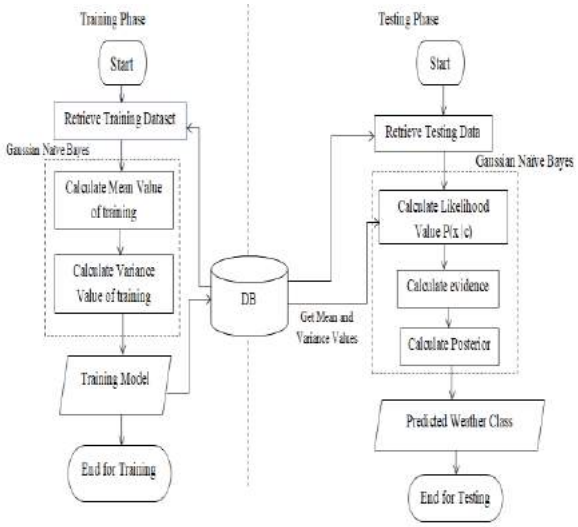


Figure 1 Overview of the Proposed System

5. Experiment Results

The informational index utilized in this paper is climate information of https://docs.open weather.co.uk/history-bulk.com from January 2015 to December 2020 which are applying the information mining process model. The trials are led to anticipate the atmospheric conditions utilizing Gaussian Naive Bayes calculation. In our gathered climate informational collection, nine class marks are anticipated variable which tells whether it will clouds, clear, etc., on a specific day or not. By applying Gaussian Naive Bayes calculation on the cleaned informational collection a model is created which is otherwise

called classifier. The level of accurately grouped cases by the classifier (model) known as order exactness gives us the presentation proportion of the classifier (model). To show system of performance, weather estimation was examined by using historical hourly data that consist of 52608 records. The training dataset are taken in from OpenWeather Organization by History Bulk API.

5.1. Performance Evaluation Measures of Model

Execution assessment of a model (classifier) is a necessary piece of model improvement process. Model assessment assists with tracking down the better model for our information and furthermore uncovers how well the picked model will act in future. To assess the exhibition of created model, various execution measure are utilized which depend on disarray network. Disarray Matrix: Confusion grid is the lattice perception of result of AI model. There are 2 types of confusion matrix such as binary class and multi-class matrix. This system is used the multi class confusion matrix which contains data about real and anticipated groupings done by a characterization calculation as displayed in table 1. Execution of such frameworks is regularly assessed involving the information in the grid.

Table 2 Multi-Class Confusion Matrix

		Predicted		
Actual		A	B	C
	A	TP_A	E_{AB}	E_{AC}
	B	E_{BA}	TP_B	E_{BC}
	C	E_{CA}	E_{CB}	TP_C

In the multi-class classification solution, as our estimation for unknown weather classes, we won't get TP, TN, FP, FN values directly as in the binary classification issues. We need to calculate these values for each class as follows: 1. True-Positive (TP): the true positive value is where the actual value and predicted value are the same 2. True-Negative (TN): the true negative value for a class will be the sum of the values of all columns and rows except the values of that class that are calculating the values for. 3. False-Positive (FP): the false positive value for a class will be the sum of values of the corresponding column except for TP. 4. False-Negative (FN): the false negative values for class will be the sum of values of corresponding rows except for TP.

In this system, we have applied the undermentioned performance measures: Precision, Recall and F1-Score. Precision is the ability to classify the class correctly. The recall is the ability of the classification to classify all of the actual class in the dataset. This means that the precision is the exactness and the recall is the completeness of the classification model. The f1-score is just a combination of the exactness and completeness of the system. We have tried the results depend on hourly data. The precision, recall, and f1-score values are calculated based on multi-class confusion matrix for training 2000, 4000, 6000, 8000 and 10000 data from random of hourly data and the calculation results for the weather conditions are obtained as shown in Table 3. The training result is in Figure 2.

Table 3 Precision, Recall, and F1-Score of Training Data

Number of Data	Precision	Recall	F1-Score
2000	0.80794586	0.861899	0.82424
4000	0.83170943	0.875011	0.844527
6000	0.85713577	0.894201	0.866447
8000	0.85685608	0.897343	0.868039
10000	0.89548122	0.900088	0.893722

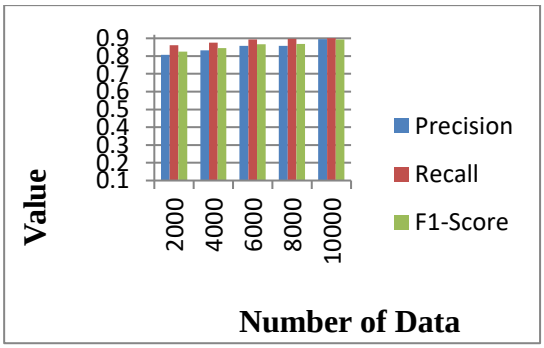


Figure 2 Precision, Recall, and F1-Score of Training Data

The precision, recall, and f1-score values are calculated based on multi-class confusion matrix for testing 2000, 4000, 6000, 8000 and 10000 data from random of hourly data and the calculation results for the weather conditions are obtained as shown in Table 4. The testing result is in Figure 3.

Table 4 Precision, Recall, and F1-Score of Testing Data

Number of Data	Precision	Recall	F1-Score
2000	0.7237776	0.751669	0.728806
4000	0.76687536	0.844085	0.783515
6000	0.7540088	0.851466	0.780371
8000	0.76927702	0.836108	0.782564
10000	0.77155028	0.850803	0.792209

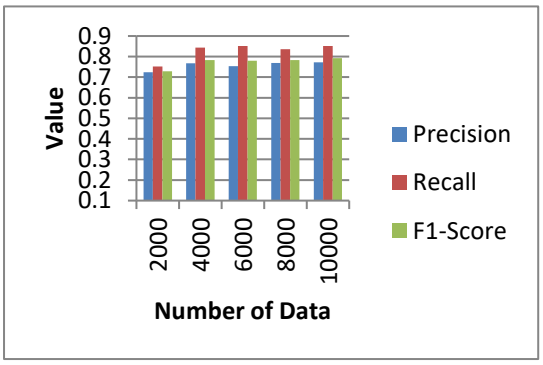


Figure 3 Precision, Recall, and F1-Score of Testing Data

According to the results for training, the rightness (precision) of weather prediction is obtained above 80% and the state of being complete and entire (recall) is more than 85%. The average F1-Score is above 82% completely and correctly can predict the weather conditions. And also for testing result, the rightness (precision) is obtained above 72% and the state of being complete and entire (recall) is more than 75%. The average F1-Score is above 72%. So, Gaussian Naïve Bayes based weather forecasting system is quite satisfactory for prediction.

6. Conclusion

Gaussian Naive Bayes is a likelihood-based characterization strategy, which expects that credits are restrictively commonly free given the class name. This framework produced for weather conditions anticipating framework by utilizing Gaussian Naïve Bayes classifier. Weather conditions gauging has dominating significance in our everyday life. Gaussian Naïve Bayes was applied to anticipate weather pattern for Yangon locale. Historical weather dataset plays a key role in the training process of this system. Weather prediction using machine learning techniques is not an easy task because predicting does not always mean predicting accurate weather. The overall accuracy of the proposed system using the Gaussian Naïve Bayes algorithm has between 70% and 80% using training and testing. Using this system can suitable using Gaussian Naïve Bayes due to its good certain results. In future, we have plan to transform to the prototype of local station weather by using the embedded multi sensors, such as Arduino Mega Board, Temperature and Humidity Sensor, Anemometer, Pressure Sensor and Raindrop Sensor, that produces more accurate results for specific local area.

References

[1] Ahmed, Bilal. "Predictive capacity of meteorological data: Will it rain tomorrow?." Science and Information Conference (SAI), 2015. IEEE, 2015.

[2] Binu Thomas, Raju G. and Sonam Wangmo, “A Modified Fuzzy C-Means Algorithm for Natural Data Exploration,” World academy of Science, Engineering and Technology 49 2009.

[3] Geetha, A., and G. M. Nasira. "Data mining for meteorological applications: Decision trees for modeling rainfall prediction." Computational Intelligence and Computing Research (ICCIC), 2014 IEEE International Conference on. IEEE, 2014.

[4] KavithaRani, B., and A. Govardhan. "Effective Features and Hybrid Classifier for Rainfall Prediction." International Journal of Computational Intelligence Systems 7.5 (2014): 937-951.

[5] Kotu, Vijay, and Bala Deshpande. Predictive analytics and data mining: concepts and practice with rapidminer. Morgan Kaufmann, 2014.

[6] Liu, James NK, Bavy NL Li, and Tharam S. Dillon. "An improved naive Bayesian classifier technique coupled with a novel input solution method [rainfall prediction]." IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 31.2 (2001): 249-256.

[7] M. Tektaş, "Weather forecasting using ANFIS and ARIMA models. A case study for Istanbul," Environmental Research, Engineering and Management, vol. 51, pp. 5-10, 2010.

[8] P. Langley, W. Iba and K. Thompson, "An analysis of Bayesian Classifiers.," in Proceedings of the Tenth National Conference on Artificial Intelligence, San Jose, CA, 1992.

[9] R. O. Duda and P. E. Hart, Pattern classification and scene analysis, John Wiley and Sons, 1973.

[10] R. Sallehuddin, et al., "Forecasting Time Series data using Hybrid Grey Relational Artificail Neural Network and Auto Regressive Integrated Moving Average," Journal of Applied Artificial Intelligence, vol. 23,

[11] Santhanam, Tiruvenkadam, and A. C. Subhajini. "An efficient weather forecasting system using radial basis function neural network." Journal of Computer Science 7.7 (2011): 962.

[12] Sawaitul, Sanjay D., K. P. Wagh, and P. N. Chatur. "Classification and prediction of future weather by using back propagation algorithm-an approach." International Journal of Emerging Technology and Advanced Engineering 2.1 (2012): 110-113.

[13] Sharma, Ashutosh, and Manish Kumar Goyal. "Bayesian network model for monthly rainfall forecast." Research in Computational Intelligence and Communication Networks (ICRCICN), 2015 IEEE International Conference on. IEEE, 2015.

[14] Tektaş, Mehmet. "Weather forecasting using ANFIS and ARIMA models." Environmental Research, Engineering and Management 51.1 (2010): 5-10.

Risk Calculation of Covid-19 for ASEAN Countries Using Backpropagation Neural Network and Fuzzy Inference System

Sabai Oo
Information Technology Engineering
Technological University (Thanlyin)
Yangon, Myanmar
sabaioo@ttu.edu.mm

Htar Htar Lwin
Faculty of Computer Systems and Technologies
University of Computer Studies
Yangon, Myanmar
htarhtarlwin@ucsy.edu.mm

Abstract

COVID-19 is a big challenge and the whole world is now facing at the present time. Predicting the number of COVID-19 patients is a crucial task in the effort to assist governments and healthcare departments respond rapidly to outbreaks. Predictive problems can be successfully solved using the backpropagation technique, a type of Artificial Neural Network (ANN). This paper proposes a prediction model to estimate the number of COVID-19 sufferers in ASEAN Countries using a Backpropagation neural network with Stochastics Gradient Descent Optimizers (SGD). And then these predicted results are used to decide the risk category of a country with Fuzzy Inference System. Root Mean Square Error (RMSE) is used to evaluate the performance of the prediction method. To evaluate the performance of the proposed method for Risk Calculation, have experimented many times using the preexisting actual data.

1. Introduction

Covid-19 is an infectious disease caused by the coronavirus. In December 2019, Wuhan, China, reported the discovery of the first case. On December 31, 2019, the coronavirus disease 2019 was initially reported to the World Health Organization (WHO). A worldwide health emergency was declared over the COVID-19 outbreak on January 30, 2020. Covid-19 is a very serious disease because it mainly attacks the lungs in the human body and can cause death for the sufferers, especially such to congenital diseases or a weak immune system. Daily life, businesses, public health, food systems, educational systems, and employment have all been quickly affected by COVID-19.

With the emergence and spread of COVID-19, different modeling, estimating, and forecasting methodologies are being used to better understand and control the pandemic. To control the spread of the virus and serve as a guide for health officials, it is crucial to predict the number of Covid-19 patients. Various prediction methods have been proposed to recognize and predict the Covid-19 outbreak.

Artificial neural network (ANN) has been widely used by many researchers to analyze traditional classification and prediction problems. ANN is one of the appropriate prediction methods and can learn from data and produce predictions or classifications as a result of their learning. There are many different types of ANN algorithms. Backpropagation is the most used neural network algorithm, and it is used in the system. Backpropagation is an important mathematical tool for improving the accuracy of predictions with good results. The optimization methods applied during training have a significant impact on its performance. The Gradient Descent method is the most fundamental and widely used optimization technique (GD).

In this paper, the model is trained by using a backpropagation algorithm with GD and SGD optimizers. The backpropagation neural network with gradient descent method is compared to the backpropagation neural network with stochastic gradient descent method in this experiment. Among them, the method that produced the better performance is used in this proposed system to get the most accurate prediction model. Firstly, the backpropagation algorithm is used to predict the number of Covid-19 sufferers and then these results from the predicted model are used to

decide the risk category of a country with Fuzzy Inference System.

2. Related Work

To predict the spread of viruses, many researchers have suggested various methods. Predictions of COVID-19 patient numbers are essential for controlling and preventing the spread of such diseases.

To estimate the mortality rate in COVID-19 patients, the paper [1] offered a prediction model using a variety of machine learning methods, including Support Vector Machine (SVM), Artificial Neural Networks, Random Forest, Decision Tree, Logistic Regression, and K-Nearest Neighbor (KNN). To thoroughly evaluate the classifiers and determine the sensitivity and specificity of the model, a confusion matrix was employed. Among them, the most effective and accurate algorithm was the Neural Network one. For the neural network, grid search was used to find the best hyperparameters and also used stochastic gradient optimizer and constant learning rate.

The number of daily confirmed Covid-19 cases was analyzed and predicted in the paper [2] using two statistical models (ARIMA and GARCH) and a deep learning model (LSTM DNN). The experiment made use of ten datasets, released by the WHO between December 31, 2019, and February 22, 2021, including nine country-specific datasets and a global dataset. Preprocessing, training, and prediction processes are the three steps that make up the prediction process of the model. The performance of the deep learning model was compared to two statistical models by using mean-square-error. LSTM DNN predicts best for all datasets, according to experimental results, while the predictions of two statistical models are dataset-dependent. Moreover, this research only used data from a limited period, which was insufficient for learning the condition.

The paper [3] considered the Backpropagation neural network with the Fletcher–Reeves method for predicting the number of COVID-19 sufferers in Malang. To optimize Backpropagation ANN using several different architectures and learning rates, the performance comparison of the Fletcher–Reeves and gradient descent methods was demonstrated. As a result, the Backpropagation neural network with the Fletcher–Reeves method achieved better results than the Backpropagation neural network with the gradient descent method.

Our research focus to calculate the risk categories of each country in ASEAN Countries using a backpropagation neural network and fuzzy inference system. The prediction model is implemented using a backpropagation neural network to predict the future number of Covid-19 sufferers. And then, these prediction results are used to calculate the risk categories of each country with a fuzzy inference system.

3. Covid-19 Dataset and Preprocessing

In the paper, Covid-19 datasets, from 22-01-2020 to 31-12-2020, which consists of the number of positive cases, the number of recoveries, and the number of deaths for ASEAN countries are used for the future predicting of Covid-19 spread. The Center for Systems Science and Engineering at Johns Hopkins University offered a GitHub repository from which the dataset was collected. The repository was largely made available by the university for the visual dashboard of the 2019 Novel Coronavirus. These data samples are time-series Covid-19 global datasets. Among them, we have only used the dataset from ASEAN countries: Brunei, Cambodia, Indonesia, Laos, Malaysia, Myanmar, Philippines, Singapore, Thailand, and Vietnam. Therefore, the number of confirmed cases, recovery cases, and death cases of each country in ASEAN countries were prepared in Table 1 and used as the input variables for the predictive model.

After manually preparing the dataset for each country from the global dataset, the data is preprocessed by using data normalization. Preparing raw data to make it usable for model construction and training is known as data preprocessing. Data preprocessing techniques include normalization. The term "normalization" refers to the rescaling of data from the original range to a new range between 0 and 1. Normalization helps speed up the learning step of the backpropagation technique. There are many methods for data normalization such as min-max normalization, z-score normalization, and normalization by decimal scaling. etc. Among

them, min-max normalization was used to preprocess the data.

Table 1. Example dataset of Covid-19 for Laos

No.	Date	Confirm Case	Recover Case	Dead Case
1	1/22/20	3	0	3
2	1/23/20	6	0	6
3	1/24/20	6	0	6
4	1/25/20	8	0	8
5	1/26/20	8	0	8
6	1/27/20	9	0	9
7	1/28/20	10	0	10
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
345	12/31/20	102589	89923	4782

4. Building and Testing the Model

The dataset is divided into two parts: 80% and 20%, after preprocessing stages. The first part is utilized as the training dataset, and the second part is used as the test dataset. And then, the predicted models are implemented by using Backpropagation neural network with the gradient descent method (GD) and the stochastic gradient descent method (SGD).

RMSE is used to evaluate how well the prediction methods perform. According to performance analysis, the root-mean-square error (RMSE) resulting from the Backpropagation neural network using the stochastic gradient descent decreases than the Backpropagation neural network using the gradient descent for several learning rates. Table 2 shows the performance evaluation of the prediction methods for several learning rates.

Moreover, the gradient descent and stochastic gradient descent were compared for ASEAN countries with a learning rate of 0.0005. The comparison of the performance of the prediction method for ASEAN countries with a learning rate of 0.0005 is shown in Figure 1.

Table 2. Evaluation of the performance of the prediction method by using RMSE for several learning rates

RMSE		
Learning Rate	GD Method	SGD Method
0.0005	0.379	0.134
0.0003	0.492	0.364
0.0002	0.558	0.468
0.001	0.178	0.137
0.01	0.631	0.184

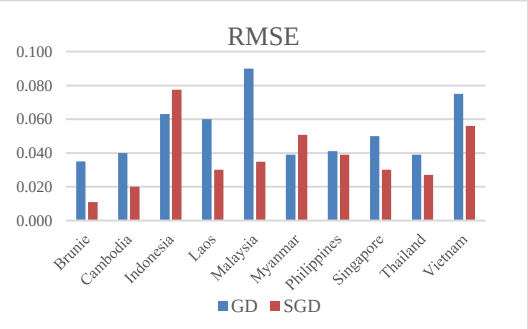


Figure 1. Comparison of the performance of the prediction method for ASEAN countries with a learning rate of 0.0005

From Table 2 and Figure 1, it can be concluded that the stochastic gradient descent method performs significantly better than the gradient descent method. So, the model was implemented using the Backpropagation neural network with stochastic gradient descent to predict the number of Covid-19 sufferers in ASEAN countries.

5. Risk Categorization

A fuzzy inference system is utilized to calculate the risk of each country based on the prediction results from the predictive model. The risk categories were defined as (1) high risk (HR), (2) medium risk (MR), and (3) low risk (LR). First, the case rate, recovery rate, and death rate were calculated. Next, the risk measurement of these parameters is represented by the fuzzy membership function. Imposing rules are established to calculate the risk categories of each country. For the fuzzy inference system, the variable, the fuzzy set, and the range are described in Table 3.

Table 3. The variable, fuzzy set, and the range for the fuzzy inference system

Variable	Fuzzy Set	Range
Case Rate	Low	[0 – 0.01]
	Medium	[0.01 – 0.1]
	High	>0.1
Recovery Rate	Low	[0 – 40]
	Medium	[40 – 70]
	High	>70
Death Rate	Low	[0 – 1]
	Medium	[1 – 1.5]
	High	>1.5

We have used 30 days ahead to calculate such risk classes. For ASEAN countries, the results of the proposed system and the results calculated using the actual data are shown in Table 4. It shows the predicted risk categories for the upcoming 30 days using Covid-19 datasets for 2020.

Table 4. Comparison of risk factors for ASEAN countries based on upcoming 30 days using Covid-19 datasets for 2020

Country	Actual Risk	Predicted Risk
Brunei	HR	HR
Cambodia	HR	HR
Indonesia	HR	HR
Laos	HR	HR
Malaysia	LR	LR
Myanmar	HR	HR
Philippines	HR	HR
Singapore	LR	LR
Thailand	LR	LR
Vietnam	HR	HR

Table 5 shows the predicted risk categories for the upcoming 30 days using Covid-19 datasets of 2021.

Table 5. Comparison of risk factors for ASEAN countries based on upcoming 30 days using Covid-19 datasets for 2021

Country	Actual Risk	Predicted Risk
Brunei	LR	LR
Cambodia	HR	HR
Indonesia	HR	HR
Laos	LR	LR
Malaysia	HR	HR
Myanmar	HR	HR
Philippines	HR	HR
Singapore	LR	LR
Thailand	LR	LR
Vietnam	HR	HR

The risk classification accuracy is calculated many times by using the preexisting actual trend data. Moreover, the system has experimented with each year's Covid-19 datasets to predict the next year's risk categories.

By using Covid-19 datasets for 2020 and 2021, the classification accuracy has been experimented with fifty times for ASEAN countries, five times per country. Among Fifty times, only produced five incorrect results.

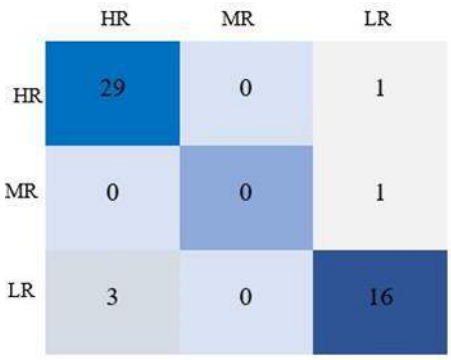


Figure 2. Confusion matrix for the three classes for the risk prediction

According to the confusion matrix, it is observed that the proposed method produces a relatively 90% high accuracy of prediction risk compared with the actual trend risk calculation.

6. Conclusion

In the paper, daily case data, an artificial neural network with backpropagation, and a combination of fuzzy inference systems were used to solve the problem of calculating the long-term risk of a country. Firstly, the prediction model is implemented on the training dataset using the Backpropagation neural network with Gradient Descent and Stochastic Gradient Descent Optimizers. And then, the experimental results show that the Backpropagation neural network with Stochastics gives better results compared to the Backpropagation neural network with the gradient descent. Therefore, the Backpropagation neural network is optimized by the Stochastics method, which is used to predict the number of COVID-19 sufferers.

The number of cases, the number of recovered, and the number of deaths achieved from the predictive model is used to calculate the risk of each country. In the paper, a new way for calculating the risk to a country and predicting the outbreak of an epidemic is presented. In addition, it presents that an Artificial neural network (ANN) can be used to solve regression problems and predict future health risks.

Acknowledgments

We would like to express our appreciation to everyone who helped with this work.

References

[1] M.Pourhomayoun, and M.Shakibi, "Predicting mortality risk in patients with COVID-19 using machine learning to help medical decision-making", Elsevier Public Health Emergency Collection, Published by Elsevier Inc., 16 January 2021.

[2] M.Kim, "Prediction of COVID-19 Confirmed Cases after Vaccination: Based on Statistical and Deep Learning Models", SciMedicine Journal, Published 01 June 2021, Vol.3, No.2.

[3] S.Anam, M.H.A.A.Maulana, N.Hidayat, I.Yanti, Z.Fitriah, and D.M.Mahanani, "Predicting the Number of COVID-19 Sufferers in Malang City Using the Backpropagation Neural Network with the Fletcher–Reeves Method", Applied Computational Intelligence and Soft Computing, Published 29 April 202.

[4] Jia, W.; Li, X.; Tan, K.; Xie, G. Predicting the outbreak of the hand-foot-mouth diseases in China using recurrent neural network. In Proceedings of the 2019 IEEE International Conference on Healthcare Informatics (ICHI), Xi'an, China, 10–13 June 2019; pp. 1–4.

[5] Hilton, J.; Keeling, M.J. Estimation of country-level basic reproductive ratios for novel Coronavirus (COVID-19) using synthetic contact matrices. medRxiv 2020.

Integrated XML Schema for Heterogeneous XML Schemas

Htun Ei Ei San, Thidar Win
University of Computer Studies, Yangon
Tuneieisan1995@gmail.com, thidarwin@ucsy.edu.mm

Abstract

With the growing popularity of the XML model and the proliferation of online XML documents, the automated matching of XML documents and databases has become a critical problem. Currently, many recent e-health records are based on XML documents. Schema matching plays a central role in a myriad of XML-based applications. There is an increasing need to develop effective matching systems to identify and discover semantic matches between XML data. XML Schema matching methods face several challenges in the form of definition, adoption, use and combination of element similarity measurements. In this system, element type conflicts, constraints, naming conflicts, and the semantic and structural information of two specific health care XML Schemas are solved.

1. Introduction

Nowadays, there is a growing demand for ubiquitous healthcare data, and healthcare providers must combine essential details so that they can be saved as an electronic health record. The term "electronic health record" refers to a collection of electronic healthcare data on specific patient that may be exchanged across several healthcare systems. Integrating healthcare data plays a critical role in enhancing patient care quality and information flow between medical systems. In order to track and maintain a patient's physical and emotional well-being, healthcare data is collected from a large variety of environment and patient sensors and actuators. Healthcare data expressed using several XML Schemas, on the other hand, presents a difficult problem when it comes to integrating the data.

Developers can construct their own eXtensible Markup Language (XML) documents that adhere to certain structure guidelines. XML Schema (XSD) and document type definitions (DTD) are commonly used to specify these structure regulations. Because of its advantages, XSD is more extensively used than DTD. It also supports data type and namespace definitions better than DTD.

2. Related Work

Many researchers do research on XML Schema integration.

H. Ahmed, A. Hamad [6] implemented a new technique for integration the heterogeneous XML Schemas, under the name XDEHD. The mediated schema that has all of the sources' ideas and relationships without duplication. The detailed technique is divided into three steps: first, decompose the schemas sources to extract all subschemas; each subschema comprises three levels: ancestor, root, and leaf. After that, the technique matches and compares the subschemas to return related candidate subschemas, and a semantic closeness function is used to determine how comparable the subschemas' concepts are modelled in the sources. The system is generated a comprehensive mediated schema between a set of heterogeneous databases by combining XML Schema sources.

X. Yang, M. Li Lee, and T. Wang Ling [16] implement a semantic approach to resolve structural conflicts in the integration of XML Schemas. The system used a data model called ORA-SS (Object Relation Attribute Model for Semi-Structured Data) to capture the implied semantics in an XML Schema. It provided a complete algorithm for integrating XML Schemas. Compared to other methods, the system adopted an n-nary integration strategy that considers the data semantics, importance of a source, and how the majority of the sources model their data when resolving structural conflicts such as attribute/object class conflict and ancestor descendant conflict. This system solved structural conflicts like attribute/object class conflicts, ancestor-descending conflicts. The proposed technique has mainly solved structural

conflicts, but most semantic conflicts have not been solved.

3. Background Theory

The integration of existing or planned schemas into a unified schema is known as schema integration. XML integration can be divided into two categories: view integration and database integration

3.1. View Integration

The goal is to develop an integrated schema from a group of independently created application views. The database structure is considerably too complex for a single designer to model in a single view for large systems. User groups operate independently in most organizations, with their own data requirements and expectations that may conflict with those of other user groups.

3.2. Database Integration

A distributed database is a collection of data that is logically part of the same system but is scattered across several computer networks. For a group of databases, database integration generates a uniform schema. A distributed database management system's global schema is a virtual representation of all databases.

3.3. Integration Processing Strategies

(1) Pre-Integration

- Choose integration processing strategies
- This governs the choice of schemas to be integrated

(2) Compare the Schemas

- The schemes are analyzed and compared in order to determine the correspondence between the concepts and to detect possible conflicts.

(3) Conform the Schemas

- Once the conflicts have been detected, an effort is made to resolve them so that the different patterns can be merged.
- Automatic conflict resolution is generally not feasible; interaction with designers is required.

(4) Merging and Restructuring

- The schemas are ready to be superimposed, giving rise to some intermediate integrated schema.

3.4. XML Data Model

The information demonstrates for XML is exceptionally basic or exceptionally theoretical, depending on one's point of view. XML gives no more than a pattern on which more complex models can be built. The reason of the information show is to characterize all passable values of expressions in XPath, counting values that are utilized amid middle of the road calculations. Each XPath expression takes as its input an occurrence of the information demonstrates and returns an occasion of the information show. When it comes to creating schemas, there are two primary approaches:

- Document Type Definitions (DTDs)
- XML Schemas (XSDs)

3.4.1 Document Type Definitions (DTDs)

DTDs are the first attempt at developing a schema for XML documents. DTDs were created before XML, with SGML's predecessor. The fundamental disadvantage of DTDs is that they are written in a complex language. When it comes to data formatting, XML offers a more structured method. Here is a very simple example of a DTD that could hold a list of basketball players on a team:

1. <!ELEMENT player_list (player) *>
2.<!ELEMENT player (name, age, school?, country)>
3. <!ELEMENT name (#PCDATA) >
4. <!ELEMENT age (#PCDATA) >
5. <!ELEMENT school (#PCDATA) >
6. <!ELEMENT country (#PCDATA) >

3.4.2 XML Schema (XSDs)

DTDs are replaced by XML Schema, which is a more powerful and easy way to create schemas for XML-based markup languages. XML Schemas are written in the XML Schema Definition language (XSD). The World Wide Web Consortium (W3C) produced XML Schema and the XSD language, which are far more powerful than DTDs when it comes to schemas. The concept behind XML Schema is to create schemas using XML as the foundation. XSD is very similar in purpose to a DTD in that it is used to establish the schema of a class of XML

documents. Here is a sample XML Schema of XSD file for patient.

```
<?xml version="1.0" encoding="UTF-8"?>
<xs:schema targetNamespace="org.di.demo.patient"
xmlns="org.di.demo.patient"
xmlns:xs="http://www.w3.org/2001/XMLSchema">
<xs:element name="PatientList">
<xs:complexType>
<xs:sequence>
<xs:element maxOccurs="unbounded"
minOccurs="0" name="Patient"
type="PatientElement"/>
</xs:sequence>
</xs:complexType>
</xs:element>
<xs:complexType name="PatientElement">
<xs:sequence>
<xs:element minOccurs="1" name="firstName"
type="xs:string"/>
<xs:element minOccurs="1" name="lasttName"
type="xs:string"/>
<xs:element minOccurs="1" name="middleName"
type="xs:string"/>
<xs:element minOccurs="1" name="ssn">
</xs:element>
<xs:element minOccurs="1" name="sex"
type="xs:string"/>
<xs:element minOccurs="1" name="dob"
type="xs:string"/>
</xs:sequence>
</xs:complexType>
</xs:schema>
```

4. Overview of the Proposed System

The proposed approaches in this system give an unique approach to creating an integrated schema that efficiently merges two disparate XML Schemas. Because relations hold domain information, this system prioritizes integration over that of individual concepts.

Figure 1 shows the overview of the proposed system. The proposed system presents semantic and structure measurement to generate an integrated schema. In order to the system flowchart, there are four main steps in the system. *Firstly*, it accepts two heterogenous XSD sources and decomposes them into subschemas from all sources. Each subschema has parent and its child elements.

The *second* step computes semantic between elements from all sources using three processes. For the previous semantic step, three phases are considered: first phases, the system finds element similarity which is complex of simple, the Occurrences of elements in sources and final phase is used to compute naming similarity using WordNet.

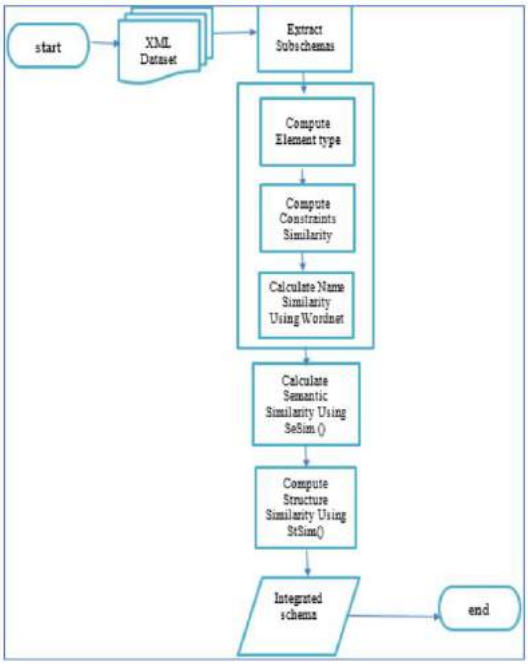


Figure 1: Overview of the Proposed System

The *next* step is to find structure similarity between element pairs that it matches the schema elements based on the similarity of their position and their nearest element.

The *final* step is to generate an integrated XML Schema.

4.1 Semantic Similarity Measurement

When it comes to integrating text documents, the semantics of an idea is crucial. The vocabularies, content model, and datatype are all part of the XML Schema semantics. XML Schema typically starts a document with the typical and the Uniform Resource Identifier associated with that namespace. The maxOccurs and minOccurs attributes in XML Schema allow the system to specify the number of possible occurrences for an element. Furthermore, the simpleType or complexType element aids in distinguishing element type similarities between two elements.

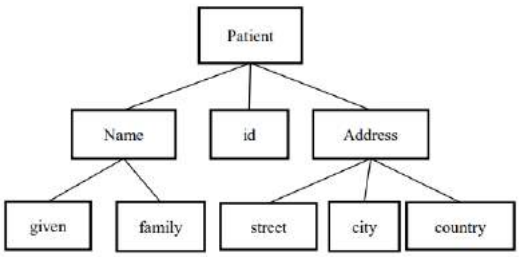


Figure 2: Subschema Tree of XSD Source 1

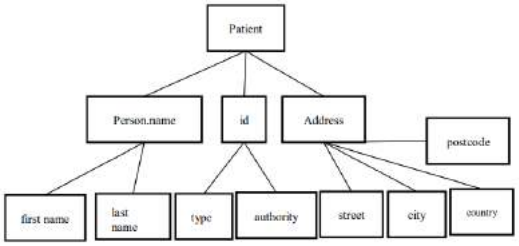


Figure 3: Subschema Tree of XSD Source 2

4.2 Element Type Similarity

Although the element name is the most essential aspect in calculating semantic similarity, other components are also considered. For example, the name similarity of *two elements id* in Figure 2 and Figure 3 is 1. This is a false matched value since the *first id* element is a complicated element and the *second id* element is a simple element. This implies that they differ in other ways. As a result, in order to exclude some erroneous matches, other characteristics must be employed to determine their semantic relatedness.

4.3 Constraint Similarity

The cardinality (occurrence) constraint is another factor that influences the semantic similarity of two items. The minimum and maximum number of occurrence times of an element in XML instances are defined by minOccurs and maxOccurs, respectively. To express the constraint similarity between two items e1 and e2, the system uses CSim(e1, e2). The following equation is used to compute cardinality constraint similarity for the specified values of minOccurs and maxOccurs.

$$CSim(e1(min,,\max),e2(min,\max) = \frac{\left(1-\frac{|e1.\min - e2.\min|}{e1.\min + e2.\min}\right) + 1 - \frac{|e1.\max - e2.\max|}{e1.\max + e2.\max}}{2} \quad \textbf{(4.1)}$$

Where min and max are short forms of minOccurs and maxOccurs, respectively, in the given equation.

4.4 Name Similarity

The linguistic similarity of two elements is the most essential factor in semantic measurement. The system applies the algorithm given in the following linguistic similarity algorithm to determine the linguistic similarity between items. The program looks for similarities between two elements, e1 and e2. The search is conducted in a breadth-first manner, starting with the element e1's synonym set on WordNet and progressing to the element e2's synonym set, and so on, until e2 is matched. If the target cannot be found, the linguistic similarity returns a value of 0, otherwise a distance of 0.9 is determined.

Algorithm 1: Name Similarity
Input: Two elements, e1 and e2
Distance = 5-level;
Output: Name similarity
If e1.name==e2.name then return 1;
else return DepthSyn(e1, {e2}, level);
{e2}=S;
Function DepthSyn(e, S, level)
Output: the synonym in depth
If (level>=distance) then return 0;
Else if (e1∈S) then return power(0.9, distance);
Return DepthSyn(e, S, distance+1);

Figure 4: Linguistic Similarity Algorithm

Definition 1: Semantic similarity measures how similar two elements' names, restrictions, and path context are. This is given by:

$$SeSim(e1,e2) = \propto * NameSim(e1,e2) + \beta * EleSim(e1,e2) + (1-\alpha-\beta) * CSim(e1,e2) \quad \textbf{(4.2)}$$

Where SeSim is the semantic similarity: ∝ and β are the weighted constants by the program. NameSim is the name similarity calculated by the procedure in the preceding Linguistics Similarity Algorithm. The cardinality constraint similarity of e1 and e2 items is represented by CSim, while the similarity between two element types is represented by EleSim.

4.5 Structure Similarity Measurement

The structural matching stage matches schema elements based on proximity to one another and the context in which position. Structure matching is influenced by the semantic similarity found in the semantic measurement. For each pair of elements, the results are needed a structure similarity coefficient, StSim. If two elements are similar in contexts, they have structural similarity.

Algorithm 2: Structure_Similarity
Input: Two schema trees S,T
Thresh_min=0; thresh_max=0.3
Output: The structure similarity
For each s ϵ S, t ϵ T where s, t are leaves
S1=post-order(S); S2=post-order(T);
for each s in S1,
  for each t in S2 tϵT
    if StSim (s,t)>=thresh_max then
      StSim(s,t) = StSim(s,t)+0.1;
    else if StSim (s,t)<=thresh_min then
      StSim(s,t) = StSim(s,t)+0.1;
Structure_Similarity(S,T)= StSim(s,t);

Figure 5: Structure Similarity Algorithm

Definition 2: The structure similarity between two elements e1 and e2 is specified as:

$$StSim(e1, e2) = \frac{\text{sum}_{\text{links(e1,e2)}} + \text{sum}_{\text{links(e2,e1)}}}{\text{leaves(e1)} + \text{leaves(e2)}} \quad \textbf{(4.3)}$$

Where leaves(e1) is the total number of leaves in the subtree-rooted at element e1; sum_links(e1, e2) is the total number of links from the leaves of element e1 to the leaves of element e2.

4.6 Generate an Integrated XML Schema

This system defines how elements can be semantically and structurally equivalent, and then produces an integrated XML Schema based on semantic and structure measurement.

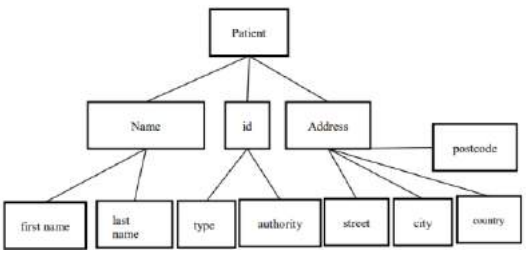


Figure 6: Integrated XML Schema

5. Performance Evaluation

This system uses Java language to be implemented element similarity. The results of several similarity measurements were gathered and summarized in the table below.

Table 1: Experiment Results of Healthcare Datasets

Similarity Measure	Number of element pairs	Number matches
XCLust	34	28
XMLSim	34	26
ESim	34	30

Table 1 shows the experiment results for this system (ESim) and other systems XMLSim, and XClust. The result is tested on same dataset.

XMLSim has the lowest matching, as indicated in the table. The algorithm computes for the next pair of nodes when one node in the first document is matched with another node in the second document. The matched value is reduced if the number of nodes in the two texts differs.

As shown in Table 1, this system's matching quality is superior to that of XMLSim and XCLust's techniques. Some element pairs have the same name but differ in type, XClust's element similarity assessment does not consider element type similarity between two elements. XMLSim was overly focused on information content similarity, ignoring datatype and cardinality constraint similarities between two items. Figure 7 shows the comparison results of this system and other systems XMLSim, and XClust for health care dataset. In the proposed approach, the percentage values of recall and precision are high.

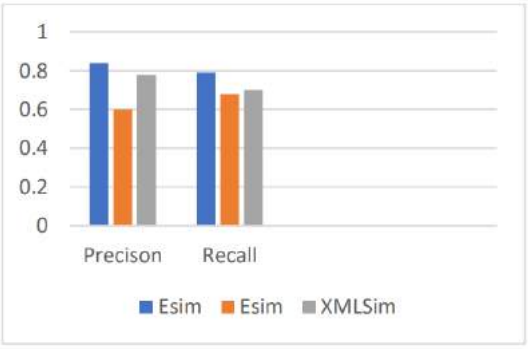


Figure 7: Comparison Results

6. Conclusion

XML is a major player in the exchange of data and information, playing a central role in health care via HL7 and CCR standards. This system is generated complete and minimal integration of schema among a set of heterogeneous XML Scheme sources. The system is presented a semantic and structure similarity approach for Healthcare XML Schemas. It is calculated element type similarity for XML elements and determining cardinality constraint resemblance between two elements. When two elements have semantic similarity, it is included not just language similarity but also element datatype and constraint compatibilities. The structure similarity method is used to display the distance between two elements. Only a structurally heterogeneous source dataset, such as the XSD format, can be used with this system. The attribute's datatype limitations are not considered.

References

[1] A. Algergawy, R. Nayak, G.Saake, "Element Similarity Measures in XML Schema Matching", 2010.

[2] A.Fernandez, A. Polleres A, S.Ossowski, "Towards Fine-grained Service Matchmaking Service Matchmaking by Using Concept Similarity", Workshop on service matchmaking and resource retrieval in the semantic web, pp 31–45, 2007.

[3] C. Batini, M. Lenzerini, S.B.Navathe, "A Comparative Analysis of Methodologies for Database Schema Integration", ACM Comput. Surv. 18 (4), 323–364, 1986.

[4] D. Aumueller, H.H.Do, S. Massmann, E.Rahm, "Schema and Ontology Matching with COMA++", Department of Computer Science, University of Leipzig Augustusplatz 10/11, Leipzig 04103, Germany,2005.

[5] D.D.Yang, M.W.David, "Measuring Semantic Similarity in the Taxonomy of WordNet", The 28th Australasian computer science conference (ACSC 2005), pp 315–322, 2005.

[6] H. Ahmed, A. Hamad, "XML-Based Data Exchange in the Heterogeneous Database (XDEHD)", Department of Information Technology, College of Computer, Qassim University, 2015.

[7] H.H.Do, "Schema Matching and Mapping-based Data Integration", Dept of Computer Science, University of Leipzig, Germany, 2006.

[8] H.H.Do, E.Rahm, "COMA: A System for Flexible Combination of Schema Matching Approaches", Proceedings of the very large data bases conference (VLDB), pp 610–621, 2002.

[9] H.Q. Nguyena, D.Taniar, J. W. Rahayua, K. Nguyena, "Double-layered schema integration of heterogeneous XML sources", a Dept. of Computer Science and Computer Engineering, La Trobe University, VIC 3086, Australia, 2011.

[10] I. Brown, A. Adams, "The Ethical Challenges of Ubiquitous Healthcare", Int Rev Inf Ethics 8(12):53–60, 2007.

[11] J.Tekli, R .Chbeir, K .Yetongnon, "A hydrid approach for XML similarity", 07 proceedings of the 33rd conference on current trends in theory and practice of computer science. Springer, Berlin, pp783–795, 2007.

[12] J. Tekli, R.Chbeir, K.Yetongnon,"An overview on XML similarity: background, current trends and future directions", Comput Sci Rev 3:151– 173, 2009.

[13] M.L.Lee, L.H. Yang, W. Hsu, X. Yang, "XCLust: Clustering XML Schemas for Effective Integration", ACM Press, New York, pp 292–299, 2002.

[14] M. M. Thu, "Clustering XML Documents using XEdge Algorithm", Conference on Parallel and Soft Computing, Yangon, Myanmar, 2015.

[15] S. Bechhofer, R. Volz, and P. Lord "Cooking the Semantic Web with the OWL API", International Semantic Web Conference (ISWC), 2003.

[16] X.Yang, M. Li Lee, and T. Wang Ling, "Resolving Structural Conflicts in the Integration of XML Schemas: A Semantic Approach", School of Computing, National University of Singapore Science Drive 2, 2003.

Classification of YouTube Comment Spam Using TF-IDF and Multinomial Naïve Bayes Classifier

Nang Mya Oo, Nang Saw Kalayar
University of Computer Studies, Yangon
nangmyaoo@ucstgi.edu.mm, sawkalayar@ucstgi.edu.mm

Abstract

Social networking such as YouTube, Facebook and others are very popular nowadays, YouTube has become a popular social media among the users. YouTube user can subscribe a channel they like and also give opinion on the comment section. It became a platform for spammer to distribute spam through the comments on YouTube [1]. This system can develop a YouTube comment spam classification framework by using term frequency-inverse document frequency (TF-IDF) and Multinomial Naïve Bayes. TF-IDF is a statistical method to measure the importance of a word in the document to the whole corpus. Multinomial Naïve Bayes classifier was used to categories the YouTube comments into the suitable category. This system intends to classify the YouTube comment as spam or ham depending on the contents contained in a comment. In order to know how the system can perform well, the system is checked the classified results with calculated accuracy (Precision, Recall, and F-measure). This system is implemented using ASP.Net programming language on Microsoft Visual Studio 2015 IDE and Microsoft SQL Server 2017 Express Version as Database Engine.

Keywords: Multinomial Naïve Bayes, TF-IDF, YouTube, spam

1. Introduction

YouTube is a video sharing site that was begun back in 2005. It was in this way purchased by Google in 2006 and is presently on of Google's auxiliaries. From that point forward YouTube has arisen as a main competitor in the video sharing space. Users of YouTube are known as channels, and YouTube permits channels to upload, rate, and share, add to top picks, report, comment on videos, and subscribe into different users. YouTube creates over 400 hours each moment and more than 1 million hours of content that are consumed by users day to day. The site has been positioned as the second most well-known site on the planet by Alexa Internet – a web traffic examination organization [3]. One of the most used elements of YouTube is its commenting framework where users can remark on videos uploaded to different channels. This component permits users to communicate with each other and share their perspectives, sentiments, and so no the video. Nonetheless, this has likewise transformed into a chance for malicious users to share special content otherwise called spam. Spam comments are frequently entirely unessential to the given video and are typically created via robotized bots to perform spam crusades – huge scope arranged posting of malicious comments has been investigated. YouTube has confronted expanding analysis about its powerlessness to direct uploaded content. An enormous user-base of YouTube comprises of kids who are frequency presented to malicious and unsafe material as comments [9]. Numerous spammers assault the YouTube user through YouTube comments fields. There are a few investigations to classify YouTube Spam, for example, proposed to classify the YouTube comment as Spam and ham by utilizing Machine Learning Techniques, for example, K-nearest Neighbor, Naïve Bayes, and so on. This framework intend to involve Naïve Bayes to classify the input textual comment as spam or ham. The YouTube users are ready to classify the YouTube spam features, they will be more mindful, and the spam spread can be diminished.

2. Related Work

Lopamudra Dey, "Sentiment Analysis of Review Datasets Using Naïve Bayes' and K-NN Classifier", Department of Computer Science and Engineering, Heritage Institute of Technology, Kolkata, India, 2016 [4]. This framework is to assess the presentation for sentiment classification

with regards to precision, accuracy and recall. This framework thought about two Naïve Bayes and K-NN for sentiment classification of the movie reviews and hotel reviews. The outcomes show that the classifiers improved results for the movie reviews with the Naïve Bayes and giving above 80% correctness and outflanking than the K-NN approach. For the hotel reviews, the accuracies are a lot of lower and both the classifiers comparable outcomes. This framework can say Naïve Bayes' classifier can be utilized effectively to examine movie reviews.

"Performance Analysis on Mail Classification with Multilayer Perceptron (MLP)", centers on email classifier with Multilayer Perceptron (MLP) approach for spam and ham mails classification. The framework is utilized term frequency and inverse document frequency (tf-idf) and fisher score feature selection method at preprocessing. These strategies permit choosing important elements and adding benefit in term of improvisation in accuracy and decreased running time to email characterization framework [7].

Paras Seth [8], "SMS spam detection and correlation of different machine learning algorithms", 2017, (IC3TSN). This framework classified the SMS spam messages as spam or ham (not spam). This framework performed tests and analysis with Naïve Bayes algorithm, Random Forest algorithm and Logistic Regression algorithm. Naive Bayes beats Random Forest and Logistic Regression and accomplished a high accuracy of 98.445%.

3. Background Theory

YouTube is one of the greatest website for users to get data on the Internet. In light of numerous spammers will deceive the YouTube user by spamming the YouTube comments. Spam is currently a pattern assault and the YouTube characterizes spam as inappropriate comments, like maltreatment of sell things. Ham can be characterized as "great comments" of YouTube liberated from spam comment. Spam can be categorized as perilous on the grounds that spam has the capability of cyber security danger for end users. The spammer utilized this chance to spread malware through comment fields, which will take advantage of weakness in the user's machines. Another aim incorporates holding onto each cash exchanges and capturing credit card and banking data. Moreover, spammer will demolish the content of web pages in general. This activity will lead guests to disturb the posted substance generally. YouTube spam comments can possibly spread malware

Spam Detection Approach: YouTube isn't avoided from malicious user who are many times found to uncover in spamming and promotional activities. There are many ways to deal with distinguish Spam, for example, utilizing Artificial Intelligent, Cryptography, Machine Learning and others, Numerous classifications are signified in binary-two class. Generally, class indicated by 0 and 1. Subsequently, it is simple for preprocessing and feature selection to perform.

Naïve Bayes classifiers is part of a series of probabilistic classifiers based on Bayes' theorem**,** is massively scalable probabilistic classifiers. The number of parameter estimation rises proportionally to the number of variables in naive Bayes classifiers. In contrast to many other forms of classifiers, maximum-likelihood training can be conducted by computing a closed-form expression, which needs linear time, rather than through time-consuming iterative approximation [5].

4. System Implementation

There are steps in this identification structures, for example, Data Collection, Data Preprocessing, Feature Selection and Extraction, Classification and Accuracy Results.

1) **Data Collection**

In this stage, the dataset for tests is download from UCI machine learning repository [10].

Table 1. Dataset for YouTube Comment Spam

Datasets	YouTube ID	Spam	Ham	Total
PSY	9bZkp7q19f0	175	175	350
KatyPerry	CevxZvSJLk8	175	175	350
LMFAO	KQ6zr6kCP8	236	202	438
Eminem	uelHwf8o7_U	245	203	448
Shakira	pRpeEdMmmQ0	174	196	370
Total		1005	951	1956

2) **Data Preprocessing**

Preprocessing is an important part of text categorization. It will be process eliminate noise or duplicate in data. For example, tokenization, stop-words removal are performed. The preprocessing dataset will be utilized for next process of feature selection and extraction.

i) Tokenization is the most common way of separating the text corpus in to individual elements.

Hello! Dear, I am on leave today.

hello	dear	i	am	on	leave	today

ii) Removing Stop Words

Stop Words are pointless word in the corpus [2]. Words, for example, in this so, and, or, the, and so on. All stop words are eliminated first. In the figure below the stop words are: you, are that, have, and the. Which are taken out by utilizing this method.

You are lucky, that you have won the cash prize

lucky	won	cash	prize

3) **Feature Selection and Extraction**

Feature determination is an interaction before grouping class. The appropriate features will be distinguished in light of the dataset.

4) **Classification**

Classification include training and testing operation. The system used 60% of dataset for training and 40% of dataset for testing. Subsequent to finishing the stage (3), assume that there is expected to be used this features as spam or ham. Hence, the dataset must be prepared in light of machine learning technique (Multinomial Naïve Bayes).

4.1. Term Weighting Schemes (TF-IDF method)

Term Frequency (TF) is term weighting in view of the words frequency that show up in a document. The higher the TF value of a word in a document; higher the impact of the term on the document. Inverse Document Frequency (IDF) is weighting technique in view of number of words that show up through all the documents. TF-IDF is one of the simplest and most grounded weighting plans to the data. TF-IDF and its algorithm version are default decision in light of its basic definition and great execution on a number of various datasets. The formulation of this method is as following:

$$W(d,t) = tf(t,d) \times \log\left(\frac{N}{n_t}\right) \quad (1)$$

where,

$W(d,t)$: Term weight in document d

$tf(t,d)$ Term frequency in document

N: Number of all documents

n_t: Number of term t in all documents

4.2. Multinomial Naïve Bayes Classifier Method

Naive Bayes is a likelihood statistical method based on Bayes Theorem with a strong independent assumption to foresee the class of a report based on its likelihood. In this framework, the Multinomial Naïve Bayes is one of the particular techniques for Naïve Bayes, dissected text documents utilizing word counts [6]. (TF-IDF) approach is utilized to describe text documents as opposed to having binary values in Naïve Bayes Classifier. A probability classifier computes the probabilities of each and every class by applying Naïve Bayes theorem. It will work out the conditional likelihood of each comment in the specific classification. The term frequencies can then be utilized to processes the maximum-likelihood estimate in light of the training data to estimate the class-conditional probabilities in the multinomial model:

Bayes Theorem is started the formula:

$$P(c \mid x) = \frac{P(x \mid c) \times P(c)}{P(x)} \quad (2)$$

where,

$P(c \mid x)$: is the posterior probability of class (c , target) given predictor (x , attribute).

$P(c)$: is the prior probability of class.

$P(x \mid c)$: is the likelihood which is the predictor probability given class.

$P(x)$: is the prior probability of predictor.

When calculating the likelihood of a document d for class c , apply the formula below:

$$P(c \mid word\ document\ d) = P(x_1 \mid c) \times P(x_2 \mid c) \times,...,\times P(x_i \mid c) \times P(c) \quad (3)$$

where,

$P(c)$: is the likelihood of class c .

x_i: is the word i in document d .

$P(c \mid word\ document\ d)$: is the likelihood of a document containing class c.

$P(x_i \mid c)$: is the likelihood of the word i is known class c.

The formula for determining prior probability in class c is:

$$P(C) = \frac{N_c}{N} \quad (4)$$

where,

N_c: Number of each class in the training dataset.

N: All documents from training dataset.

The following equation is used to calculate the probability of the word i

$$P(x_i \mid c_j) = \frac{\sum tf(x_i, d \in c_j) idf(x_i) + \alpha}{\sum N_{d \in c_j} + \alpha.V} \quad (5)$$

where,

x: A term from a given sample's feature x.

$\sum tf(x_i, d \in c_j) idf(x_i)$: The sum of TF-IDF all words x_i from all documents in the training samples which correspond to class c_j

$\sum N_{d \in c_j}$: The sum of all term frequencies in the class c_j training data.

α: An additive smoothing parameter ($\alpha = 1$ for Laplace smoothing).

V: All words from training dataset.

4.3. The Process Flow

The following figure shows the brief process flow of the proposed framework. The processing phases of the framework can be separated as Training Phase and Testing Phase. In both stage, Data pre-processing (Tokenization and Removing Stop Words) will happen prior to computing weight or classification. The brief explanation of pre-processing steps are depicts as above segment and the remaining steps are include feature selection (TF-IDF) and Multinomial Naive Bayes.

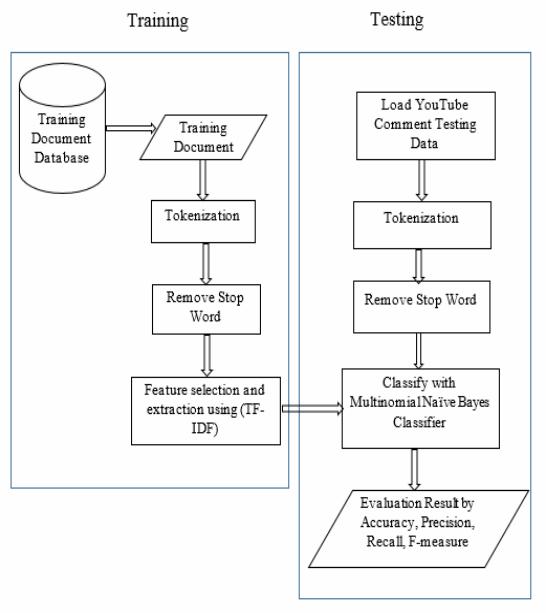


Figure 1. The Process Flow

5. Evaluation Metrics

Five evaluation metrics, which are precision, recall, F-measure, accuracy make used to evaluate the effectiveness of the system. These are calculated by using the following equations:

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN}$$

$$\Pr ecison = \frac{TP}{TP + FP}$$

$$\mathrm{Re}call = \frac{TP}{TP + FN}$$

$$F - measure = \frac{2 \times \Pr ecison \times \mathrm{Re}call}{\Pr ecsion + \mathrm{Re}call}$$

where,

TP: In this situation, the classifier predicted ‘spam’ but the comment really ‘spam’.

TN: In this situation, the classifier predicted ‘not spam’ but the comment really ‘not spam’.

FP: A situation in which the classifier predicted ‘spam’ but the comment was ‘not spam’.

FN: A situation in which the classifier predicted ‘not spam’ but the comment was spam.

5.1. Experimental Result

In this YouTube Spam and Ham classification system, experiments are made for 7 times. In each analysis of 7 different training dataset and 6 different testing dataset pairs are used (Test 1: used 100 documents (YouTube comments) as training dataset and 50 documents (YouTube comments) as testing dataset; Test 2: used 150 documents (YouTube comments) as training dataset and 50 documents (YouTube comments) as testing dataset; Test 3: used 400 documents (YouTube comments) as training dataset and 100 documents (YouTube comments) as testing dataset; Test 4: used 700 documents (YouTube comments) as training dataset and 150 documents (YouTube comments) as testing dataset; Test 5: used 1000 documents (YouTube comments) as training dataset and 200 documents (YouTube comments) as testing dataset; Test 6: used 1300 documents (YouTube comments) as training dataset and 250 documents (YouTube comments) as testing dataset; Test 7: used 1654 documents (YouTube comments) as training dataset and 300 documents (YouTube comments) as testing dataset. This system made the experiment result's performance evaluation based on Accuracy, Precision, Recall, and F-measure of each analysis. The analysis results of 7 different dataset are shown in table 1 and figure 2.

Table 2. Experimental Results of Analysis

Testing No	Training and Testing data Proportion	Accuracy Result	Precision Result	Recall Result	F-Measure
Test 1	100/50	52%	50%	83.33 %	62.50%
Test 2	150/50	62%	57.14%	83.33 %	67.80%
Test 3	400/100	71%	57.89%	62.86 %	60.27%
Test 4	700/150	81.33%	93.88%	64.79 %	76.67%
Test 5	1000/200	90%	91.84%	88.24 %	90%
Test 6	1300/250	91%	91.05%	92.42 %	91.73%
Test 7	1654/300	93.67%	93.67%	94.27 %	93.97%

Base on the analysis, the system can give better classification result if the more trained data can feed to this system.

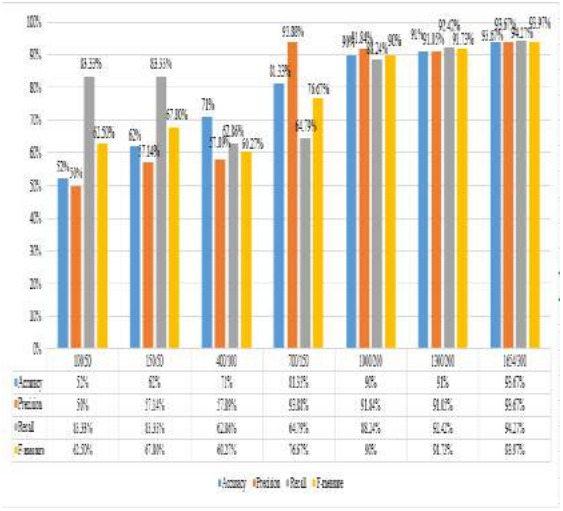


Figure 2. The Experiment Results with Different Dataset

6. Conclusion

YouTube is one of the most popular video-sharing websites, and it is developing rapidly. Its popularity attracts many types of spammers, who post undesired spam comments. A system is classification the YouTube Comment Spam that is Ham is Spam. The system is focused on implementing the classifier for YouTube comment spam classification using Multinomial Naïve Bayes approach with feature selection (TF-IDF) methods. Multinomial Naïve Bayes is a text classification algorithm that is both effective and frequency used. The experimental results show that the more training data utilize and the system performance is better. This can be seen by the performance evaluation result which sufficient. Multinomial Naïve Bayes requires a huge dataset to define the data so it can be accurate. This system used YouTube Comment Spam dataset from UCI machine learning repository. The system performance is evaluated by the percentage of accuracy, recall, precision, f-measure that is shown in section 5.

References

[1] Aqliima Aziz, Cik Feresa Mohd Foozy, Palaniappan Shama, Zurinah Suradi "YouTube Spam Comment Detection Using Suppoet Vector Machine and K-Nearest Nighbor" Indonesioan Journal of Electrical Engineering and Computer Science, Vol .5, No. 3, March 2017.

[2] C.Ramasubramanian, R.Ramya " Effective Pre-Processing Activities in Text Mining using Improved Porter's Stemming Algorithm" International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 12, December 2013.

[3] Hayoung Oh "A YouTube Comments Detection Scheme Using Cascaded Ensemble Machine Learning Model" accepted October 17, 2021.

[4] Lopamudra Dey, Sanjay chakraborty, Anuraag Biswas, Beepa Bose, Sweta Tiwari "Sentiment Analysis of Review Datasets using Naïve Bayes' and K-NN Classifier" Department of Computer Science & Engineering, Heritage Institute of Technology, Kolkata, India, 2016.

[5] https://en.wikipedia.org/wiki/Naive_Bayes_classifier

[6] Mehr-Un-Nisa Manjotho, Tariq Jameel Salfullah Khazada, Liaquat Ali Thebo, Ali Asghar Manjotho "Improving Performance of Mobile SMS Classification Using TF_IDF and Mutinomial Naïve Bayes Classifier" Engineering and Science Technology International Research Journal, Vol.2, No.1, March, 2018.

[7] New Ni Hlaing "Performance Analysis on Mail Classification with Multilayer Perceptron (MLP)" M.C. Sc 2017, University of Computer Studies Yangon.

[8] Paras Sethi, Vaibhav Bhandari, Bhavna Kohli "SMS spam detection and copmparison of various machine learning algorithm" 2017 International Conference on Computing and communication Technologies for Smart Nation (IC3TSM).

[9] Rishpbh Kaushal, Srishty Saha, Payal Bajaj, Ponnurangam Kumaraguru "Kids Tube: Dectection, Characterization and Analysis of Child Unsafe content & Promoters on YouTube" on 04 October 2016.

[10] UCI Machine Learning Repository , YouTube Spam Collection Dataset-[Accessed: 28 Nov 2013].

An Efficient Email Spam Detection Using Multinomial Naïve Bayes Algorithm

Nwe Nwe Aye, Amy Aung
University of Computer Studies (Magway)
Nwenweaye134@gmail.com, amyaung@ucsmgy.edu.mm

Abstract

Nowadays, from business to education, emails are used in almost every field. Emails have two subcategories. A necessary contribution to messaging providing email via the internet. An email has two categories. They are ham email and spam email. Various models and techniques detect spam emails automatically. Spam emails are defined as junk emails and these are unsolicited messages. For email, Spam detection and filtration are important and massive problems. This paper mainly describes an efficient email spam detection using Multinomial Naïve Bayes classification using an SMS spam collection dataset from Kaggle.com. Before preprocessing step, the dataset analyses cleaning the text, removing digits, and case folding steps. The system describes firstly pre-processing steps. These steps are removing punctuation, tokenization, detection of stop-words, and lemmatization. After the preprocessing steps, we calculate the pos of tagging. To detection of spam email messages was approached with sentiment analysis techniques using Multinomial Naïve Bayes classifier in proposed thesis paper. The calculation of MNB is based on a bag of words in sentiment analysis. The experimental results of the proposed method can calculate the Accuracy, F1-Score, Precision, and Recall. The experimental results of the proposed method are detected spam emails with 84.0 accuracy on the Multinomial Naïve Bayes classifier. For actual implementation of this system using python with NLTK.

1. Introduction

Now, a part of routine life, email messages become increasing popularity, and it is essential for many communications. An email has three components. There are header, body and attachment. The header consists of the address of the sender, the address of the recipient and the subject. The message is included in the body. Documents, images, and audio or video files consist of attachments. Usually, spam email is unwanted, uninvited or unprompted messages. There are many techniques for email detection. That mail is spam mail or ham mail. Multinomial Naïve Bayes classifier is a simple and efficient method, and this method requires a little amount of dataset [2]. For sentiment classification and opinion mining applications, SENTIWORDNET 3.0 is used. The result of automatically annotating all the synsets of WORDNET is displayed in SENTIWORDNET 3.0. Each synset a is associated with three numerical scores.

These scores are Pos(a), Neg(a), and Obj(a). The range of each score exist in the interval [0:0; 1:0]. For each synset, the total sum is 1:0. Explanation of above the mean, for all the three synset scores may have nonzero scores.

In this proposed system, the incoming email is been divided into the ham or spam emails with sentiment analysis and MNB. Sentiment classification is used different resources at different levels. To develop sentiment analysis systems, there are many challenging tasks for tokenization, feature selection and stemming, etc. Sentiment analysis is commonly used in several areas that include products, movies, politicians, and companies. Sentiment classification is used different resources at different levels. There are many classification algorithms for sentiment analysis such as SVM, Neural Network, Naïve Bayes, Bayesian Network, and Maximum Entropy in machine learning.

In sentiment classification, the Multinomial Naïve Bayes method has determined more effectiveness and performs better than any other classification method because this method is described simplicity in text classification. This paper describes as follows: section 2, addresses related work. In section 3, we discuss the proposed methodology. The experimental results

will be presented in section 4. Finally, section 5 describes the conclusion.

2. Related Work

In thesis paper, some references are used from previous proposes papers. Alazani Rayan, Ahmed I. Talabani proposed Detection of Email Spam using Natural Language Processing Based Random Forest Approach. The author describes the Random Forest approach (NLP-RF), and this method was found easily in spam emails and temporary emails. After that, it prevents exposing the private data of the users effectively. That is used Multinomial Naïve Bayes classifier [1]. K sai Prasanthi, T Deepika, S Anudeep, M Sai Koushik discussed An Efficient Email Spam Detection using a Support Vector Machine. The author discussed the spam message required for manufacturing and a proficient framework. This system used a Naïve Bayes classifier for spam email detection. [2]. Priyanka Sao, Pro. Kare Prashanth explained E-mail Spam Classification Using Naïve Bayesian Classifier. The author discusses the performance or accuracy of the Naive Bayes Algorithm is Better than others. Naïve Bayes Classifier Algorithm classifier is a statistical classifier, and that will efficiently classify the email messages into spam or ham [3]. Jeremy J. Eberhardt proposed Bayesian Spam Detection. Author analysis the content of spam detection using the Naïve text classification method [4]. Prachi Gupta, Ratnesh Kumar Dubey, Dr Sadhna Mishra proposed Detection Spam Emails/Sms Using Naïve Bayes and Support Vector Machine. The author proposes SMS Spam filtering using a machine learning technique. This technique is based on Naïve Bayes algorithms and a Support vector machine. Performance of accuracy is analyzed on two methods. The accuracy of Naive Bayes has 99.49 and the accuracy of SVM has 86.45. Naïve Bayes method has performed than SVM [5]. SIMRAN GIBSON 1, BIJU ISSAC et: discussed Detecting Spam Email with Machine Learning Optimized with Bio-Inspired Metaheuristic Algorithms. Author research on five algorithms. These algorithms are Support Vector Machine, Multinominal Naïve Bayes, Random Forest, Decision Tree and Multi-Layer Perceptron. In this algorithm, the Multinomial Naïve Bayes (MNB) algorithm has performed better than all the other algorithms [6].

3. Proposed Methodology

This paper proposes an effective email spam detection using an SMS spam collection dataset from Kaggle.com using Multinomial naïve Bayes classifiers. In this paper, the first step is done preprocessing step of a dataset. Preprocessing steps include removing punctuations, tokenization, removing stop words, and lemmatization. The proposed model is described as following Figure 1.

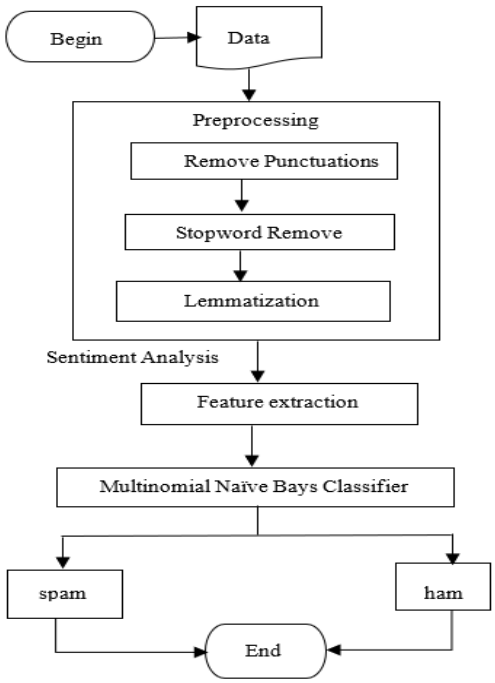


Figure 1. Proposed model

3.1. Preprocessing

To develop textual data quality, pre-processing is an important step in NLP. Data Preprocessing is a process technology that is not only used to low- quality data into high-quality data but also reduces of dimension a dataset. For this research, all pre-processing steps were applied to the SMS spam collection dataset from Kaggle.com. In this proposed system, before preprocessing, the SMS spam collection dataset has analyzed the cleaning of the text, removing digits, and case folding. Preprocessing steps are included following steps. There is removing punctuation, removal of stop words, lemmatization, and tokenization. In the first step,

we preprocessed the data displayed by removing the different punctuations in a sentence. The second step is the removal of stop words. Stopword removal is the process of removing, that does not remove important words and often appears on documents. To desire an effective classification process, it can eliminate stop words such as “which", "the", "and". After that, Lemmatization is the process that is grouping the different inflected forms of a word, it can be analyzed as a single item. An example of lemmatization is:

“Better” ⟶ “Good”

Base Word: “Good”

Another example of the use of lemmatization is:

applied ⟶ apply

saw ⟶ see

Stemming is a process of reducing the size of the corpus by converting words to their root word form. Stemming is removing the suffix from a word and reducing it to its root word. Examples of Stemming are:

“plays” ⟶ “play”

“played” ⟶ “play”

“player” ⟶ “play”

“playing” ⟶ “play”

Root Word: play

After the preprocessing step for the dataset, an example of the pos of tagging is calculated in the sentence. In a sentence, ‘ok’ (JJ), ‘joking’(NN), ‘ oni’ (RB), and ‘wkly’ (VBD), might convey a different meaning. In different NLP applications, one of the most commonly used preprocessing steps are removal of stop words, it is removing not desired words on documents such as "which", "the", "and". Text tokenization and segmentation are the first part of text classification. Examples of SMS spam collection datasets in Kaggle.com are described in table 1.

Table 1. Spam Dataset

Text Review	Label
Ok lar... Joking wif u oni...	ham
Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005.	Spam
U dun say so early hor... U c already then say...	ham
Did you hear about the new \Divorce Barbie\"? It comes with all of Ken's stuff!"	ham

Table 2. Preprocessing process

No	Process	Input	Output
1	Remove Punctuation	Ok lar... Joking wif u oni... Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005.	ok lar joking wif u oni free entry in a wkly comp to win fa cup final tkts
2	Stopword Removal	ok lar joking wif u oni free entry in a wkly comp to win fa cup final tkts	ok lar joking wifi free entry comp win cup final
3	Lemmati-zation	ok lar joking wifi free entry comp win cup final	ok lar joking wifi free entry comp win cup final

3.2. Feature Extraction

In a machine learning algorithm, to extract features, a bag of the word is a way with text used for modelling. A Bag of words is included in the text of a sentence or document that is also a representation. All the vocabulary contains in each sentence. In the sentence, calculating frequency occurrence or the number of occurrences of words is described. Feature selection is the finding and selecting processes of more useful features in a dataset. A Bag of words is a way to extract features with text used for modelling like a machine learning algorithm. In a machine learning algorithm, to extract features, a bag of the words is a way with text used for modelling. The selected features will be stored in the bag of words after pre-processing. An example of a Bag of words in a text document is despite in following table 3.

Text document1: This is a self-published book

Text document 2: I spent an evening with the book

Table 3. The example of a bag of words

	This	is	a	self-published	book	I	spent	an	evening	with	the
Text Doc 1:	1	1	1	1	1	0	0	0	0	0	0
Text Doc 2:	0	0	0	0	1	1	1	1	1	1	1

By using the feature selection method, it can produce better prediction accuracy and better efficiency.

3.3. Multinomial Naïve bayes

The Multinomial Naive Bayes algorithm is mostly used in Natural Language Processing (NLP) and it is a probabilistic method. In many texts classification, Multinomial Naive Bayes (MNB) is widely used from a Bayesian approach. To solve text classification problems, MNB is an extension of the Naïve Bayes considered. Naïve Bayes classifiers are based on the Bayes theorem. Bayes theorem calculates probability $\beta(\mu|\infty)$ where μ is spam or ham and ∞ is ∞_1, ∞_2, ∞_3 ... ∞_n from an upload email. Naïve Bayes algorithm calculate as follow:

$\beta(\mu|\infty) = \boldsymbol{\beta(\alpha|\mu) * \beta(\mu) / \beta(\alpha)}$

$\beta(\mu|\infty)$ = posterior probability

$\boldsymbol{\beta(\alpha)}$= prior probability

$\boldsymbol{\beta(\alpha|\mu)}$ = likelihood

The probability of a μ is calculated from the bag of words.

Equation for Multinomial Naïve Bayes can be denoted as Follow:

$$P(g\backslash n) \propto P(g) \prod_{1 \le k \le nd}$$

where;

nd= number of tokens

n=number of emails

P (*tk* |*g*) is the conditional probability for MNB and is calculated as

$$P(t_k|g) = \frac{count(t_k|g) + 1}{count(t_p) + |a|}$$

tk = spam term in email

1 and | a | = unique word (constant)

Spam classification is described using the Multinomial Naïve Bayes as following algorithm.

Algorithm

Br s = Training subset of "s" and

Hr s = Testing subset

Br= Training and Hr= Testing

P (tk |g) = conditional probability

Initial= input variables;

t =number of documents;

s = datapoints;

y =desire inputs;

```
for i = 0;  i < Br s; i=i+1 do
    if (i,y) = Spam then
            i = Spam;
    else
            i = Ham;
for testing
do
    for j in sd do
        s_test and y_test = testing size;
        s_train and y_train = training size;
        for i = 0; i < Hr s;
    i ++
            Calculate _ P (tk |g);
            Calculate the Accuracy; return tk;
return tk;
```

Calculation and classification will be explained using the MNB algorithm. Examples of datasets were depicted as the following table 4.

Table 4. The Example of MNB calculation

Doc	Message	Class
Training (Doc1)	ok lar joking wif	ham
Training (Doc2)	free entry wkly comp win facup final may	spam
Testing (Doc3)	free entry joking |	spam

The calculation can be seen in prior probability:
P(spam) =1/2
P (ham) =1/2
Calculate of Likelihood probability:

$P(\text{joking}|\text{ham}) = \frac{1+1}{4+12} = \frac{2}{16} = 0.125$

$P(\text{wif}||\text{ham}) = \frac{1+1}{4+12} = \frac{2}{16} = 0.125$

P (free |ham) = $\frac{0+1}{4+12} = \frac{1}{16}$ = 0.0625

P (entry |ham) = $\frac{0+1}{4+12} = \frac{1}{16}$ = 0.0625

P (free| spam) = $\frac{1+1}{8+12} = \frac{2}{20}$ =0.1

P (entry| spam) = $\frac{1+1}{8+12} = \frac{2}{20}$ =0.1

P (joking |spam) = $\frac{0+1}{8+12} = \frac{1}{20}$ =0.05

Calculate of Posterior probability

P(spam|Doc3) = ½*0.1*0.1*0.05 =0.00025

P(ham|Doc3)=½*0.625*0.0625*0.125= 0.000244

Above the calculation, the class of spam probability value is greater than the class of ham probability value, so document 3 becomes the spam class.

4. Experimental Results

The experimental result in the analysis is done by using an intel i7-7500U CPU with a 2.7 GHz processor along with 4 GB of RAM and Python programming language along with Natural Language Toolkit (NLTK). The Proposed system runs five experiment times and chooses the average value as the result. The detection of spam emails can be evaluated by Multinomial Naïve Bays Algorithm. Confusion Matrix is used to detect the emails for models. Below table 5 represents a confusion matrix:

Table 5. Confusion matrix

	Spam	Ham
Spam	True Positive (PN)	False Positive (PF)
Ham	False Negative (NF)	True Negative (PT)

Performance evolution of spam detection can be described by using Accuracy, F-Score, Cross-entropy, Recall, and Precision. In this experiment, the highest accuracy for detecting the emails correctly is ham and spam. Proposed thesis paper has used dataset of SMS messages from the Kaggle.com. The dataset is labeled with two classes such as spam and ham. It contains a total of 5,573 messages, of which 4,831 are ham, and 742 are spam message. As the result, 85-15 proportions are better result for accuracy. So, proposed thesis paper considers 85-15% of training and test messages by different classifiers. Accuracy shows the following equation.

Accuracy = $\frac{PN+PT}{PN+PT+PF+NF}$

Where-, PN is True Positive, PT is True Negative, PF is False Positive, and NF is False Negative.

Precision:

The Precision calculate the total number of emails. That emails are predicted as total number of emails positive prediction.

Precision can be calculated as follow:

P = $\frac{PN}{PN+PF}$

Recall:

The measurement of recall describes the calculation of predicted spam from the total number of spams emails.

This measurement was provided Recall.

Recall can be calculated as follow:

R = $\frac{PN}{PN+NF}$

A test's accuracy is measured F1- Score. It depends on the Precision and Recall of the test result. F1-Score can be calculated as follow:

F1-Measure= [2 * (Precision * Recall) / (Precision + Recall)

Table 6. Evolution results for divided dataset

No	Divided Dataset	Accuracy	Precision	Recall	F-Score
1.	65-35	80%	80%	92%	85%
2.	75-25	81%	82%	94%	88%
3	85-15	84%	85%	97%	90%

Proposed system uses example of total dataset 100. From above table, there are divide three dataset such as 65-35, 75-25, and 85-15. In evolution results, 85-15 dataset has received better accuracy result than another divided

dataset. Performance evolution of divided dataset is depicted as following figure 2.

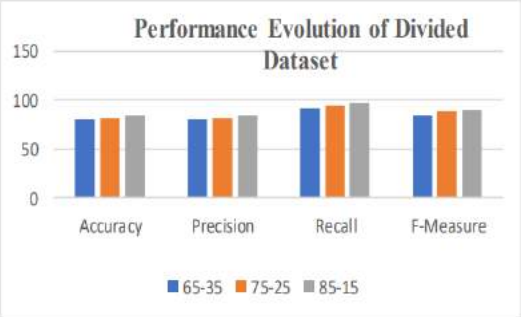


Figure 2. Performance evolution result

5. Conclusion

This paper proposes an efficient email spam detection of SMS spam collection dataset from Kaggle.com using Multinomial naïve Bayes classifiers. Multinomial naïve bayes classifiers use multinomial distribution for each one of the features on data. The calculation of MNB is based on a bag of words in sentiment analysis. The proposed system calculated the Accuracy, F1-Score, Precision and Recall. Using the Multinomial Naïve Bayes Algorithm is able to detect spam emails and ham emails, and that produces an accuracy of 84.00% for spam emails. The thesis paper clearly provides spam detection of text and reduce execution time. Thus, this is an efficient method. For future work, we will try to compare with detection of spam emails using Naïve Bayes and Support Vector Machine results.

References

[1] Alanazi, Rayan & Taloba, Ahmed. (2021). "Detection of Email Spam using Natural Language Processing Based Random Forest Approach." 10.21203/rs.3.rs-921426/v1.

[2] K sai Prasanthi, T Deepika, S Anudeep, M Sai Koushik," An Efficient Email Spam Detection using Support Vector Machine", International Journal of Innovation Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-9 Issue-2, December 2019.

[3] Priyanka Sao, Pro. Kare Prashanthi, "E-mail Spam Classification Using Naïve Bayesian Classifier" IJSDR1906001, International Journal of Scientific Development and Research (IJSDR) www.ijsdr.org.

[4] Jeremy J. Eberhardt, "Bayesian Spam Detection", University of Minnesota, Morris Undergraduate Journal Volume 2 Issue 1 Article 2 March 2015.

[5] Prachi Gupta, Ratnesh Kumar Dubey, Dr. Sadhna Mishra," Detection Spam Emails/Sms Using Naïve Bayes And Support Vector Machine", NTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH VOLUME 8, ISSUE 11, NOVEMBER 2019 ISSN 2277-8616 1 IJSTR©2019 www.ijstr.org

[6] SIMRAN GIBSON 1, BIJU ISSAC 1, (Senior Member, IEEE), LI ZHANG 1, (Senior Member, IEEE), AND
SEIBU MARY JACOB2, (Member, IEEE), "Detecting Spam Email With Machine Learning Optimized With Bio-Inspired Metaheuristic Algorithms", Digital Object Identifier 10.1109/ACCESS.2020.3030751.

Frequent Pattern Mining on Online Judge Education Web Log Data Using Eclat Algorithm

Poe Myat Zin, Daw Aye Aye Maw
University of Computer Studies, Yangon
poemyatzin1995@gmail.com,ayeayemaw@ucsy.edu.mm

Abstract

The internet has been increasingly widespread in recent years, and it is often regarded as one of the most significant inventions. Web mining is a strong technique in data mining and an emerging study field in which numerous strategies have been employed to tackle various difficulties linked to analysing the pattern of web usage of users available in the web server. Web Usage Mining is a kind of web mining that extracts web users' behaviour from log files. Data pre-processing and pattern discovery are the steps of the proposed system. Noise and contaminants may be presented in raw web log data. Noise should be eliminated utilizing some data pre-processing techniques. Data cleaning, user identification, page identification and session identification algorithms are used in pre-processing stage of the proposed system. In the purposed system, the Eclat algorithm is used to identify frequent patterns in web log data. The proposed system tests the data set from Online Judge Web Log Data of one of the University. Therefore, the most interesting and visiting web pages can be collected in the server.

Keywords: Preprocessing, User Identification, Session Identification, Frequent Pattern

1. Introduction

The World Wide Web is one of the primary data sources for millions of people all over the world, obtaining information from a huge amount of data, including advertisements, consumer information, e-commerce, education, financial management, government, news, and a variety of other services. Web mining is a type of data mining technique that analyzes web contents, structure, and usages to automatically identify and extract information. Utilizing web content mining, meaningful data can be gained from web document content. Finding structure-related information from the web is called web structure mining. Web usage mining is used to discover interesting patterns from web log data [10]. By using association rules, relationships between groups of users who have similar interests can be defined. By providing linkages between pages that are frequently visited together, this information can be utilized for restructuring the website. Association rule is a methodology that identifies specified types of data associations [7]. Before generating frequent pattern, firstly we need to preprocess raw data. First step of preprocessing, data cleaning means removing the non-access of log records or useless requests of log file. In the proposed system, user identification, page identification and session identification are used as the pre-processing step. Moreover, in this system, Eclat algorithm is used to identify frequent patterns in web log data. In this paper, web server log data from one of the University is used to find frequent pattern.

This paper is organized as follows: the related work is described in section 2. The section 3 describes design of the system. The section 4 shows the nature of web log and the phases of the proposed system. The section 5 shows experimental result and finally section 6 describes the conclusion and future work.

2. Related Work

In this paper, they addressed the problem of Web usage mining, i.e., finding relationships between data stored, user frequent patterns from one or more Web servers were mined and paired particular attention to the interesting new patterns. Apriori algorithm was adapted for matching interesting new patterns and measuring of interesting patterns by applying support and confidence, to this particular context. [1]. Web mining describes the process of Web data mining in detail: source data collection, data pre-

processing, pattern discovery, pattern analysis and cluster analysis. Servers are able to collect and store huge amount of data by the help of advanced information technologies. To study the customers' behaviour was the main purpose of this paper and they used the Web mining techniques and tested its application in e-commerce [3]. To discover the hidden knowledge and to identify the behaviour of the user on the web, this paper used the web data sources. The web access log file collected from the organization was used to examine the user pattern by using the updated Web Log Expert application. This enhanced tool attempted to carry out web mining in a domain-agnostic manner. There were three parts to this algorithm: 1. Given an input item, extract a set of IP addresses and visor lists and rank them by comparison. 2. Identifying and summarizing the competitive data that specifies the organization's strength and 3. Identifying and summarizing the domains in which the specified entity participates. The Web Log Expert tool had been used to implement the entire analytic process. The findings of the experiment made it easier to traverse the website and improve its design architecture [2]. In this paper, they would learn how to use R algorithms to find common patterns, association rules, and correlation rules. Then, using benchmark data, they would assess all of these ways to see how fascinating the common patterns and rules are. They came to the conclusion that the APriori algorithm, which was the first efficient mining method for mining common patterns, is the source of many variants [7]. This system aimed to utilize an intelligent technique to deliver personalized web service for accessing related web pages more rapidly and effectively, so that it could be determined which web pages the user is most likely to visit in the future. For forecasting user behaviour, this system incorporated two intelligence algorithms: FP Growth and Eclat. These methods solved the existing system's time and space problems. In addition to the frequent pages pattern, Direct and Indirect Association Rules were developed, and a ranking is assigned to pages based on these rules, which would aid the recommendation engine in recommending related search pages.

3. Overview Design of Proposed System

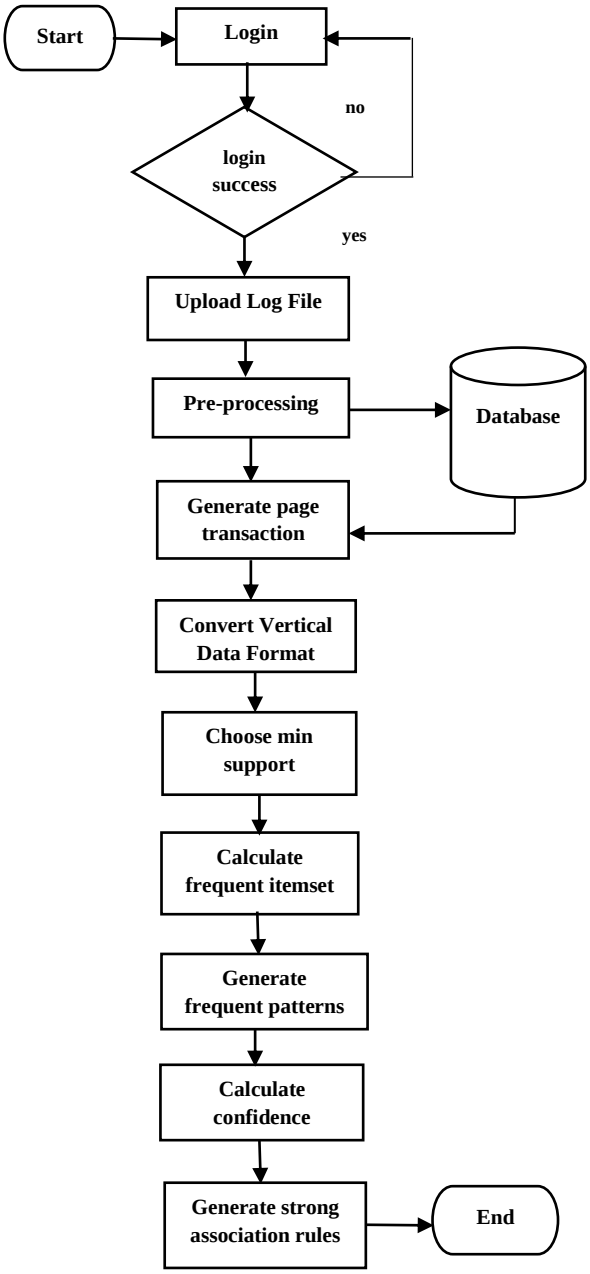


Figure 1. Flow Diagram of the Proposed System

In overview design of the system (Figure 1), the main process is to generate frequent pattern means that association rules. Firstly, the user input web log file then the system checks log file and removes unnecessary data as data pre-processing algorithms. Secondly, the system stores clean data to database and selects data to change the vertical data format and to generate frequent patterns using ECLAT Algorithm. Frequent pattern mining also used to find information like set of pages repeatedly accessed together by users. Thirdly, the result from

mining frequent pattern is measured performance which means measures the confidence on frequent pattern. Finally, it generates the output that is strong association rules.

4. Phases of the Proposed System

4.1. Web Log Data Nature

In this proposed system, web server log file is taken as a data source which is one of the Online Judge web log data. This normally includes client's IP address, the request date and time and the requested page and HTT code and status code sent by the server. Our application is implementing for Online Judge website of one of the University that intended to know the most visited pages and user navigational pattern. Firstly, data pre-processing is done on log file and then categorized the file for generating user navigation pattern.

Table 1. Web log Data Format

No	Web log entries	Description
1	10.128.2.1	IP address of the client and make request of the server
2	29/Nov/2017 06:59:03	The request date and time
3	GET /home.php	The requested page and GET is the method used
4	HTTP/1.1	The protocol used by client
5	200	This is status code sent by server

4.2. Pre-processing

Pre-processing is required because log files contain noisy, irrelevant, and unambiguous data that may affect the mining process's results. Before applying any web mining algorithm, it is critical to filter and organize appropriate data. The goal of data pre-processing is to improve data quality and accuracy during the mining process [9].

4.2.1. Data Cleaning

In the proposed methodology, we eliminate log file records that are not relevant for the study in order to improve the quality of the data. The goal of data cleaning is to have a clear picture of the needs or behaviors of online users; as a result, it was necessary to remove files with suffixes like jpeg, jpg, gif, cgi, etc., as well as error codes like 401, 404, which are irrelevant and useless for mining. The following algorithm is used as data cleaning in the proposed system.

Data Cleaning Algorithm

Step 1: Read LogRecord from Web Server Log

Step 2: **IF** ((LogRecord.status== (200,302)

AND((LogRecord.method='GET'))

AND((LogRecord.stem (php,css,js)))

THEN Insert LogRecord into LogDatabase

End of IF Condition

Step 3: Repeat the above two steps until endof (Web Server Log File)

Step 4: Stop the process.

Figure 2. Data Cleaning Algorithm

4.2.2. User Identification

User identification means the identifying individual users by using their IP address. If there is new IP address then there is a new user. The following algorithm is used as the user identification.

User Identification Algorithm

Step 1: Initialize IPList =0; UsersList = 0;

No-of-users = 0;

Step 2: Read Record from LogDatabase

Step 3: **IF** Record.IP address is not in IPList

THEN Add new Record.Ip address into IPList increment count of No-of-usersinsert new user into UserList

End of IF

Step 4: Repeat the above steps 2 to 3 until endof(Log Database)

Figure 3. User Identification Algorithm

4.2.3. Page Identification

The page identification step identifies individual page by using their URL name. If there is new URL, there will be new page. In this proposed system, it will define each URL name

as new page. The page identification algorithm used in the system is shown in the following.

Page Identification Algorithm

Step 1: Initialize PageList =0; No-of-Page=0;

Step 2: Read Record from LogDatabase

Step 3: **IF** Record.URL name (Web Page) is not in PageList

THEN

Assign Page Id for new Web Page

Insert Page id into PageList

increment count of No-of-Page

Step 4: Repeat the above steps 2 to 3 until endof(Log Database)

Step 5: Stop the process.

Figure 4. Page Identification Algorithm

4.2.4. Session Identification

Session means the duration of the user's spent on a web page. Session identification is used to divide the page access of each user into different sessions. In this paper session ids are created by using depends on time, which is calculated by the difference between two-time stamps of the same user. If there is a new user, there is a new session. If the time between page request exceeds a certain limit, it is assumed that the user is starting a new session. The session identification algorithm used in the system is shown.

Step 1: Initialize SessionList = 0;

No-of-Sessions = 0;

Step 2: Read LogRecord from User Database

Step3: **IF** LogRecord.UserID. equal

(NewLogRecord.UserID) **AND**

(NewLogRecord.timetaken-LogRecord.timetaken>5))

THEN

Get Url address of corresponding Session

Insert into SessionList

increment No-of-Sessions

End of IF

Step 4: Repeat the above steps 2 and 3 untill endof (Logrecord)

Step 5: End of process.

Figure 5. Session Identification Algorithm

4.3. Pattern Discovery Phase

In this pattern discovery, choosing necessary user patterns should be done after every pre-processing stage. The process of discovering patterns is also known as frequent pattern mining. After pre-processing phase, the pattern discovery method should be applied. After data pre-processing from server log file, we get the access page transaction lists of each user. And then we will find the user's behaviour or access page links from these transaction lists using one of the frequent itemset mining algorithm. A lot of algorithms have already been designed for generating frequent itemset. One of the algorithms that we have used is Eclat algorithm.

4.4. Eclat Algorithm

Frequent itemset are those items which are frequently occurring in the transaction. Eclat algorithm uses vertical dataset and bottom-up approach for searching items in database [4]. Eclat algorithm finds the items from bottom like depth first search. Eclat algorithm is very simple algorithm to find the frequent item sets. Apriori is a very basic algorithm. But it takes a lot of time for calculations of frequent itemsets. In this algorithm, we need to calculate support and confidence. We need to scan to database again and again. So, FP growth algorithm has been developed. FP growth stands for frequent pattern growth. In this algorithm, database scanning is required two times [6]. Time consumption is more. To remove these limitations, Eclat algorithm was developed. Eclat needs to scan the database only once. All the data is stored in vertical form. Bottom-up approach is used for searching items in the database.

Input: F_k = {I1, I2, ..., In} // cluster of frequent k-itemsets.

Output: Frequent l-itemsets.

Bottom-Up (F_k) {

1. for all $I_i \in F_k$
2. $F_{k+1} = \emptyset$;
3. for all $I_j \in F_k$, $i < j$
4. $N = Ii \cap Ij$;
5. if N .sup>= min_sup then
6. $F_{k+1} = F_{k+1} \cup N$;
7. end
8. end
9. end
10. if $F_{k+1} \neq \emptyset$ then
11. Bottom-Up (F_{k+1});
12. end
13.}

Figure 6. ECLAT Algorithm

In Bottom-Up (F_k), for loop is defined in which all items belong to database F_k under first

step. In step2, take F_{k+1} as empty database. In step3, check that item exists in the database F_k or not. If item exists in the database F_k then calculate support of item in step4 where support means how many times item occurs in database F_k. In step5, compare support of items individually with minimum decided support. In step6, put those items whose support is more than then minimum support in database F_{k+1}. In step7, check that database F_{k+1} is empty or not. If it is not empty then start the same procedure for another database.

5. Experimental Result

In this table describe that we found frequent itemset using ECLAT algorithm. In this proposed system, we mainly find the frequent itemset on web log means that find the most visiting page links or user behaviour patterns. When finding frequent itemset, it needs to choose min-support. According this system, the user can choose min-support 2 or 3 or 4. As this result, it displays frequent itemset according to each min-support and total records or amount of frequent itemset.

Table 2. Frequent itemset of each min-support

Frequent itemset	Min _ support	Total Frequent Itemset
P00,P02,P03,P11,P12	2	204
P00,P02,P05,P06,P08		
……		
P08,P11,P12,P13,P14		
P00,P02,P05,P06,P09	3	46
P00,P02,P05,P06,P11		
P00,P02,P05,P06,P14		
……..		
P06,P08,P09,P11,P14		
P00,P02,P05,P11,P14	4	9
P00,P02,P06,P09,P11		
P02,P05,P06,P09,P11		
……		
P05,P06,P09,P11,P14		

5.1. Algorithm for Confidence

The strength of a given association rule is measured by confidence. Confidence refers to the amount of times a given rule turns out to be true in practice. By traversing the frequent itemset and computing associated confidence levels, explicit association rules can be generated quickly. Confidence is the proportion of the transactions containing item A which also contains item B, and is calculated as

$$\text{Confidence } (A \Rightarrow B) = \frac{\text{Support } (A \cup B)}{\text{Support } (A)} \quad (1)$$

In table (3) describe strong association rules which calculate by using confidence. When finding confidence, the system calculates for each the final or fifth frequent itemset. After calculation for frequent itemset, we can see pages strongly associated each other.

Table 3. Calculation Confidence

Frequent itemset	Support (AUB)	Support (A)	Confidence
P00 =>P02,P05,P11,P14	4	26	15.38 %
P02 =>P00,P05,P11,P14	4	52	7.69 %
……	……..	……..	……..
P00,P05 =>P02,P11,P14	4	4	100%
P00,P11 =>P02,P05,P14	4	8	80%
……	…….	……..	……..
P00,P02,P05 =>P11,P14	4	4	100.00 %
P00,P02,P11 =>P05,P14	4	7	57.14 %
……	…….	……..	……..
P00,P02,P05,P11 =>P14	4	4	100.00 %
P00,P05,P11,P14 =>P02	4	4	100.00 %

6. Conclusion and Future Work

In this proposed system, to gain user interested pages or user behavior is the main target. In data pre-processing phase, data cleaning and defining the user pattern is most important fact. This system intends to help data cleaning and categorizing the web site for understanding the user interest and for improving the satisfactory of users' requirements. The contribution of the paper is to introduce the process of web log mining, and to show how frequent pattern discovery tasks can be applied on the web log data in order to obtain useful information about the user's navigation behaviour. In the future, we plan to use this proposed system on different web sites and to test the various web logs. It is also need to add the security features.

7. Acknowledgements

We sincerely thank to our supervisor Daw Aye Aye Maw for her useful discussion and valuable suggestion.

References

[1] S.VijayaKumar, "*Frequent Pattern Mining in Web Log Data using Apriori Algorithm*", International Journal of Emerging Engineering Research and Technology Volume 3, Issue 10, October 2015.

[2] K. Dharmarajan1 and M.A. Dorairangaswamy, "Mining The Customer Behavior Using Web Usage Mining In E-Commerce" Indian Journal of Science and Technology, Vol 9(42), Nov 2016.

[3] Mahendra Pratap Yadav, Mhd Feeroz, Vinod Kumar Yadav,"Mining The Customer Behavior Using Web Usage Mining In E-Commerce" IEEE 2012.

[4] Bina kotiyal,ankit kumar. Bhaskar pant, R.H. goudar, shivaji chauhan and sonam junee, "*User behavior analysis in web log through comparative study of Eclat and Apriori*" IEEE 2012

[5] Asfiya Khatoon1, Kuldeep Jaiswal, "*Web Page Ranking using Web Usage Mining*", International Journal of Advanced Research in Computer and Communication Engineering ISO 3297:2007 Certified Vol. 6, Issue 4, April 2017

[6] R. Krishnamoorthi, K.R. Suneetha, "*Extracting users pattern from web log data using decision tree and association rule*", Int. J. Business Performance and Supply Chain Modelling, Vol. 2, No. 2, 2010

[7] Yo unghee Kim, Won Young Kim and Ungmo Kim "*Mining frequent item sets with normalized weight in continuous data streams*". Journal of information processing systems. 2010.

[8] Dr. S. Vijayarani, Ms. P. Sathya "*An Efficient Algorithm for Mining Frequent Items in Data Streams*" in International Journal of Innovative Research in Computer and Communication Engineering Vol. 1, Issue 3, May 2013.

[9] K.sudheer reddy, m. kantha reddy,v,sitaramula,"*An effective data preprocessing method for web usage mining*" IEEE conference 2013.

[10] http://www.ijcce.org/papers/128-B047.pdf

Shopping Assistant System using Multi-Attribute Utility Theory (MAUT)

Aye Myint Khine, Si Si Mar Win
University of Computer Studies, Yangon
ayemyintkhine1994@gmail.com, sisimarwin@ucsy.edu.mm

Abstract

With the widespread use of the Internet, electronic commerce (e-commerce) becomes very popular and one of the most important fields. In today's fast-changing era, it is extremely important to be able to respond to the needs of customers or buyers in the most effective and timely manner. So, most business websites make it easy to help both the customer needs and business requirements. This paper presents a shopping assistant system for searching laptop computers matched with user preferences in a variety of several alternatives from seventeen computer shops. An attractive theoretical multi-attribute utility theory (MAUT) is used in order to evaluate which laptop(s) performs best match with user preferences.

Keyword: e-commerce, user preferences, MAUT

1. Introduction

Electronic commerce or e-commerce is the activity of buying or selling of products over the Internet. One of the popular examples of e-commerce is online shopping that is the activity of buying goods or services online. Online shopping has grown in popularity over the years because people find it convenient and easy to buy various items comfortably from their office or home and eliminates the need to wait in long lines or search from store to store for a particular item. In shopping assistant systems, the buyers have to provide their product preferences through product attributes. Based on this, the relevant products are displayed to the buyers. The system will develop the design of searching relevant products based on users' desired product characteristics and displaying the best match product(s) to the users by comparing product information using MAUT.

The rest of the paper is organized as follow: Section 2 presents the closely related work to this paper. Section 3 explains the methodology that we used in shopping assistant system. Section 4 describes the architecture of the proposed system. Section 5 expresses the calculation details, the results of the experiments performed. Section 6 describes the conclusion and the future work of the proposed system.

2. Related Work

Multi agents-based model for Shopping Assistant system was described in 2009. Their system proposed shopping assistant agents as sale representatives for both shoppers and stores to negotiate for desired products based on shopper preferences. The main agent used in their system was buyer agent in order to help customers or buyers in finding the products as they desire. Buyer agent will communicate with other seller agents to complete the product selection phase [1]. The other authors also implemented the Agent Based Online Shopping Assistant System as a shopping assistant system using multi-agent technology. In buying products with multiple attributes and multiple prices, the users might have difficulty in finding the right product they want within their budget. Their shopping assistant system can find the set of relevant items within the buyers' budgets [2]. By using online shopping system over internet, users can save their time and effort. So, this paper presents the shopping assistant system for both customers and shops by using effective multi-attribute utility theory.

3. Background Theory

This system implements shopping assistant system for the individuals who wish to choose a recent configuration Laptop computer with respect to his budget, using MAUT as the decision-making method.

3.1. Shopping Assistant

Shopping assistant is the process of serving customers or buyers. The capabilities of shopping assistant include: (1) helping the buyers decide what product to buy, e.g., by listing what products of certain type are available, (2) finding the specifications and reviews of them, (3) making recommendations and (4) comparison shopping to find the best price for the desired product.

3.2. Multi-Attribute Utility Theory

Multi-Attribute Utility Theory (MAUT) is a normative method for the evaluation of items involving multiple competing attributes. MAUT is one of the Multiple Criteria Decision Making (MCDM) methods. It was introduced by Fishbum (1965, 1970), Keeney (1969, 1971, 1973) and Raiffa (1969) who proposed a decision-making technique designed for taking decisions under risk. It is a decision-making method used when the decision maker has to take multiple objectives into account. It is also an evaluation method used by many systems for evaluating the interests and preferences of the users and supporting them in configuring the desired product(s). It evaluates not just one user's but several users' preferences and it computes the degree of interest (or utility) of the products regarding the user preferences.

MAUT is especially a structured methodology designed to handle the tradeoffs among incomparable and conflicting multiple objectives, captured by multiple attributes. Example, better performance and lower price of computer. The basic goal of MAUT is to substitute information with an arbitrary measure called utilities and it evaluates the best match product(s) by normalized utility functions for attributes and by weights for expressing the relative importance of attributes. This method is recommended when prospective alternatives must be evaluated to determine which alternative(s) performs best.

The aim of MAUT is to help the decision makers who face very complex problems choosing between the different possible alternatives, taking into account their preferences. MAUT has been widely used in situations where the decision making depends on multiple factors and the utility calculation of decision alternatives is based on multiple attributes. Therefore, MAUT has been extensively used for decision making of economic and financial like the application area of e-commerce.

There are different utility functions in multi-attribute utility theory (MAUT). But most commonly used MAUT functions are: additive utility function and multiplicative utility function.

3.3. Additive Utility Function

In additive utility function of MAUT, the overall utility of an alternative is calculated by the weighted sums of its measures (i.e. evaluation criteria). It is described by the following equation:

$$U(x_1, \dots, x_n) = \sum_{i=1}^{n} k_i U_i(x_i)$$

where,

$U(x_1,\dots,x_n)$ = the overall utility score of each alternative

$U_i(x_i)$ = the utility function of the ith attribute

k_i = the weight of the i^{th} attribute

$0 \le U(x_1,\dots,x_n), U_i(x_i) \le 1$

$k_1 + \dots + k_n = 1$

3.4. Multiplicative Utility Function

In multiplicative utility function of MAUT, the overall utility of an alternative is achieved by multiplication of the utility factors for all attributes. It is described by the following equation:

$$U(x_1, \dots, x_n) = I^{-1}\{\prod_{i=1}^{n}[1 + Ik_iU_i(x_i)] - 1\}$$

Where,

$U(x_1,\dots,x_n)$ = the overall utility score of each alternative

$U_i(x_i)$ = the utility function of the ith attribute

k_i = the weight of the ith attribute

I = a scaling constant

$0 \le U(x_1,\dots,x_n)$, $U_i(x_i) \le 1$

If $k_1 + \dots + k_n < 1$ then $I > 0$

If $k_1 + \dots + k_n > 1$ then $-1 < I < 0$

Among these functions, additive utility function will be used to find the best match with user requirements in this system because it ignores attribute interaction which is a situation

in two or more attributes that act upon one another.

3.5. Processes of MAUT

The processes of multi-attribute utility theory (MAUT) involve these stages:

- Identify the attributes, which collectively describe the overall utility of all relevant decision options.
- Identify the alternatives or options to calculate.
- Weight the attributes in terms of their importance.
- Transform the attribute scores, measured in different units, into commensurate (similar measurable standard) units.
- Define aggregate utility function, which combines the transformed scores and weights to measure the overall utility of each option.

4. The Proposed System Architecture

This system is implemented as a shopping assistant system. In this system, the users have to provide their desired characteristics (preferences) to the system. Then the list of relevant products based on user preferences are displayed to the users. Among them, the best match product(s) are evaluated using multi-attribute utility theory (MAUT). Therefore, the users can reduce searching time for shop by shop and save effort to find the right product. The overall architecture of the proposed shopping assistant system is described in Figure 1.

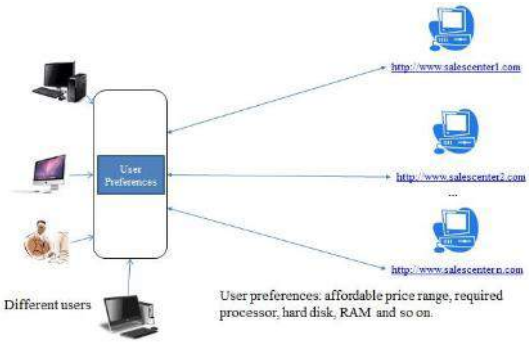


Figure 1. Overall System Architecture

4.1. Implementation of the System

The proposed system provides searching and retrieving laptop computers based on user preferences and consists of the following processes. The users have to enter the laptop attributes (user preferences) to the system. The system surfs around the seventeen computer shops. When it gets the relevant laptops with user desired characteristics from the shops, compares the laptop information using MAUT in order to determine which laptop(s) is best match with user preferences. Finally, the best match laptop(s) is displayed to the users. The system flow diagram is described in Figure 2.

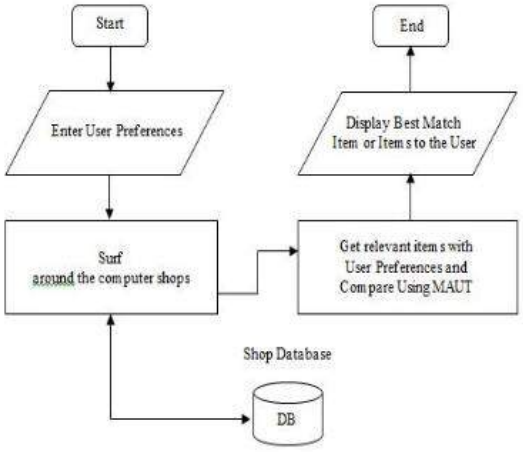


Figure 2. System Flow Diagram

5. Experimental Setting and Results

In MAUT, weight values and utility values of the attributes are needed to define.

5.1. Defining Weight Values of the Attributes

Rank Sum Weight Method is used to define the weight value of each attribute. According to Rank Sum Weight Method,

$$k_i = \frac{2(n+1-r_i)}{n(n+1)}$$

Where, k_i = Weight of each attribute
n = Number of the attributes
r_i = Rank of each attribute

There are nine attributes specified in this system. They are Processor, Generation, RAM, Hard Disk, Graphics Card, Screen Size, Battery, Price and Brand. Since it is difficult to decide which Brand is good or not, only eight attributes

except Brand attribute are used in calculation. Moreover, the relevant laptops matched with user preferences are searched from seventeen computer shops, ADT, Asia Tech, Chan Myay, Citicom, EASTERN, IDEALINK, KING EMPIRE, King Power, KSW, KMD, Life Mark, MEDIA LINK, RIT, Technoland, Tha Pyay Pwint, TOP POWER and Unique. So, the number of attributes, n is 8 and the detailed information of the attributes and their types are shown in Table 1.

Table 1. Defining Attributes and Attribute Types

Attributes	Attribute Types
Processor	Core i7, Core i5, Core i3, Pentium, Celeron
Generation	7th Gen, 6th Gen, -
RAM	8GB, 4GB, 2GB
Hard Disk	1TB, 500GB
Graphics Card	4GB, 2GB, 1GB, -
Screen Size	15.6", 14", 13.3", 11.6"
Battery	6 Cells, 4 Cells, 3 Cells, 2 Cells
Price	Numeric Price Value
Brand	Asus, Acer, Dell, Hp, Lenovo, Msi

5.2. Defining Utility Values of Each Attribute

The utility values of each attribute are needed to define. Processor, Generation, RAM, Hard Disk, Graphics Card, Screen Size and Battery are qualitative attributes. So, the utility value of such each qualitative attribute is defined between 0 and 1. The worst score on each attribute is defined as a utility of 0, the best score is defined as a utility of 1 and the remaining scores are defined according to their quality. The predefined utility values of processor used in this system are described in the Table 2, the utility values of generation are described in Table 3, and the utility values of RAM are shown in Table 4 and so forth. Finally, the predefined utility values of Battery are also shown in Table 8.

Table 2. Defining Utility Values of Processor

Processor	
Attribute Types	Utility Values
Core i7	1
Core i5	0.75
Core i3	0.50
Pentium	0.25
Celeron	0

Table 3. Defining Utility Values of Generation

Generation	
Attribute Types	Utility Values
7th Gen	1
6th Gen	0.5
-	0

Table 4. Defining Utility Values of RAM

RAM	
Attribute Types	Utility Values
8GB	1
4GB	0.5
2GB	0

Table 5. Defining Utility Values of Hard Disk

Hard Disk	
Attribute Types	Utility Values
1TB	1
500GB	0

Table 6. Defining Utility Values of Graphics Card

Graphics Card	
Attribute Types	Utility Values
4GB	1
2GB	0.70
1GB	0.35
-	0

Table 7. Defining Utility Values of Screen Size

Screen Size	
Attribute Types	Utility Values
15.6"	1
14"	0.70
13.3"	0.35
11.6"	0

Table 8. Defining Utility Values of Battery

Battery	
Attribute Types	Utility Values
6 Cells	1
4 Cells	0.70
3 Cells	0.35
2 Cells	0

5.3. Defining Utility Values of Price

The utility values of Price are also needed to define in this system. Since price is a quantitative attribute, the utility value of price can be computed as follow:

$$U = \frac{x - Worst}{Best - Worst}$$

Where,
x = Price defined in each shop
Worst = Maximum price defined in the system,
Best = Minimum price defined in the system

In this system, the maximum price and minimum price are fixed such that the value of Worst as 2,500,000Ks and best as 300,000Ks. Moreover, the price must be selected in this system because it has a significant effect on the buying behavior of consumers because the higher a product is priced, the fewer units are sold. Additionally, the price of a good is also determined by the point at which supply and demand are equal to each other. Normally, the demand dictates the price. For (purely) inelastic demand, the price is entirely set by demand and price is the dependent variable.

Suppose that user wants the laptop with the following attributes based on the following user preferences:
Processor = Core i7, Core i5, Generation = 7th Gen, RAM = 8GB, Hard Disk = 1TB, Graphics Card = 4GB, 2GB, Screen Size = 15.6", Battery = 2 Cells, Brand = Asus and Price is >=500,000Ks.

There are five relevant laptops that matched with user preferences received from "17" computer shops and the results are shown in Table 9.

Table 9. Relevant Laptops with user preferences

Laptop	Shops	Processor	Generation	RAM	Hard Disk	Graphics Card	Scree Size	Battery	Price (Ks)	Brand
L33	Technolnd	Core i5	7th Gen	8GB	1TB	4GB	15.6"	2 Cells	845,000Ks	Asus
L27	Unique	Core i7	7th Gen	8GB	1TB	2GB	15.6"	2 Cells	936,000Ks	Asus
L18	Asia Tech	Core i7	7th Gen	8GB	1TB	2GB	15.6"	2 Cells	937,000Ks	Asus
L25	Cititom	Core i7	7th Gen	8GB	1TB	2GB	15.6"	2 Cells	938,000Ks	Asus
L14	KMD	Core i7	7th Gen	8GB	1TB	2GB	15.6"	2 Cells	938,000Ks	Asus

By Calculating MAUT, the utility values of each laptop relevant to the user preferences are described in Table 10.

Table 10. Relevant Laptops with Utility Values

Laptop	Shops	Processor	Generation	RAM	Hard Disk	Graphics Card	Scree Size	Battery	Price(Ks)	Brand	Utility Values
L27	Unique	Core i7	7th Gen	8GB	1TB	2GB	15.6"	2 Cells	936,000Ks	Asus	0.90646464
L18	Asia Tech	Core i7	7th Gen	8GB	1TB	2GB	15.6"	2 Cells	937,000Ks	Asus	0.90642677
L25	Cititom	Core i7	7th Gen	8GB	1TB	2GB	15.6"	2 Cells	938,000Ks	Asus	0.90638889
L14	KMD	Core i7	7th Gen	8GB	1TB	2GB	15.6"	2 Cells	938,000Ks	Asus	0.90638889
L33	Technolnd	Core i5	7th Gen	8GB	1TB	4GB	15.6"	2 Cells	845,000Ks	Asus	0.19602273

In the proposed system, the top three best match laptops relevant to the user preferences are displayed from "17" computer shops. Since there can be many best match laptops relevant to the user preferences with the same specification and same price, the top three best laptops are displayed in this system. These best match laptops relevant to the user' preferences are shown in Table 11. According to the results in this table, the first best match laptop is from "Unique", the second is from "Asia Tech" and the third best match relevant laptop is from "Cititcom".

Table 11. Top Three Best Match Laptop(s)

Laptop	Shops	Processor	Generation	RAM	Hard Disk	Graphics Card	Scree Size	Battery	Price(Ks)	Brand	Utility Values
L27	Unique	Core i7	7th Gen	8GB	1TB	2GB	15.6"	2 Cells	936,000Ks	Asus	0.90646464
L18	Asia Tech	Core i7	7th Gen	8GB	1TB	2GB	15.6"	2 Cells	937,000Ks	Asus	0.90642677
L25	Cititom	Core i7	7th Gen	8GB	1TB	2GB	15.6"	2 Cells	938,000Ks	Asus	0.90638889

6. Conclusion

Considering the multiple attributes of the shopping item and its various parameters, the item information can be described in various ways and the approximate cost of these items can be estimated. The proposed system is implemented as a shopping assistant system for buying laptop computers. The additive utility function of MAUT is used to evaluate the best match product(s) based on user preferences. This system can assist human buyers by searching the products according to their preferences within their budgets. For the users, it is difficult to find the right product shop by shop. By using this, they can save time and effort in searching required information and are able to choose the

best match laptop(s) from the most suitable ones.

References

[1] Aung Zaw Htet, “Agent based Online Shopping Assistant System”, University of Computer Studies, Yangon, 2009

[2] Yin Su Thwin, “Agent based Model for Shopping Assistant”, University of Computer Studies, Yangon, 2009

[3] Su Myat Kyaw Lin, “Implementation of Multiple Attribute Reverse English Auciton using Multi-Attribute Utility Theory (MAUT)”, University of Computer Studies, Yangon, 2014

[4] Bing Xu, Zhi-geng Pan and Hong-wei Yang, "Agent-based Model for Intelligent Shopping Assistant and its Application", State Key Lab of CAD&CG, Zhejiang University, Hangzhou, 310027, P.R.China

[5] Maw Min and Nyein Nyein Oo, “Mobile Agent based Information Retrieval for Shopping Assistant”, Yangon Technological University, Yangon, 2015

[6]Churee Theetranont, Peter Haddawy and Donyaprueth Krairit, “Integrating Visualization And Multi-Attribute Utility Theory For Online Product Selection”

[7] M.Shanmuganathan, K.Kajendran, A.N.Sasikumar, M.Mahendran, " Multi Attribute Utility Theory – An Over View", Faculty, Dept of C.S.E, Panimalar Engineering College, Chennai, TamilNadu, India

[8] Yan Liu, “Multi-attribute Utility Theory (MAUT)”, Department of Biomedical Industrial and Human Factors Engineering Wright State University.

[9] Mark Velasquez and Patrick T. Hester, " An Analysis of Multi-Criteria Decision-Making Methods", Department of Engineering Management and Systems Engineering, Old Dominion University, Norfolk, VA USA

[10] Michael R. Middleton, School of Business and Management, University of San Francisco, "Sensitivity Analysis For Multi-Attribute Utility Using Excel".

Classification of Mushroom in Myanmar Using Naive Bayesian Classifier

Khaing Ei Ei Zaw, Thin Lai Lai Thein
University of Computer Studies, Yangon
khaingeieizaw @ucsy.edu.mm, tllthein @ucsy.edu.mm

Abstract

Mushrooms are the most recognizable scrumptious food which is sans cholesterol as well as plentiful in nutrients and minerals. Numerous types of mushrooms have been figured out all through the earth. Distinguishing palatable or harmful mushrooms through the unaided eye is very difficult, so mushroom species should have to arrange eatable and noxious. This framework will be arranged the sort of mushroom by utilizing Naive Bayesian classifier to foster helpful subset of mushroom highlights for characterization task. The Naive Bayesian classifiers have been perhaps the most loved approaches as premise of numerous grouping technique both hypothetically and basically. This system can classify the edible and poisonous mushrooms from mushroom dataset by using Naive Bayes Classifier. In this system, performance comparison of the two algorithms are used Naïve Bayesian classifiers and K-Nearest neighbor (KNN) by using confusion matrix. The evaluation result of Naïve Bayesian classifier is better than K-Nearest neighbor (KNN).

Keywords: Naive Bayesian (NB), K-nearest neighbor (KNN), Mushroom Classification, supervised learning

1. Introduction

Mushroom is beefy and consumable natural product groups of a few types of parasites individuals from Basidiomycetes that normally fill in ground surface or substrate of different plants like straw and wood. Myanmar is ordered as one of the agrarian nations and known as the stockroom of unmistakable mushroom in the agricultural nation. The million types of mushroom, by and large, can be partitioned into two kinds, specifically consumable and harmful mushrooms. The Family of Agaricus and Lepiota ridiculously live in the open spaces; both with different shapes, varieties, and qualities which are not realized by many individuals are toxic. The Family of harmful Agaricus and Lepiota can cause ailment for one who consumes and furthermore can cause passing. The Family of Agaricus and Lepiota that are living fiercely can be consumed and, surprisingly, utilized as meds.

Recognizing the edibility of mushroom physically is a too troublesome undertaking. Due to the greater part of the noxious mushrooms seem as though eatable mushroom attributable to variety and shape. Thus, computerization is vital in this field to diminish time and work. There are numerous arrangement approaches exist in AI. Different Authors are utilized characterization strategies, where Decision Tree ID3, CART and Neural Network classifier calculations have been utilized to order mushroom. This framework arranges the kinds of mushroom by utilizing Naive Bayesian Classification. The mushroom datasets were gathered from Ministry of Agriculture, Livestock and Irrigation and papers from on the web. 46 types of mushrooms are recorded. There are 16 credits. The properties use in datasets there are class, cap tone, cap shape, cap surface, cap umbonate, gills/pores tone, gills/pores connection, gills/pores dividing, stipe tone, stipe shape, stipe, annulus or ring, spore tone, spore shape, spore surface, spore size and last developing territory of mushroom. For the quantity of each class, comprising of 192 information remembered for the food mushroom class and 802 information remembered for the harmful mushroom classification, so the complete number of information utilized was 994 information.

The primary reason for this framework is to arrange mushroom edible and poisonous. In this framework, mushroom datasets are divided into two classes, a training class and testing class. 70% of the information are assigned to the

training set and 30% is dispensed to the testing set.

This paper is organized as follows: Section 2 discusses related works; Section 3 explains the Bayesian Classifier. The corresponding system design and the implementation of this system are described in Section 4. Finally, Section 5 presents experimental result and conclusion.

2. Related Work

Yuhan Zhang, [1] have proposed "Neural Network classification on mushroom dataset with feature selection using evolutionary algorithm and auto– associative network". In This System, the result is comparison of prediction accuracy. Prediction accuracy of Neural Network with feature selection is 77%.Neural Network without feature selection is 70%.

In [2], Roshna Chettri, Shrijana Pradhan and Lekhika Chettri have presented Comparative Study on Classification Algorithms (k-NN, Naive Bayes and Case based Reasoning)". In This System, the result is comparison of prediction accuracy. The accuracy for k-NN, Naive Bayes and Case based Reasoning are 72%, 85% and 92 % respectively.

B.Lavanya and G.R.Preethi, [3] have described "Decision tree for classification of mushroom dataset". In This System, the result is comparison of prediction accuracy. The accuracy for ID3, CART and HoeffdingTree (HT) are 69%, 90% and 100% respectively.

3. Naïve Bayes Classification

The Naïve Bayes classification depends on the Bayes hypothesis, and is especially fit when the dimensionality of the information sources is high-priced. Regardless of its effortlessness, Naive Bayesian classification can frequently accomplish tantamount execution with some refined order techniques, for example, choice tree and chose brain net classifier. Gullible Bayesian classifiers have likewise displayed high-priced exactness and quickness when practiced to enormous datasets. Here part, we will momentarily survey Bayes' hypothesis, then give an outline of Naive Bayesian classification and its utilization in AI, particularly record characterization.

A generally involved system in grouping is given by the straightforward hypothesis of likelihood well-known as Bayes hypothesis or standard. Fore we present Bayesian Theory, let first survey two principal laws of likelihood hypothesis in the accompanying structure:

$$p(X) = \sum_{Y} p(X,Y) \tag{3.1}$$

$$p(X,Y) = p(Y|X)p(X). \tag{3.2}$$

where the first mathematical statement is the *sum law*, and the next mathematical statement is the *produce law*. This $p(X, Y)$ is a join probability, the amount $p(Y|X)$ is a conditional probability, and the amount $p(X)$ is a marginal probability. These twice easy laws form the base for all of the probabilistic theorem.

Depended on the result law, all together with the similarity proprietary $p(X, Y) = p(Y, X)$, it is simple to obtain the next Bayesian theory,

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}, \tag{3.3}$$

which assumes a focal part in AI, particularly order. Utilizing the total rule, the denominator in Bayes' hypothesis can be communicated as far as the amounts showing up in the numerator.

The denominator in Bayes' hypothesis can be viewed just like the standardization consistent expected to guarantee that the amount of the contingent likelihood on the left-wing part of mathematical statement (3.3) over all upsides of Y approaches one.

Allow to think the straightforward guide toward all the more likely figure out the essential ideas of likelihood hypothesis and the Bayes' hypothesis. Assume us have two packs. They are one pink and one yellow, and in the pink pack we have two oranges, four apples and six lemons, and in the yellow pack have three oranges, six apples and one lemon. Presently guess us haphazardly pick one of the containers and from that case we arbitrarily collect a thing, and have seen which kind of thing it is. All the while, we supplant the thing in the case from which it came, and we could envision rehashing this cycle commonly. Allow us to assume that us take the pink pack 40% and the yellow pack 60% of the time, and that when us collect a thing from a container we are similarly prone to choose some things in the crate.

Let us specify randomly variable Y to represent the pack us make choice, then us have

$$p(Y = p) = 4/10 \text{ and}$$
$$p(Y = y) = 6/10,$$

where $p(Y = p)$ is the marginal probability that we make choice the pink pack, and $p(Y = y)$ is the minor likelihood that pick the yellow pack. Assume that we take a crate indiscriminately, and afterward the likelihood of choosing a thing is the negligible part of that thing given the chose pack, which can be composed as the accompanying contingent probabilities

$$p(X = o|Y = p) = 2/12 \quad (3.4)$$
$$p(X = a|Y = p) = 4/12 \quad (3.5)$$
$$p(X = l|Y = p) = 6/12 \quad (3.6)$$
$$p(X = o|Y = y) = 3/10 \quad (3.7)$$
$$p(X = a|Y = y) = 6/10 \quad (3.8)$$
$$p(X = l|Y = y) = 1/10. \quad (3.9)$$

Note that these probabilities are normalized so that

$p(X= o|Y = p)+ p(X = a|Y = p)+ p(X = l|Y = p) = 1$ and

$p(X= o|Y = y)+ p(X = a|Y = y)+ p(X = l|Y = y) = 1.$

Presently guess a thing has been chosen and it is an orange, and we might want to realize which box it came from. This expects that we assess the likelihood conveyance over packs adapted on character of the thing, though the probabilities in mathematical statement (3.4)- (3.9) represent the circulation of thing molded on personality of the case. In view of Bayes' hypothesis, we can compute the back likelihood by turning around the restrictive likelihood.

$$P(Y=p|X=o) = \frac{p(X=o|Y=p)p(Y=p)}{P(X=o)}$$
$$= \frac{2/12\times4/10}{37/150}$$
$$= 10/37$$

where the total probability of deciding an orange $p(X = o)$ can be computed by applying the sum and produce laws.

$$P(X=o)=p(X=o|Y=p)p(Y=p)+p(X=o|Y=y)p(Y=y)$$
$$= 2/12\times4/10 +3/10\times6/10$$
$$= 37/150$$

From the aggregate rule, it then, at that point, sees that $p(Y = p|X = o) = 1-10/37= 27/37$. Overall reasons, we are keen on the probabilities of the groups known the information tests.

Assume we utilize irregular variable Y to signify the group name for information tests, and arbitrary variable X to address the component of information tests. We can decipher $p(Y = Ck)$ as the earlier likelihood for the group Ck, which addresses the likelihood that the group name of an information test is Ck before we notice the information test. When we notice the component X of an information test, we can then utilize Bayes hypothesis to process the comparing back likelihood $p(Y|X)$. The amount $p(X|Y)$ can be communicated as how plausible the noticed information X is for various groups, which is known as the probability.

Memo that the probability isn't a likelihood dispersion over Y, and its essential regarding Y doesn't be guaranteed to rise to one. Considering this meaning of probability, we can express Bayes hypothesis as back $\propto$ probability $\times$ earlier. Since we have presented Bayes hypothesis, in the following subsection, we will see the way Bayes hypothesis is utilized in the Naive Bayesian classification.

4. Proposed System

4.1. System Design

This system is to classify the prediction mushroom dataset based on Naive Bayesian classifier.

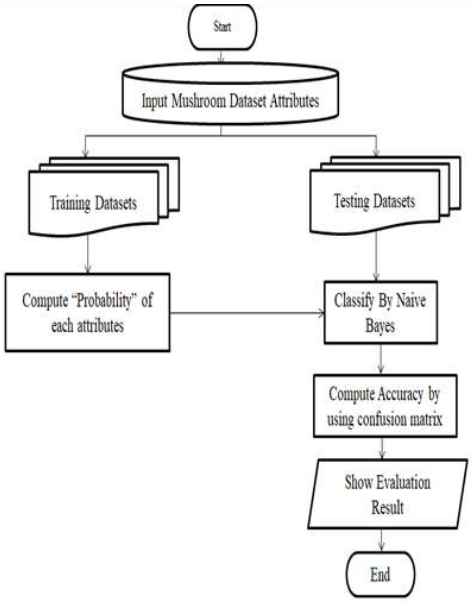


Figure 1. System Flow Diagram

In Figure 1, the mushroom datasets are divided into training dataset (70%) and Unknown data (or) testing dataset (30%). Compute the "Probability" of each attributes by using Naive Bayesian classifier. Testing datasets are used to estimate the classifier accuracy by using confusion matrix. And then, this system show evaluation result for accuracy, precision, recall and f-measure.

4.2. Attributes Information and Sample Dataset

The mushroom dataset contains 16 attributes. The two classes are edible and poisonous. There are consisting of 192 records contained in the food mushroom group and 802 records included in the poisonous mushroom group, so that the total number of data used was 994 records. The following table1 expresses name of attributes and description of these attributes (features).

Table 1. Name and Description of Attributes

No	Attribute Name	Description
1	cap-color	white to pale gray, white, orange-red, leaden-brown, pale-orange, dark-brown, orange-brown, grey-brown, dull green, red, gray, silver, brown, buff yellow, brownish yellow, greenish yellow, golden, golden brown, grayish brown, pink, orange, pale, brick-red, yellow, bright yellow, purple
2	cap-shape	campanulate, expanded, convex, convex to depressed, convex with depression, globose, ear-shaped, bell, flat, depressed, puffball, round, umbrella-shaped, funnel-shaped, lobed, kidney-shaped, skirt, pattern, conical, ball, shell, fan, irregular
3	cap-surface	Fertile, flat scales, smooth, waxy, powdery, velvety, fibrous, rough, dry, hard, silky
4	cap-umbonate	present, slightly present, Slightly, Absent
5	gill-color	grayish-brown, white, yellow, pale-yellow, creamy-white, golden-yellow brown, dark-brown, black, pale pink, cinnamon brown, pink, red, chocolate, purple-gray, cream, absent
6	gill-attachment	free, adnate, adnate to decurrent, decurrent, attached, adnexed, absent
7	gill-spacing	close, crowded, distant, absent
8	stipe-color	white, yellow, reddish brown, orange-brown, red, brown, gray, cream, pink, black, absent
9	stipe-shape	slender, equal, unequal, conical, fan, cup, curved, club, cylindrical, fusiform, rhizoids, fibrous, asymmetrical, flat, bulbous, elliptic, tubular, straight, absent
10	stipe	hollow, solid, short, long, fleshy, truncated, thin, thick, dry, absent
11	annulus or ring	absent, present, double
12	spore-color	dark-brown, white, pink, olive-brown, brown, brownish black, rosy, purple-brown, red, yellow, cinnamon, pale, gray
13	spore-shape	elliptic, globose, fusiform, oblongoid, cylindrical, round, subelliptic, tellipsoid, broadly elliptic, angular, club, curved, tender, tropical, conical, bean, amygdaloid,
14	spore-texture	smooth, rough, smooth apical germ pore, spring with faint reticulum, sordid, ovate, fibrous, amyloid, meaty, Jelly-like
15	spore-size	6-7.2×4.8-4.8 µm, 8.4-11.4×6-7.2 µm, 6-7.2×6-6 µm, 8.4-12×6-8.4 µm, 8.4-10.8×4.8-6 µm, 6-7.2×3.6-4.8 µm, 7.2-9.6×4.8-6 µm, 6-8.4×6-7.2 µm, 7.2-7.2×7.2-7.2 µm, 6-7.2×4.8-6 µm, 10 - 12µm, 10 - 15 ×4. 0 - 6.0 µm, 6.0-7.5 ×3.0 - 5.0 µm, 4 -6 × 2 - 3 µm, 9.0 - 10.0 x 6 - 7µm, 12-13µm×3-5µm, 4-6 x 3-4 cm, 12-15x10-12µm, 4-6 × 1-1.5 µm, 8-12 × 7-9 µm, 10-12 × 9-10µm, 7-8 µm, 3 - 6×3 - 5µm, 7-9× 4-5µm, 8.9 ×4.6 µm, 6.8-5-6 µm, 5.9–6.8 µm by 4.2–5.1 µm , 4-6x2.5-3µm, 7.5–11 × 3–4 µm, 6-9 by 2–3.5 micrometers, 5-8 / 4-6 µm, 9 –13 ×6 –8 µm, 8-9× 6-7µm, 8–10 µm, 12-14× 5.5-6.5µm, 8.8-11.0× 5.5-8.0µm, 9-13 by 6.5-9 µm, 7–10 µm, 6.5-8 x 3.5-4.5 µm, 4-5.5 x 2-4.5µm, 10-13 x 5.5-7µm, 6-7× 3-4µm, 6-7× 3-4µ, 14-15× 3-3.5µm, 3.5-5.5×0.75-1µm,17-25× 6-8µm, 3.2-4.3µm, 9-13 to 5-7µm, 25-35/3-5µm, 5-10µm, 8-12.5 by 3.5-5µm, 6.8-9.3µm, 8-10 by 5.5-7µm, 6.5-9× 2.8-3.5µm, 5-10× 2-.5µm, 3.5µm, no
16	growing habitat	Decay woods, woods of deciduous trees, soil, grasses, bamboo, bush, under small tree, paddy straw, logs, houses, oak, woodlands, tree stumps, fields, broad-leaf trees, softwoods, hardwoods, dry trees, underneath the soil
17	class	edible, poisonous

Table 2. Sample Mushroom Dataset

cap color	cap shape	cap surface	cap umbonate	gills color	gills attachment	gills spacing	stipe color	stipe shape	stipe	annulus	spore color	spore shape	spore texture	spore size	growing habitat	class
gray	convex	smooth	slightly	white	free	close	white	cylindrical	hollow	absent	white	elliptic	smooth	10 - 12µm	woods	edible
gray	convex	smooth	slightly	red	free	close	white	cylindrical	hollow	absent	white	elliptic	smooth	10 - 12µm	woods	poisonous
white	ball	smooth	absent	white	attached	closed	absent	absent	absent	absent	white	globose	smooth	5-6 x 5.5-6 µm	woods	edible
white	ball	smooth	absent	white	attached	closed	absent	absent	absent	absent	white	globose	rough	5-6 x 5.5-6 µm	woods	edible
gray	shell	smooth	absent	white	adnate	crowded	white	straight	short	absent	gray	cylindrical	smooth	6.5-9× 2.8-3.5µm	woods	edible
gray	shell	smooth	absent	white	adnate	crowded	red	straight	short	absent	gray	cylindrical	smooth	6.5-9× 2.8-3.5µm	woods	poisonous
brick-red	Flat	smooth	slightly	cream	adnate	distant	gray	fibrous	long	absent	white	ovoid	smooth	5.5µm × 7.2µm	soil	edible
brown	Flat	smooth	slightly	cream	adnate	distant	gray	fibrous	long	absent	white	ovoid	smooth	5.5µm × 7.2µm	soil	edible
white	Flat	smooth	slightly	cream	adnate	distant	gray	fibrous	long	absent	white	ovoid	smooth	5.5µm × 7.2µm	soil	edible
gray	Flat	smooth	slightly	cream	adnate	distant	gray	fibrous	long	absent	white	ovoid	smooth	5.5µm × 7.2µm	soil	edible
gray	Flat	smooth	slightly	red	attached	broad	gray	fibrous	short	absent	white	ovoid	smooth	5.5µm × 7.2µm	hardwoods	poisonous
gray	Flat	smooth	slightly	red	attached	broad	red	fibrous	short	absent	white	ovoid	smooth	5.5µm × 7.2µm	hardwoods	poisonous
brown	Flat	smooth	slightly	red	attached	broad	gray	fibrous	short	absent	white	ovoid	smooth	5.5µm × 7.2µm	hardwoods	poisonous
brown	Flat	smooth	slightly	red	attached	broad	red	fibrous	short	absent	white	ovoid	smooth	5.5µm × 7.2µm	hardwoods	poisonous
yellow	Flat	smooth	slightly	green	attached	broad	pink	fibrous	long	absent	white	ovoid	smooth	5.5µm × 7.2µm	hardwoods	edible

4.3. Implementation of the System

The proposed mushroom classification system can only be used for registered user. So, the user who want to use this system must be registered (Sign Up) first and can be enter by registered user information. After the authentication is successful, the user can get the main page of the system as shown in figure 2.

Figure 2. Login Page of System

This is first phase for the classification process. The user must be loaded the training data excel sheet from the system supported open dialog box and data are loaded to the system and then stored in the system database. After successfully upload the training dataset, the message will be shown in Figure 3.

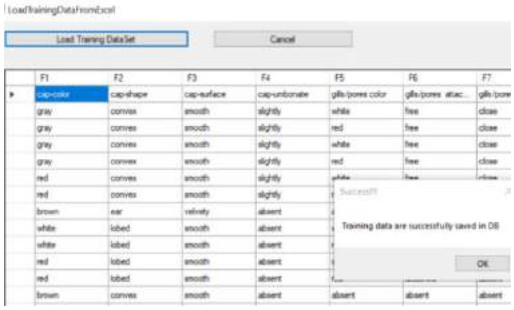


Figure 3. Training Data of mushroom dataset

For Naive Bayes Classification, the user must be loaded testing data via the system support "Load Testing Data" button from the page shown in figure 4.

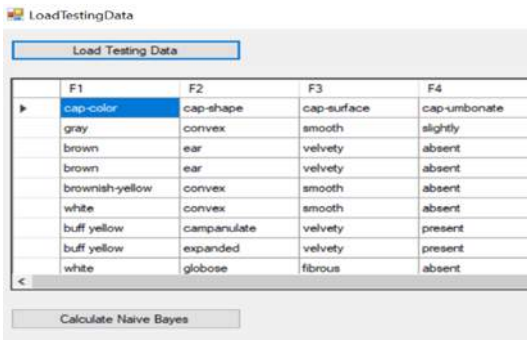


Figure 4. Testing Data of mushroom dataset

Then, "Calculate Naive Bayes" button is support to proceed the classification process by Naive Bayes Classifier. Classification observation of each testing records are as shown in figure 5.

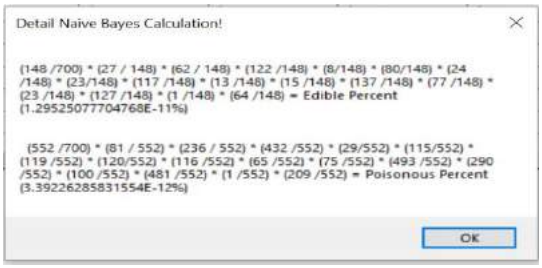


Figure 5. Classification Observation

And then, result for all testing datasets calculate by Naive Bayes as shown in figure 6.

Figure 6. Result for testing record 1

5. Experimental Result

The best level of accuracy between the two algorithms can be determined by comparison. Naive Bayesian and K-Nearest Neighbors are compared by calculation accuracy, precision, recall and f-measure to get more reliable and good performance in classification. The Naive Bayesian Result is better than K-Nearest Neighbors base on the experiment of Accuracy, Precision, Recall and F-measure by using confusion matrix. 70% of training datasets are assigned to the train set and 30% is dispensed to the test set.

The experimental result can be seen that Naive Bayes gave the highest test accuracy better than K-Nearest Neighbors as shown in figure 7.

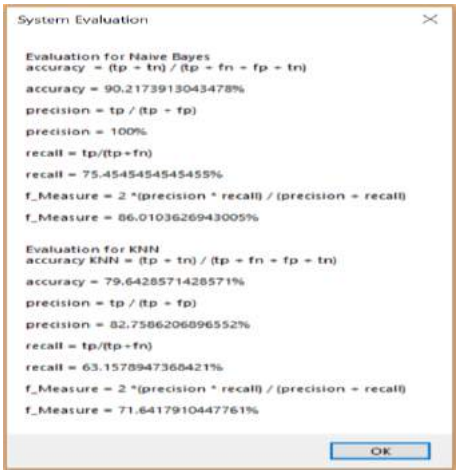


Figure 7. Performance Comparison of Naive Bayesian and K-Nearest Neighbors

6. Conclusion

The Naive Bayesian classifier uses the naive Bayesian formula to calculate the probability of each class given the prices of all attributes. The Bayes Classifier is based on Bayes theorem of posterior probability. The system support users in classifying edible and poisonous mushroom based on the mushroom attributes. The classification based on mushroom datasets by using Naive Bayesian Classifier. For the performance comparison of accuracy, the two algorithms are used Naive Bayesian classifiers and K-Nearest neighbor (KNN) by using confusion matrix. The mushroom identification method that has been done used Naive Bayes and KNN algorithms with the prediction accuracy of 90.21% and 79.64%.

References

[1] Agung Wibowo, Yuri Rahayu, Andi Riyanto, Taufik Hidayatulloh; "Classification Algorithm for Edible Mushroom Identification", International Conference on Information and Communications Technology (ICOIACT), March 2018.

[2] Balika J. Chelliah, S. Kalaiarasi, Apoorva Anand, Janakiram G, Bhaghi Rathi, Nakul K. Warrier; "Classification of Mushrooms using Supervised Learning Models", Undergraduate, SRM Institute of Science and Technology, Ramapuram, Chennai, Tamil Nadu, India, April 2018.

[3] Bandana Garg, "DESIGN AND DEVELOPMENT OF NAÏVE BAYEs CLASSIFIER", North Dakota State University of Agriculture and Applied Science, June 2013.

[4] Kanchi Tank; "A Comparative Study on Mushroom Classification using Supervised Machine Learning Algorithms", International Journal of Trend in Scientific Research and Development (IJTSRD) Volume 5 Issue 5, July-August 2021.

[5] Kasarapu Ramani, "MACHINE LEARNING TOOLS FOR DATASET CLASSIFICATION", Department of IT, Sree VidyanikethanEngg College (Autonomous), Tirupati, India, January 2018.

[6] Ramazan Sener, "DETERMINATION OF POISON MUSHROOM USING NAÏVE BAYES ALGORITHM", Firat University, February 2020.

Matriculation Students' Result Prediction System for Rakhine State

Nwet Nwet Zin, Zaw Tun
University of Computer Studies, Yangon
cu.nweinweizin@gmail.com, zawtun78@gmail.com

Abstract

These days, huge measure of information put away in various data sets and is expanding quickly. These data sets contain educational information that can be valuable for anticipating understudies' scholarly presentation and will help the academic educational environment for improvement. Educational Data Mining is used to study the available data stored in the educational database and create new knowledge out of it. C4.5 (J48) classification algorithm is applied to create a decision tree model that will predict the academic performance of Information technology students of the Rakhine State. The decision tree result anticipated the potential understudies who will get the opportunity to pass or don't get the opportunity in light of their memorable information and this will assist the educator with giving fitting contributions to help the faltering understudies. This system is implemented using C#.Net programming language with Microsoft SQL Server 2017 Express version database engine.

Keywords: C4.5, Decision Tree, Student Performance, Information Technology, Rakhine State.

1. Introduction

While data set advancement has given us the key instruments to capable limit and question of gigantic instructive assortments, the issue of how to help individuals understand and separate colossal gatherings of data stays trouble. To deal with the data excess, one more period of sharp gadgets for mechanized data mining and data exposure is required. This thought and procedures are extensively used in displaying, course, especially in enlightening investigation.

Educational Data Mining (EDM) is a procedure in information mining that is broadly utilized for instructive examination, this will assist with distinguishing designs that is valuable for anticipating the scholastic presentation of the understudies. Understudies' scholastic exhibition is crucial for instructive establishments wherein it is utilized for key intending to improve and keep up with the nature of investigations of the understudies.

Hence, this study will introduce the idea of information mining (C 4.5) to anticipate the scholarly exhibition of the Information Technology understudies of Rakhine State.

2. Related Work

R. R. Kabra, R. S. Bichkar [2] directed a concentrate on foreseeing understudy execution utilizing information mining approach in which the specialists utilized three regulated calculations. They directed a few tests to decide the expectation exactness, accurately and mistakenly characterized examples and the learning time.

Las Johansen B Caluza and Jostens Keneth D. Trecene [4] driven a focus on predicting the presentation improvement of planning students of VBS Purvanchal University, Jaunpur. It was assumed that students were likely going to bomb considering the students' associated factors in first year planning test and it was seen that ID3 and C4.5 decision tree in the best computation.

Sohajbir Singh Ubha, Gaganpreet Kaur Bhalla [5] guided a focus on planning students to evaluate the show by using data mining systems and applying 3 methods like connection, request and packing. The variables used are given out, support, Sessional Marks, GPA, and current last grade from the data base organization system course. That is the very thing that the result expected if a student is poor in support and assignment, the grades are poor. The experts acknowledge that data mining is valuable in high level training, especially in planning students wherein new data is found.

Khaledun Nahar, Boishakhe Islam Shova, Tahmina Ria, Humayara Binte Rashid [6], (2021) broke down the achievement of the students in the PC planning division at Karabuk University using the actions, for instance, direction, mature, sort of auxiliary school graduated, and whether the student is focusing on in distance learning or standard preparation. Decision tree computation made better assumption result using a 10 overlay holdout dataset. The result in like manner uncovered that as the age of the student fabricates the accomplishment score reduces. Plus, the experts moreover sorted out that students in standard preparation have a higher accomplishment rate than in distance learning.

3. Background Theory

The prediction of the student's academic execution is significant because it helps increment pass rates by suitably directing understudies, directing changes in college scholarly strategies, illuminating educational works on, looking at proficiency and viability of getting the hang of, giving significant criticism to educators and students and altering learning conditions. A high forecast precision of the understudies' exhibition is useful to distinguish the low presentation understudies toward the start of the growing experience. In any case, to accomplish these targets, enormous volume of understudy information should be examined and anticipated utilizing different AI models.

The shifting forecast level by different AI models might be because of contrasts in financial. It might likewise be vital to take note of that understudy's scholastic exhibitions are impacted by many elements, as financial variables of understudies like family pay, parental degree of schooling and work status of understudies or guardians however are not thought about while testing the precision of different AI models in anticipating understudies' presentation. Besides, the different AI models didn't distinguish the most fitting AI model in working on understudies' outcome [1].

3.1. C4.5 Decision Tree Algorithm

C4.5 algorithm is a derivation of ID3 algorithm, that generate a decision tree from training datasets and is typically used in machine learning and natural language processing domain. Information gain is used for attribute selection. Gain ratio is also used to select the attribute which has the smallest entropy value. Confusion matrix is used to evaluate the accuracy of classifier.

The C4.5 calculation is a renowned calculation in Data Mining. The C4.5 calculation goes about as a Decision Tree Classifier. C4.5 is an information mining calculation and it is utilized to create a choice tree. The C4.5 calculation is exceptionally useful to create a helpful choice that depends on an example of information. C4.5 is given a bunch of information addressing things that are now grouped. At the point when we create the choice trees with the assistance of C4.5 calculation, then, at that point, it tends to be utilized for order of the dataset, and that is the fundamental explanation because of which C4.5 is otherwise called a factual classifier.

For each trait X, find the standardized data gain proportion by parting between X. Assume that X is a trait with the most elevated standardized data gain. Make a choice hub that parts on property X. Rehash it on the sub records acquired by parting the characteristic X, and add these hubs as offspring of the hub.

Expected information required to classify a tuple in D

$$\text{Info (D)} = -\sum_{i=1}^{m} Pilog2(Pi) \quad (1)$$

Expected information needed to classify a tuple in D if the tuples are partition according to A is,

$$\text{Info}_A\text{ (D)} = \sum_{j=1}^{v} \frac{|Dj|}{|D|} * Info(Dj) \quad (2)$$

The gain in information from such a partitioning,

$$\text{Gain(A)} = \text{Info(D)} - \text{Info}_A\text{(D)} \quad (3)$$

If information gain measure is biased (value continuous), it applies a kind of normalization to information gain using a split information value defined with Info(D) as

$$\text{Split Info}_A\text{ (D)} = -\sum_{j=1}^{v} \frac{|Dj|}{|D|} * \log2\left(\frac{|Dj|}{|D|}\right) \quad (4)$$

Then,

Gain Ratio (A) = Gain (A) / Split $Info_A(D)$ (5)

The attribute of maximum gain ratio is chosen as the **splitting attribute [3].**

3.2. Advantages of C 4.5 Algorithms

The calculation is exceptionally useful in moderating the overfitting in light of the fact that C4.5 innately utilizes the Single Pass Pruning Process. This can work with discrete information and can likewise work with Continuous Data. This is exceptionally useful in settling the issues of information deficiency. Further, it is critical to realize that this isn't the most ideal calculation in all cases, however it is extremely helpful in certain circumstances [7].

4. The Proposed System

In this framework, the entire dataset is haphazardly isolated into two fundamentally unrelated sets, a planning set and a really look at set. Commonly, 66% of the information are utilized as the planning set, and the lingering 33% is utilized as the actually take a look at set. The irregular is finished with k cycles. As indicated by irregular technique, planning set is brought into the framework. The imported planning set is prepared to fabricate the choice tree model to get the greatest rule length, how much guidelines made, and the absolute number of condition checks to group the entire arrangement set as results from the inferred model by C4.5 calculation. And afterward the check put is consumed to check the model together to get exactness assessment. The outcomes from every cycles are found the middle value of.

Trial results from both planning and checking stages are utilized to look at the changed information scope of informational index with C4.5 calculation. As the planning stage results, the complete number of leaves of choice tree addressing the absolute number of rules created by C4.5 calculation, the most extreme profundity the choice tree addressing the greatest length of the principles, the all-out number of hubs addressing the condition check expected to arrange the entire readiness set and handling time addressing the time expected to construct the model are acquired from every emphasis.

Every one of arrangement results is determined to make execution assessment. As the checking stage, the typical precision assessment from every cycles are utilized to think about execution of each testing. For examination reason irregular cycle runs are performed on each dataset. An emphasis run for one dataset is proceeded as follows:

Setp1: Initial dataset is haphazardly isolated into two fundamentally unrelated sets, a readiness set and a really take a look at set. Commonly, 66% of the realities are utilized as the readiness set, and the excess 33% is utilized as the actually look at set.

Setp2: The framework carries out the calculation on the readiness set to develop the model and keeps an eye on the really taking a look at set to gauge precision.

Setp3: For testing calculation, the framework produces arrangement results from the model in readiness stage and partitions the quantity of accurate grouping to the component size of the actually look at set to assessment precision in really looking at stage.

Setp4: Step 2 and stage 3 must be finished until each part is allowed for really looking at set.

After this multitude of iteration, the readiness results and the exactness assessments acquired from cycles are found the middle value of. The subsequent typical outcomes are utilized to look at each testing. The cycle stream of the framework is displayed in figure 1.

In the system implementation: the training dataset are loaded first and iteratively calculate the gain values of each attributes. Then, iteratively calculate the split info of each attributes to get the gain ratios of all attributes to sort the root, internode and leaf priority for rule generation. Equations to calculate the gain value, split info value and gain ratios are described in section 3.1. Based on the descending of gain values, split info values and gain ratio values, the feasible classification rules are generated to be perfectly predicted the testing data. The sample testing result of the system is shown in figure 2.

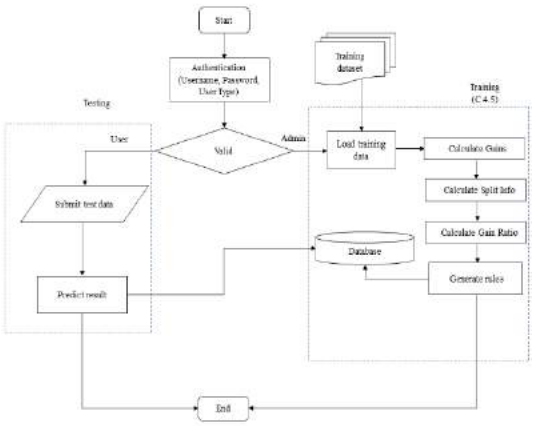


Figure 1. The System Flow

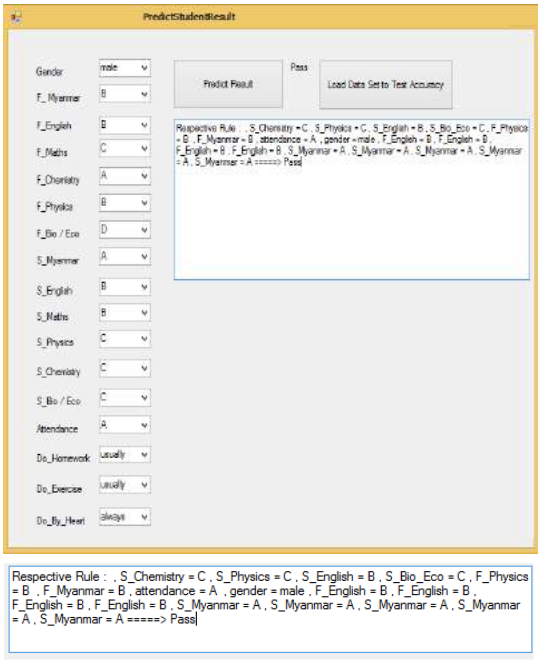


Figure 2. Sample Testing Result with Respect to Generated Rule

4.1. Dataset of Student Data in Sittway, Rakhine State

In this paper, there are 1 class labels and 17 attributes in the classification process. The student activities dataset contains 1000 instances and 18 attributes. Each of the characteristics is assigned value as shown in table 2. Attributes and values used in the preparation datasets are shown in Table 1.

This system has 17 attributes (gender, F_Myanmar, F_English, F_Maths, F_Chemistry, F_Physics, F_Bio_Eco, S_Myanmar, S_English, S_Maths, S_Physics, S_ Chemistry, S_Bio_Eco, attendance, do_Homework, do_exercise and by_heart_skill) to calculate the gain values and recursive iteration will be made to get the optimal decision trees.

Table 1. Attributes Name and Values

No.	Attribute Names	Attribute Values	Data types
1.	Gender	male, female	Categorical
2.	Pretest condition (Six Subjects)	A, B, C, D, E	Categorical
3.	First test condition (Six Subjects)	A, B, C, D, E	Categorical
4.	Attendance	A, B, C, D, E	Continuous
5.	Do homework	usually, often, rarely	Categorical
6.	Do exercise	usually, always, never, often	Categorical
7.	By_heart	usually, always, never, often, rarely	Categorical
8.	Status	Pass, Fail	Categorical

Sample code development of "Gender" attribute's calculation for the requirements of rule generation is shown in figure 3.

```
adp.Fill(dsstudentcount.StudentTable);
if(dsstudentcount.StudentTable.Rows.Count>0)
{
    studentrowcount = dsstudentcount.StudentTable.Rows.Count;

    //Gender
    MasterDataSet dsGender = new MasterDataSet();
    adp.FillByGender(dsstudentcount.StudentTable, "male");
    male = dsstudentcount.StudentTable.Rows.Count;
    adp.FillByGender(dsstudentcount.StudentTable, "female");
    female = dsstudentcount.StudentTable.Rows.Count;

    MasterDataSet dsGenderStatus = new MasterDataSet();
    adp.FillByGenderStatus(dsGenderStatus.StudentTable, "male", "pass");
    malePass = dsGenderStatus.StudentTable.Rows.Count;
    adp.FillByGenderStatus(dsGenderStatus.StudentTable, "male", "fail");
    maleFail = dsGenderStatus.StudentTable.Rows.Count;
    adp.FillByGenderStatus(dsGenderStatus.StudentTable, "female", "pass");
    femalePass = dsGenderStatus.StudentTable.Rows.Count;
    adp.FillByGenderStatus(dsGenderStatus.StudentTable, "female", "fail");
    femaleFail = dsGenderStatus.StudentTable.Rows.Count;

    info_gender = (male / studentrowcount) * (-(malePass / male) * Math.Log((malePass / male), 2)) - (maleFail / male) *
    gain_gender = 1 - info_gender;
```

Figure 3. Sample C# Code Development for "Gender" Gain Value Calculation

```
label3.Text = info_gender.ToString();
label4.Text = gain_gender.ToString();
splitinfo_gender = -((male / studentrowcount) * Math.Log((male / studentrowcount), 2)) - ((female / studentrowcount)
GainRatio_gender = gain_gender / splitinfo_gender;
```

Figure 4. Sample Code Development for "SplitInfo" and "GainRatio" for Gender Attribute

In calculation of C4.5, the SplitInfo and GainRatio are needed to be calculated iteratively as require as possible to get better result. The sample c# code development of SplitInfo and GainRatio for the "Gender" attribute is shown in figure 4.

4.2. Classifier Accuracy Measure

Utilizing the planning set to spring a classifier or expert and approximating the precision of the ensuring learned model can bring about unclear overoptimistic evaluations in light of overspecialization of the focusing on estimation to the data. The precision of a classifier on a given check set is the degree of check set tuples that are precisely requested by the classifier.

The disorder structure is an impetus gadget for examining how well a classifier can perceive tuples of dissimilar classes. Given m classes, a disorder framework is a table of at smallest size m by m. A part, CMi, j in the essential m lines and m portions relegates the amount of tuples of class I that were set apart by the classifier as class j. For a classifier to have moral precision, preferably generally noticeable of the tuples would be connoted along the leaning of the confusion organization, from entry CM1,1 to segment CMm, m, with the other entryways don't being near anything. Figure 5 addresses disarray grid for multi-classes.

		Predicted Class	
		2	4
Actual Class	2	True Positive (TP)	False Negative(FN)
	4	False Positive (FP)	True Negative (TN)

Figure 5. Confusion Matrix for Accuracy Test

Given two classes, genuine up-sides allude to the positive tuples (tuples of the principal class of interest) that were properly set apart by the classifier, while certified negatives are the negative tuples that were right named by the classifier. False certain tuples are the negative tuples that were mess up named. Additionally, deceptive negatives are the positive tuples that were mistakenly stamped. These terms are sensible while surveying a classifier's wellness.

If how well the classifier can recognize the positive tuples and how well it can see the negative tuples would have the choice to be gotten to, then, the survey, exactness, f-measure and explicitness measures can be used independently.

Precision is the rate degree of exactly mentioned cases for all occasions. These activities are characterized as

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

Recall is the proportion of positive case that is accurately ordered and it tends to be determined by the accompanying condition.

$$Recall = \frac{TP}{TP+FN}$$

Precision is the correctly classified instances for those instances that are classified as positive of the following equation.

$$Precision = \frac{TP}{TP+FP}$$

F-measure is the consolidated lattice of accuracy and review, i.e, it is consonant mean of both. It shows how exact the classifier is and furthermore the way that well the classifier is vigorous of the accompanying condition.

$$F - measure = \frac{2 \times Recall \times Precision}{Precision+Recall}$$

Figure 6. Evaluation Results with Different Data Size

5. Conclusion

Data mining is the most common way of investigating information according to alternate points of view and summing up it into helpful data. Grouping is an information mining procedure which resolves the issue of building a prescient model for a class quality given the upsides of different properties and a few instances of records with known class. Decision tree is one of the most deeply grounded arrangement techniques. This theory presents a C4.5 decision tree calculation for understudy result expectation.

This system presents a helpful prediction system for students who will sit the final matriculation examination in Rakhine State. Based on the prediction and analysis result, the educational level of student will be trained and raised for future better result. If the school maintains more accurate data, the system's performance will also increase. This system can help partly Rakhine matriculation students' needs to pass examination.

5.1. Advantages of the System

This system presents the study of C4.5 decision tree algorithm. The main advantages of this system is performing the comparative study on three different amount of data set, building the decision trees with C4.5 algorithm. It helps the prediction on the student information in deciding to raise the educational sector of Rakhine state.

5.2. Limitations and Further Extensions

The study has few limitations. This comparative study is only for academic student result Classification based on 18 variables. More algorithms may also be implemented in this system presents only the classification and hence student information features can be 33 added in its further extension.

References

[1] Ramanathan L, Saksham Dhanda, Suresh Kumar D, "Predicting Students' Performance using Modified ID3 Algorithm", International Journal of Engineering and Technology (IJET), pp. 2491-2497, vol. 3, no. 5, 2013.

[2] R. R. Kabra, R. S. Bichkar, "Performance Prediction of Engineering Students using Decision Trees", International Journal of Computer Applications (IJCA), vol. 36, no. 11, 2011.

[3] H. Yuliansyah and M. Wibowo, "Predicting Students Graduate on Time using C4.5 Algorithm", Journal of Information Systems Engineering and Business Intelligence, vol.7, no. 1, 2021.

[4] Las Johansen B. Caluza and Jostens Keneth D. Trecene, "Predicting Academic Performance of Information Technology Students using C4.5 Classification Algorithm: A Model Development", International Journal of Information Sciences and Application, pp. 7-21, vol. 10, no. 1, 2018.

[5] Sohajbir Singh Ubha, Gaganpreet Kaur Bhalla, "Data Mining for Prediction of Students' Performance in the Secondary Schools of the State of Punjab", International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE), vol. 4, no. 8, 2016.

[6] Khaledun Nahar, Boishakhe Islam Shova, Tahmina Ria, Humayara Binte Rashid, A. H.M Saiful Islam, "Mining Educational Data to Predict Students Performance", Springer, 2021.

[7] Las Johansen B. Caluza and Jasten Keneth D. Trecene, "Predicting Academic Performance of Information Technology Students using C4.5 Classification Algorithm: A Model Development", International Journal of Information Sciences and Application (IJISA)

Prediction of Students' Academic Performance Using Multiple Linear Regression

Chan Myae Myint Zu, Kyi Lai Lai Khine
University of Computer Studies, Yangon
chanmyaemyitzu@ucsy.edu.mm, kyilailaikhine@ucsy.edu.mm

Abstract

Nowadays, the education system has been changed from teacher-centered approach to learner-centered approach. Thus, students' related features need to analyze to predict students' academic performance to give the active learning in educational system. In this system, Multiple Linear Regression (MLR) is applied to predict the students' academic performance on the UCI's student performance dataset. The purpose of this system is to predict the students' final grade based on previous grades and relevant features. Moreover, feature selection method (Chi-Square) is utilized to reduce the number of input variables which are important to a Multiple Linear Regression model. Four different measures: accuracy, precision, recall and F-measure are used to evaluate the proposed system. This system is implemented using C# programming language with Microsoft Visual Studio IDE and Microsoft SQL Server.

Keywords: Multiple Linear Regression (MLR), Chi-Square, UCI

1. Introduction

Education is a pivotal component in our general public. It is a critical variable for accomplishing a drawn out economic process. Further developing student's scholastic accomplishment began with working on way of behaving and urges the understudy to take part in the homeroom.

The quality of education and students' mental, cognitive, emotional and behavioral responses to the educational experience as well as to in-class/out-of-class scholarly and social exercises are critical to accomplish fruitful learning results. The separated information will assist schools with upgrading student's scholarly achievement and to assist executives with further developing learning frameworks. Two significant regions presently being worked on because of which are situated towards the incorporation and investigation of huge information capacities in the instructive climate are Educational Data Mining (EDM) and Learning Analytics (LA).

Predicting Students' Academic Performance (SAP) is one of the significant examination regions in Higher Learning Institutions. A powerful prescient model requires great information (boundary), reasonable Data Mining techniques and devices for the information investigation.

Feature selection is a method for diminishing the quantity of elements and consequently decrease the computational intricacy of the model. Ordinarily highlight choice turns out to be exceptionally valuable to defeat with overfitting issue. It helps us in deciding the littlest arrangement of elements that are expected to anticipate the reaction variable with high exactness.

2. Related Work

Alaf.A.A, Thair.H and Ibrahim.A [1] proposed further developing understudies' presentation utilizing the dataset from an e-Learning framework included three classes: segment, scholastic foundation and social highlights. This framework used Artificial Neural Network, Naïve Bayesian, Decision Tree classifiers and WEKA instrument to assess the proposed arrangement models and examinations assessment. The trial results demonstrate that the solid impact of student conduct on understudy's scholastic accomplishment.

Paulo.C and Alice. S [3] investigated to Predict Secondary School Student Performance utilizing information mining that incorporate understudy grades, segment, social and school

related highlights. The framework applied Decision Trees, Random Forest, Neural Networks, Support Vector Machines and RMiner device. The outcomes demonstrate the way that a decent prescient exactness can be accomplished when the first as well as 2nd school time frame grades are accessible.

R.R.Rajalaxmi, P.Natesan, N. Krishnamoorthy ,S.Ponni, [4] introduced a model which predicts the presentation of the understudies in Engineering Discipline. The free factors of the model contained how long spent on the web and ward variable is the forecast of end semester assessment grades CGPA (Cumulative Grade Points). Different measures are utilized to work out and validate the models.

Sana, Isma. F. S and Qasim .A. A, [5] portrayed breaking down understudies' scholastic execution utilizing various classifiers, for example, choice tree, credulous bayes and fake organization and to look at the impact of understudies' highlights on scholarly execution. The dataset comprises of 500 understudy records and 17 distinct elements. The exactness accomplished up to 10% to 15% is improved when conduct highlights are incorporated.

Oyerinde O. D. furthermore, Chia P. A [2] proposed a system for anticipating understudies' scholastic exhibitions utilizing Multiple Linear Regression, utilized optional information and was separated from the software engineering division, University of Jos. With the guide of the Statistical Package for Social Sciences (SPSS) investigation apparatus and the outcomes in light of the Mathematics characteristics of understudies that anticipate the understudies' exhibition on software engineering subject.

3. Background Theory

3.1. Multiple Linear Regression (MLR)

Multiple linear regression (MLR), also known simply as multiple regression, is a measurable strategy that utilizes a few informative factors to foresee the result of a reaction variable. The objective of different straight relapses is to display the direct connection between the informative (free) factors and reaction (subordinate) factors. Basically, various relapses is the augmentation of Ordinary Least-Squares (OLS) regression because it involves more than one explanatory variable.

Multiple Regression Multiple Tasks: A multiple regression thinks about the impact of more than one illustrative variable on some result of interest. It assesses the general impact of these illustrative, or free, factors on the reliant variable while holding the wide range of various factors in the model consistent.

The equation of multiple linear regression is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon$$

Here, Y is the dependent variable to be estimated
X_i's are the independent variables
β_0 is the constant
β_i's are the regression coefficients
ε is the error term

3.2. Feature Selection

In AI and measurements, highlight determination, otherwise called variable choice, characteristic choice or variable subset determination, is the most common way of choosing a subset of significant elements (factors, indicators) for use in model development. Include determination methods are utilized because of multiple factors:

- improvement of models to make them simpler to decipher by scientists/clients,
- more limited preparing times,
- to stay away from the scourge of dimensionality,
- work on information's similarity with a learning model class,
- encode inborn balances present in the information space.

The focal reason while utilizing a component determination procedure is that the information contains a few highlights that are either excess or unimportant, and can in this manner be eliminated without causing a lot of deficiency of information [9]. Repetitive and immaterial are two unmistakable thoughts, since one applicable component might be excess within the sight of one more pertinent element with which it is firmly associated.

Chi-Square Feature selection: Chi-square is utilized in information comprise of individuals disseminated across classes, and to know whether that appropriation is not the same as what might expect by some coincidence. Chi-Square is one method for showing a connection between two clear cut factors. There are two sorts of factors in measurements: mathematical factors and non-mathematical factors. The worth can be determined by utilizing the given noticed recurrence and anticipated recurrence. A tiny Chi-Square test measurement implies that your noticed information accommodates your normal information incredibly well. A very large Chi-Square test statistic means that the data does not fit very well.

$$X^2 = \sum(O - E)^2/E$$

Where,

O = Observed frequency
E = Expected frequency
$\sum$ = Summation
χ^2 = Chi-Square value

Feature selection techniques ought to be recognized from include extraction. Include extraction makes new highlights from elements of the first elements, while include determination returns a subset of the elements. Highlight determination procedures are much of the time utilized in spaces where there are many elements and relatively couple of tests (or data of interest).

4. The Proposed System

The aim of the system is to build the prediction model of students' academic performance by using multiple linear regressions (MLR). The proposed system examines the role of satisfaction on students' academic performance and investigates the relationship between satisfaction of students and academic performance and explores other factors that contribute academic performance using Multiple Linear Regression Method. The proposed system's overall system flow is shown in figure 1.

The dataset collected from UCI that consists of achievement of student and their related features in the secondary school of education. The dataset consists of 649 student's records with 33 features. The features are students' grades (G1, G2, G3), demographic features (school, sex, age, address, etc.), and social features (internet, romantic, freetime, goout, etc.). After data collection process, some preprocessing techniques are applied on dataset. Data cleaning is used to solve irrelevant missing part such as ignore the tuple or fill the missing values, inconsistent data and to deal with incomplete values. Data transformation is used to transform the raw data in a useful and efficient format so that it can be easily accepted and used by multiple linear regression method.

After preprocessing, data are stored in students' training database. This data are used MLR method to predict the student academic performance. And then, results are evaluated by four evaluation methods (accuracy, precision, recall and F-measure) and then show the result of prediction as shown in figure 2 as MLR Result.

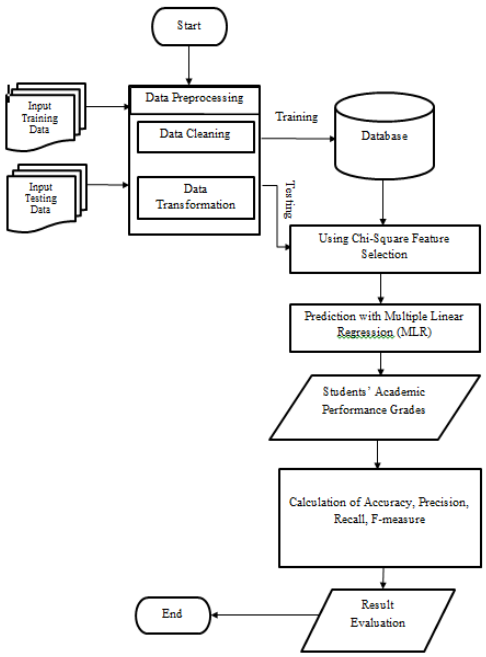


Figure 1. The proposed system architecture

Moreover, the data in student training are used for feature selection method (Chi-Square) and results are evaluated by four evaluation methods. The results show the students' grades and which features are related to improve students' academic performance.

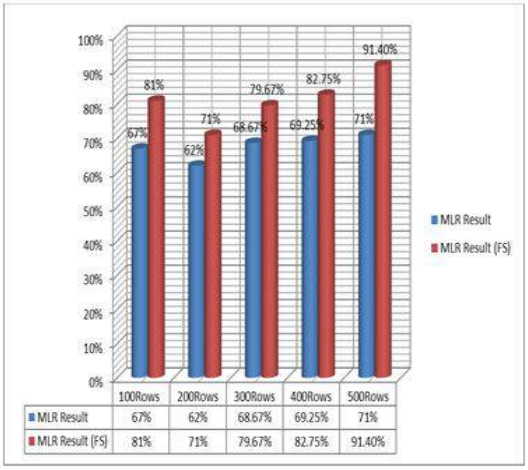


Figure 2. Calculation Results of MLR (with 32 variables) and MLR (with feature selection)

And then, testing data are used by MLR method and MLR with feature selection calculation and the results are shown in figure 3.

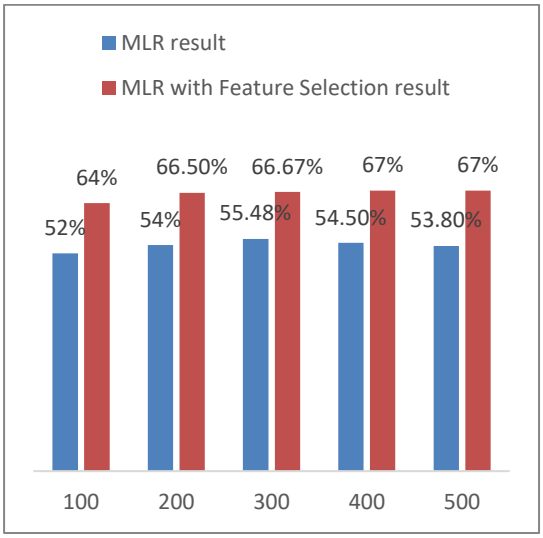


Figure 3. Calculation results of MLR and MLR with Feature Selection for testing data

According to the experimental results especially with training and testing data, applying multiple linear regression with feature selection approach offers better results than applying multiple linear regression only for the system.

4.1. Experimentation of the system

Assuming all of the assumptions for a multiple linear regression have been met, this can be done by generalizing to unseen data. The test data should be set aside and not looked at until ready to determine how well proposed regression model is generalizing. The train data is used to fit proposed model. The following figure 4 shows the performance analysis of the system with different data size.

Results are evaluated by four evaluation methods (accuracy, precision, recall and F-measure) called confusion matrix.

$$\text{Accuracy} = \frac{TP+TN}{TP+FN+FP+TN} \quad (1)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (3)$$

$$\text{F- measure} = 2\frac{Precision*Recall}{Precision+Recall} \quad (4)$$

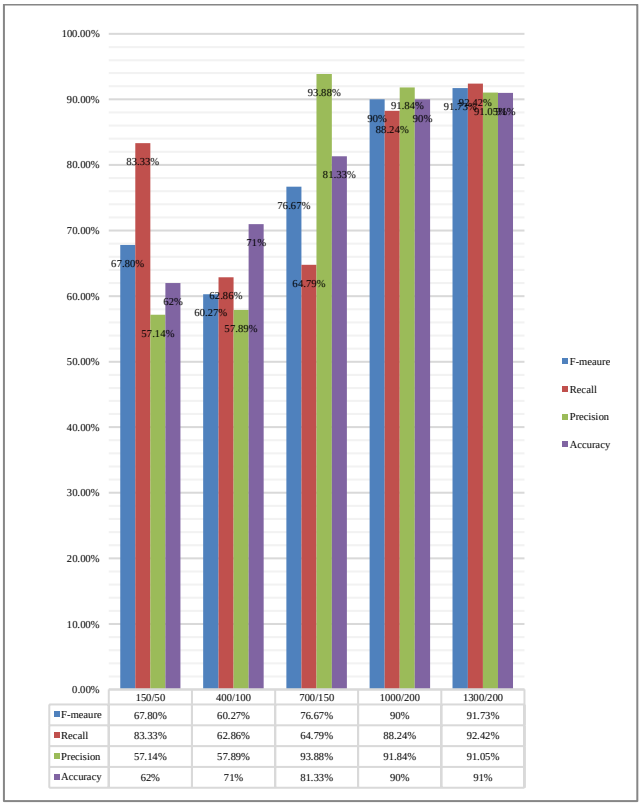


Figure 4. Evaluation of the system with Different Data Size

5. Conclusion

The various data mining techniques are used in analysing the academic performance of students and one of them is Multiple Linear Regression (MLR). This study detailed the prediction of students' performance of final grades by using their first grades, second grades and their related features. The developed system improved when Multiple Linear Regression Method with feature selection method is used and

this system will help the teachers and school management system to predict the students' academic performance and that can be achieved with various factors of students. The factors considered on UCI focus on demographic attributes and school performance over past years. Several studies have used business intelligence (BI)/ data mining (DM) methods to improve the quality of education and enhance school resource management.

Predicting students' academic performance has long been an important area of research in education. Most existing literature have made use of traditional statistical methods that run into the problems of over fitted models, inability to effectively handle large numbers of participants and predictors, and inability to pick out non-linearity that may be present. Regression-based ML methods that can produce highly interpretable yet accurate model for new predictions are able to provide some solutions to the aforementioned problems.

References

[1] Alaf.A.A, Thair.H and Ibrahim.A, "Analyzing Students' Academic Performance Through Educational Data Mining", IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT), 2015.

[2] Oyerinde O. D. and Chia P. A., "Using data mining to predict secondary school student performance", International Journal of Computer Applications (0975 – 8887) Volume 157 – No 4, January 2017.

[3] Paulo.C and Alice. S, "Predicting Students' Academic Performances – A Learning Analytics Approach using Multiple Linear Regression", www3.dsi.uminho.pt, 2008.

[4] R.R.Rajalaxmi, P.Natesan , N. Krishnamoorthy ,S.Ponni, "Preprocessing and Analyzing Educational Data Set Using X-API for Improving Student's Performance", 2015 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT), 2015.

[5] Sana, Isma. F. S and Qasim .A. A, "Regression Model for Predicting Engineering Students Academic Performance", 3C Tecnología. Glosas de innovación aplicadas a la pyme. Special Issue, May 2019.

[6] Scott. B. H., "Multiple Linear Regression Analysis: A Matrix Approach with MATLAB, Auburn University Montgomery", Alabama Journal of Mathematics, Spring/Fall 2009.

[7] Warey and Gregory, "Multivariate Analysis of Variance (MANOVA): I. Theory". Retrieved March 22, 2011.

[8] W. Rencher, Schaalje G. Bruce, "Linear Models In Statistics", Department of Statistics, Brigham Young University, Provo, Utah. second edition.

[9] Z. Smyth, "Nonlinear regression", Encyclopedia of Environmetrics (ISBN 0471 899976), 2002, Volume 3, pp 1405– 1411.

[10] Z. Rolph, Tatham L. R. and Black C.W, "Multivariate Data Analysis", fifth edition, chapter 6, pp.326- 352.

Web Page Category Classification Using Decision Tree Classifier and Recommendation of Related Links

Phyu Phyu Thant, Amy Aung
University of Computer Studies (Magway)
pphyuthant1o@gmail.com, amyaung@ucsmgy.edu.mm

Abstract

Today, there has been an exponential growth in the number of electronic documents and pages in the web that is needed accurate automated classifiers of machine learning method. The purpose of the web data mining is to find the useful knowledge or information from web contents, web usages and hyperlinks. In this paper, the web page category classification system is proposed. The background theory of this system is web data mining. This system is tested a collection of hyperlinks in the computer science domain and ten categories of class. In this paper, it is seen web preprocessing for extracting contents. For classification, the system uses TF-IDF features extraction and decision tree classifier. The result shows that the proposed system produces the category of classified page according to predefined class category and related links of its.

Keywords: Web Page Classification, TF-IDF, C4.5, Web Content Mining

1. Introduction

People have been using the Internet extensively in recent years. Today, people are interested in learning, teaching and doing business on the internet, besides searching on the internet for whatever they want. Therefore, the number of web pages has increased exponentially. As the number of sites growing, the web is necessary several ways to organize the web pages, to improve the web directory and for the performance of the search engine. There are massive amount of different things about the web, such as technology researches, challenges, explorations, inventions, and so on. Although web data mining is a group of data mining, pure traditional data mining is not used because the data on the web is semi-structured and unstructured. There are three different kinds of techniques: web structure mining, web content mining and web usage mining [3].

Web content mining is the analysis of content, knowledge and information on the web for a variety of purposes. This system is proposed as the web page category classification, which is the subset of web content mining. There are many mining tasks such as from hyperlinks, html documents, customer reviews, consumer opinions and comments, etc. This system uses hyperlinks for preprocessing of web page classification. This system uses TF-IDT features extraction and decision tree classifier for classification. The system aims to maintain and organize web directories, display related links and try to focus on crawling. By networking with this system in library, it is optimized to group the category on computer science domain for library. Therefore, this system is proposed as the powerful and useful system for students, researchers and interested person in IT.

2. Related Work

K. Thangairulappan and A. D. Kanagavel (2016) presented an improved Term Weighting technique for automatic and effective classification of web pages. In this system, the proposed method is to extract and select the most important features reducing the great dimensionality problem of classification. It is showed the better performance than most of the existing term weighting techniques in the result [4].

Dimpleveer Singh and Sumit Malhotra (2018) presented the system that is classified categories of Intra News from BBC news dataset. In this paper, sports category has been selected for intra class classification. There are five different sports have been provided named as tennis, rugby, football, cricket and athletics in sport category. Feature extraction has been carried out N-gram

Term Frequency-Inverse Document Frequency (TF-IDF) and the classification method has been performed decision tree classifier. Experiment result shows about 96% accuracy in intra news classification in the paper [5].

A. M. James Raj, F. Sagayraj Franics and P. Julian Benadit (2016) proposed an approach that is composed of three steps for classification of the web. The first phase is to extract unique word and features by using TF-IDF (weighting calculation function) for retrieving informative contents. And then, the system takes not only terms but also HTML tags together when extracting contents. The decision tree algorithm (J48) is used as an information learning method for extraction rules. Finally, Hybrid Fire Fly algorithm based Naïve Bayes Classifier (FA-NBC) is used for classification to predict results [7].

3. Web Page Preprocessing

Some preprocessing tasks are usually carried out before the web pages in a dataset are operated for classification. Firstly, web preprocessing is needed to get the content from hyperlink for classification. In the web preprocessing phase, HTML tag, Ads, footer, and unnecessary part are removed to get willing text.

Text is unstructured data and is organized useful and useless data. So, text needs preprocessing step to clean noise. Text preprocessing are tokenization, removing stopword, stemming, and handling of digits, and Transforming cases of letters [2, 3].

- Tokenization: In a phrase, sentence and document, tokenization is the work of cutting it up into pieces.
- Removing stopword: It is the method of removing unimportant or unnecessary words that is not pointed any contents. Prepositions, articles, some pronouns and conjunctions are stopwords that point to natural.
- Stemming: It is changing the root form or common base form of word.
- Handling of digits: In this step, digits are removed or filtered in traditional IR systems. Example of digits are dates, number, times, and other regular expressions, etc.
- Transforming cases of letters: It is transformed all letters to either the lower or upper case.

4. Web Page Classification

Web page classification is used as content-based assignment of one or more predefined categories (topics) to web pages. The automatic web page classification consists of two phases that are learning phase and classification phase. Automatic web page classification tasks can be divided into three sorts: supervised web page classification where some external mechanism (such as human feedback) provides information on the correct classification for web pages, unsupervised web page classification, where the classification must be done entirely without reference to external information, and semi-supervised web page classification, where parts of the web pages are labeled by the external mechanism [6].

5. Feature Extraction based on TF-IDF

TF_IDF is the well-known algorithm used in text mining research to calculate the weight of each term. TF stands for term frequencies and IDF stands for Inverse document frequency [9]. Term frequencies (TF) mean how often a term appears in document. The inverse document frequency (IDF) is calculated the inverse probability of finding a word in a document collection. Equation 1 is a calculation for TF-IDF.

$$W_{ij} = tf_{ij} \times \log \frac{N}{df_i} \quad (1)$$

In equation (1), w_{ij} is the weight of the term i in the document j, N is the total number of documents, tf_{ij} is the duplicates of the term i in the document j, and df_i is the quantity of documents containing the term i [8].

6. C4.5 Decision Tree Classifier

The decision tree is a tree structure in which the internal node represents the test of the attribute and the branch refers to the test result, and the leaf node holds the class label [1]. Types of decision tree classifier are ID3, C4.5, CART,

CHAID, MARS, and conditional inference trees. Decision tree is simple to use, understand and explain due to generate rules through them. It requires very little efforts to prepare and produce. It provides strategic answers to uncertain situations. But training data using decision tree is very high-cost to improve efficiency. Decision tree C4.5 is used for classification in this system. C4.5 is added split info and gain ratio to ID3 (Iterative Dichotomiser). C4.5 algorithm is applied to hold discrete and continuous values. The C4.5 algorithm is as follow:

Algorithm: C4.5 Decision Tree
Input: An attribute-valued dataset *D*
1: Tree = { }
2**: if** *D* is "pure" OR other finishing condition met **then**
3: **stop**
4: **end if**
5: **for all** attribute a € D **do**
6: Compute information-theoretic condition if we split on a
7: **end for**
8: a_{best} = Best attribute according to above calculated condition
9: Tree = Create a decision node that tests a_{best} in the root
10: D_v = persuaded sub-datasets from *D* based on a best
11: **for all** D_v **do**
12: $Tree_v$ = C4.5 (D_v)
13: Attach $Tree_v$ to the corresponding branch of Tree
14: **end for**
15: **return** Tree

The normalized information gain is applied to find the test attribute of every node in the tree. This function is as follows:

$$\mathrm{Info(D)} = -\sum_{i=1}^{m} P_i \ \mathrm{Log_2}\ (P_i) \quad (2)$$

$$\mathrm{Gain(A)} = \mathrm{Info(D)} - \mathrm{Info_A(D)} \quad (3)$$

$$\mathrm{Info_A(D)} = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \mathrm{Info(D_j)} \quad (4)$$

$$\mathrm{SplitInfo(A)} = -\sum_{i=1}^{m} \frac{|C_i|}{C} \log 2 \frac{|C_i|}{C} \quad (5)$$

$$Gainratio = \frac{Gain(A)}{SplitInfo(A)} \quad (6)$$

P_i is the probability of arbitrary tuple in separation D. Info(D) is the average amount of information required to classify the class label of a tuple in D. $|D_j|/|D|$ acts as the weight of the j^{th} partition. $Info_A(D)$ is the expected information required to identify a tuple from D based on the separation by A. C_i is the objects in class C that have value A of A_i. SplitInfo(A) is the information due to the split of class C on the basis of the value of the categorical attribute A. The attribute A with the greatest information gain, Gain ratio (A), is chosen as the splitting attribute at Node N [10].

7. Proposed System Design

Figure 1 is described about system design of web page category classification system. There are four main parts in this system. The four steps are pre-processing, calculation weight, classification and displaying related hyperlinks that are given more information or knowledge for user.

The preprocessing phase involves web preprocessing and text preprocessing. In web preprocessing phase, the contents are extracted from hyperlinks to get only text.

Every page is not simple to get contents because of different type. There are two different types (static page and dynamic page) of pages generally. Static page is easy to extract than dynamic page. Static page is applied in languages like HTML, CSS and JavaScript, etc. and remained the same until someone changes it manually. Dynamic page is not simple because it is applied in languages like AJAX, CGI, ASP, ASP.NET, etc. Dynamic web pages are used the information is changed frequently.

This system is implemented python programming language. For crawling, selenium and beautifulsoup are used in this system. Beautifulsoup and selenium are python libraries for web crawling. Beautifulsoup can apply HTML pages and it ignores javascript. Selenium can make available the data from javascript link. Therefore, the combination of selenium and beautifulsoup can handle the dynamic pages to get desired contents. Scrapy can be used instead of selenium. After crawling, text preprocessing is performed. In text preprocessing step, tokenization, filtering stopword, stemming,

removing of digits, and transforming cases of letters are active.

And then, the weights of words are calculated using TF-IDF features extraction. In the classification step, the system is used decision tree classifier C4.5 to produce their class category.

Next, the system is displayed class category of tested page and showed list of hyperlinks deal with its category to get more information and knowledge for user. Finally, a tested link is added to existing list of its category for next testing.

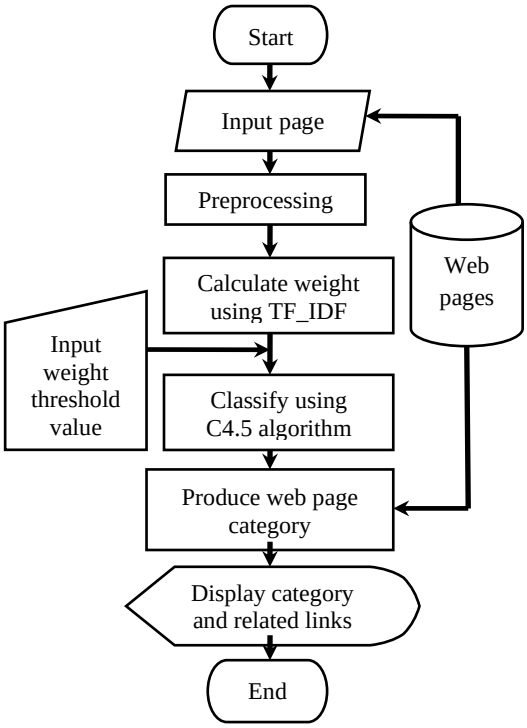


Figure 1. Proposed System Design

8. Class Categories for Classification

There are ten class categories for this system. They are

- Database management system,
- Data structure,
- Distributed system,
- Artificial intelligence,
- Cloud computing,
- Web data mining,
- Software engineering,
- Operating system,
- Web engineering and
- Natural language processing.

9. Data Collection

In this system, the numbers of pages for training and testing data are 300 pages (including 30 pages of each category). The data resource is collected from the Institute of Electrical and Electronics Engineers (IEEE), GreeksforGreeks and other sites under the computer sciences domain.

IEEE is involved conference papers, journals and research documents for technology information, computing, and engineering throughout the globe and this system is collected 182-pages from it. GreeksforGreeks contains courses, tutorial, services for learning and jobs in IT field. Therefore, the system is collected 101-pages from GreeksforGreeks. The remaining 17-pages are gathered from other sites under computer science domain.

10. Explanation of the System

For explanation, this system tests seven web pages form computer technology domain. Firstly, this approach carries out the tokenization and removing stopwords process. And then, keywords are extracted from each web page. Sample web pages are presented in Table (1).

Table 1. Training Data

Name	Content Description	Class
Web page 1	Digital image processing (DIP) consists of noise reduction, segmentation and feature extraction. Digital image processing (DIP) also consists of recognition tasks.	Digital image processing
Web page 2	Students learn and test Matlab. Then, they apply and test image by implementing and investigating image processing algorithms in Matlab.	Digital image processing
Web page 3	Statistic data mining is the process of discovering patterns from database. In large data sets, methods are involved at the intersection of machine learning, statistics and database systems.	Data mining

Web page 4	Data mining is a subfield of computer science and statistics with an overall goal to extract information from a data set. Statistics is a component of mining.	Data mining
Web page 5	Data analysis is used to test models and hypothesis on the data set. So, model-based testing is essential.	Data mining
Web page 6	Cryptography is the technique for secure communication in the presence of third parties. Cryptography is the study for secure communication.	Cryptography
Web page 7	Cryptography is the study for secure communication. According to Matlab, it is about constructing and analyzing protocols that prevent third parties or the public from reading private messages. Matlab is useful for cryptography.	Cryptography

After extracting keywords, this system calculates the weight of each keyword by using TF-IDF methods. Sample weight results from web page 1 is described in Table (2).

Table 2. Weight Results from Web Page 1

Name	Keyword Name	TF	IDF	Weight
Web page 1	Digital	1	0.845	0.845
	image	1	0.544	0.544
	process	1	0.368	0.368
	DIP	1	0.845	0.845
	consist	1	0.544	0.544
	noise	0.5	0.845	0.423
	reduct	0.5	0.845	0.423
	segment	0.5	0.845	0.423
	feature	0.5	0.845	0.423
	extract	0.5	0.544	0.272
	recognise	0.5	0.845	0.423
	task	0.5	0.845	0.423

For feature selection, this system uses 0.5 for the weight threshold value. This system extracts features that have equal to or more than threshold value. These extracted features are used as the training data for C4.5 classification. These data are shown in Table (3).

Table 3. Training Data for C4.5 Classification

Digital	Image	DIP	Consist	Matlab	…	cryptography	secure	Class
1	1	1	1	0	…	0	0	digital image processing
0	1	0	1	1	…	0	0	digital image processing
0	0	0	0	0	…	0	0	data mining
0	0	0	0	0	…	0	0	data mining
0	0	0	0	0	…	0	0	data mining
0	0	0	0	0	…	1	1	cryptography
0	0	0	0	1	…	1	0	cryptography

To choose the root node for the decision tree, this approach computes gain ratio of all features. Table (4) shows the gain ratio results that are obtained from the first iteration.

Table 4. Gain Ratio Result from First Iteration

Keyword	Gain Ratio Result
digital	0.5173
image	1 **(Root Node)**
DIP	0. 5173
consist	0. 5173
Matlab	0.3368
statistic	0.5441
database	0.3351
min	0.3351
test	0.2132
model	0.3351
cryptography	1
secure	0. 5173
commun	0. 5173

According to the gain ratio results, this system chooses the feature that has greatest gain ratio value as the root node. Then, this system continues to calculate the gain ratio for the next node selection. This system stops the gain ratio calculation until each leaf node has different classes. Table (5) shows the gain ratio results that are obtained from the second iteration.

Table 5. Gain Ratio Result from Second Iteration

Keyword	Gain Ratio Result
digital	0
DIP	0
consist	0
Matlab	0.4459
statistic	0.43305
database	0.2366
min	0.2366
test	0.2366
model	0.2366
cryptography	1 **(Root Node)**
secure	0.4459
commun	0.4459

If each leaf node has different classes, this system generates the decision tree. Decision tree is presented in Figure.2.

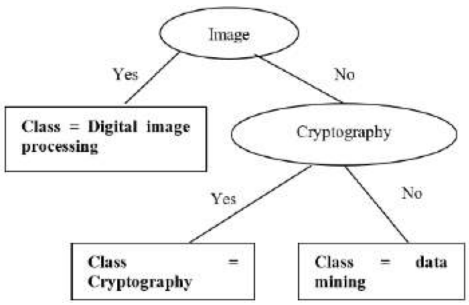


Figure 2. Decision Tree

Then, this system generates the decision rules according to the decision tree. The following decision rules are obtained from the seven training web pages. The decision rules are as follows:

- **IF** "Image" = Yes **Then** Class = "Digital image processing"
- **IF** "Image" = No **AND** "Cryptography" = Yes **Then** Class = "Cryptography"
- **IF** "Image" = No **AND** "Cryptography" = No **Then** Class = "Data mining"

Finally, this system can classify the other web pages by using the generated decision rules. The more web pages are trained, the more decision rules are obtained.

11. Experimental Result of the System

The dataset used for this result work is collected from IEEE, GreeksforGreeks and other sites under the computer sciences domain. There are ten class categories for this research.

The proposed web page classification system has been considered for performance evaluation as below. In Table (6), 210 pages are training and 90 pages are testing. This table is described the precision, recall, f1-score and support for each class category. The accuracy result is 87% for classification using decision tree classifier. Figure.3 is Described the result of precision, recall, F1-score for each category.

Table 6. Gain Ratio Result from Second Iteration

Web Page	Precision	Recall	F1-score	Support
database management system (C1)	0.86	0.67	0.75	9
data structure (C2)	1.00	0.71	0.83	7
distributed system (C3)	0.88	0.88	0.88	8
natural language processing (C4)	1.00	0.89	0.94	9
web engineering (C5)	0.89	0.89	0.89	9
operating system (C6)	1.00	0.75	0.86	8
cloud computing (C7)	0.89	1.00	0.94	8
software engineering (C8)	0.65	1.00	0.79	17
web data mining (C9)	1.00	0.83	0.91	6
artificial intelligence (C10)	1.00	0.78	0.88	9
Accuracy = 87				90

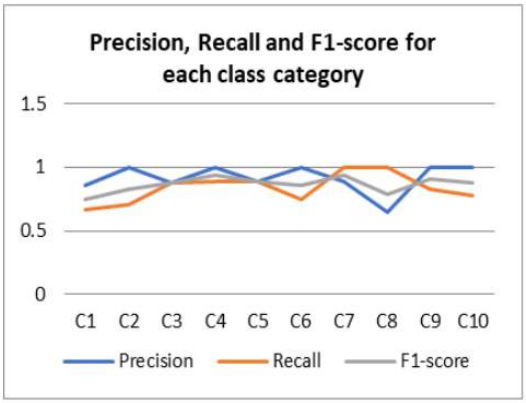


Figure 3. Precision, Recall and F1-score for each class category

Table (7) is showed the experimental result for this classification. In 210 pages training and 90 pages testing of this classification, the rate of correction is 87% and the rate of error is 13%. Correct and error rate are shown in Figure.4.

Table 7. Experimental Result of the System

Correct rate	87%
Error rate	13%

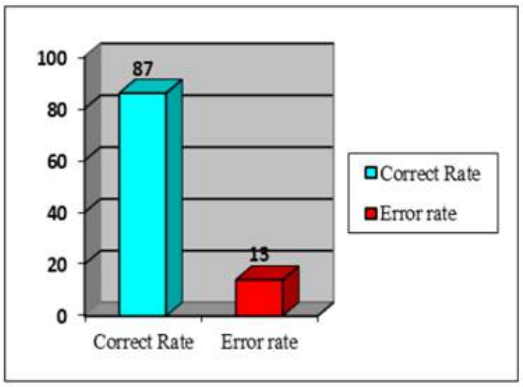


Figure 4. Correct and Error Rate of the System

In figure.5, the accuracy by using threshold value (0.2) for filtering keywords is 87%. The accuracy for threshold value (0.3) is 48%. The accuracy result of threshold value (0.4), (0.5) and (0.6) are 48%, 49% and 43%, respectively. Therefore, the threshold value (0.2) is optimized for accuracy in this system.

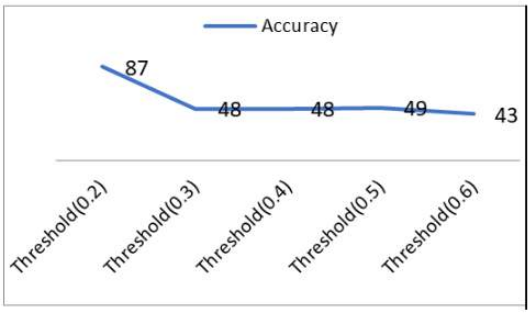


Figure 5. The accuracy result for various threshold for classification

12. Conclusion

With the widening of the web pages, it is required to distinguish the group of the pages for the development of web directories and the improvement of search engine performance. The system is intended to solve the requirements mentioned above due to the growth of web pages. In this system, decision tree classification has been used. Precision, recall, f-measures were evaluated for performance. This system can be changed other feature extraction method instead of TF-IDF and replaced classification method similarly. For further extension, web page classification can be used ontology based semantic search. Besides, PDF file classification on the web can be added in this system and the performance of web crawling can be developed more and more.

References

[1] H. Jiawei and K. Micheline, Data Mining Concepts and Techniques, Simon Fraser University, United States of America, 2001.

[2] Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, An Introduction to Information Retrieval Cambridge University Press Cambridge, England, 2009

[3] Bing Liu, Web Data Mining second edition, University of Illinois, Chicago, 2011 Springer

[4] K. Thangairulappan and A. D. Kanagavel, "Improved Term Weighting Technique for Automatic Web Page Classification", Journal of Intelligent Learning Systems and Applications, vol. 8, pp. 63-76, 2016.

[5] S. Dimpleveer and M. Sumit, "Intra News Category Classification using N-gram TF-IDF Features and Decision Tree Classifier", IJSART – Volume 4 Issue 3 – MARCH 2018.

[6] Hnin Pwint Myu Wai, Phyu Phyu Tar and Phyu Thwe, "Ontology based Web Page Classification

System by using Enhanced C4.5 and Naïve Bayesian Classifiers", 2018 IEEE.

[7] A. M. James Raj and F. Sagayraj Franics, "Optimal Web Page Classification Technique Based on Informative Content Extraction and FA-NBC", Computer Science and Engineering, Scientific & Academic Publishing, vol. 6, no. 1, pp. 7-13, 2016.

[8] W. Zhang, T. Yoshida, and X. Tang, "A comparative study of (TF-IDF), LSI and multi-words for text classification", Expert Systems with Applications, vol. 38, pp. 2758-2765, 2011.

[9] A. A. Hakim, A. Erwin, K. Eng, M. Galinium, and W. Muliady, "Automated document classification for news article in bahasa Indonesia based on term frequency inverse document frequency (TFIDF) approach", ICITEE, pp. 1-4, 2014.

[10] A. M. Mahmood, N. Satuluri, and M. R. Kuppa, "An Overview of Recent and Traditional Decision Tree Classifiers in Machine Learning", International Journal of Research and Reviews in Ad Hoc Networks, Vol. 1, No.1, 2011.

Diagnosis Classification Soybean Disease Using Machine Learning Techniques

Hnin Nwe Phyo[1], Dr. Myo Khaing[2]
University of Computer Studies (Maubin)[1], University of Computer Studies (Mandalay)[2]
hninnwephyo13@gmail.com[1], myokhaingucsm@gmail.com[2]

Abstract

The prevention of disease transmission in plants is largely dependent on early detection of pathogen infection. Plant diseases can be identified using machine learning techniques before they fully manifest their symptoms. The more problem has been solved; the more reliable system has been built. This paper will help the agricultural development. Machine learning is a new area of study for agricultural analysis. The uses of machine learning techniques in the sector of agriculture are the main topic of this study. Different machine learning techniques are in use, such as k-Nearest Neighbors (k-NN), J48 Decision Trees, Naïve Bayes and Decision Table for very recent applications of data mining techniques in agriculture field. This paper, properly classifies the problem of soybean diseases. For this purpose, different types of machine learning techniques were evaluated on soybean disease data sets. This paper discusses the development of an expert system to diagnose soybean disease using machine learning techniques.

Keywords: k-Nearest Neighbors (k-NN), J48 Decision Trees, Naïve Bayes, Decision Table

1. Introduction

Machine Learning is the study and development of algorithms that can extract information from a sample dataset and use that information to generate data-driven predictions or choices on fresh data. Machine Learning requires the development of computer systems that adjust or learn when presented with fresh data, it is comparable to data mining. Machine learning uses data to detect patterns in data and adjusts program actions as appropriate, whereas data mining extracts data for human comprehension.

Data or observations, direct experience, or instruction are all used in machine learning. Machine learning generally focuses on discovering ways to improve future performance based on current experiences. The goal is to create learning algorithms that can learn on their own, without having the need for human involvement. Machine learning creates the means for the computer to design its own program based on examples provided, as opposed to merely programming it to solve the problem. Artificial intelligence's fundamental subfield is machine learning. These things are otherwise impossible to complete.

In sentiment classification, the experiments are performed to determine the classification accuracy of four algorithms in terms of which is the better predictive algorithm of user's decision making. This paper describes as follows: section 2 is addresses related work, section 3 includes the theoretical background of machine learning, section 4 includes supervised learning, section 5 includes regression, section 6 includes classification and section 7 is datasets for four methods comparison. In section 8, the proposed of methodology is discussed. The experimental results will be presented in section 9. Finally, section 10 describes the conclusion.

2. Related Work

In this paper, some references are used from previous proposes papers. R.S, Michalski and R.L. Chilausky proposed “An Experimental Comparison of the Two Methods of Knowledge Acquisition in the Context of Developing an Expert System for Soybean Disease Diagnosis”. The author described Comparison of Expert Derived and Inductively Derived Rules the research was supported in part grant from the National Science Foundation and in part by grant [11]. Vinita Shah, Prachi Shah introduced

"Groundnut Crop Yield Predication Using Machine Learning Techniques". The author discussed Comparison of four different algorithms are used multiple linear Regression, Regression Tree, k-nearest neighbor, and artificial neural network. This system crop yield prediction is an important area of research, which helps in ensuring food security all around the world [13]. Minarni, Indra Warman, Yuhendra Teknik Informatika, Institut Teknologi Padang, Indonesia presented "Implementation of Case-Based Reasoning and Nearest Neighbor Similarity for Peanut Disease Diagnosis". The author discussed this study produces an expert system diagnoses the peanut diseases using case-based reasoning inference and nearest neighbor similarity [9].

3. Background Theory

In essence, machine learning converts data into knowledge. It is impossible to get information or understanding from raw data by merely looking at it. For instance, the user cannot tell if an email is spam by looking at the frequency of a single phrase; instead, must look at the frequency of a number of words, the length of the email, and other criteria. Statistics are also employed in machine learning, which may be used to any problem requiring the interpretation of data and subsequent action. The facts discovered can then be utilized to choose a new collection of data. Static programs are typically employed to tackle deterministic issues with clear solutions, however for nondeterministic problems that lack sufficient data, apply a method known as machine learning, because there were insufficient datasets to train the algorithms on the beginning, it was challenging to use machine learning to make sound conclusions. However, with the rise of sensors and their capacity to connect to the Internet, the true challenge today is to effectively sort through the voluminous free data that is accessible and use it to train machine learning algorithms [2].

3.1. Introduction to the Machine Learning

The creation of efficient general-purpose algorithms is the main objective of machine learning research. In the context of learning, one should be concerned not just with time and space efficiency but also with the amount of data that the learning algorithm requires. Learning algorithms should solve issues in a general way that makes them easily applicable to a variety of learning challenges, like those mentioned above. A prediction rule that is accurate when making predictions on new data should be the learning process' principal goal. Machine learning algorithms are data driven and capable of analyzing vast amounts of data, they have a major benefit over static programming in that the results are frequently more accurate with machine learning than with static programming results. Figure 1 shows the general process involved in a typical machine learning model [5].

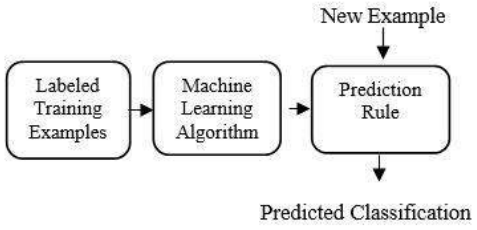


Figure 1. An illustration of a typical Machine Learning Process

3.2. Machine Learning Algorithms

Machine learning algorithms can model each problem differently depend on the input data. Based on their preferred learning styles, machine learning algorithms can be divided into four groups. They are

- Supervised Learning
- Unsupervised Learning
- Semi supervised Learning and
- Reinforcement Learning

Traditional classifications of learning techniques include supervised, unsupervised, semi-supervised, and reinforcement learning. The fields of signal processing, optimization, control, modeling and identification, and pattern recognition all make extensive use of supervised learning. The main applications for unsupervised learning systems include pattern recognition, clustering, vector quantization, signal coding, and data analysis [7].

4. Supervised Learning

Supervised learning is based on a direct comparison between the actual network output

and the desired output. Network parameters are adjusted by a combination of the training pattern set and the corresponding errors between the desired output and the actual network response. Supervised earning is a closed-loop feedback system, where the error is the feedback signal. The trained network is used to emulate the system.

The test dataset is used to assess the algorithm's effectiveness after it has been trained using the training dataset. Next, a procedure known as cross-validation is employed to compare various feature selection, dimensionality reduction, and learning algorithm combinations. The most popular one is k-fold cross-validation, which involves splitting the training dataset into k subsets (k-1 subset is used for training and 1 subset can be used for testing). This splitting can assist in estimating the average error rate once the learning process is completed. Since each feature can have a distinct range of values when learning and making judgments, normalization is done to give equal weight to every feature in the dataset. It must be done on both training and test data. There are numerous learning algorithms in this work. In the next sections k-Nearest Neighbor, J48, Decision tables, and Naive Bayes will discuss. In the post-processing phase, they would test the algorithm's accuracy using test data. If the accuracy didn't meet our expectations, they could always restart the process by giving the algorithm access to more abundant and accurate data. They can also improve the way the user gather and prepare their input datasets to get better results. The user can use the algorithm to forecast actual data once it gets the anticipated accuracy [1].

There are two types of algorithms for supervised learning. They are

- Regression and
- Classification

4.1. Advantages of Supervised Learning

Supervised learning allows collecting data and produces data output from previous experiences. The usages of supervised learning experience to assist optimize performance criteria. Supervised machine learning helps to solve various types of real-world computation problems.

4.2. Disadvantages of Supervised Learning

Supervise learning can be challenging when it comes to big data classification. The computing time required for training supervised learning is significant. Unwanted data reduces productivity. Pre-processing data presents a significant hurdle. Constantly in need of updating supervised algorithms are easily over fit by anyone.

5. Regression

Finding correlations between dependent and independent variables is the process of regression. It aids in the forecast of continuous variables like market trends, house values, and other things.

Regression Algorithm Types are

- Simple Linear Regression
- Multiple Linear Regression
- Polynomial Regression
- Support Vector Regression and
- Decision Tree Regression.

6. Classification

Finding a function to divide the dataset into classes based on several parameters is the process of classification. In classification, data is divided into various classes by a computer program that has been trained on the training dataset [10].

ML Classification Algorithm Types:

The following categories of classification algorithms exist. They are

- Logistic Regression
- K-Nearest Neighbors
- Support Vector Machines
- Kernel SVM
- Naïve Bayes classification
- Decision Tree Classification
- Random Forest Classification
- Evaluation of classifiers
- Rule induction
- Classification using association rule
- Naïve Bayesian Naïve Bayes for text classification and
- Ensemble methods: Bagging Boosting

This paper, four important classification algorithms are

- k-Nearest Neighbors

- J48 Decision Tree
- Naïve Bayes and
- Decision Table

6.1. k-Nearest Neighbors

In this section, the first classification algorithm will be discussed: k-Nearest Neighbors. Compared to other machine learning methods, it is simple to comprehend and straightforward to implement. The letter "K" stands for the number of closest neighbors to a new unknown variable that needs to be forecasted or categorized. The k-nearest neighbors (k-NN) algorithm is a simple, supervised machine learning algorithm that can be used to solve both classification and regression problems [6].

6.1.1. Advantages

The k-NN algorithm does not require training prior to produce predictions, the new data can be added without affecting the system's accuracy. Implementing k-NN is fairly simple. k-NN implementation just needs two parameters: the value of K and the distance function (e.g. Euclidean or Manhattan etc.). The main importance of using k-NN is that it's easy to implement and works well with small datasets.

6.1.2. Disadvantages

k-NN does not perform well with large dataset: In large datasets, the algorithm's speed suffers due to the high cost of computing the distance between each new point and each current point.

k-NN Does not perform well with high dimensional data: The k-NN algorithm performs poorly with high dimensional data because it becomes challenging for the algorithm to calculate the distance in each dimension as the number of dimensions increases.

k-NN Feature scaling is required: Before applying the KNN method to any dataset, feature scaling (standardization and normalization) must be completed. In the absence of this, KNN could produce inaccurate predictions.

6.2. J48 Decision Tree

A decision tree is a supervised learning technique that has a pre-defined target variable and is often used in classification problems. Building a model of classes from a set of records that have class labels is the process of classification. Decision Tree Algorithm is to find out the way the attributes vector behaves for a number of instances [3]. The classes for the freshly generated instances are also being discovered on the basis of the training instances. The rules for the target variable's prediction are generated by this algorithm. With the help of tree classification algorithm, the critical distribution of the data is easily understandable. An extension of ID3 is J48. Accounting for missing values, decision tree pruning, continuous attribute value ranges, the development of rules, etc. are further characteristics of J48.

6.2.1 Advantages

The main benefits of decision trees are that they are computationally affordable and that people can readily interpret the data. Additionally, the construction of a decision tree is not significantly impacted by missing values in the data. Technical teams and stakeholders may easily understand a decision tree model because it is so simple.

6.2.2 Disadvantages

The decision tree's structure can drastically change in response to little changes in the data, which might cause instability. Sometimes calculations for a decision tree can be significantly more complicated than for other methods. Decision trees usually require more time throughout the model training process.

6.3 Naïve Bayes

Naive Bayesian classifier is based on Bayes' theorem with the independence assumptions between predictors. A Naive Bayesian model is simple to construct and does not require time-consuming iterative parameter estimation, making it especially beneficial for very large datasets. Naive Bayes is a machine learning algorithm to solve classification problems. It is one of the simplest yet powerful ML algorithms in use and finds applications in many industries. They would make the best prediction and give it a probability under the Naive Bayes model [8].

6.3.1 Advantages

The main benefit of Naive Bayes' is that it can accommodate several classes and work with less data. Evaluation of the conditional probability is simple. Naive Bayes' is very quick - no iterations necessary because the probabilities may be calculated directly. Therefore, this method is helpful in situations where training speed is crucial.

6.3.2 Disadvantages

Naive Bayes assumes that all predictors (or features) are independent, rarely happening in real life. If the test data set has a categorical variable of a category, it is not present in the training data set. The Naive Bayes model will assign it zero probability and will not be able to make any predictions in this regard [4].

6.4 Decision Table

Decision tables are used to model complicated programming logic. They can make it simple to understand that all potential conditions have been taken into account. The decision table are composed of four parts: conditions, actions, condition alternatives and actions for the rules. A decision table that includes every conditional statement.

Decision tables are used to lay out in tabular form all possible situations which a business decision may encounter. A matrix of causes and effects is listed in a decision table. A different combination is represented by each column. The goal of decision tables to structure logic.

6.4.1 Advantages

Decision tables are very helpful in test design technique. It aids testers in investigating the results of various input combinations and other software states that apply business rules. It offers a consistent manner of expressing intricate business rules, which is advantageous to both developers and testers.

6.4.2 Disadvantages

The decision table is not equivalent to complete test cases containing set-by-step instructions of what to do in what order. If the user has a lot of combination, it may not possible or sensible to test every combination.

7. Datasets for four Methods Comparison

This soybean disease dataset was obtained from the R. S. Michalski's research while affiliated with University of Illinois, (Donor:MingTan&JeffSchlimmer(Jeff.Schlimmer % cs.cmu.edu)).

The soybean disease dataset which has 638 instances, 35 attributes and 19 classes. In this work, out of 683 soybean disease dataset, 70% are used as training data and 30% are used as testing data.

Date, plant-stand, precip, temp, hail, crop-hist , area-damaged, severity, seed-tmt, germination, plant, leaves, leafspot-halo, leafspot-marg, leafspot-size, leaf-shread, leaf-malf, leaf-mild, stem, lodging, stem-cankers, canker-lesion, fruiting-bodies, external decay, mycelium**,** int-discolor, sclerotia, fruit-pods, fruit spots, seed, mold-growth, seed-discolor, seed-size, shriveling and roots represents 35 attributes of the soybean disease dataset.

The types of the diseases are diaporthe-stem-canker**,** charcoal-rot, rhizoctonia-root-rot, phytophthora-rot, brown-stem-rot, powdery-mildew, downy-mildew, brown-spot, bacterial-blight, bacterial-pustule, purple-seed-stain, anthracnose, phyllosticta-leafspot, alternarialeaf-spot, frog-eye-leafspot, diaporthe-pod-&-stem-blight, cyst-nematode, 2-4-d-injury and herbicide-injury [12].

8. Proposed Methodology

This paper proposes to compare k-Nearest Neighbors, Decision Trees, Naïve Bayes and Decision Table classification algorithm based on classification of datasets such as soybean disease. Moreover, the performance of each method with each dataset is also compared and analyzed.

There are four steps in data preprocessing. Step one is importing the row dataset. Step two verifies the values which are missing. View the categorical values are described in step three. Dividing the data set into a training and test set are described in step four.

Data normalization is generally considered the development of clean data. It increases the

cohesion of entry types leading to cleansing, lead generation, segmentation and higher quality data. In this paper, data are normalized by data normalization process.

The overview flow of the system is illustrated in figure 2.

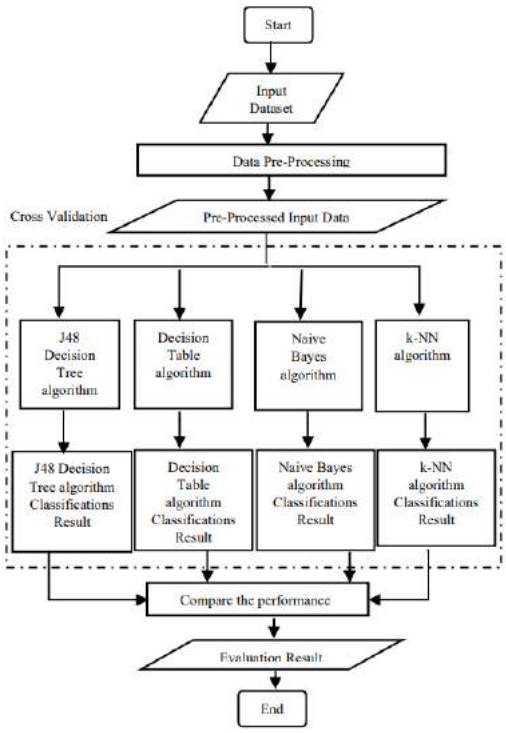


Figure 2. Overview Flow of the System

9. Experimental Results

The experimental result in the analysis is done by using an Intel(R) Core (TM) i7-5500U CPU with @ 2.40GHz 2.40 GHz processor along with 4 GB of RAM and Apache NetBeans IDE 12.6 programming language along with Java version 11.0.1. It will show the comparison between the result of k- Nearest Neighbors, J48, Naïve Bayes and Decision Table for soybean disease dataset that depend on iteration.

The evaluation result used training dataset with J48 is 96.3397%. If Decision Table is used, the result is 87.2621%. If Naive Bayes is used, the result is 93.7042%. If k-NN is used, the result is 99.8536%. Among then k-Nearest Neighbors algorithm (k-NN) algorithm is high dimension to classify the observation. As a result, it performs the best k-Nearest Neighbors algorithm (k-NN) for prediction model and it gives better results than other three techniques.

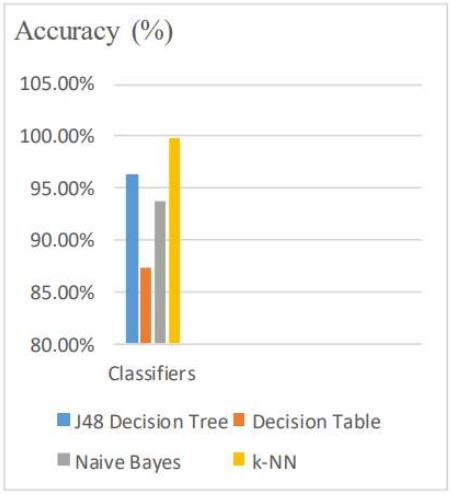


Figure 3. The compare result four different algorithms are used training dataset

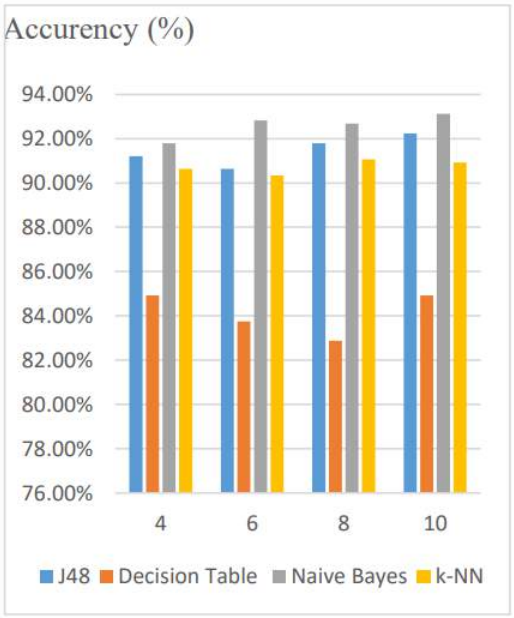


Figure 4. The compare result four different algorithms are used cross-validation

In the cross-validation result, there are 4-cross validation. If J48 is used, the result is 91.22%. If Decision Table is used, the result is 84.92%, If Naive Bayes is used, the result is 91.80%. If k-NN is used, the result is 90.63%. On the other hand, there are also 6-cross validation. If J48 is used, the result is 90.6296%. If Decision Table is used, the result is 83.7482%. If Naive Bayes is used, the result is 92.8258%. If k-NN is used, the result is 90.3367%. In this section, there are also 8-cross validation. If J48 is used, the result is 91.8009%. If Decision Table is used, the result is 82.8697%. If Naive Bayes is used, the result is 92.6794%. If k-NN is used, the result is 91.0688%. In 10-cross validation result, If J48 is

used, the result is 92.2401%. If Decision Table is used, the result is 84.9195%. If Naive Bayes is used, the result is 93.1186%. If k-NN is used, the result is 90.9224%.

Therefore, the evaluation findings may vary depending on the dataset.

10. Conclusion

There are many real-life use cases to create unique machine learning projects. If the users are still having trouble coming up with an actual use case, they look for something unique and useful, such as a machine learning project where they can compare several machine learning classification algorithms.

Classification in machine learning refers to training a model to identify the category to which a given entry belongs. Since there are so many different categorization methods in machine learning, it would be a fantastic and original machine learning project for a novice if the user could present a thorough comparison of these techniques. To complete this job, they must first select a classification-based issue statement and list all possible classification algorithms.

Four kinds of algorithm are used in this paper. They are k-Nearest Neighbors (k-NN), J48 Decision Trees, Naïve Bayes and Decision Table. Among them Naïve Bayes Algorithm is the best by according the results.

This paper has been performed the experiments in order to determine the classification accuracy of four algorithms in terms of which is the better predictive algorithm of user's decision making, with the help of an attractive data mining tool known as Python, and Apache Net Beans IDE. The system will give diagnosis for respective diseases to famers.

References

[1] AK Jain, RPW Duin, Jianchang Mao Statistical pattern recognition: a review. IEEE Trans Pattern Analysis and Machine Intelligence - 2000. 22(1):4–37.

[2] Bezdek,J.(1974), “Fuzzy mathematics in pattern classification” Ph.D. thesis. Ithaca,NY; Cornel University.

[3] Bernhard Pfahringer, Geoffrey Holmes and Richard Kirkby, “Optimizing the Induction of Alternating Decision Trees”, Proceedings of the Fifth Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining.2001,pp.477487.http://en.wikipedia.org/wiki/Alternating_decision_tree.

[4] C. Andrew, “Buiding Decision Trees with the ID3 Algorithm”, Dr. Dobbs Journal, Jun 1996.

[5] C. Aggarwal, “Towards Effective and Interpretable Data Mining by Visual Interaction”, ACM SIGKDD Exploration Vol., pp.11-12. 2002.

[6] Cover TM, Hart PE. Nearest neighbor pattern classification. IEEE Transactions on Information Theory. 1967;13(1):21–27.

[7] EE I9900 - Master’s Thesis Submitted in partial fulfillment of the requirement for the degree Master of Engineering (Electrical) Spring 2017 At The City College of New York “Brief Study of Classification Algorithms in Machine Learning” Ramesh Sankara Subbu CUNY City College.

[8] Kaur, G., & Chhabra, A. (2014). Improved J48 classification algorithm for the prediction of diabetes. International Journal of Computer Applications, 98(22).

[9] Minarni, Indra Warman, Yuhendra Teknik Informatika, Institut Teknologi Padang, Indonesia “Implementation of Case-Based Reasoning and Nearest Neighbor Similarity for Peanut Disease Diagnosis” to cite this article: Minarni et al 2019 J. Phys.: Conf. Ser. 1196 012053.

[10] Ramesh Sankara Subbu “Brief Study of Classification Algorithms in Machine Learning” EE I9900 - Master’s Thesis Submitted in partial fulfillment of the requirement for the degree Master of Engineering (Electrical) Spring 2017.

[11] R.S, Michal ski and R.L. Chilausky “An Experimental Comparison of the Two Methods of Knowledge Acquisition in the Context of Developing an Expert System for Soybean Disease Diagnosis” Vol.4, No.2, 1980.

[12] UCI Machine Learning Repository Irvine, CA: University of California, School of Information and Computer Science. Accessed Online from http://www.ics.uci.edu/~mlearn/ML Repository.html.

[13] Vinita Shah, Prachi Shah “Groundnut Crop Yield Predication Using Machine Learning Techniques” © 2018 IJSRCSEIT | Volume 3 | Issue 5 | ISSN: 2456-3307.

Analysis of Teaching and Learning Assessment to Support Internal Quality Assurance (IQA) System in Higher Education

Ei Ei Phyo Maung[1st], Nan Saw Kalayar[2nd], Moe Moe Hlaing[3rd]
University of Computer Studies (Taunggyi)
eieiphyomaung@ucstgi.edu.mm, sawkalayar@ucstgi.edu.mm, moemoehlaing@ucstgi.edu.mm

Abstract

Myanmar stays to development on the route of a main political, economic, education, health and social transformation. In the information age, Myanmar requires to reform and develop its nationwide higher education system. So, statistical analysis of examination results supports the theoretical basis for teaching quality and management in Higher Education. This paper used the item analysis theory to check the multiple-choice questions (MCQs) quality. It supports to classify questions which can be revised or discarded when constructing a quality MCQs set. The input Dataset is taken from Mid - term Examination questions and exam marks from First Year (CST-1112) subject in University of Computer Studies (Taunggyi), Myanmar for last 2018-2019 Academic Year by calculating with Difficulty Index($DifId$) and Discrimination Index (DisId). Difficulty Index examines the level of difficulty of the question and Discrimination Index investigate which question is acceptance, which one need to be revise, which one need to discard.

Keywords: DifId , DisId, MCQs, IQA, CST-1112,

1. Introduction

Universities are a major field that produces subject matter experts who can greatly contribute to the socio-economic development of not only their own country but also the international community. In the present time, the Internal Quality Assurance (IQA) System has been widely used or learning and assessment activities in the teaching system of universities. Checking the teaching and examination questions quality of universities for the Internal Quality Assurance (IQA) System includes many factors. Self-assessment of an academic program is a part of the IQA procedure. It includes teaching-learning processes assessment, institutional facilities, process control, computer labs, program mission, objectives, and outcomes. In this paper, examination results are analyzed doing and it supports the theoretical base for teaching quality and management in Higher Education. In this paper, the system supports not only for the learners but also for teachers who prepare the questions.

The statistical analysis, which is used for selecting and rejecting the items of the test on the basic of their Difficulty Index (DifId) and Discrimination Index (DisId). Item analysis is a process of collecting, summarizing and assessing the quality of questions in examination such as full form question, True or False question, choose the correct answer question, by using information from students' responses. Student's responses to individual test items are examined in order to assess the quality of those questions and the test on a whole. It is based on the responses given by the examinees. The most difficult questions or the easiest questions need to be discarded or revised. The decision to retain the questions is based on good Difficulty Index (DifId) and Discrimination Index (DisId). The quality of the questions is examined by applying Difficulty Index (DifId) and Discrimination Index (DisId).

Based on the examination results, the quantitative analysis for several parameters including difficulty and discrimination are performed. By analyzing examination results, two factors are using as the following [1].

1. The teachers need to know how much knowledge the students have obtained.
2. By using the statistical analysis as the feedback, the teacher could be well prepared the questions for the next examination.

It makes the standard of the exam paper better.

The input data to statistical analysis is taken from the first year, mid-term examination mark of subject (CST-1112) in the last 2018-2019 academic year. Three MCQs question types are used such as (i) write the full form of the followings questions (ii) True or False question, and (iii) choose the correct answer questions types. The data are collected from the 154 students. To examine the statistical analysis, the following steps must be done.

1. The exam marks are sorted in descending order (highest to lowest).
2. Taken the top (upper) 27% and below (lower) 27% of the examinees.
3. Count the number of examinees in the upper group (P_U) and lower group (P_L) who got each question correct.
4. Compute the Difficulty Index (DifId) of each item. (It is measured the proportion of examinees who answered the questions correctly.)
5. Compute for the discrimination Index (DisId). (This is measured of how well a question is able to distinguish between examinees that are knowledgeable.)
6. Finally output the decision to retain and revise the questions. (The most difficult questions or the easiest questions need to be discarded or revised. The decision to retain the questions is based on good difficulty index and discrimination index.)

2. Related Work

R. L. Ebel suggested that formation of "high" and "low" groups comprising only of the top 27% (PT-performing top) and the last 27% (PB-performing bottom) of all the students ranked in order of merit [3]. Taken the upper and lower 27% gives the best analysis result. If 50% was used, the two groups would be of maximum size but since the basis of ranking is not absolutely accurate, certain students in the top group would really belong to the bottom group, and vice versa. 1/4 or 25% and 1/3 or 33% are the good item analysis to get accurate decision.

W. Yuan, C. Deng, H. Zhu, J. Li described that the results indicate the distribution of examination scores approximate to normal distribution that the design of the examination paper was good and dependable [4]. Similarly, another example illustrated that the difficulty index obtained is 0.54 which indicates the exam moderate. Thus, it was concluded that the design of the examination paper was moderate and dependable. L. R. Sharma, evaluated that the criteria of acceptable difficulty and good discrimination which means the MCQs selected were of good quality [2].

3. System Design and Methodology

3.1. System Design

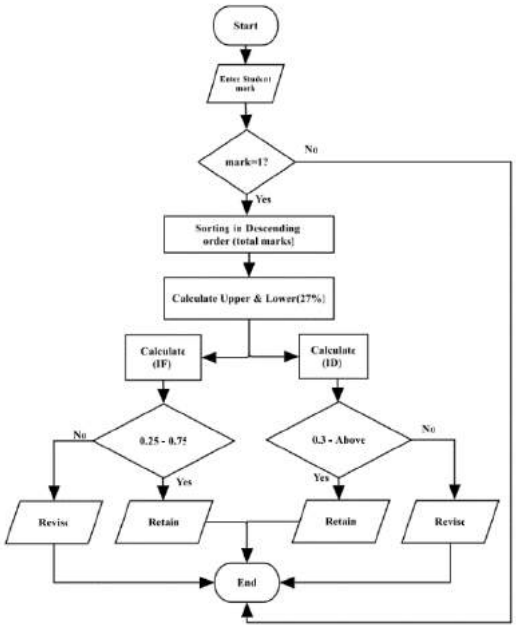


Figure 1: System Design of Teaching and Learning Assessments using Item Analysis

When the system starts, the student exam marks (which are getting from three questions, writing the full form questions, the true or false questions, and choosing the correct answer question) are input to the system. The corrections for one question get one mark (1 mark). Take all the student's marks and sort this examination mark in descending order by using the bubble sort algorithm. To calculate the statistical analysis, take the upper 27% and lower 27% from the sorted mark [3] and calculate the difficulty index (DifId), Discrimination index (DisId) and output the decision to revise or retain the questions at the end of the system.

3.2 Methodology

3.2.1. Bubble Sort Algorithm

In this paper, when the system starts, the input mark must be entered and then sorting it in descending order using the following bubble sort algorithm.

```
procedure bubbleSort( list : array of items )
  loop = list.count;
  for i = 0 to loop-1 do:
    swapped = false
    for j = 0 to loop-1 do:
          if list[j] < list[j+1] then
            swap( list[j], list[j+1] )
            swapped = true
      end if
    end for
    if (not swapped) then break
    end if
  end for
end procedure return list
```

After the sorting process, taken the top (upper) 27% and below (lower) 27% of the examinees [3]. The difficulty index (DifId) was determined by the percentage of students who passed in the MCQs questions in the mid-term examination. Discrimination index (DisId) is measured, how the 'good' students are answering the questions and how the 'poor' students answering a particular question.

3.2.2. Difficulty Index (DifId)

The Difficulty Index is the proportion or probability that students answered a test item correctly. Generally, more hard questions have a lower percentage. Difficulty Index is also equal to the item mean. The item difficulty index ranges from 0.00 to 1.00 the higher the value, the easier the question. It can be indirect proportion of the index and the decision that the higher the difficulty index result, the easier the question decision be done.

$$DifId = \frac{P_u + P_l}{Ts}$$

where, $DifId$ is Difficulty Index of examination score. P_u is average summarize highest students group score for 27% of total students. P_l is average summarize lowest students group score for 27% of total students. T_s are total students of examinees in upper and lower group.

Table 1. Rule for Difficulty Index (DifId)

Rule	Range	Level of Difficulty
Rule 1	0.76 - 1.00	VERY EASY
Rule 2	0.25 - 0.75	AVERAGE
Rule 3	0.00 - 0.24	VERY DIFFICULT

3.2.3. Difficulty Index (DifId)

The Discrimination Index is determined by subtracting from the number of students in the upper group to the lower group. Then, the answer is divided by the number of students in each group. A basic measure of the validity of an item is called discrimination index.

$$DisId = \frac{P_u - P_l}{T_{egs}}$$

where, $DisId$ *is* Discrimination Index of examination score. P_u is average summarize highest students group score for 27% of total students.P_l is average summarize lowest students group score for 27% of total students. T_{egs} are total students of examinees in each group.

Table 2. Rule for Discrimination Index (DisId)

Rule	Range	Level of Discrimination
Rule 1	0.40 - Above	VERY GOOD
Rule 2	0.30 - 0.39	REASONABLY GOOD
Rule 3	0.20 - 0.29	MARGINAL ITEM
Rule 4	0.19 - Below	POOR ITEM

4. Dataset and Index Evaluation

The input data to statistical analysis is taken from the first year, mid-term examination mark of subject (CST-1112) in the last 2018-2019 academic year in University of Computer Studies (Taunggyi).

4.1 Dataset

Three MCQs questions types are used. They are (i) write the full form of the followings questions (ii) True or False question, and (iii) choose the correct answer questions types. The data are collected from the 154 students. The numbers of questions are:

Question I. Write the full form of the followings question type. (10–items)

Question II. Write down True or False of the following statements. (10–items)

Question III. Choose the correct answer of the followings. (10–items)

4.2 Index Evaluation

When analyzing the Full Form questions, the Difficulty Index (DifId) level is AVERAGE level, so it is considered that 10 questions are appropriate questions and the decision can be set as RETAIN. The examination result will be shown in the following Table 3.

Table 3. Difficulty Index (DifId), Item analysis of Full Form question type results

Test Question	Upper Group (P_u)	Lower Group (P_l)	Difficulty Index (DifId)	
1	36	17	0.34	AVERAGE
2	38	13	0.33	AVERAGE
3	36	11	0.31	AVERAGE
4	37	11	0.31	AVERAGE
5	41	9	0.32	AVERAGE
6	41	11	0.34	AVERAGE
7	36	9	0.29	AVERAGE
8	40	10	0.32	AVERAGE
9	41	19	0.39	AVERAGE
10	38	16	0.35	AVERAGE

When analyzing the Full Form questions, the Discrimination Index (DisId) level is VERY GOOD, so the students' answer results are considered to be very good and the decision can be set as RETAIN. The calculating results are shown in the following Table 4.

Table 4. Discrimination Index (DisId), item analysis of Full Form question type results

Test Question	Upper Group (Pu)	Lower Group (Pl)	Discrimination Index (DisId)	
1	36	17	0.45	VERY GOOD
2	38	13	0.60	VERY GOOD
3	36	11	0.60	VERY GOOD
4	37	11	0.62	VERY GOOD
5	41	9	0.76	VERY GOOD
6	41	11	0.71	VERY GOOD
7	36	9	0.64	VERY GOOD
8	40	10	0.71	VERY GOOD
9	41	19	0.52	VERY GOOD
10	38	16	0.52	VERY GOOD

Table 5. Difficulty Index (DifId), item analysis of True or False question type results

Test Question	Upper Group (P_u)	Lower Group (P_l)	Difficulty Index (DifId)	
1	39	32	0.46	AVERAGE
2	38	24	0.40	AVERAGE
3	39	26	0.42	AVERAGE
4	37	20	0.37	AVERAGE
5	41	26	0.44	AVERAGE
6	38	25	0.41	AVERAGE
7	40	30	0.45	AVERAGE
8	41	24	0.42	AVERAGE
9	42	26	0.44	AVERAGE
10	41	29	0.45	AVERAGE

When analyzing the True or False questions, the Difficulty Index (DifId) level is AVERAGE level, so it is considered that 10 questions are appropriate questions and the decision can be set as RETAIN. The demonstration results are shown in the above Table 5.

When analyzing True or False questions, according to the Discrimination Index (DisId) level, question numbers (1,7,10) are POOR ITEM and MARGINAL ITEM level, so the students' ability to answer is considered as questions to be re-examined and the decision is REVISE. The rest of question numbers (2,3,4,5,6,8,9) is REASONABLE GOOD and VERY GOOD, so the answer result is considered moderate and excellent, and the decision is RETAIN. The analyzing results are shown in the Table 6.

Table 6. Discrimination Index (DisId), item analysis of True or False question type results

Test Question	Upper Group (P_u)	Lower Group (P_l)	Discrimination Index (DisId)	
1	39	32	0.17	POOR ITEM
2	38	24	0.33	VERY GOOD
3	39	26	0.31	REASONABLE GOOD
4	37	20	0.40	VERY GOOD
5	41	26	0.36	REASONABLE GOOD
6	38	25	0.31	REASONABLE GOOD
7	40	30	0.24	MARGINAL ITEM
8	41	24	0.40	VERY GOOD
9	42	26	0.38	REASONABLE GOOD
10	41	29	0.29	MARGINAL ITEM

In examine choose the Correct Answer questions, according to the Difficulty Index (DifId) level, it got the AVERAGE level, so it is considered that 10 questions are suitable questions, and the decision is set as RETAIN. The examination results are shown in the following Table 7.

Table 7. Difficulty Index (DifId), item analysis of Correct Answer question type results

Test Question	Upper Group (P_u)	Lower Group (P_l)	Difficulty Index (DifId)	
1	42	29	0.46	AVERAGE
2	41	22	0.41	AVERAGE
3	42	29	0.46	AVERAGE
4	42	29	0.46	AVERAGE
5	40	24	0.42	AVERAGE
6	41	29	0.45	AVERAGE
7	42	33	0.49	AVERAGE
8	41	27	0.44	AVERAGE
9	39	27	0.43	AVERAGE
10	41	26	0.44	AVERAGE

Table 8. Discrimination Index (DisId), item analysis of Correct Answer question type results

Test Question	Upper Group (P_u)	Lower Group (P_l)	Discrimination Index (DisId)	
1	42	29	0.31	REASONABLE GOOD
2	41	22	0.45	VERY GOOD
3	42	29	0.31	REASONABLE GOOD
4	42	29	0.31	REASONABLE GOOD
5	40	24	0.38	REASONABLE GOOD
6	41	29	0.29	MARGINAL ITEM
7	42	33	0.21	MARGINAL ITEM
8	41	27	0.33	REASONABLE GOOD
9	39	27	0.29	MARGINAL ITEM
10	41	26	0.36	REASONABLE GOOD

When analyzing choose the correct answer questions, according to the Discrimination Index (DisId) level, question numbers (1,2,3,4,5,8,10) is REASONABLE GOOD and VERY GOOD, so the answer result is considered fair and excellent, and the decision is set as RETAIN, and question numbers (6, 7, 9) are at the MARGINAL ITEM level, so the students' ability to answer is set as questions to be re-examined, and the decision is set to REVISE as the above Table 8.

5. Conclusions

Examining the answered mark of the questions using the difficulty index (DifId) and Discrimination Index (DisId) in item analysis methods support both teachers who preparing the questions and students who answering the question. According to the results, it will be possible to change the teaching style of the subjects with weak learning ability of the students. The result of analyzing the questions according to this statistical technique is used in preparing the exam questions and this assessment system helps a lot in Internal Quality Assurance in Higher Education of Myanmar.

References

[1] G. Ramya, G. Swetha, M. Ramya, “The Statistical Analysis and Evaluation of Examination results using R”, IJERCSE, Vol 4, Issue 11, Nov, 2017, pp. 316-320.

[2] L.R. Sharma, “Analysis of Difficulty Index, Discrimination Index and Distractor Efficiency of Multiple Choice Questions of Speech Sounds of English”, IRJMMC, Vol.2 Issue 1, Feb, 2021.

[3] R.L. Ebel, “Measuring Educational Achievement”, Michigan State University, 1965.

[4] W. Yuan, C. Deng, H. Zhu, J. Li, “The Statistical Analysis and Evaluation of Examination Results of Materials Research Methods Course”, Vol.3, Published Online in SciRes, Dec, 2012, pp. 162-164

Covid-19 Vaccine Data Management System for Taunggyi Township

Nang Thida Aye, Cherry Phyo Wai
University of Computer Studies (Taunggyi), Myanmar
nangthidaaye@ucstgi.edu.mm, cherryphyowai@ucstgi.edu.mm

Abstract

COVID-19 has caused millions of deaths around the world since 2019. Vaccination has been shown to contribute to reducing deaths and severe illness from COVID-19, and to reducing the transmission of COVID-19. The system in this paper provides the total number of people who have gotten vaccinations (first, second, and booster dose) in Taunggyi Township. The number of vaccinated people could be searched by age and by NRC. Records of the person who needs to take the next dose and when to take the next dose could be searched through this system by entering their NRC number. In this system, Bubble Sort is used to sort the database with NRC numbers and age. Also, Binary Search algorithms is used to search vaccinated people information. The system administrator can manage the information such as insert, update and delete information in the Vaccinated people database. This system also provides information of vaccine types, how to use masks, how to use PPE, and so on.

Keywords: COVID-19, NRC, PPE

1. Introduction

COVID-19 can be severe and has caused millions of deaths around the world, as well as lasting health problems in some who have survived the illness. The coronavirus can be spread from person to person. It is diagnosed with a test. The best way to protect ourselves is to get vaccinated and boosted when we are eligible, follow testing guidelines, wear a mask, wash our hands, and practice physical distancing. For a person to get vaccination on time depends on the information recorded for that person, which includes the frequency of dose, the date that dose was taken, and the date that the next dose must be taken. The system in this paper enables to maintain the mentioned information and provide information regarding the vaccinated people when it is necessary. The system could allow the user to search for people's information easily by entering an NRC number, also by entering age. In addition, the system administrator can manage the vaccinated people information such as insert in the new record, update the record, and delete the record of information in the vaccinated people database.

2. Related Work

D. Knuth invented another search algorithm, uniform binary search [1.] It saves the index of the middle element rather than the lower and upper bounds. It also saves the difference in the middle element between two consecutive iterations (current and next). However, this method is only faster in cases where calculating the middle point is inefficient, such as decimal computers [2].

Another search technique developed in 1976 by J. Bentley and Andrew A.C. Yao is explosive search [3]. Determining the upper bound, or the first element whose index is also a power of two and higher than the goal number, is the first step in the process. The greatest time needed for the search is x log2x, where x is the location of the desired value. Only when the goal value is close to the start of the array does explosive search exceed binary search.

Other search technique that was presented by P. Kumar in March 2013 is quadratic search [6]. It was a step up from binary search. Instead of relying on the center element, it takes on the middle, $1/4^{th}$, and $3/4^{th}$ elements of the sorted array and determines whether or not one of these matches the element being looked for. After checking numerous scenarios, if one of these elements is the target element, it quickly returns that element; otherwise, it narrows the search space.

3. Methodology and Dataset

In today's world, the communication network is growing very quickly. Businesses are turning digital to increase management efficiency. The amount of information collected on the internet is increasing, and as a result, datasets are getting more complicated. It is essential to organize, manage, access, and analyze the data carefully and effectively. The system in this paper supports to search the vaccination people list very quickly and maintains the digital records of vaccinated list. It helps a lot in the national health care sector.

In computer science, data structures are the foundation of abstract data types (ADT), where ADT is the logical representation of the data types. The physical design of data types is established using data structures. Different types of data structures are utilized for different types of applications; some are specialized in particular activities. Data structures are defined as a collection of data values and their relationships, as well as functions and operations that apply to the data. In this system, users can efficiently access and update information.

Data structures support in the management of large amounts of Covid-19 data, such as massive databases. Efficient data structures are the foundation for efficient algorithms. Data structures are responsible for the efficient retrieval of data from stored locations in addition to efficient storage. It has an array, Sorting, Searching, Graph, Programs, Linked List, Pointer, Stack, Queue, Structure, and so on.

Sorting is the process of arranging data in a specific format. A sorting algorithm describes how to arrange data in a specific order. The most common orders are numerical or lexicographical. The importance of sorting is due to the fact that data searching can be greatly improved if data is stored in a sorted manner. Sorting could also be used to make data more readable. In this paper, Covid-19 vaccinated people NRC numbers are sorted by using Bubble Sort.

The bubble sort algorithm is used in this paper because it not only works by making passes, from right to left (or left to right) over the array but also takes just a few lines of code. Moreover, it is easy to read, and can be plugged in anywhere in program.

Searching in data structures refers to the process of finding the location of an element in a sequence. The effectiveness of a search for an element improves the effectiveness of any algorithm. In this paper binary search technique are used to search Covid-19 vaccinated people information by entering their NRC number because this search algorithm is more efficient for using large lists.

3.1.1. Bubble Sort

A bubble sort, also known as a sinking sort, is a fundamental sorting algorithm that steps through the list to be sorted repeatedly, compares each pair of adjacent items, and swaps them if they are out of order. The list is searched indefinitely until no swaps are required, at which point the list is said to be sorted.

The amount of repetitions needed is usually proportional to the length of the list; that is, the more items to be sorted, the more passes are required to complete the sorting successfully. The algorithm, a comparison sort, is named after the way smaller elements "bubble" to the top of the list. It can be relevant if the input is typically in sort order or is small, but it can be ruinous if the list is already sorted in the reverse manner as the one required.

The dataset using in this system is vaccinated people information only for Taunggyi Township. As the dataset is small, bubble sort algorithm is using in this system because it is a very simple algorithm to implement on a computer. Most of the NRC number in the dataset are already in order, by reason of using the bubble sort is the efficient way to check if a list is already in order and it is not used too much memory.

3.1.2. Algorithm for Bubble Sort

```
Start
    Data: input array A []
    Result: sorted A []

    int i, j;
    N= length (A);
    for j= 1 to N do
        for i= 0 to N- 1 do
            if A[i] > A [i+1] then
                temp = A [i];
                A[i] = A [i+1];
                A [i+1] = temp;
            End
        End
    End
End
```

Now the stage and symbols utilized in this algorithm are explained. First sorted the input array using the bubble sort.

The bubble sort algorithm begins with the first component (index 0 in array *[]*). Then determine whether or not the following element (index 1 in array *[]*) is bigger than the one currently using in this array. If the current element (index 0 in array *[]*) is greater than the following element (index 1 in array *[]*), the element will be swept. If the current element is smaller than it, the following element in the array will be moved.

3.1.3. Binary Search

The fundamental method compares the search element to the element in the center of the search space, limiting further searching to the relevant half of the search space (this can be done because the search space is sorted). Once the search element has been found or has not been found to compare, the procedure is repeated, cutting the remaining search space in half at each step, until the element has either been located or has not been located to compare and it was not in the search space. After sorting the objects in an array according to the key in either ascending or descending order, a method called binary search can be used to achieve significantly better performance. As the complexity of linear search is O(n) operation but the complexity of binary search is O (ln n) operations. So, it is a much faster algorithm then other search such as linear search.

3.1.4. Algorithm for Binary search

```
Start
    A ← sorted array
    n ← size of array
    x ← value to be searched
    Set lb = 1
    Set ub = n
    while x not found
      If ub < lb
          EXIT : x does not found
          Set mid=lb + (ub -  lb) / 2
          if A[mid] < x
          set lb = mid + 1
          if A[mid] > x
          set ub = mid- 1
           if A[mid] = x
          EXIT : x  found at location mid
    end while
end
```

3.1.5. Dataset

The dataset using in this system is vaccinated people information only for Taunggyi Township. In this system, 300 vaccinated people information's (record) are used and the dataset include the attributes such as name, father's name, NRC, phone number, gender, date of birth, address, job, vaccine center, vaccine type, and medical team as shown in Figure 4.

4. System Design and Implementation

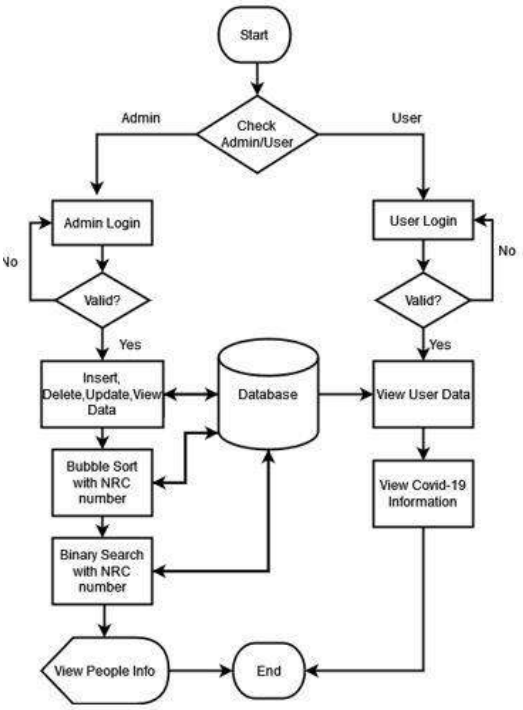


Figure 1. System flow diagram of searching vaccinated people information by NRC number

The system in this paper is implemented for the administrator site. For Data Security, the system administrator must enter the name and password correctly. If the name and password are correct, the administrator can Insert, Update and Delete people information in the database. The NRC numbers in the database be sorted using the Bubble sort algorithm. Searching information on COVID-19 vaccine recipients using the NRC number could be done by using binary search technique and then the information of the person could be viewed. If the person who is being searched is not in the database, the system will end.

In user view, the user can view information about the logged-in user and the information from COVID that concerns healthcare.

The system in this paper intends to help the people who get the vaccine for next dose and to save the record of vaccinated people data digitally. Searching Information could be done easily in a short time by using this system. The system flow diagram is shown in Figure 1.

Figure 2. Log in Form for System Administrator

For system security, the system administrator and user must fill the login form correctly as shown in Figure 2. The administrator and user need to enter the name and password, and then click the login button. If the information is correct, the administrator can enter to the system.

Figure 3. Home Form

Figure 3 shows the menu that could be done in this system. The system administrator can do the processes, such as viewing all record forms, searching information forms, adding new record forms, updating forms, deleting forms in the select field.

Figure 4 shows the dataset of vaccinated people information. The user can search information by age in this form.

Figure 4. People who have received vaccination Form

Figure 5: Search Record Form by Entering NRC Number

Figure 5 shows the search form for a person who comes to take a vaccination. The administrator could search for the information of the person by putting their NRC numbers in the search box. The information related to the search could be seen as in Figure 5.

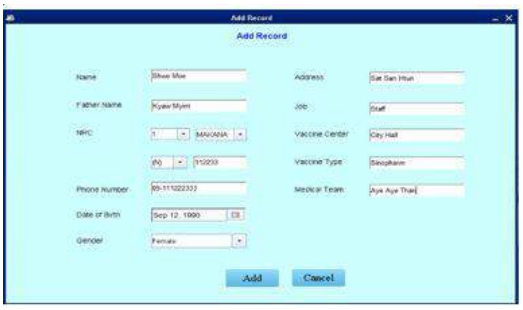


Figure 6. Add Record Form

The system administrator can insert a new record to the dataset for newly vaccinated people by using the form shown in Figure 6. The new vaccinated people's information such as name, father's name, NRC, phone number, gender, date of birth, address, job, vaccine center, vaccine type, and medical team are entered.

Figure 7. Update Record Form

The system administrator can edit the records in the dataset such as name, father name, NRC, phone number, gender, date of birth, address, job, vaccine center, vaccine type, and medical team using the Update Record Form shown in Figure 7. After editing the data, the update button must be clicked to update the data in the database.

Figure 8. Delete Record Form

The admin can delete the record by using the Delete Record Form shown in Figure 8. The admin can delete the data by inserting the specific NRC number and then clicking the Delete Button.

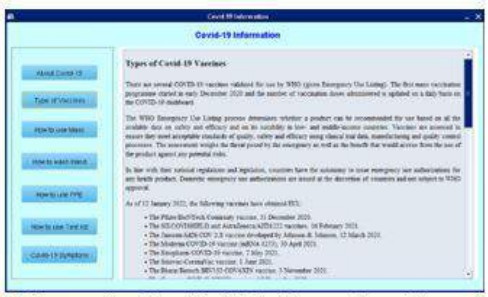


Figure 9. Covid-19 Information Form

Figure 9 shows the information provided under the View Information Menu. This Menu consists of information about COVID-19, vaccine types, how to use a mask, how to wash your hands, how to use PPE, how to use a test kit, and COVID-19 symptoms.

5. Conclusion

This system intends to use not only to save vaccination information records but also to search for the vaccination information and manage it easily. This system supports people who want to know information about vaccination frequency. The system administrator can do searching, c, and Deleting information easily. This system supports vaccine administration data of Taunggyi Township. This will support a lot as one corner in health care sector of Myanmar.

6. Limitation and Further Extension

This system is implemented with English language. So as the further extension, it can be implemented with Myanmar Language and be search with Myanmar Language. The bubble sort is not suitable for large dataset (Example for vaccinated people information for the whole Myanmar). So, the other sorting methods will be used when the dataset is large.

References

[1] D.E. Knuth, "The Art of Computer Programming", Vol. 3: Sorting and Searching, Addison Wesley, 1973.

[2] H.V. Schmid, "Decimal Computation (1 st ed)", J. Wiley & Sons Ins., NY, USA, 1974.

[3] J. L. Bentley, A. C. Yao, "An almost optimal algorithm for unbounded searching", Vol. 5, Issue 3, Information Processing Letters, doi: 10.1016/0020-0190(76)90071-5, ISSN 0020-0190, 1976, pp. 82-87.

[4] O. Appiah, and E.M. Martey, "Magnetic Bubble Sort Algorithm", International Journal of Computer Applications (0975 – 8887) Volume 122 – No.21, July 2015, pp 24-28.

[5] P. Kumar, "Quadratic Search: A New and Fast Searching Algorithm (An extension of classical Binary search strategy)", International Journal of Computer Applications, Vol. 65, Hamirpur Himachal Pradesh, India, March 2013.

[6] R.M. Fitrian, I. Taufik, M.S. Ramadhan, N. Mulyani, J. Hutahaean, A.S. Sitio and H.T. Sihotang, "Digital Dictionary Using Binary Search Algorithm" Journal of Physics: Conf. Series 1255, 012058, 2019.

Analysis of Economic Growth Using Multiple Linear Regressions

Phyu Sin Htwe[1st], Htwe Htwe Lin[2nd]
University of Computer Studies (Taunggyi)
phyusinhtwe@ucstgi.edu.mm, htwehtwelin@ucstgi.ecu.mm

Abstract

The economic growth analysis process is an important role in developing countries. The economic growth of a country is measured by the Gross Domestic Product (GDP). The economic growth analysis system analyzes the influence of the agriculture sector, the manufacturing sector, and the service sector of the economy on the gross domestic product (GDP) that makes up the economic growth. In this system, multiple linear regression model was used to analyze relationship and the effect of the agriculture sector, manufacturing sector, and service sector on how they influence GDP growth. In this system, data are collected from the Pindaya township planning department in the period from 2016-2017 to 2020-2021. According to the analysis result, it was discovered that agricultural value added has a strong influence on the growth rate of GDP than manufacturing and service sector value-added. In this system used Java Programming language and apply multiple linear regression.

Keywords: Gross Domestic Product, Agricultural Sector Value Added, Manufacturing Sector Value Added, Service Sector Value Added, Multiple Linear Regression

1. Introduction

Economic growth corresponds to an increase in the total market value of goods and services produced within a country, compared from one period of time to another. Gross domestic product (GDP) or gross national product (GNP) is the most important indicator to measure the economic growth of a country, state, and township. It can be measured in nominal and real terms including inflation by the increase in the percentage of gross domestic product GDP growth rate. Economic growth can be negative or positive. Positive growth could be an increase in the economy and negative growth can occur an economic depression and economic recession.

Many factors have an effect on economic growth and development, and many other factors also influence on the gross domestic product. Gross Domestic Product (GDP), represents the total market value of all final goods and services produced within the geographic boundaries of a country during a specified period of time, normally a year. It counts the goods and services produced within the country and hence does not consider the products that the country imports from another country.

Gross domestic product (GDP) includes three main sectors, they are primary sector also called the agriculture sector (agriculture, livestock & fishery and forest), the secondary sector as the manufacturing sector (mining, industry, electric, construction) tertiary sector as service sector (transportations, communication, financing, social, other services, trade) that make up the economic growth. This paper investigates (examines) which variables of the agriculture sector, manufacturing sector, and service sector have a strong influence on the gross domestic product (GDP).

2. Literature Review

According to Sipiwe Ahihana & Douglas Kunda, 2017 [1], regression analysis, trend and time series analysis, and data mining techniques were applied to find out the economic factors affecting the performance of the gross domestic product of the Zambia economy. From these three techniques analyzed, service value added variables were the most prominent variables which showed a strong influence on GDP growth. The author found that from these three techniques, regression analysis is easy to use for the predicament influence of independent variables on dependent variables.

According to Suraj Gaudel, 2015[2] was applied multiple linear regression analysis finds

the strength of the long-run relationship between the response variable and independent variables. The author targeted the variables that influence GDP growth. According to the author's analysis result of the study demonstrates a strong positive relationship between GDP growth rate and variation in agriculture, industrial and service sectors of the economy.

Ligia DUDU & Aaluca Georgiana MOSCU 2016 [3] applied the effective_possibility to use the multiple linear regression models in the analysis of the gross domestic product (GDP). This analysis determined the function that describes the best relationship between the indicators undergoing analysis, observed the link that is established between them, and estimated an econometric statistically significant valid model. Authors applied a multiple linear regression model to analyze the GDP as the dependent variable and independent variables of households consumption (CP), public consumption (CPL), gross investment (INVB), changes in inventories (VS), variation in net export (EXN) and disposable income (VND). According to the analysis, there has influenced households' consumption, public consumption, gross investments, change in inventories, net exports, and gross disposable income on GDP growth.

Numerous factors have affected on economic growth and country development, and one of the most important indicators is gross domestic product (GDP). On the other hand, many factors also influence the gross domestic product (Stanko Stanic & Zeljko V.Racic 2019 [4]). They applied the multiple regression analysis models to evaluate the effects of macroeconomic factors of foreign direct investment, import, export, growth rate, unemployment, and inflation as independent variables to the gross domestic product as the dependent variable, and based on theoretical and methodological research. The result of quantification of the relationship between macroeconomic factors and GDP used a multiple regression model. According to the analyzed result of independent variables, import has a major effect on GDP, followed by FDI and Export. The selection of predictors defined in the multiple regression model is significant for Bosnia and Herzegovina's GDP development.

3. Methodology and System Design

3.1. Methodology

Regression analysis is a statistical method used to estimate the relationship between one or more explanatory variables and a response variable. Regression analysis is useful for many other research studies, such as engineering, business, finance, medical and other fields to find the relationship between one dependent variable and more than one independent variable. This paper used the multiple linear regression model to analyze agriculture, manufacturing, and service sector of the economic sectors that influence on GDP growth in Pindaya township.

3.1.1. Model specification

Multiple linear regression analysis is associated with finding a relationship between one or more independent variables (x_1 and x_2 and x_3) and one dependent variable of (y) that would have interrelationships with each other.

The statistical model was estimate as:

$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \ldots.. b_nx_n + \mu$ where n represent the number of observation and b_o, b_1, b_2, b_3 are the coefficient of the independent variables to be estimated and μ is the residual error term that represents the factors that are not mentioned in the model.

Rewriting the model as:

$y_i = b_0 x_{i,0} + b_1x_{i,1} + b_2x_{i,2} + b_3x_{i,3} + \mu_i$ Equation (1)

- Where we defined $x_{i,0} = 1$ for all observations $i = 1,2,\ldots n$ (n= number of year)
- $x_{i,1}$ = Agricultural Sector Value Added Growth Rate(AVAGR)
- $x_{i,2}$= Manufacturing (Industry) Sector Value Added Growth Rate (MVAGR)
- $x_{i,3}$ = Service Value Added Growth Rate(SVAGR)
- y_i = Gross Domestic Product Growth Rate(GDPGR)
- μ_i = error term or disturbance error term.

Matrix notation can help for the calculation and manipulations. Matrix form of regression are follow. To find the regression coefficient, used matrix notation:

$$y = xb + \mu \quad (2)$$

$$b = (x'x)^{-1}(x'y) \quad (3)$$

$$y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{bmatrix}, b = \begin{bmatrix} b0 \\ b1 \\ b2 \\ b3 \end{bmatrix}, x = \begin{bmatrix} 1 & x_{11} & x_{12} & x_{13} \\ 1 & x_{21} & x_{22} & x_{23} \\ 1 & x_{31} & x_{32} & x_{33} \\ 1 & x_{41} & x_{42} & x_{43} \\ 1 & x_{51} & x_{52} & x_{53} \end{bmatrix}$$

where y is the column vector, x is a n (k+1), b is a column vector, and μ is the residual.

Residual **(μ)** is a measure of how far away a part is vertically from the regression line. Simply, it is the error between a predicted value and an observed actual value.

$$\mu_i = y_i - y_i\hat{}\ , \quad i=1,\ldots n \qquad (4)$$

3.1.2. Data Collection Method

Data are collected from a secondary source from Pindaya's township planning department. Data are analyze for five years from 2016-2017 to 2020-2021 (1 October to 31 March). Analysis for gross domestic product growth rate and economic factors which are the agriculture sector, manufacturing sector, and the last sector is service sector at constant price.

3.2. System Design

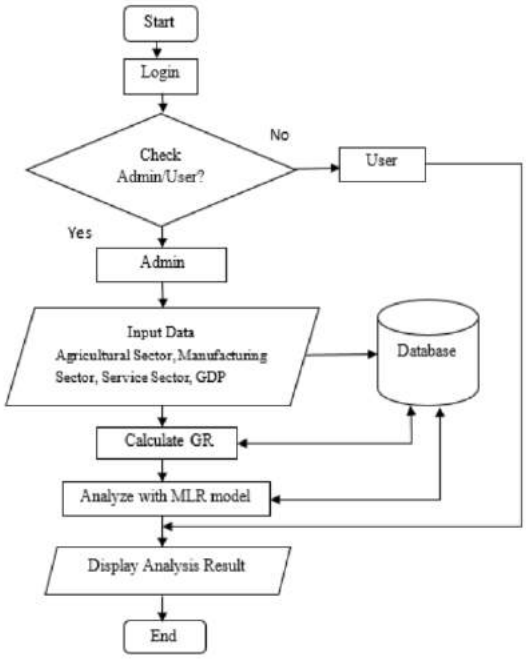


Figure 1. System Design

From figure 1 above, the system is implemented by the Admin and User. Admin can insert into the database and update data, view all data of GDP and also calculates the AVA, MVA, SVA, and GDP growth rate. Based on these growth rate results data, the system analyzes with the multiple linear regression model to get the regression coefficient and residual that display the final result and end of the program. User can only view all data of GDP information in the system.

3.3. Algorithm for the system

Begin
Step 1 Check Admin or User
Step 2 If (Admin) then
Begin
Step 2.1 Input AVA, MV, SVA GDP
Step 2.2 Calculate GDPGR, AVAGR, MVAGR, SVAGR
$GR = (Y_t - Y_{t-1}) / Y_{t-1} \times 100$
Where, GR= Growth Rate, Y_t = present year, Y_{t-1} = previous year.
Step 2.3 Output GR result
Step 2.4 Based on step 2.3 and analyse with MLR model. Calculate the coefficient of regression.
$b= (x'x)^{-1} (x'y)$
Step 2.5 Output of regression coefficient (b_0, b_1, b_2, b_3)
Step 2.6 Calculate predicted value $y_i\hat{}$
$y_i\hat{} = b_0 x_{i,0} + b_1 x_{i,1} + b_2 x_{i,2} + b_3 x_{i,3}$
Step 2.7 Output of $y_i\hat{}$
Step 2.8 Calculate Residual,
Residual $(\mu_i) = (y_i) - (y_i\hat{})$
Step 2.9 Output Residual μ_i
Step 2.10 Display Result
End
Step 3 else
Begin
3.1 User can only view all the data
End
End

4. Interpretation Result and Discussion

GDP at constant price and value of economic factors which is Pindaya township for 5 years, from FY 2016-2017 to FY 2020-2021 has been presented see in Table 1. GDP growth rate at constant price and growth rate of economic factors of Pindaya for 5 years, from FY 2016-2017 to FY 2020-2021 has been presented see in Table 2.

Table 1: GDP and Major Economic Sectors (Kyat in million)

Year	AVA sector	MVA Sector	SVA Sector	GDP
2016-2017	39201.6	24246.2	23149.7	86597.5
2017-2018	47275.4	24527.8	24299.9	96103.1
2018-2019	44192.0	27545.7	25670.6	97408.3
2019-2020	46193.2	27354.9	26431.1	99979.2
2020-2021	44580.0	27531.1	25750.4	97861.5

Table 2: GDP Growth and Major Economic Sectors Growth (In Percentage)

Year	AVA GR	MVA GR	SVA GR	GDP GR
2016-2017	-4.4	10.6	0.7	0.8
2017-2018	20.6	1.2	5.0	11.0
2018-2019	-6.5	12.3	5.6	1.4
2019-2020	4.5	-0.7	3.0	2.6
2020-2021	-3.5	0.6	-0.25	-2.1

Multiple linear regression coefficient's output results have been shown in Table (3), it explains relationship between a dependent variable and independent variables by using equation (3).

Table 3. Output of coefficient

Variables	Coefficients
Intercept	-0.02878
AVAGR	0.46332
MVAGR	0.25384
SVAGR	0.23681

Based on these regression coefficients, the system can calculate the predicted value (y^) and residual (μ). To get the residual value (μ), the predicted value (y^) was subtracted from the actual value(y), the results was shown in see Table 4.

Table 4. Output of Residual

Year	Predicted Value	Residual
2016-2017	0.78903	0.01097
2017-2018	11.00426	-0.00426
2018-2019	1.40793	-0.00793
2019-2020	2.58990	0.01110
2020-2021	-2.09012	-0.00988

4.1. Interpretation Result

The coefficient of agriculture value added (AVAGR) is 0.46 was found that the influence of MVA and SVA growth rates are held constant, as AVA growth rate increases by 1 percent, the GDP growth rate in a year will increase by 0.46. The coefficient was positive and statistically significant. This means that, if agriculture value added increases by 1 percent, the GDP growth rate will increase by 46 % as shown in Table 3.

The partial regression coefficient of the manufacturing value-added growth rate was found at 0.25, which means that the influence of AVA and SVA growth rates are held constant, as MVA increases by 1 unit; the GDP growth rate in a year will increase by 0.25 unit. Economically means that, if the manufacturing sector continues to produce goods for GDP growth rate will continue will increase by 25 % as shown in Table 3.

Service value added of regression coefficient was found 0.24, which means that the influence of AVA and MVA held constant, as SVA increases by 1 unit; GDP growth rate in a year will increase by 0.24. This means that, if service value added increases by 1 percent, the GDP growth rate will increase by 23 % as shown in Table 3.

The agriculture sector, manufacture sector and service sector are the important indicator of economic growth of Pindaya township. Thus the government need to promote investing AVA, MVA and SVA sector respectively.

4.2. Residual

A residual is a measure of how far away a point is vertically from the regression line. Simply, it is the error between a predicted value and the observed actual value. In figure 2, 3, and 4, the residual plot show a good random pattern. This random pattern means that a multiple linear

regression model used in this system is supports good fit to the data of AVA, MVA and SVA as shown in figures 2,3, and 4. Figure 2 shows the agriculture value add residual plot. These residual plot are normally distributed. Figure 3 shows the Manufacturing value added residual plot that are normally distributed. Figure 4 shows the Service value added residual plot that are normally distributed.

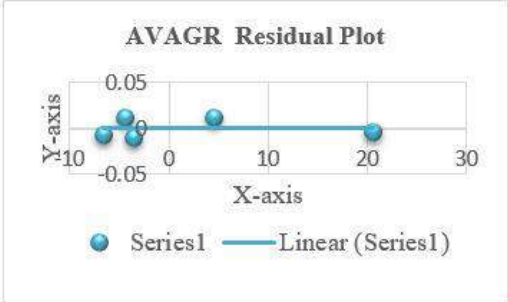


Figure 2. Residual Plot of AVA

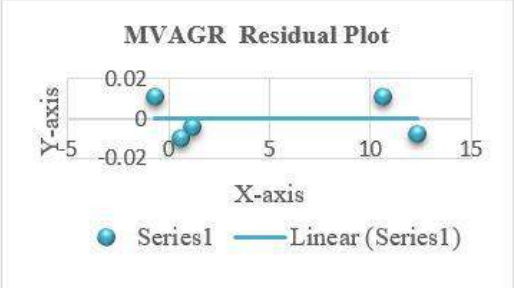


Figure 3. Residual Plot of MVA

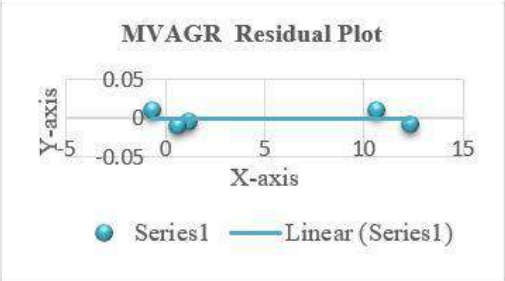


Figure 4. Residual Plot of SVA

5. Conclusion

Regression techniques are useful statistical methods that can be leveraged to estimate the degree to which independent variables are affecting the dependent variables of gross domestic product (GDP). In this system, the multiple regression technique help to find out the relationship between GDP (dependent variable) and agricultural, manufacturing, and service sectors (independent variables) growth rate. According to the analysis result, agriculture sector has a stronger influence on GDP than the manufacturing and service sector in Pindaya township. So that, agriculture sector need to promote than others because our country is based on agriculture. This analytical tool is useful in analyzing the value of the agricultural manufacturing and service sector of the economy. Planners and policymakers can be seen easily in economic information by using this system.

References

[1] D.Kunda, S.Shihana, "Analysis of Value added Services on GDP Growth Rate using Data Mining Techniques", Database System Journal Volume.3, 2017.

[2] S.Gaudel, "Influence of Difference Economic Sector on Gross Domestic Product (GDP) of Nepal", The International Journal of Business & Management Volume 3, Issue 3, 2015.

L.DuDu[1], R.G.MOSCU[2], "Practical Use of The Linear Regression in the Complex GDP Analysis", Knowledge Horizons-Economics, Volume 8, No.2, pp.74-79.

[4] S.Stanic, Z.V.Racic, "Analysis of Macroeconomic Factors Effect to Gross Domestic Product of Bosnia and Herzegovina Using the Multiple Linear Regression Model", NS Global Economic Volume 7, No.2, 2019.

[5] D.I.Siregar, R.R.Alhempi, "Analysis of the Regression and Correlation of Gross Domestic Product (GDP) of Indonesia Against Rupiah Exchange Rate (IDR-USD", Journal of Economics Volume 6, Issue 4, 2018.

[6] Rinkesh Jain, Divakar Singh) ,"Data Mining and Analysis of Economic Data" , International Journal of Advanced Research in Compuer Science and Software Engineering, Volume 3, Issue 8,August 2013.

Decision Support System of Civilize Agriculture in Southern Shan State

Hlaing Lwin Moe[#1], War War Khaing[#2]
University of Computer Studies (Taunggyi)
hlainglwinmoe@ucstgi.edu.mm, warwarkhaing@ucstgi.edu.mm

Abstract

This system uses Linear Programming Model (LP) to optimize which crop to be grown to get the maximize profit. In this system a linear programming model (simplex method) is applied to determine the optimum land allocation to 3 major crops or 5 major crops by using agriculture data. The input data for this system is crop types such as total land (L), total budget (Y), total cost per acre of each crop (C_i), profit per acre of each crop (Z). A basic method called linear programming is suitable for determining the best distribution of land among the main crops in the research area. This system is solved by simplex method. This system is written by C# programming language. The output of the system shows the result to which crop to be grown to get maximum profit and acre of cropland of each crop. Other agriculture information will be shown in the system such as a broker information, crops details information, market information etc. Future day, farmer will be used agricultural technologies that farmer plants crop.

Keyword: land allocation, Linear programming, Computer-based system, simplex method

1. Introduction

About 70% of Myanmar's population depends on agriculture. Crop production by smallholder farmers supports the economy of Myanmar. Myanmar is developing agricultural country. Land allocation is very important for our country. In Shan State, there is farmland 5.7 million acres (cultivated land – 2 million acres and not cultivated land 3.7 million acres). The agriculture sector contributes to 37.8pc GDP. Agriculture planning is important in recent times because of population increase.

Crop planting is influenced by a variety of elements, including soil types, cropping methods, market prices, climate, labor, and money. Agriculture is essential to Myanmar's economy. The geographical setting of Myanmar is exceptional for agriculture since it offers a number of beneficial conditions. Rural farmers are mostly confronted with the problem of allocating scarce land and labor resources in a crop. This system supports to enhance the farm income of the rural farmers. The overall goal of farmers to maximize the farm income with optimum resource allocation allows the use of the LP technique as an appropriate decision making tool in the analysis. For agricultural planning utilizing a linear programming model, it is crucial to know the best crop pattern for maximizing productivity and profit.

This kind of issue can be resolved using optimization techniques including linear programming, dynamic programming, goal programming, and integer programming. A Linear programming model (simplex method) is more popular because of the proportionate characteristic of the allocation problems. Comparison is made between the results obtained from the use of the LP model and the traditional method of planning. The results of profit obtained using the LP model are more superior to the ones obtained from using traditional methods.

2. Literature Review

Nowadays, all types of businesses use LP to decide on various production, distribution, land allocation, marketing, and policy-related options as well as the creation of new industries. Many problems in the actual world are formulated using linear programming [1]. The Dantzig simplex method methodology begins with a primary feasible basis and employs pivot operations to maintain the basis's viability and ensure the monotonicity of the objective value.

Winhui Hua proposed the LP technique (simplex method, Revised method) to determine

optimal allocation of a firm's limited resources such as land, labor and capital among alternative crop and livestock enterprisesm [9]. A LP crop-mix model for a finite-time planning horizon under limited available resource such as budget and land acreage, the crop-mix planning model was formulated and transformed into a multi-period LP problem by Nordin Hj. Radhakrishnan and Raj Krishna proposed the linear programming model (simplex method) for determining the optimal farm planning. "Linear programming-based cropland allocation to enhance performance of smallholder crop production" was published in November 2018 [5]. Authors are Meselu Tegenic Millaku, Travis W, Reynolds and Jeshale. The Solver add-in for Microsoft Excel was used to solve the LP models used in this paper. To allocate agricultural resources, Felix and Judith Majeke employed an LP model [2]. They compared the outcomes of employing the LP model and the conventional approach to planning and found that the LP model's outcomes were superior to the conventional approach.

This system is mainly intended for farmers. It supports for the farmers to produce more agricultural product and earn more profit. This paper compares the results obtained from the traditional method and the use of LP model. In Myanmar, most farmers are using conventional farming methods. In the cultivation of land, chemical fertilizers and chemical pesticides that make crops grow quickly are widely used. So, it can damage the agricultural land and lead to toxic crops. The system will also provide the necessary information for farmers.

3. Background Theory

3.1. Material and Method

The approach of linear programming is used to determine the best crop to plant in order to maximize profit. Linear programming function consists of linear equality and inequality constraints. Linear Programming Model (LP) is used in many farm planning system, transportation, telecommunication and manufacturing. The goal of an LP model is to optimize (maximize or minimize) the objective function.

The objective function can be defined as the mathematical equation that is a linear function of a set of variables that needs to be optimized. It is the mathematical expression that represents the aim of the system. In most cases, the objective is to maximise resources or profits and minimise the time or cost. All different types of businesses use linear programming (LP) to decide on various production, distribution, marketing, and policy-related options as well as to build new industries. The most significant and thoroughly researched optimization problem is likely linear programming (LP). Many issues in the real world can be expressed as linear programming issues.

The farmer should use Linear Programming (LP) technique to cultivate of their crops. Because, Linear Programming (LP) technique not change the farming method of the farmer, but what kind of crop should be planted and the allocation of acre of land will make more profit for the farmer.

3.2. Profit Maximizing Objectives

This study's primary goal is to maximize profits for smallholder agriculture production system. The LP problem was to estimate the amount of land Li, i = 1,2, 3…., n given n crop options used in the study region, productivity per unit of land qi, i = 1,2, 3……, n and total land size L. We modeled the objective function profit (Z) as given the market prices Pi, i = 1,2, 3…., n per kilogram or viss of each crop and the cost of production per acre of land for each crop Ci, i = 1,2, 3…..., n.

$$Z = \sum_{i=1}^{n} (P_i L_i q_i - C_i L_i) \quad (1)$$

3.2.1. Ecological Constraint

The entire amount of land Li, i = 1,2, 3..., n that is accessible for agriculture production could not be greater than the sum of land L allotted to the chosen crops.

$$\sum_{i=1}^{n} L_i \leq L \quad (2)$$

3.2.2. Aggregated Crop Budget Constraint

Result of total cost (Ci) crop production and acre of cropland of farmer (Li) could not exceed total budget of each crop (Y).

$$\sum_{i=1}^{n} C_i L_i \leq L \quad (3)$$

3.2.3. Constraint on total food crop production for self-sufficiency

The model explicitly took into account the minimum aggregated food crop production requirement for each crop Qi, i= 1,2, 3, n, which is the total of the minimum food crop production requirements.

$$q_i L_i \geq Q_i \quad (4)$$

The highest profit that might theoretically be realized while utilizing linear programming was finally estimated using an LP-based farmland allocation model with the purpose of profit maximization subject to ecological, financial, and food production limitations.

$$Z = \sum_{i=1}^{n} (P_i L_i q_i - C_i L_i)$$

Subject to:

$\sum_{i=1}^{n} C_i L_i \leq L$ (Aggregated Crop budget)

$q_i L_i \geq Q_i$ (Aggregated Food Crop Production Requirement)

$\sum_{i=1}^{n} C_i L_i \leq L$ (Aggregated Land)

3.3. Linear Programing Model

The mathematical model for a linear programming issue with "n" choice variables and "m" constraints is (Taha, Zeleny , Wiston, Higleetal).

$$Maximize\ Z = c_1x_1 + c_2x_2 + \ldots\ldots\ldots + c_n x_n$$

$$Subject\ to\ condition(s\ to\ c)$$

$$a_{11}x_1 + a_{12}x_2 + \ldots\ldots\ldots + a_{1n} x_n \leq b_1$$

$$a_{21}x_1 + a_{22}x_2 + \ldots\ldots\ldots\ldots + a_{2n} x_n \leq b_2$$

$$a_{m1}x_1 + a_{m2}x_2 + \ldots\ldots\ldots + a_{mn} x_n \leq b_m$$

$$x_j \geq 0,\ j = 1,2, \ldots\ldots\ldots, n$$

This can write this as,

$$Max\ Z = C\ ^tX$$

$$Subject\ to\ condition(\ s\ to\ c\ )$$

$$AX \leq b,\ X \geq 0$$

From the aforementioned model, C and b are vectors of known coefficient matrices, while X represents the vector of variables (to be determined). The objective function (in this case, C^t) is the expression that must be maximized.

The constraint that specifies a convex polytope over which the objective function is to be optimized is the equation $AX \leq b$. The unit returns for each production process, represented by the coefficients C_1, C_2..., C_n, are X_1, X_2, X_3..., X_n.

4. System Architecture

4.1. System Flow Diagram

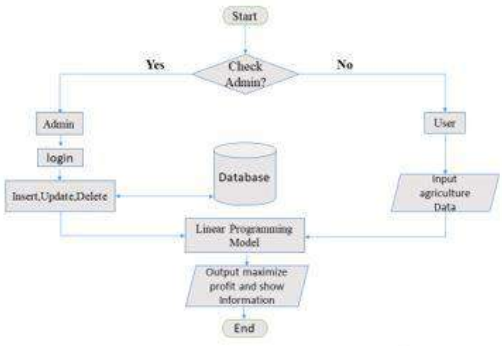


Figure 1. System Architecture

Figure 1 shows the design of the system. In this system, firstly, the user (farmer) needs to press start button to enter the system and user can search agriculture information. Beside, user calculates to maximize profit using agriculture

data added by the farmer. Administrator also needs the password to enter the system because it controls the system. The administrator can add, delete and update this system.

4.2. System Architecture

This system is written by C# Programming Language. This system is divided into two parts, the first part is the administrator part and the second part is the user part.

4.3. Administrator

Administrator needs login name and password to enter the system. The administrator will be able to insert, update, and delete data when needed in this system. The administrator is the person who plans to search and view the information by the farmers. Admin uses Linear Programming Model technology to get the maximum profit when planting agricultural crops.

4.4. User (farmer)

User (farmer) also needs to press start button to enter the system. This system is divided into two parts for farmers. The first part is that farmers can search for the agricultural information they want. The second part is the getting maximum profit in framing crops using Linear Programming Model technology to get the maximum profit in farming crops.

This system will determine what kind of crop should be planted on the land owned by the farmer and how many acre of land should be planted. The profit will change depending on the amount of money that the farmer will spend.

4.5. Input processing

These are the inputs for agriculture data for a farmer.

(1) Crop types (e.g. Rice, Maize, Wheat, Potatoes, and Tomato).
(2) Total Land (e.g. A farmer holds 15 acres).
(3) Total cost of each crop per acre.
(4) Profit of each crop per acre.
(5) Total budget of total land for using each crop.

4.6. Output Processing

The result shows that not only increase profit but also the best crop types and allocation of cropland after the system has processed the data.

5. System Implementation

5.1. Performance of the farmland allocation's current profits (Scenario#1)

Table 1. Scenario#1 Traditional Method Output Summary

Crops	Acreage	Cost per Acre	Profit per Acre	Total budget	Total profit
Rice	19211	747900	462100	14367906900	8877403100
Maize	65472	569900	320100	37312492800	20957587200
Sunflowers	1698	170000	251200	28866000	426537600
Garlic	3379	700000	680000	2365300000	2297720000
Soybean	1816	349500	448500	634692000	814476000
Total	91576			54969051700	33373723900

Table 1 shows profit result of traditional method of agricultural land in Taunggyi township from a total land area of 91576 acres and data is collected from the 2021- 2022 growing season.

5.2. Profit Performance under Scenario#2

Table 2. Linear Programming Model Output Summary (Scenario#2)

Crops	Acreage	Total budget	Total profit	Difference in Gross income
Garlic	65515.6624	4586096370 0	4455065050 0	
Soybean	26060.3375	9108087960	1168806140 0	
Total	91576	5496905170 0	5623871190 0	2286498800 0

Table 2 shows profit result of Linear Programming Model of agricultural land in Taunggyi township from a total land area of 91576 acres and data is collected from 2021 – 2022 growing season. In table2, the total budget will not change, but this system supports to choose the best crops and allocates cropland. The best profitable crops are Garlic and Soybean for Taunggyi township in 2021-2022 growing season. Cultivation land 91576 acres,

22864988000Kyats more profit will be found on the table2.

5.3. Result and Discussion

Agriculture is an important business for a country, it will be important for many years to come. Farmers need to be familiar with agricultural techniques. Farmers are expected to use agricultural technology to produce more crops and make a lot of profit. Agriculture production needs to be healthy crops.

In the future, it is expected that the agricultural land owned by the farmer can be cultivated for a long time. This system is expected to reduce the poverty of the farmers which support sustainability development goal. An extension of the sensitivity analysis is expected in the future work of this system.

6. Conclusion

Using linear programming, one may get the most efficient sources of output. In this essay, a decision support system for civilized agriculture in Southern Shan State utilizing linear programming to implement approach. One of the most used optimization models is linear programming. It deals with distributing resources among competing activities in the best way possible. Using the LP technique, the optimal solution that maximizes the profit for crops production of the farmer are examined in this research.

Comparison is made between the results obtained from the LP model and the traditional method of planning. The results obtained from using the LP model are more superior to the ones obtained from using traditional methods. Using agricultural data, a linear programming technique is used to calculate the best distribution of land among the key crops in terms of characteristics including land use, labor in man-days, seeds, fertilizers, and crop production for the years 2021–2022. The suggested linear programming model is suitable for determining the best distribution of land among the principal crops in the research area.

Land use rights are managed effectively and that individuals and communities to which they belong are able to enjoy their land use rights. This system is divided into two parts. The first part is for the farmer to see the information such as brokers information, crops details information, market information etc., the second part is that by entering their agricultural data into this system, farmers will be able to calculate the acreage of agricultural land. Finally, this system aims to promote rural development because the agriculture sector contributes to 37.8pc of GDP and to reduce the poverty of the farmers which support sustainability development goal.

References

[1] F. Majeke and J. Majeke, “A farm resource application problem”, 2010.

[2] F. Majeke and J. Majeke. “A farm resource allocation problem case study of small scale-commercial farmers in Zimbabwe”.

[3] Hemannatha, P.W.Jayasuriya and R. Das. “Agricultural resource management through a linear programming: A case study on productivity optimization of crop-livestock farming integration”.

[4] K.C.Lgwe and C.E Oeyenweaku, “A Linear Programming Approach to Food Crops”.

[5] M. T. Mellaku, T. W. Reynolds and T. Woldeamanuel, “Linear programming-based cropland allocation to enhance performance of smallholder crop production”, November 2018.

[6] N. Patel, M. Thaker, C. Chaudhari. “Agricultural land allocation to the major crops through linear programming model”

[7] R. S. Arabia, “A review of application of linear programming to optimize agricultural solution”, April 2021.

[8] Shreedhar, “Multi crop optimization using linear programming model for maximum net benefit”, 2018.

[9] W. Hua, “Application of the revised simplex method to the farm planning model”, 1988.

Clustering of Countries based on Number of COVID-19 Cases by using DBSCAN Algorithm

Min Khant Htway, Hay Mar Soe Naing
University of Computer Studies, Yangon, Myanmar
minkhanthtway@gmail.com, haymarsoenaing@ucsy.edu.mm

Abstract

Clustering is now an important technique for data analysis. There are numerous clustering algorithms available. Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is one of them that is useful in the medical domain. Therefore, this system will implement the clustering of COVID-19 statistic data by using DBSCAN. It is a density-based clustering algorithm that finds clusters of arbitrary shape and size by growing regions with sufficiently high density point into clusters. This system clusters each country that occurs the similar number of COVID-19 cases. In experiments, three distance measuring methods, namely, Euclidean, Manhattan and Minkowski are used to calculate the distance between each country and evaluate the effectiveness of clustering performance in DBSCAN. The silhouette coefficient is used to measure the goodness of clustering quality. According to the experiments, using DBSCAN with Euclidean distance achieved the superior result.

Keywords: Clustering, COVID-19, DBSCAN, Euclidean, Manhattan, Minkowski

1. Introduction

Clustering is a crucial data analysis task that divides data points into reasonable homogeneous groups so that data points in the groups have similar properties and data points in different groups have different properties in some characteristics. Image processing, data analysis, business applications, the medical domain, and pattern recognition all benefit from clustering techniques. There are five different techniques for clustering such as Partitioning, Hierarchical, Density-based, Grid-based and Model-based. Among them, this system uses the density-based approach.

DBSCAN (density-based spatial clustering of application with noise) algorithm performs well in clustering as an outstanding representative of density-based clustering algorithms. Density-based clustering algorithms are based on the idea that a cluster in a data space is a contiguous region of high point density and are independent of prior knowledge of the number of clusters. DBSCAN data clustering works on the assumption that clusters are dense regions in space separated by regions of lower density, and effectively high density into clusters and discovers clusters of arbitrary shape in spatial databases with noise. The DBSCAN algorithm identifies clusters of counties that are responsible for the cause COVID-19 disease. Using the proposed system, the user can clearly see which countries are in the worst case and which countries have similar cases to other countries around the world.

The outline of this paper is organized as follows. The literature reviews of clustering on statistical COVID-19 cases are described in Section 2. Overview of Clustering is explained in Section 3. The detail explanation of Density-based Spatial Clustering of Applications with Noise (DBSCAN) is shown in Section 4. Then, the proposed system design is described in Section 5 and 6. In Section 7, we will conclude our proposed work.

2. Related Work

In 2021, V. Crnogorac, M. Grbic and M. Dukanovic [1] compared the BIRCH, K-means, agglomerative clustering. European Countries are cluster by number of European COVID-19 patients. The clustering is based on publicly available data published on the website of the European Centre for Disease Prevention and Control and is carried out using three different clustering methods. Silhouette Coefficient value used to calculate the clustering performance of

clusters. The findings of this study may be useful to public health officers and practitioners in dealing with COVID-19 challenges.

In 2020, R. Pung, C. J. Chiew and B. E. Young [2] investigated three cluster of COVID-19 in Singapore linked to tour group from China, church and company conference were identified in February, 2020. By assuming that each of the three clusters had a single primary case that was unknown, they were also able to determine the likely number of transmissions that would result from each verified COVID-19 case.

3. Clustering

Clustering is data mining technique that deals with finding similar objects in a collection of unlabeled data. It is the process of organizing objects into different groups whose is based on the similarity of the data [3]. The result of clustering represents a data concept, where a cluster represents basically comprising a set of abstract objects into groups of similar objects and dissimilar objects. In medical domain, cluster analysis provides a systematic, formalized method for data exploration and defining groups with clinical similarities [4].

4. Density-Based Spatial Clustering of Applications with Noise (DBSCAN)

The DBSCAN clustering method finds clusters with noise and arbitrary forms. A cluster is defined by DBSCAN as the highest collection of densely connected points. It consists of different characteristics. The "ϵ" is the maximum radius of the neighborhood. The user identifies the ϵ value by using K-distance graph. To draw the K-distance graph, find minimum values of distances to the nearest *MinPts* and sort distances ascending. "ϵ" is corresponding to the critical change (strong bend) in curves on K-distance graph. "*MinPts*" is the minimum number of points in an Eps-neighborhood of that point. *MinPts* imports user-specified data ranging from 2 to 2 x dimensions. *MinPts* is the number of countries in a cluster. The user can specify the minimum number of countries in a cluster. A core object is one that is surrounded by other objects that are at least *MinPts* away from it and *Eps* away from it. Noise object that is neither a border object nor a core object. Density-reachable, "*p*" object is density reachable from "*q*" objects, "*q*" not from *p*. Density-connected, if there is a third object "*o*" to which both object "*b*" and object "*q*" are density reachable but "*q*" is not density reachable from object "*p*", then object "*b*" and object "*q*" are connected [5].

4.1. DBSCAN Algorithm

The step-by-step processing of the DBSCAN algorithm [6] is shown in Figure 1.

Algorithm: DBSCAN
Input:
D: a data set containing n objects
ϵ : the radius parameter, and
MinPts: neighborhood density threshold
Output: A set of density-based clusters.
Method:
(1) mark all objects as unvisited;
(2) do
(3) randomly select an unvisited object *p*;
(4) mark *p* as visited;
(5) if the ϵ-neighborhood of *p* as at least *MinPts* objects
(6) create a new cluster *C*, and add *p* to *C*;
(7) Let *N* be the set of objects in the ϵ-neighborhood of *p*;
(8) for each point *p'* in N
(9) if *p'* is unvisited
(10) mark *p'* as visited;
(11) if the ϵ-neighborhood of *p'* has at least *MinPts* points, add those points to *N*;
(12) if *p'* is not yet a member of any cluster, add *p'* to *C*;
(13) end for
(14) output *C*;
(15) else mark *p* as noise;
(16) until no object is unvisited;

Figure 1. DBSCAN Clustering Algorithm

4.2. Advantages of DBSCAN

The DBSCAN algorithm has the following benefits:

- Ability to arbitrarily shaped clusters.
- It finds cluster completely surrounded by different clusters.
- It is robust towards outliers (major) detection.
- Setting the two parameters just takes two points, which are simple for someone who is familiar with the dataset.

4.3. Limitation of DBSCAN

Some limitation of DBSCAN algorithm are as follows:

- Sensitive to clustering parameters *MinPts* and *Eps* (ϵ) for an unknown dataset
- Sampling affects density measures
- Result depends highly on distance metric
- Datasets with altering densities are tricky

4.4. Distance Measure Methods

The Manhattan distance measure is defined as

$$d(a,b) = |x_{a1} - y_{b1}| + |x_{a2} - y_{b2}| + \cdots + |x_{an} - y_{bn}| \quad (1)$$

The Euclidean distance measure is defined as

$$d(a,b) = \sqrt{(x_{a1} - x_{b1})^2 + (x_{a2} - x_{b2})^2 + .. + (x_{an} - x_{bn})^2} \quad (2)$$

The Minkowski distance measure is defined as

$$d(a,b) = \sqrt[h]{|x_{a1} - x_{b1}|^k + |x_{a2} - x_{b2}|^k + \cdots + |x_{an} - x_{bn}|^k}$$

where k = 3 (3)

The formula defines data objects "a" and "*b*" with a number of dimension equal to "*n*". *d (a, b)* is the distance between two data objects. x_{an} is the measurement of object "a" in dimension "*n*" [7].

4.4. Evaluation of Clustering Quality

Silhouette coefficient based on the cohesion and separation of each sample point, is a metric used to calculate the clustering performance of clusters. *a(o)* stands for average intra-cluster distance, which is determined by averaging the distances between each item in the cluster to which *o* belongs. The *b(o)* is average inter-cluster distance that means the minimum average distance from *o* to all clusters to which *o* does not belong.

$$a(o) = \frac{\sum_{o' \epsilon C_i, o \neq o'} dist(o,o')}{|C_i| - 1} \quad (4)$$

$$b(o) = \min_{C_j : 1 \leq j \leq k, j \neq i} \left\{ \frac{\sum_{o' \epsilon C_j} dist(o,o')}{|C_j|} \right\} \quad (5)$$

$$s(o) = \frac{b(o) - a(o)}{\max\{a(o), b(o)\}} \quad (6)$$

Silhouette source ranges from -1 to 1. When the value is near to 1, the cluster members are clearly distinct from one another and are spaced widely apart. Close to 0 are indifferent or have negligible distances between them. If value is close to -1, clusters have been incorrectly assigned.

5. Proposed System Design

This system uses the DBSCAN algorithm for clustering of countries based on the number of COVID-19 cases. In this system, the user firstly imports statistical COVID-19 cases dataset for DBSCAN process. Then, the user inputs the *MinPts* value that is the nearest neighbor value. To calculate the distance between each object, user can choose the distance measuring methods, namely, Euclidean, Manhattan and Minkowski. According to the user selected distance method, this system calculates the distance between each object (each country) based on confirmed cases, death cases, recovered cases, active cases, new cases, new death cases and new recovered cases. And then, this system identifies the ϵ value based on the resultant k distance graph. The system flow diagram is illustrated in Figure 2.

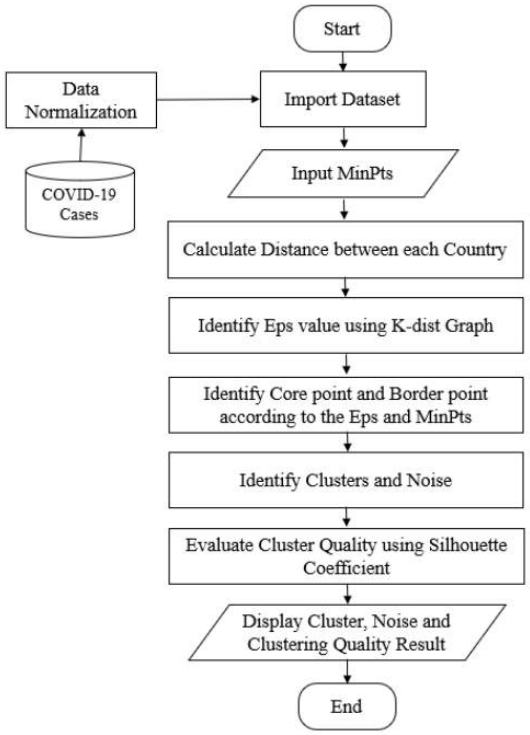


Figure 2. System Flow Diagram

By using the given ϵ and *MinPts* values, this system identifies the core objects. According to the DBSCAN process, this system clusters each

core object until all objects have been processed. If an object is a core object, the DBSCAN determines all points expect this point is density-reachable from it and forms a cluster. After clustering each core object, this system identifies noise object that is not included in some clusters. To identify the clustering quality about each country, this system calculates the Silhouette score. The cluster quality results are different based on the selected similarities calculation methods. This system compares each cluster quality of each country using silhouette scores. Finally, this system displays each cluster, noises and cluster quality results.

6. Experimental Results

In this system, COVID-19 dataset is used to produce clusters. This system groups the countries which faces the similar number of confirmed cases, deaths cases, recovered cases, active cases, new cases, new deaths and new recovered into each cluster. This dataset includes the "35,157" statistic records that are obtained from "22/1/2020" to "27/7/2020". These records are related with "187" countries from the WHO region. Since there is a huge difference in population size and also in number of tested cases among countries, this system is not considered on the original absolute numbers of COVID-19 cases. Because of this, we considered data on the number of reported COVID-19 cases per 10 million people for each nation. **Rate of Cases** from each country are calculated based on 10 million by using the following equations.

$$\text{Rate of Cases(country)} = \frac{\text{Cases}}{Population} * 10\ million \quad (7)$$

Last but not least, we used the dataset in this study, which includes the cumulative number taken for each country and every day. As a sample, this system is tested on the date 1/5/2020. The first 10 statistic records are shown in Table 1.

After importing COVID-19 dataset, the user inputs the *MinPts*. In this example, the *MinPts* = 3 is inputted from the user. Then, the user can choose either the Euclidean or Manhattan or Minkowski to calculate the distances between points. After the user has chosen, this system calculates the distance between each data record. Then, find the k distance metric between each record. One record represents one WHO country.

Table 1. Covid-19 Statistics Data

Country	Confirmed	Deaths	Recovered	Active	New Cases	New Deaths	New Recovered
Afghanistan	602.6	17.5	80.0	505.1	42.3	1.0	12.9
Albania	2716.7	107.7	1695.3	913.7	31.3	0	62.5
Algeria	950.8	103.7	416.8	430.4	33.9	0.7	9.6
Andorra	96452.6	5567.0	60590.4	30295.2	0	129.5	0
Angola	9.2	0.61	3.37	5.2	0.92	0	1.2
Antigua and Barbuda	2557.2	306.9	1534.3	716.0	102.2	0	409.1
Argentina	1004.6	49.9	286.4	668.4	23.1	1.6	7.9
Armenia	7251.5	111.4	3298.3	3841.8	276.8	3.4	162.0
Australia	2664.4	36.6	2270.1	357.7	4.7	0	12.9
Afghanistan	602.7	17.6	80.01	505.1	42.3	1.0	12.9
…	..	..	..	..	..	..	..

After calculating distance between each object, the user identifies the ϵ value by using K-distance graph for three distance methods. The optimal Epsilon value (ϵ) is corresponding to the critical change (strong bend) in curves on K-distance graph. Therefore, we choose Epsilon value is 11,000 for Euclidean distance. Figure 3 shows the K-Distance graph for Euclidean distance method.

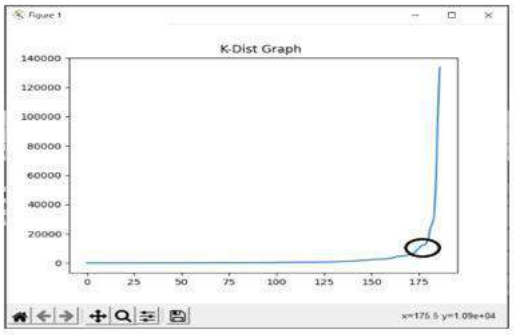


Figure 3. K-Distance Graph for Euclidean

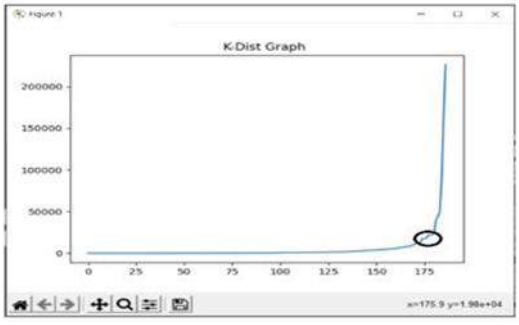


Figure 4. K-Distance Graph for Manhattan

As the same way, this system finds the K-distance graph and choose the ϵ value for other distance methods as shown in Figure 4 and 5. The value 20,000 is used for Manhattan and 9,000 is used for Minkowski distance as shown in Table 2.

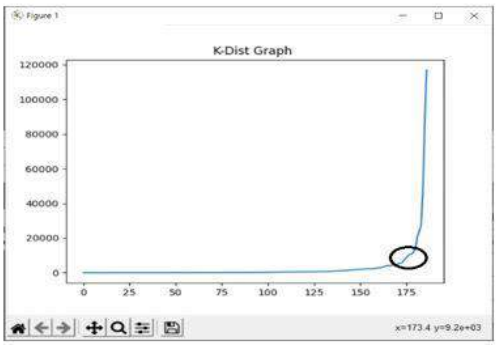


Figure 5. K-Distance Graph for Minkowski

Table 2. Epsilon (ϵ) values and MinPts

Distance	MinPts	Epsilon (ϵ)
Euclidean	3	11000
Manhattan	3	20000
Minkowski	3	9000

This system identifies core countries according to the corresponding *Eps* and *MinPts* values. And find border countries that are density- connected to core countries. This system forms clusters with core countries using DBSCAN algorithm. Also, this system defines the noise object that does not include in any cluster. With the use of Euclidean distance, the system produced four clusters. In Cluster 1 contains 135 countries, Cluster 2 contains 7 countries, Cluster 3 contains 5 countries, Cluster 4 have 3 countries and 37 countries are noises. While using Manhattan distance, it made four clusters. There are 127 countries in Cluster 1, 8 countries in Cluster 2, 3 countries in Cluster 3, and another 3 countries contains in Cluster 4. This turn, the system gave 36 countries as noises. Using Minkowski distance method, the system also generated four clusters. Cluster 1 have 127 countries, Cluster 2 have 7 countries, Cluster 3 have 8 countries and Cluster 4 have 4 countries. The rest 41 countries are denoted as noises. Finally, the Silhouette scores of each country is calculated to know the clustering quality as expressed in Table 3.

Table 3. Cluster Quality Results

Country	Silhouette Scores		
	Euclidean	Manhattan	Minkowski
Afghanistan	0.79544256	0.79230895	0.78472838
Albania	0.7421843	0.71399929	-0.0802103
----	----	----	---
Germany	0	0	0
Ghana	0.79989379	0.79701839	0.78954697
Greece	0.77571712	0.75489585	0.14346144
Greenland	0.83311619	0.84922168	0.66598382
----	----	-----	----
Yemen	0.834	0.8292	0.8207
Zambia	0.834	0.8293	0.8306
Zimbabwe	0.8398	0.8492	0.8349

Positive values indicate that the clusters are well separated and distinct. Negative indicates that clusters have been assigned incorrectly. So, we can say that the clusters are well apart from each other as the silhouette score is closer to 1. In testing on the date 5/1/2019, Euclidean distance with ϵ = 11,000 and *MinPts* = 3 is the best performance compared to the others. In other words, the combination of Minkowski with ϵ = 9,000 and *MinPts* = 3 is the worst case as illustrated in Figure 6.

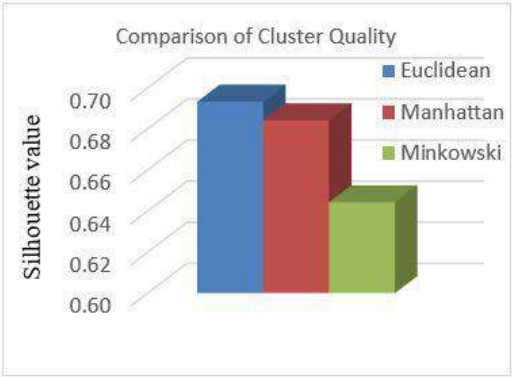


Figure 6. Compare the Clustering Quality

7. Conclusion

Clustering algorithms are attractive for the task of class identification in medical database. In this paper, DBSCAN algorithm is presented for large COVID-19 database. It requires only two parameters (ϵ and MinPts) and supports the user can choose ϵ value on K-Distance Graph. It was suggested to use DBSCAN, which is quite effective at finding clusters with any shape and noise in the database. This system is applied to cluster countries that suffer COVID-19 cases.

Three distance calculation methods are compared in DBSCAN. According to the testing results (5/1/2020), Euclidean distance is the best performance. Moreover, based on the 16 trials experiments, Euclidean distance achieved the highest clustering performance compared to the others. By using this system, people can clearly know which country has the similar amount of COVID-19 suffered cases.

References

[1] V. Crnogorac, M. Grbic and M. Dukanovic, "Clustering of European Countries and Territories based on Cumulative Relative Number of COVID19 Patients in 2020", IEEE, International Symposium Infoteh-Jahorina, 2021.

[2] R. Pung, C. J. Chiew and B. E. Young, "Investigation of Three Clusters of COVID-19 in Singapore Implications for Surveillance and Response Measures", Elsevier, 2020.

[3] P. Kalyani, "Approaches to Partition Medical Data using Clustering Algorithms", International Journal of Computer Applications, vol. 49, no. 23, pp. 7-10, 2012.

[4] K. Kameshwaran and K. Malarvizhi, "Survey on Clustering Techniques in Data Mining", International Journal of Computer Science and Information Technologies, vol. 5, no. 2, pp. 2272-2276, 2014.

[5] E. Martin, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise", University of Munich, Germany, 1996.

[6] A. Moreira, M. Y. Santos and S. Carneiro, “Density-based clustering algorithms – DBSCAN and SNN”, University of Minho – Portugal, 2005.

[7] A. Doroshenko, "Analysis of the Distribution of COVID-19 in Italy using Clustering Algorithms", IEEE, 3rd International Conference on Data Stream Mining & Processing, 2020.

The Detection of Fake Job Posts by Using K-Nearest Neighbor (KNN)

Khin Mar Htay, Yu Yu Than
University of Computer Studies, Yangon
khinmathtay@ucsy.edu.mm , yuyuthan@ucsy.edu.mm

Abstract

Every day, there are issues on many job-calling websites on the internet. Some of the jobs advertised online are actually fake jobs that lead to theft of sensitive information. So it needs to be identified. The system is developed with Fake jobs posts detection by using term frequency – inverse document frequency (TF-IDF) and K-Nearest Neighbor (KNN) algorithm. TF-IDF method is one of the statistical techniques for calculating the importance or score of each word in a document. In order to extract features, TF-IDF is employed. The purpose of this system is to classify real or fake jobs by using the KNN classifier. The system evaluates the results with the accuracy (precision, recall, and F- Measure).

Keywords: Fake job, Preprocessing, TF-IDF, K-Nearest Neighbor

1. Introduction

The development of the internet has made the recruitment process a much quicker process. Additionally, the present pandemic has had a big impact on the way that people are hiring these days. Online recruitment has made it possible to find more applicants and streamline the recruiting process, and it has been very helpful in connecting the gap between employers and potential candidates. On the internet, job seekers can now quickly and easily apply to a wide range of jobs according to their specialization with simply a click of a button. With the aid of E-recruitment, businesses utilize a variety of internet-based solutions. Users can expand their job searches and find the best applicants by using online recruitment. This makes it easier for them to connect with qualified applicants from all over the world. When the client relies on online recruitment, it ends up hiring the best applicant. Companies can choose the most qualified candidates and improve efficiency by using tools like pre-employment screening, personality assessments, and tests for candidate screening. Therefore, there is very little human influence in this process. In terms of communication, online recruitment has a cost-effective advantage.

But, some of these advertised positions are only fake jobs used as traps to take potential data instead of actual jobs. When candidates apply for these jobs, it's possible that their personal information will be stolen, or in some situations, their computers may be hacked to steal important information. Cybercriminals combine the victim's data and either resells it on the dark web for use by another, or they continue to use for years. This fake job post detection attracts considerable attention for developing an automated solution for classifying fake jobs and reporting them to people to avoid applying for such jobs [5].

In the proposed system, the fake jobs can be detected and data theft can be avoided using Term Frequency-Inverse Document Frequency (TF-IDF) and K-Nearest Neighbor algorithm. This system consists of three steps: preprocessing, feature extraction, and classifying fraudulent job by the KNN algorithm. Lastly, this system evaluates the accuracy with precision, recall and f-measure with K value changes.

2. Related Works

Nasser et al. illustrated a number of machine learning classifiers, including Multinomial Naive Bayes, Support Vector Machine, Decision Tree, K Nearest Neighbors, and Random Forest in a text categorization problem,. The data contains both real and fake job postings. For feature extraction, the cleaned and preprocessed data were applied to TF-IDF. The evaluation metrics utilized are accuracy, precision, recall, and f-measure, with one attribute for description. Finally, the system made decision that Random

forest classifier and K-Nearest Neighbor classifier achieved the highest recall [6].

Alghamdi et al. created a model for identifying fraudulent job postings in online job ad systems. The authors had utilized the Employment Scam Aegean Dataset (EMSCAD) dataset on several machine learning algorithms. The methodology consists of three steps: preprocessing, feature selection, and classifying fraudulent by the classifier. In this stage, tags and one unnecessary noise are removed from the data and added to the general text. Selective features are chosen with the use of a support vector machine and random forest classifier to reduce extraversion features that are underutilized. According to reports, the classification accuracy for detecting fraudulent job postings was 97.4% [3].

Van Huynh et al. used deep neural network models that have been retrained using text datasets. The classification of IT-related jobs was done. Text CNN, Bi-GRU CNN, and Bi-GRU-LSTM CNN were the models utilized. There are layers of convolution and pooling in the text CNN model, which is fully associated (Mujtaba et al., Mujtaba & Ryu). Layers were used during the training (convolution and pooling). This model uses the Softmax function for classification, and an ensemble classifier was used to increase accuracy. According to text CNN, the reported accuracy was 66%. BiGRU-LSTM CNN accuracy is 70% [9].

3. Background Theory

3.1. Preprocessing

Preprocessing data is a phase that includes text cleaning and text conversion into a format that is suitable for the classification method. Text is unstructured data that is separated into useful and useless data. So, preprocessing is necessary to remove noise from text. Text preprocessing are tokenization, removing stop word, and stemming [5].Table 2 of preprocessing step instance 1 is displayed in Table 1.

Tokenization: It is the process of dividing a string of characters into smaller units, such as words, phrases, symbols, and other so-called tokens. Tokens may consist of a single word, a few words, a phrase, or even an entire sentence. In this step, digits are removed or filtered. Examples of digits include dates, numbers, times, and other regular expressions and others. All letters are changed to either lowercase letters or uppercase letters.

Removing stop words: It is the method of removing unimportant or unnecessary words that is not pointed any contents. The stop words that point to natural are prepositions, articles, some pronouns, and conjunctions. Stop words in English include "the," "a," "an," "so," "what," etc.

Stemming: It is changing the word's root or common base form. The basic goal of stemming is to reduce a word to its root form from any other grammatical forms or word forms it may have, such as its noun, adjective, verb, or adverb.

Table 1 Sample Text for Instance 1

Original Text
We place highly qualified governors, govern esses, nannies and private tutors into VIP family homes across the world.
After Tokenizing
we/place/highly/qualified/governors/governesses/ nannies/and/private/tutors/into/vip/family/homes/ across/the/world/
After Remove Stopwords
place/highly/qualified/governors/governesses/nan nies/private/tutors/family/homes/across/world/
After Stemming
place/high/qualifi /governor/gover /nanni /privat /tutor /famili/ home /across /world/

3.2. Feature Extraction based on TF-IDF

TF-IDF is the well-known algorithm utilized in text mining methods to calculate the weight of each term [1]. TF stands for term frequencies and IDF stands for Inverse document frequency. It brings attention to a specific problem that might not be mentioned frequently in our corpus but is yet quite important. The TF-IFD score increases proportionally to a word's frequency in a document and decreases as the number of documents in the corpus that use the word raises. Equation 1 is a calculation for TF-IDF.

$$W(t,d) = tf(t,d) * log\frac{N}{nt} \quad (1)$$

Where, W (t, d) is term weight in document d, t f (t, d) is term frequency in document, N is the total number of document and n_t is the number of documents that have term t [5].

TF-IDF is one of the simplest and most efficient weighting schemes for the data. TF-IDF and its algorithm version are the standard choice in text categorization because of it is simple formulation and effective performance on a variety of different data sets.

3.3. Classification

Classification is a method for predicting the class of given data points. Targets, labels, and categories are all terms that can be used to describe classes. Jobs detection can be classified as a classification issue. There are only two categories of real and fake jobs. Using some training data, a classifier can determine how certain input variables relate to a certain class. The training data in this case must consist of both real and well-known fake jobs. When the classifier is accurately trained, it can be used to recognize fake jobs [2]. The most popular types of classification algorithms include support vector machines (SVM), decision trees, random forests, and K-nearest neighbors.

K-Nearest Neighbor (KNN) Algorithm: The KNN algorithm, one of the simplest in machine learning, is based on the supervised learning method. It is a non-parametric order calculation. Training data can be used during the testing phase of the algorithm, which does not require training data to accomplish classification. KNN is based on identifying the objects from sample groups that are the most similar in terms of the mutual Euclidean distance. This KNN algorithm is illustrated as the following.

Step 1: Start: Load the training and test data
Step 2: Choose the value of K
Step 3: For each point in test data:
- 3.1: Find the Euclidean distance between testing data and each of the training data
- 3.2: Store the Euclidean distances in a list and sort it
- 3.3: Choose the first k points
- 3.4: Assign a class to the test point based on the majority of classes present in the chosen points

Step 4: End

This algorithm assumes that similar data points can be found near each other. The purpose of this algorithm is to evaluate the performance of fake or real job based on attributes and training samples. It is simple and easy to implement [4]. In Figure 1 is shown for sample classification of KNN.

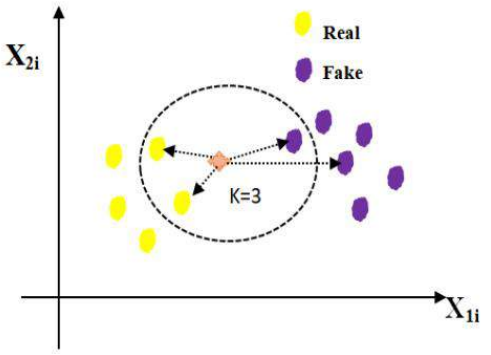


Figure 1. Sample Classification of KNN

$$dist(X1, X2) = \sqrt{\sum_{i,j=1}^{n}(x1i - x2j)} \quad (2)$$

Euclidean distance of formula is shown in Equation 2. The formula defines dist (X1, X2) is the distance between two documents. n is the amount of distinct words in the documents collection. x1i is a weight of the term r in document x1, x2i is a weight of the term r in document x2.

4. Proposed System Design

The overview design of the proposed system is shown in Figure 2. In this proposed system, there are two phases: training and testing. In both phases, two main stages are essential. In the first step, text preprocessing is performed for tokenizing, remove stop words and stemming. In the second step, the weights of words are calculated using TF-IDF features extraction. In the testing phase of the classification step, the system is used K-Nearest Neighbor classifier (KNN) to identify the class category. In the final step, the performance evaluation is measured by using confusion matrix: precision, recall and f-measure.

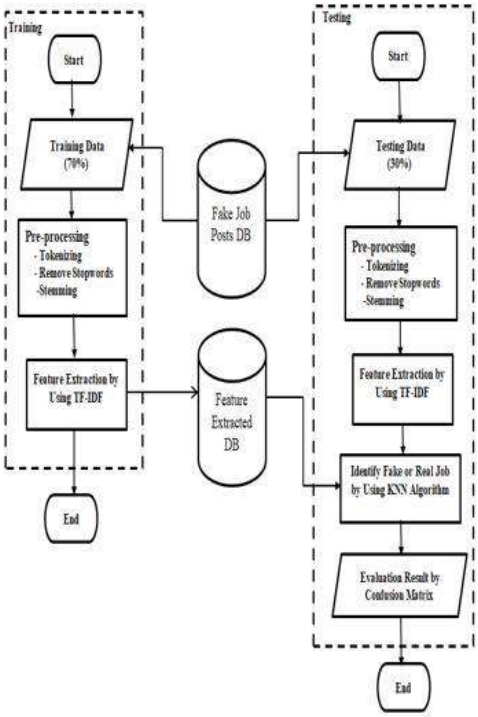


Figure 2. Proposed System Design

5. Explanation of the System

For explanation, this system tests eight job posts from company profile attribute. Firstly, this approach carries out the tokenization; remove stop words and stemming process. And then, words are extracted from each job post. Example jobs are presented in Table 2.

Table 2. Examples of Training Data

Name	Company Profile	Class
Instance 1	We place highly qualified governors, governesses, nannies and private tutors into VIP family homes across the world.	Real
Instance 2	Growing event production company providing staging, scenic, and drapery primarily in the state of Florida. We have a secondary location in Las Vegas and will soon be adding a third location in Southeast Florida. We are a small team passionate about creating high quality events and providing excellent customer service, both on show and in the office.	Real
Instance 3	The most well-liked children's soccer training program in the nation is called Super Soccer Stars. For over a decade, we have provided outstanding instruction for thousands of children in 400+ locations in NY, NJ, CT, MA, CA, FL, IL, Washington, DC, and London, UK! Super Soccer Stars was founded in 2000, and since its inception, it has been providing outstanding soccer development instruction for children aged 2 and up.	Real
Instance 4	UAH is a multi-divisional, full-service private investment firm. With 41 locally managed offices around the nation, we offer private equity funds, residential and commercial clients individualized real estate development, management, and investment services. UAH is an employer that values diversity.	Real
Instance 5	The most interesting items for your life can be found on Fab. Everybody may create their lives and exhibit their unique sense of style with the help of our contemporary, urban-inspired products. Always distinctive, expertly crafted, and of the greatest caliber. Great. Grins. Guaranteed.	Fake
Instance 6	Systems Technology International is a full-service, cutting-edge professional services company that works with almost every sector of the economy. Technical/ engineering, IT (Information Technology), creative and marketing, business analyst, accounting, and office assistance are among the professional services provided by STI.	Fake
Instance 7	At Command, we genuinely care about hiring the right people for the right jobs. We have more than 50 locally-	Fake

	managed branches around the United States that act as trusted partners to businesses and job seekers.	
Instance 8	Eaton's Cooper Notification division offers an extensive range of goods and services, ensuring the safety and security of industrial facilities. These include emergency communications systems, explosion-proof and hazardous location notification appliances, and more.	Fake

Data Collection: The dataset name is Real or Fake Job Posting Prediction from the website Kaggle [10]. This dataset has seventeen attributes. These are “job id”, “title”, “location”, “department”, “salary range”, “company profile”, “description”, “benefits”, and “telecommuting”, “has the company logo”, “has questions”, “employment type”, “required experience”, “required education”, “industry” and “function”. The label is binary for the particular domain of the problem; the real is "0" and the fake is "1". The attribute types are Boolean and Text. Boolean attributes are salary range, and telecommuting, has the company logo, has questions. Some attributes of the description are the same as the text (e.g., department, required experience, required education, industry, and function). Therefore, there are only seven attributes used.

After the extracting words, this system calculates the weight of each word by using TF-IDF methods. Sample weight results from real instance 1 are described in Table 3.

Table 3. Weight Results from Instance 1

Name	words	TF	IDF	Weight
Instance 1	place	0.058	0.4259	0.0251
	high	0.058	0.6021	0.0354
	qualifi	0.058	0.9031	0.0531
	governor	0.058	0.9031	0.0531
	gover	0.058	0.9031	0.0531
	nanni	0.058	0.9031	0.0531
	privat	0.058	0.6021	0.0354
	tutor	0.058	0.9031	0.0531
	famili	0.058	0.9031	0.0531
	home	0.058	0.9031	0.0531
	across	0.058	0.9031	0.0531
	world	0.058	0.9031	0.0531
	world	0.058	0.9031	0.0531

6. Experimental Results

In this fake job detection system, experiments are performed for four times. Table 4 uses four separate training dataset and testing dataset combinations for each analysis (Test1: Training Data 210 records and Testing Data 90 records; Test 2: Training Data 350 records and Testing Data 150 records; Test 3: Training Data 700 records and Testing Data 300 records; Test 4: Training Data 1306 records and Testing Data 560 records). This system used the Precision, Recall, and F-measure of each analysis to evaluate the performance of the experiment results. The analysis results of four different datasets are shown in Figure 3.

Table 4. Precision, recall and f-measure of each test

Testing No	Training and Testing data Proportion	Precision Result	Recall Result	F-Measure
Test 1	210/90	81%	70%	67%
Test 2	350/150	80%	67%	63%
Test 3	700/300	80%	65%	61%
Test 4	1306/560	81%	68%	66%

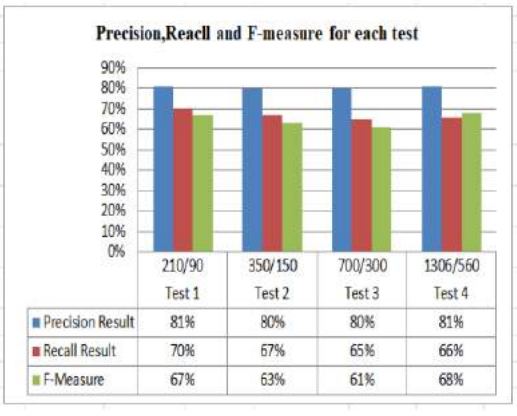


Figure 3. Precision, Recall and F-measure for each test

Table 5 is showed the sample example of 1866 jobs for experimental results with different K value by using cross validation method.

Table 5. Sample example for testing results

Performance Measure Metric	K=3	K=5	K=7
Precision	81%	82%	82%
Recall	68%	69%	69%
F-measure	66%	67%	67%

7. Conclusion

This system is intended to prevent the stealing of personal information when looking for jobs. This system mainly focuses on implementing the classifier for Fake Job Posts detection using the K-Nearest Neighbor (KNN) classifier with feature extraction (TF-IDF) method. The KNN algorithm is one of the most extensively used and successful text categorization methods. The KNN algorithm is used to evaluate the performance of the proposed system for real or fake job posts. According to the experimental results, the KNN algorithm with Euclidean Distance has better performance. The performance accuracy results were calculated with K value changes using the cross validation method.

References

[1] A. A. Hakim, A. Erwin, K. Eng, M. Galinium, and W. Muliady, "Automated document classification for news article in bahasa Indonesia based on term frequency inverse document frequency (TFIDF) approach", ICITEE, pp. 1-4, 2014.

[2] Ariruna Dasgupta Asoke Nath, “Classification of Machine Learning Algorithms”, International Journal of Innovative Research in Advanced Engineering (IJIRAE) ,March (2016).

[3] Bandar Alghamdi, Fahad Alharby, “An Intelligent Model for Online Recruitment Fraud Detection”, Journal of Information Security, (2019).

[4] Bruno Trstenjaka ,Sasa Mikacb , Dzenana Donkoc, “KNN with TF-IDF Based Framework for text Categorization”, 24th DAAAM International Symposium on Intelligent Manufacturing and Automation, (2013).

[5] C.S. Anita ; P. Nagarajan; G. Aditya Sairam; P. Ganesh; G. Deepakkumar, “Fake Job Detection and Analysis Using Machine Learning and Deep Learning Algorithms”, February ,(2022).

[6] Ibrahim M. Nasser and Amjad H. Alzaanin, “Machine Learning and Job Posting Classification: A Comparative Study”, International Journal of Engineering and Information Systems (IJEAIS), September, (2020).

[7] G. Alandjani, “ Online Fake Job Advertisement Recognition and Classification Using Machine Learning”, 3C TIC. Cuadernos de desarrollo aplicados a las TIC, 11(1), 251-267. https://doi.org/10.17993/3ctic.2022.111.251-267(2022).

[8] Vaishali Kalra, Dr. Rashmi Aggarwal, “Importance of Text Data Preprocessing & Implementation in RapidMiner”, Proceedings of the First International Conference on Information Technology and Knowledge Management. pp. 71–75 (2018).

[9] Van Huynh, T., Van Nguyen, K., Nguyen, N. L. T., & Nguyen, A. G. T. , “Job prediction: From deep neural network models to applications” , In RIVF International Conference on Computing and Communication Technologies (RIVF) (2020). IEEE. https://ieeexplore.ieee.org/document/9140760

[10] S. Bansal, “[Real or Fake]: Fake Job Posting Prediction”Kaggle.https://www.kaggle.com/shivamb/real-or-fake-fakejobposting-prediction (accessed Jun. 03, 2020).

[11] W. Zhang, T. Yoshida, and X. Tang, "A comparative study of (TF-IDF), LSI and multi-words for text classification", Expert Systems with Applications, vol. 38, pp. 2758-2765, 2011.

The Analysis of COVID-19 Immunization Data in Rakhine State Using KNN Algorithm

Khin Myat Thu, Thaung Myint Htun
University of Computer Studies, Yangon, Myanmar
khinmyatthu@ucsy.edu.mm, thaungmyinttun@ucssittway.edu.mm

Abstract

Data mining involves the searching of large information of data or records to discover patterns and utilize these patterns in the prediction the future events. Classification is one of the methods in data mining for categorizing a particular group of items into targeted groups. The main goal of classification is to predict the nature of an item or data based on the available classes of items. The construction of the classification model is always defined by the available training data set. In this system, an analysis of COVID-19 immunization results in Rakhine State was carried out using the k-Nearest Neighbor (k-NN) classification algorithm in data mining. The data set about COVID-19 immunization details are collected from General Administration Department, Rakhine State. The primary objective of this system is to evaluate algorithm in the prediction of COVID-19 immunization finishing rate and analysis result of Rakhine state. k-Nearest Neighbor algorithm is utilized to carry out for the prediction of COVID-19 immunization results.

Keywords: k-Nearest Neighbor classifier, COVID-19 immunization, Classification, Data Mining

1. Introduction

COVID-19 is an infectious disease, a type of virus that can spread rapidly through the air. The infection is easy and fast and can be spread from person to person immediately. It spread to many countries worldwide, and the number of infected people increased rapidly daily [5][6].

The COVID-19 disease is caused by the coronavirus and has caused concern and fear in almost all countries, including Myanmar. It caused panic. The rate of spread was also speedy. The virus is not fixed, and the symptoms are similar, but there are differences. There were differences. Especially parts of the body related to the respiratory tract. In particular, the damage to the lung organ is very severe and may lead to loss of life [5][6].

Ministry of Health and Sports confirmed that the coronavirus was first detected in Myanmar on March 23, 2020. In the first wave, 374 cases were released in Myanmar. Six people lost their lives. It was announced that the infection was detected in Myanmar on July 16, 2020[5][6].

The infection prevention related to COVID-19; control and safety measures have been implemented in all regions and states of Myanmar. Guidelines were issued for the public to follow. We should not go out of our homes unless necessary and wear a mask if we go outside. Thus, warnings begin with educating people not to live in densely populated areas. Vaccination of students, employees, people, against COVID-19 is being carried out on a rotating basis, which are imported from abroad [7].

Making better plans, vaccination stations in towns and villages for the prevention of COVID-19 are organized and vaccinated in rural areas. Age-specific vaccination schedules; For the first time in a specified period, the second time, A third vaccination program is underway. In addition, children from primary schools, those working in government departments, residents in ward and village are also vaccinated against COVID-19 by setting up alternate programs [5][6][7][8][10].

On August 16, 2020, it was announced that the infection was detected in Rakhine State. In terms of the number of people affected by the disease, the second wave was more than the first wave. Relevant health workers and local authorities work together to administer vaccinations, but in Rakhine State, transportation and transportation are limited. For the people who live in areas where communication is complicated, it is a rare opportunity to inject the COVID-19 vaccine into the country a specified number of times. Therefore, it can be said that it is necessary to do a

specific business in combination with IT skills to collect accurate lists and meet the criteria for these immunization activities [5][6][7][8][10].

Data mining technology collects much information to obtain prediction results, anomalies patterns and relationships. It helps study in many areas, including health care and education. Therefore, data mining is a research activity. It can also be said that it is a functional modern and innovative technology that combines critical technologies to achieve local benefit projects [5][6][7][8][10].

In Rakhine State, the immunization services started on 27th January 2021 until now. There were five types of vaccines for COVID-19 immunization, Sinopharm, Sinovac, Covid shield, COVAX in, and Astra Zeneca. In this paper, we will evaluate the classification using COVID-19 immunization data sets from the General Administrative Department, Rakhine state [5] [6] [7] [8][10].

2. Related Works

Thirunavukkarasu K. et al. took the iris dataset and used the K-Nearest Neighbors (KNN) classification Algorithm. The model can recognize the iris species automatically. The dataset had 150 samples and three classes, each containing 50 samples— the choice of performance metrics as measured by the algorithms and compared [4].

Seda Çamalan and Gökhan Şengül classified the images using K-Nearest Neighbors (KNN) and Discriminant Analysis (DA)methods. Then, their performances according to the LBP parameters were compared. Also, classification methods' parameters were changed, and the comparison results were shown [3].

Prasannavenkatesan Theerthagiri et al. presented a predictive disease analysis. A study was conducted to achieve prediction on four types of classification algorithms. As a dataset, predictions were calculated using classification techniques on the dataset related to COVID-19. As a result of the analysis, the KNN algorithm produced the slightest error in accurate covid-19 disease prediction than other algorithms. Nevertheless, it is a good predictor of disease risk [2].

Hnin Yu Maw et al. took data sets related to heart disease diagnosis from the UCI machine learning repository and evaluated KNN, one of the most robust classification techniques. The similarity of the symptoms of patients suffering from heart disease was calculated by Euclidean distance. Performance results were also presented. The system is made by the person making the potential decision on behalf of the doctor [1].

3. Methodology

Data mining is the study of new theories and approaches to analyze large datasets. It is an emerging field of computer intelligence that provides new technologies and tools. It is a process that uses meaningful recording techniques, as well as numerical and mathematical techniques, to analyze large amounts of data in the storage and discover new patterns and correlations. Data mining refers to extracting knowledge from a large amount of data. It is an essential step in the knowledge discovery process in databases [9].

Classification and prediction are two forms of data analysis that can be used to extract models describing important data classes or to predict future data trends. There are several classification methods, like Decision Tree, Naive Bayes, k - Nearest Neighbor (k NN) classifier is a very simple algorithm that is nonetheless probably good to classify data. Data preparation may be one of the most difficult steps in any data mining project. The reason is that each dataset is different and highly specific to the project. Nevertheless, there are enough commonalities across predictive modelling projects that a loose sequence of steps can be defined and subtasks that are likely to perform. This process provides a context in which the data preparation required for the project can be considered, informed both by the definition of the project performed before data preparation and the evaluation of data mining algorithms performed after. In this system, the data sets have numeric, binary, ordinal and nominal data types. On a predictive modelling project, such as classification or regression, raw data typically cannot be used directly. Thus, the data sets have to prepare the same data type of number to make the evaluation.

3.1. K-Nearest Neighbor (k NN)

The k-Nearest Neighbor (k-NN) algorithm is one of the most widely used classification algorithms due to its simplicity and easy implementation. It is also used as the baseline

classifier in many domain problems. The k-NN algorithm is a conventional non-parametric classifier usually used for classification and regression problems. The learning and prediction analysis is performed based on the given problem or dataset. In the k-NN classification model, the prediction is purely based on neighbor data values without any assumption on the dataset. In k-NN, the 'k' represents the number of nearest neighbor data values. Based on 'k', i.e., the number of nearest neighbors, the decision is made by the k-NN algorithm on classifying the given dataset. The k-NN model directly classifies the training dataset. It means the prediction of a new instance is made by searching the similar 'k' neighbor instances in the entire training set and classifying them based on the class of highest instances. A similar instance is determined using the Euclidean distance formula. Euclidean distance between two points or tuples, say, X1= (x11, x12, …, x1n) and X2= (x21, x22, …, x2n), is

$$dist(X_1, X_2) = \sqrt{\sum_{i=1}^{n} (x_{1i} - x_{2i})^2}. \quad (1)$$

In other words, for each numeric attribute, take the difference between the corresponding values of that attribute in tuple X1 and in tuple X2, square this difference, and accumulate it. The square root is taken from the total accumulated distance count [7][11].

Typically, the values of each attribute are normalized before using Eq.1. This helps prevent attributes with initially large ranges from (e.g., income) from outweighing attributes with initially smaller ranges (e.g., binary attributes). Min-max normalization, for example, can be used to transform a value v of a numeric attribute A to v 0 in the range [0, 1] by computing

$$v' = \frac{v - min_A}{max_A - min_A}, \quad (2)$$

where *minA* and *maxA* are the minimum and maximum values of attribute *A* [12].

3.2. An Example Calculation

Let X be a data sample whose class label is unknown. Each data sample is represented by an n-dimensional feature vector, $X = (X_1, X_2, \ldots, X_n)$. In this sample, X = (Housing = "430", Household = "426", L18M = "738", L18F = "699", O18M = "821" only with 5 attributes. Calculate the Euclidean distance of the new sample for each row of the training dataset. First, find the square difference for each attribute. Second, calculate the sum of the square differences. Finally, find the square root of the sum.

ID	Housing	Household	L18_M	L18_F	O18_M	Finishing Rate
	x1	x2	x3	x4	x5	Class
1	0.151633	0.151633	0.257387	0.263608	0.103421	Yes
2	0.095645	0.110420	0.069207	0.048989	0.195179	Yes
3	0.575428	0.609642	0.405910	0.365474	0.962675	Yes
4	0.206065	0.207621	0.073872	0.079316	0.350700	Yes
5	0.569207	0.578538	0.338258	0.370918	1.000000	Yes
6	0.521773	0.572317	0.361586	0.365474	0.979782	Yes
7	0.298600	0.322706	0.380249	0.392691	0.581649	Yes
8	0.099533	0.099533	0.160964	0.146190	0.078538	Yes
9	0.059876	0.056765	0.055210	0.048212	0.132970	No
10	0.000778	0.000000	0.007776	0.005443	0.006998	No
11	0.007776	0.006221	0.016330	0.013219	0.034992	No
12	0.041991	0.039658	0.049767	0.061431	0.076205	No
13	0.017885	0.020218	0.034992	0.027216	0.059876	No
14	0.090980	0.087092	0.096423	0.090202	0.163297	No
15	0.094868	0.083204	0.130638	0.118196	0.177294	No
New Data	0.320373	0.317263	0.559876	0.529549	0.624417	Yes

Figure 1. Training Datasets with 5 Attributes and New Data for k-NN Evaluation

ID	(X1-X1)^2	(X2-X2)^2	(X3-X3)^2	(X4-X4)^2	(X5-X5)^2	Sum	Distance	Neighbors	Prediction
1	0.0285	0.0274	0.0915	0.0707	0.2714	0.4896	0.6997		Yes
2	0.0505	0.0428	0.2408	0.2309	0.1842	0.7492	0.8656		
3	0.0651	0.0855	0.0237	0.0269	0.1144	0.3156	0.5618	Yes	
4	0.0131	0.012	0.2362	0.2027	0.0749	0.5389	0.7341		
5	0.0619	0.0683	0.0491	0.0252	0.1411	0.3455	0.5878		
6	0.0406	0.0651	0.0393	0.0269	0.1263	0.2981	0.546	Yes	
7	0.0005	0	0.0323	0.0187	0.0018	0.0533	0.2309	Yes	
8	0.0488	0.0474	0.1591	0.147	0.298	0.7003	0.8368		
9	0.0679	0.0679	0.2547	0.2317	0.2415	0.8636	0.9293		
10	0.1021	0.1007	0.3048	0.2747	0.3812	1.1635	1.0787		
11	0.0977	0.0967	0.2954	0.2666	0.3474	1.1039	1.0507		
12	0.0775	0.0771	0.2602	0.2191	0.3005	0.9344	0.9667		
13	0.0915	0.0882	0.2755	0.2523	0.3187	1.0263	1.0131		
14	0.0526	0.053	0.2148	0.193	0.2126	0.726	0.8521		
15	0.0509	0.0548	0.1842	0.1692	0.1999	0.659	0.8118		

Figure 2. k-NN Evaluation and Prediction

4. System Design and Datasets Description

The system is implemented for those who want to know whether the specific region meets the COVID-19 immunization finishing rate 90% or not and the analysis report for that region by allowing digitized software. The following figure 3 illustrates system design overview of the proposed system.

This dataset contains information concerning the finishing rate of COVID-19 immunization services in Rakhine State. The data was collected from the General Administration Department, Rakhine State. Before training and classifying, a number of pre-processing decisions had to make. The attributes which are important for making a prediction of whether the finishing rate of 90% immunization services in Rakhine State has 18 attributes (including class attribute). The attribute

information and sample dataset of this system is described in section 4.1 and 4.2.

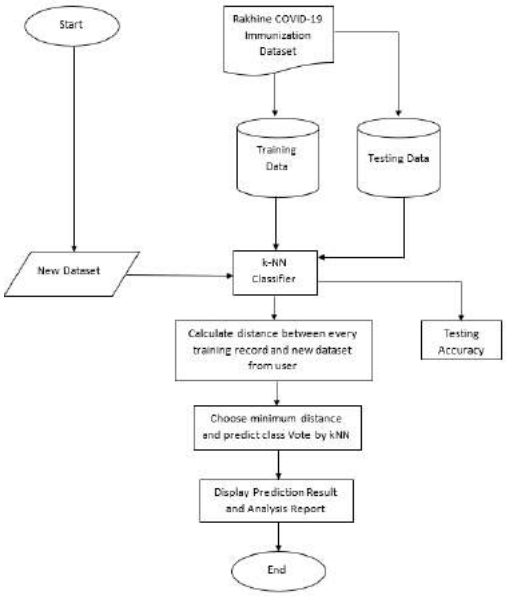


Figure 3. System Flow Diagram

4.1 Attribute information

The attributes information and descriptions are shown below:

1. Housing = Number of housing (integer)
2. Household = Number of household (integer)
3. L18M = Number of males under 18 (integer)
4. L18F = Number of females under 18 (integer)
5. O18M = Number of males over 18 (integer)
6. O18F = Number of females over 18 (integer)
7. Eligible = Number of eligible people in the region (integer)
8. Quarter, Village tract Or Small Village = Is the region is whether quarter, village tract or small village (Yes, No)
   value1: Yes (1)
   value2: No (0)
9. Ethnic Type = Type of ethnic in the region (Citizen, Bengali, Mix)
   value1: Citizen (1)
   value2: Bengali (2)
   value3: Mix (3)
10. Distance From City = The condition of the region's distance from City (Near, Fair, Far)
    value1: Near (1)
    value2: Fair (2)
    value3: Far (3)
11. Transport Congestion = The condition of the region's transport (Easy, Normal, Difficult)
    value1: Easy (1)
    value2: Normal (2)
    value3: Difficult (3)
12. Rules Of Law = The condition of the rules of law in the region (Good, Fair, Bad)
    value1: Good (1)
    value2: Fair (2)
    value3: Bad (3)
13. District Hospital = Is there district hospital in the region? (Yes, No)
    value1: Yes (1)
    value2: No (0)
14. BES (Basic Education School) = Is there basic education schools (Yes, No)
    value1: Yes (1)
    value2: No (0)
15. Stable Living = Are people in the region stable living (Yes, No)
    value1: Yes (1)
    value2: No (0)
16. Natural Disaster Area = Is the region natural disaster area (Possible, Rarely)
    value1: Possible (1)
    value2: Rarely (0)
17. Internet/Phone Connection = The condition of the internet/ Phone connection (Normal, Often, Never)
    value1: Normal (1)
    value2: Often (2)
    value3: Never (3)
18. Finishing Rate = Does the region meet the 90 % immunization services (Yes, No) (Class)
    value1: Yes (1)
    value2: No (0)

4.2 Sample Dataset

There are 900 sample datasets in this system. 600 for training and 300 for testing datasets. We present five of them in this section.

337,384,510,572,819,869,1688, Yes, Mix, Near, Easy, Good, Yes, Yes, Yes, Rarely, Normal, Yes

196,207,126,137,313,404,717, Yes, Mix, Near, Easy, Good, Yes, Yes, Yes, Rarely, Normal, Yes

650, 720, 941, 835, 1114, 1303, 2417, Yes, Mix, Near, Easy, Good, Yes, Yes, Yes, Rarely, Normal, Yes

95, 91, 89,80,189,211,400, No, Bengali, Far, Normal, Good, No, No, Yes, Rarely, Normal, No

135,130,142,134,228,225,453, No, Citizen, Far, Normal, Good, No, No, Yes, Rarely, Normal, No

5. Implementation

Figure 4. Main Form

Figure 4 shows the main form of the system. The sub-menus like 'system' and 'COVID-19' of the File menu explain the system and COVID-19. When the "Login" sub-menu is clicked from the File menu to management as admin role, the login form appears as shown in Figure 4.3. The 'Exit' sub-menu is to leave the system, and the "Prediction" menu is to check the prediction result and analysis from the user level. Figure 7 shows the training dataset and to manage the data as admin and figure 8 shows the testing data set and accuracy result of the system after login successful.

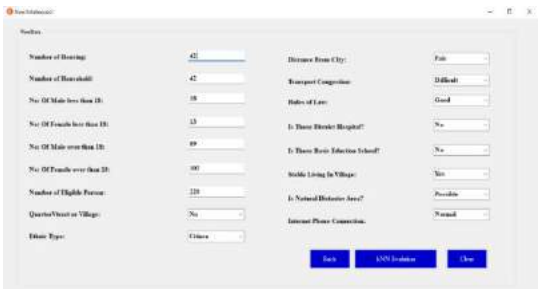


Figure 5. New Data Request Form

Figure 5 shows when the prediction menu is clicked, and the user fills the data required. KNN evaluation menu is clicked the distance and prediction result of KNN calculation and analysis report after analysis button is clicked as shown in Figure 6.

Figure 6. Prediction and Testing Result generated from k-Nearest Neighbor

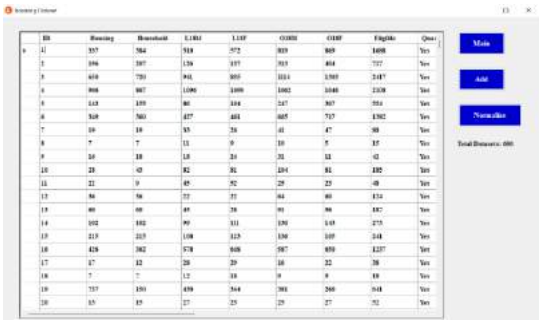


Figure 7. Training Dataset

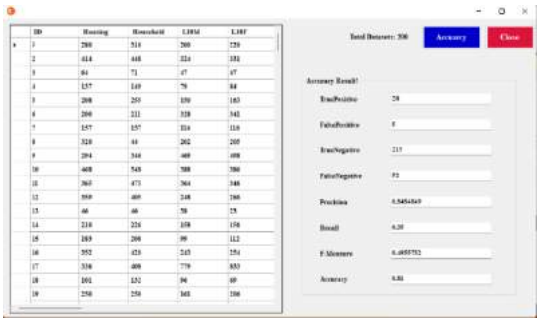


Figure 8. Testing Dataset and Accuracy Result

6. Classifier Accuracy

Estimating the classifier accuracy is important, as it allows one to evaluate how accurately a given classifier will label future data or data on which the classifier has not been trained. Accuracy estimates also help in comparing different classifiers. Using the training data to derive a classifier and estimate its accuracy can produce misleading, over-optimistic estimates.

In this system, about 900 COVID-19 immunization data points from Rakhine State's records are used as a dataset to evaluate the performance of the classification system. The holdout method estimates the classifier performances in this system. The given data are randomly partitioned into two independent sets a training set and a test set. Typically, two-thirds of the data are allocated to the training set, and the remaining one-third is allocated to the test set.

The classifier evaluation measures include accuracy (also known as recognition rate), sensitivity (or recall), specificity, precision, and F1-score (or F-Measure). To evaluate the classification accuracy, two main matrices have been computed for the k-NN classifier in terms of the correct classification rate (%) in both the training and the testing phases. A sensitivity of 100% means that the test recognizes all 90% finishing rates that COVID-19 immunization

requires. Thus, in a high-sensitivity test, a negative result is used to rule out the finishing rate of COVID-19 immunization in Rakhine State. A specificity of 100% means that the test recognizes all 90% finishing rates of COVID-19 immunization meets as 90% finishing rate meets. Thus, a positive result in a high specificity test can be used to confirm the finishing rate. Precision may also be used to access the percentage of samples labelled, for example, "Yes" of COVID-19 immunization 90% finishing rate that is actual "Yes" COVID-19 immunization 90% finishing rate samples.

Using training data derived from the classifier or predictor to estimate the accuracy of the resulting learned model can result in misleading, over-optimistic estimates due to the over-specialization of the learning algorithm to the data. The accuracy of a classifier on a given test set is the percentage of test set tuples correctly classified by the classifier.

Table 1. A Confusion Matrix for Positive and Negative Tuples

		Predicted class	
		Class 1	Class 2
Actual class	Class 1	True positives	False negatives
	Class 2	False positives	True negatives

$$\text{Sensitivity} = \frac{t_{pos}}{pos} \quad (3)$$

$$\text{Specificity} = \frac{t_{neg}}{neg} \quad (4)$$

$$\text{Precision} = \frac{t_{pos}}{t_{pos} + f_{pos}} \quad (5)$$

Where

t_{pos} = the number of true positives ("Yes" of finishing rate 90% samples that were correctly classified as such),

pos = the number of positives ("Yes" of finishing rate 90% samples),

t_{neg} = the number of true negatives ("No" of finishing rate 90% samples that were correctly classified as such),

neg = the number of negatives ("No" of finishing rate 90% samples),

f_{pos} = the number of false positives ("No" of finishing rate 90% samples that were incorrectly labelled as "Yes" of finishing rate 90%),

$$\text{F-Measure} = (2*\text{Precision}*\text{Recall})/(\text{Precision} + \text{Recall}) \quad (6)$$

Show that accuracy is a function of sensitivity and specificity:

$$\text{Accuracy} = \text{sensitivity}\ \frac{pos}{pos + neg} + \text{specificity}\ \frac{neg}{pos + neg} \quad (7)$$

The true positives, true negatives, and false positives are also useful in assessing the cost and benefits that have been computed and are associated with a classification model of this classification [12].

There are 900 records in this system. 600 datasets have been trained and 300 datasets are used as testing datasets. There are numerous methods to check accuracy which is used to evaluate the performance of k-nearest neighbor classification. Among them, hold-out method is used in this system. In accordance with table 2, testing accuracy has obtained 81%.

Table 2. Accuracy Result

True Positive	28
False Positive	5
True Negative	215
False Negative	52
Precision	0.8484849
Recall	0.35
F-Measure	0.4955752
Accuracy	0.81

7. Conclusion

The system is implemented to predict the COVID-19 immunization finishing rate 90% in Rakhine State by using the k-Nearest Neighbor classification. After that, the analysis report for the region is presented, that report is based on only 5 attributes which are selected as per instructions and calculation. As the future work, this system can be used to predict all regions in Rakhine state and analyze why the region in Rakhine state meets the immunization finishing rate 90% or does not. This system can know whether a region has reached 90% immunization services coverage or not and can support those who are working in COVID-19 defense areas.

References

[1] Hnin Yu Maw et al, "Evaluation of Symptoms in Heart Disease Patients by using k – Nearest Neighbor Classification".

[2] Prasannavenkatesan Theerthagiri et al, "Prediction of COVID-19 Possibilities using KNearest Neighbour Classification Algorithm ", DOI:10.31782/IJCRR. 2021.SP173, Jan 2021.

[3] Seda Çamalan and Gökhan Şengül, "Gender Prediction by Using Local Binary Pattern and K Nearest Neighbor and Discriminant Analysis Classifications"

[4] Thirunavukkarasu K. et al, "Classification of Iris dataset using Classification based KNN Algorithms in Supervised Learning", 2018 4th International Conference on Computing Communication and Automation (ICCCA) 2018 4th International Conference on Computing Communication and Automation (ICCCA).

[5] https://covid19.who.int/region/searo/country /mm

[6] https://www.ijisme.org/wp-content/uploads/papers/v5i1/A1056085117.pdf

[7] https://www.researchsquare.com/article/rs-70985/v1.pdf

[8] https://www.safeatworkca.com/safety-articles/covid-19-emergency-regulations/

[9] https://www.talend.com/resources/what-is-data-mining/

[10]https://www.unicef.org/myanmar/press-releases/more-1-million-doses-vaccines-and-10000-covid-19-test-kits-arrive-myanmar

[11] https://ivypanda.com/essays/data-mining-essay/

[12] Jiawei Han, Micheline Kamber, Jian Pei, "Data Mining Concepts and Techniques" , Third Edition.

Classification of Psychological Illnesses Using Naïve Bayes

Kay Khaing Soe, Win Lai Hnin
Information Technology Supporting and Maintenance, Faculty of Computer Science
University of Computer Studies, Hinthada
kaykhaingsoe187@gmail.com, winlaihnin.84@gmail.com

Abstract

Physical and psychological illness are important for people in their daily life. If people face risks, conflicts and difficulties, they suffer from psychological illness problems such as depression, bipolar disorder, schizophrenia, and anxiety. Psychological illness professionals ask the patient concerned with the present conditions (risk and symptoms) and mood or feelings. So, the patients caused with the risks or symptoms are classified (psychological illness Yes or No) according to their cases. There are many techniques of data mining applied for performance of many application areas. This system contributes to classify psychological illness by using one of Data mining techniques, Naïve Bayes. The system's performance is evaluated in terms of accuracy with accuracy, precision, and recall.

Keywords: Mental Health Domain, Classification, Bayes Classifier, Naïve Bayes, Confusion Matrix.

1. Introduction

In Data mining, there is a variety of techniques to identify suggest of information or decision-making knowledge in the database and extracting these in a way such as decision support, and predictions. In the late of 1960s, model-oriented decision support system became practical. Decision support system is to help complete decision makers with managerial decisions based on personal intuition and experience.

Naive Bayes classifier is based on Bayes theorem. This classifier algorithm used conditional independence, means it assumes that an attribute value on a given class is independent of the values of other attributes.

Psychological illness, called mental health disorder, is the feelings, thoughts and behavior in mind. Psychological illness concern when ongoing conditions and symptoms cause frequent stress and affect the ability of people to function. A mental illness can make you miserable and can cause problems in your daily life, such as at school or work or in relationships. People join and take the help of Mental Health professionals and mental health professionals ask their conditions. In this system, the patients are classified that he/she suffers or not mental health based on the patients' answers with the mental conditions using Naïve Bayes.

2. Methodology

2.1. Naïve Bayes

Naïve Bayesian classifiers is one of Data Mining Techniques. This algorithm works quickly and can save a lot of time. Naive Bayes is the most suitable classifier for solving multi-class prediction problems and for categorical input variables than numerical variables. Naïve Bayesian classifiers (Formula):

$$P(c|x) = \frac{P(x \mid c)P(c)}{P(x)} \quad (1)$$

where,

P(c) = The prior probability of class
P(x) = The prior probability of predictor
P(c|x) = The posterior probability of class (target) given predictor (attribute)
P(x|c) = The likelihood which is the probability of predictor given class

X can be written as follow:
$X = (x1, x2, x3, \ldots, xn)$

In Figure 1, the relationship between C(class) and X (predictor) are described.

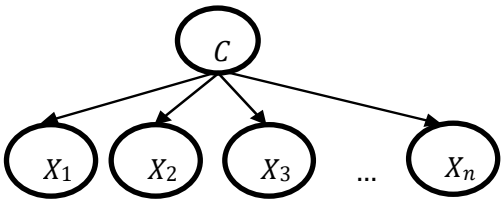


Figure 1. Naïve Bayes

$$P(C|x_1, x_2, \dots, x_n) = \frac{P(C)P(x_1, x_2, \dots, x_n|C)}{P(x_1, x_2, \dots, x_n)} \quad (2)$$

In this formula, with the substitution of X, the Bayes formula can be printed as follows complex factors of probability values which is nearly impossible to analyze one by one. As a consequence, the calculation develops hard to do.

$$P(C|x_1, x_2, \dots, x_n) = P(C) \prod_{i=1}^{n} P(x_i|C) \quad (3)$$

In Naïve Bayes Classifiers essential to maximize the probability value of individually class, which is expressed as the Hypothesis Maximum aa Posteriori (HMAP):

$$H_{map} = argmaxP(C|x_1, x_2, \dots, x_n) = argmaxP(C) \prod_{i=1}^{n} P(x_i|C) \quad (4)$$

In Naïve Bayes Classifier, by means of Equation can forecast which classes can be used in Naïve Bayes Model. But, if the attribute X in Equation has quantitative types, then the probability will be very minor such that the value cannot be used to find the value. So we essential to use other technique such as normal (Gaussian) distribution.

$$P = \left(X_i = x_i \middle| C = c_j\right) = \frac{1}{\sqrt{2\pi\sigma_{ij}}} exp\left(-\frac{\left(X_i - \mu_{ij}\right)^2}{2\sigma_{ij}^2}\right) \quad (5)$$

Where:
: Opportunity
: The attribute
: The attribute value
: Class
: The sub class
: The mean of all attributes
: Standard deviation

2.2. Using Naive Bayes Algorithm

Naive Bayes classifier makes statistical estimation and is based on Bayes Theorem.
A and B are random numbers;

P (A | B) = P (B | A) P (A) / P(B) (6)

P (A): The independent probability of event A is the primitive probability.
P (B): Independent probability of event B.
P(A|B): Probability of B event (conditional probability) when it is known that A event occurs.
P(A|B): Probability of event A (aftershock probability) when event B is known.

The Naive Bayes concept can be explained as follows: X is considered an instance of unknown class membership. Example X = {x1, x2, …, xn} consists of attribute values. In this example class, it is assumed to be n class. C1, C2, …, Cn class values are accepted. The following possibilities are calculated for the sample that will determine the class.

P (〖X|C〗 _i) = (P(X | C_i) P (C_i))/(P(X)) (7)

Simplification is made for the probability of P (X | Ci) to reduce the processing load in the calculation. Assuming that the Xi values of the sample are independent of each other, the following relation is used.

P (X|C_i) = ∏_(k=1)^n 〖P(X_k | C_i) 〗 (8)

In order to classify the unknown example X, it is sufficient to compare only the numerator values since the denominators in P (Ci | X) are equal to each other. The class of the unknown instance is considered to be the same as the class of the largest of these values.

(_C_i^argmax) {P (X|C_i) P (C_i)}

The above equation, post-probabilities, is also known as the largest post-classification method (MAP).

C_MAP=(_C_i^ argma {(k=1)^n 〖P(X_k | C_i) 〗 } (9)

2.3. Confusion matrix

Confusion matrix is used to portion the presentation of a classification algorithm. The terminology connected to the confusion matrix can be rather confusing, but the matrix itself is simple to understand in Table 1.

Table 1. Confusion Matrix

Actual True Positive (TP)	False Positive (FP)
False Negative (FN)	True Negative (TN)

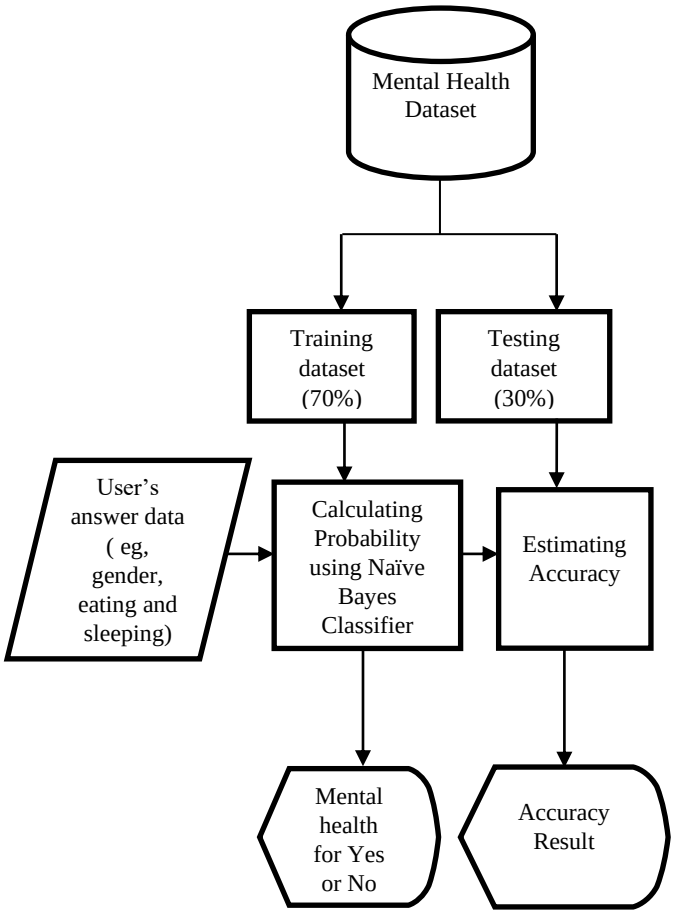


Figure 2. System Flow Diagram

In table 1, 'True' or 'False' indicates if the class is properly predicted or not, while 'Positive' or 'Negative' indicates the prediction of the class of people causes mental illness or not. From the confusion matrix, accuracy, precision, and recall. Where the formula of each of these things are:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN} \quad (10)$$

2.4. System Flow Diagram

The flow diagram of the proposed system is shown in Figure 2. The proposed system is implemented to classify where the patients or users causes psychological illness or not using Naïve Bayes classifier. Mental health dataset 159 instances are used for this system from Kaggle.com. There are 28 attributes, Prediction status (Yes) or (No) are defined as Class.

While the system is started, the 159 users' answers for 28 questions assigned attributes for the system are collected as dataset. Dataset is divided into two datasets: training (70%) and testing (30%).

70% training dataset is calculated to get probability result by using Naïve Bayes algorithm. And then the user is classified where he or she causes mental illness problems or not according to these probability result.

And the performance of the system is evaluated with accuracy result, precision and recall.

Finally, the system displays the user causes mental illness or not and evaluation results.

3. Implementation of the proposed system

3.1. Data Source

Mental health dataset: 159 instances are used for this system from Kaggle.com. 159 instances

of Mental Health are described as dataset, 28 attributes: questions, and Prediction_status: Class (Yes) or (No) are described in Table 2.

Table 2. Questions and Answers as Attributes and Their Values

Attributes Name	Attributes Value
1.Gender	Female, Male
2.above 30 years?	Yes, No
3.Employment	Student, Unemployed, Employed
4.today feeling?	Fine, Good, Sad, Depressed
5.Eating and Sleeping	Yes, No, Maybe
6.(If sad) for the past few days	Yes, No, Maybe
7.Is your sadness for a long time?	For some time, Significant time, Not sad, Long time
8.At what time of the day low?	Evening, Morning, Afternoon
9.sudden and huge change in your life?	Yes, No, Not sure
10.Your stress?	Personal, , None, Financial, Work
11. little pleasure or interest in the activities usually enjoy?	Very Often, Sometimes, Never, Often
12.confident you feel in your capabilities recently.	1,2,3,4,5
13.supported you feel by others around your friends, family.	Highly supportive, Little bit, Satisfactory, Not at all
14. frequently have you been doing things?	Very Often, Sometimes, Never, Often
15.If you have a mental health condition, do you feel that it interferes with your work?	Yes, No, Maybe
16.How easy is it for you to take medical leave for a mental health condition?	Not so easy, Very easy, Difficult, Easy
17.you use substance abuse (e.g. smoking, alcohol)?	Never, Often, Sometimes, Very Often
18.you take medication in the near past for mental health?	Yes, No, Maybe
19.Having trouble concentrating on things?	Yes, No, Maybe
20.Do you feel bad about yourself or your family down?	Yes, No, Maybe
21.hours you spend per day on watching mobile phone?	1-2 hours, 5-10 hours, More than 10 hours, 2-5 hours
22.appointment with a psychologist for your current mental state?	1, 2, 3, 4, 5
23.COVID-19 pandemic affected your mental ?	Yes, No, Not sure
24.How often do you get offended or angry or start crying?	Never, Sometimes, Often, Very often
25.feel yourself vulnerable or lonely?	1, 2, 3, 4, 5
26.comfortable about your mental health?	1, 2, 3, 4, 5

4. Performance Analysis of Proposed System

Table 3. Accuracy, Precision, and Recall

	Accuracy (%)	Precision (%)		Recall (%)	
Naïve Bayesian	95.21%	Yes	No	Yes	No
		0.981	0.973	0.988	0.961

The accuracy, precision, and recall of implementation results are shown in Table 3. The precision and recall for Class1 (Yes) are 0.981 and 0.988. And the precision and recall for Class (No) are 0.973 and 0.961. According to these results, the exactness and completeness of classification of mental illness system is over 95%. So, this system gives good performance.

4.1. Advantages

The proposed system serves user-friendly for diagnosing the health problem and users or patients to classify where the mental illness causes or not depending upon their conditions, accurate results to the users and efficient and effective decision as psychiatrists' decision.

5. Conclusion

This paper presented the Classification of Psychological Illness Using Naïve Bayesian Classifier. Dataset:159 instances with 28 different attributes of patients are analyzed and classified that the patient causes psychological illness or not. The performance of the system is evaluated with the result of the system in terms of accuracy, precision and recall. The Naïve Bayes classifier to be very efficient and effective in psychological illness.

References

[1] Abiyoga, Arya Wicaksana and Ni Made Satvika Iswari, "Decision Support System for Choosing an Elective Course Using Naive Bayes Classifier", pp. 97-110, 2020. https://www.researchgate.net/publication/335366488_Decision_Support_System_for_Choosing_an_Elective_Course_Using_Naive_Bayes_Classifier

[2] Aditiya Hermawan, "Implementation of Naïve Bayes Algorithm for Classification of Mental Health of Social Media Users", Vol.4, No.2, December 2021. https://jurnal.kdi.or.id/index.php/bt/article/view/282

[3] E. Chandra Blessie , Bindu George, "A Survey on Data Mining Algorithms in Prediction of Psychiatric Disorders", International Journal of Science and Research (IJSR), pp. 783-786. https://www.ijsr.net/get_count.php?paper_id=SR21817180245

[4] Mrs.G.Subbalakshmi (M.Tech), Mr. K. Ramesh, Mr. M. Chinna Rao, "Decision Support in Heart Disease Prediction System using Naive Bayes", Indian Journal of Computer Science and Engineering (IJCSE), Vol. 2 No. 2 Apr-May 2011, pp. 170-176. http://ijcse.com/docs/IJCSE11-02-02-56.pdf

[5] Mubarik Ahmad, Vitri Tundjungsari, Dini Widianti, Peny Amalia, Ummi Azizah Rachmawati, "Diagnostic Decision Support System of Chronic Kidney Disease Using Support Vector Machine", 2017. https://www.researchgate.net/publication/322945479_Diagnostic_decision_support_system_of_chronic_kidney_disease_using_support_vector_machine

[6] Nilda N. dela Cruz1, Ricardo Q. Camungao," Decision Support System for Predicting Cardiovascular Diseases Using Naïve Bayesian Algorithm", nced Trends in Computer Science and Engineering, 9(3), May – June 2020, 3178 – 3183 3178 Figure 1: Conceptual Framework of the Research ISSN 2278-3091 Volume 9, No.3, May - June 2020 International Journal of Advanced Trends in Computer Science and Engineering, pp. 3178-3183. https://www.academia.edu/43610885/Decision_Support_System_for_Predicting_Cardiovascular_Diseases_Using_Na%C3%AFve_Bayesian_Algorithm?from_sitemaps=true&version=2

[7] Ojaswi Borate, Snehal Nibe, Sayali Manikwar, Shivani Gund, "Psychological Illnesses Revealing Method Using Machine Learning", 2020 JETIR May 2020, Volume 7, Issue 5, pp. 48-52. https://www.jetir.org/view?paper=JETIR2005310

[8] Sakshi Kapoor, Rabina Verma, Surya Narayan Panda, "Detecting Kidney Disease using Naïve Bayes and Decision Tree in Machine Learning",International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-9 Issue-1, November 2019, pp. 498-501. https://www.ijitee.org/wp-content/uploads/papers/v9i1/A4377119119.pdf

Improving the Accuracy of CNN by Applying Random Sampling Methods on NSL-KDD Dataset

Aye Thawta Sann, Zin Thu Thu Myint
University of Computer Studies, Yangon
whiteboard.star@gmail.com

Abstract

Nowadays, an Intrusion Detection System (IDS) performs the monitoring and detection of various security threats in hardware and software on the network. Although there have many existing IDS but still, we face challenges in improving accuracy in detecting security vulnerabilities, not enough methods to reduce the level of alertness and detecting intrusion attacks. Machine learning methods can detect data from past experience and differentiate normal and abnormal data. In this system, the Convolutional Neural Network (CNN) in deep learning method is used for solving the problem of identifying intrusion in a network. NSL – KDD dataset is used to train the data with the CNN algorithm. The system implementation is performed for balanced and unbalanced nature of NSL – KDD dataset. This system achieves accuracy of 80% in unbalanced dataset and 83% in balanced dataset.

Keywords: Intrusion Detection System; Convolutional Neural Network; Deep Learning;

1. Introduction

Cyber security is a detection and prevention method for malicious attacks in networks and computer systems such as firewalls, anti-virus software, etc. The development of intrusion detection system is performed for the improvement in network security. The intrusion detection system is widely utilized in monitoring, analysis, and detection of various attacks in hardware and software on networks and computer systems. For the forecasting model development to the prevention of attacks, learning models for intrusion system are utilized to divide the connection as “bad” for abnormal and “good” for normal. The intrusion detection system is critical for disrupting network security as the rapid development of internet. The intrusion detection system was introduced for solid protection of the equipment among dangerous software attacks, such as denial of service (DoS), with the patterns analysis of data collection [8]. This system has the capability of attacks detection; the prevention or denial of illegitimate traffic is performed as the network traffic acts as denial of service (DoS). The intrusion detection system is the procedure for solving the classification issue. One of the challenges in some existing intrusion detection system is that their detection accuracy is low. Another challenge is that the depending on the signatures of known attacks, which means they have inability for the detection of malicious attacks.

An intrusion detection system is most powerful and robust thing for the detection of malicious attempts in manipulation, accessing, or disrupting the computer system across the network traffic. The monitoring of outgoing and incoming network traffic is taken for the detection in unauthorized and dangerous activities that disrupts the security of system. Therefore, the intrusion detection system takes as a critical thing to network administrators in the conditions of deviceless as the impossibility of the inspection in the huge amount of data travelling the network every second. The characterization of the intrusion detection system is anomaly intrusion detection system and a misuse intrusion detection system [2]. In misuse detection system, the comparison of the user’s activity and known attack signatures is done, and if there is a match the respective action is known as an attack where in the anomaly detection system, any variation from normal behavior is considered as an attack.

The new kinds of attack detection cannot be solved by anomaly intrusion detection system. In such conditions, the detection rate of anomaly intrusion detection system is not greater than misuse intrusion detection system. In addition, the system evaluation of intrusion detection

system must require the enough amount of intrusion detection data so that the system deployment can be performed after the system performance analysis. Therefore, the utilization of model testing and training must be taken by many researchers. The currently most common datasets are Botnet, UNSW-15, NSL-KDD, and KDD-99. The misjudgment, lack of real time response and false detection are the important issues of intrusion detection system.

For handling these issues, traditional machine learning approaches have been applied for the classification of various attack types. But many traditional machine learning approaches are not learning approaches in deeply that focus on the feature engineering and extraction. In addition, these approaches cannot provide the effective result to huge volume of intrusion data prediction problems became by massive volume of the traffic in network application. Traditional machine learning approaches are not suitable in forecasting and analysis in the conditions of the requirement of high dimensions data learning by the huge amount data. However, deep learning outperforms than traditional machine learning approaches in the extraction of good representations in better models' creation. Therefore, various researchers are trying in the introduction of intrusion detection system by deep learning.

In this paper, the Convolutional Neural Network (CNN) in deep learning method is used for solving the problem of identifying intrusion in a network. In this proposed work, Convolution Neural Network (CNN) is used as a learning model for classification in IDS. NSL – KDD dataset is used to train the data with the CNN algorithm. The system implementation is performed for balanced and unbalanced nature of NSL – KDD dataset. In this paper, the next section 2 describes about the related works of the proposed system and then section 3 presents about the background theory of the proposed system. In section 4, the proposed system is presented and the experimental results of the proposed system are discussed in section 5. Finally, the system is concluded in section 6.

2. Related Works

In this paper [1], an improved intrusion detection system according to hybrid feature selection and the ensemble of two-step classification proposed. This feature selection approach contains: ant colony, genetic, and particle swarm optimization methods. These methods are applied for the reducing of feature in training events. In this system, UNSW-NB15, and NSL-KDD are used as input data sources. The selection of features is done according to the efficiency of reduced error pruning tree. After, the ensemble of two-step classification: bagging and rotation forest is performed. This system achieved the sensitivity of 86.8%, the detection rate of 88%, and the accuracy of 85.8% in the NSL-KDD dataset, and the state of the art of accuracy, detection rate, and sensitivity in the UNSW-NB15 dataset. However, this paper did not consider effective classification for balanced dataset in intrusion detection system.

The authors introduced a comparative analysis of the intrusion detection system efficiency with random forest in terms of false alarm rate and accuracy [7]. In this system, UNSW-NB15, GPRS, and NSL-KDD are used as input data sources for the implementation. The consideration in the ensemble of various tree types are done whereas another most optimum parameters are provided by applying grid. The system implementation evaluated that random forest is the best outperforming for intrusion detection system as the significant performance of this classifier in terms of k-cross validation with the other ensembles of naïve bayes and neural network, and naïve bayes and random forest. However, this paper did not consider unbalanced data nature for achieving desired accuracy in intrusion detection system.

The authors proposed the deep learning approach, nonsymmetric deep autoencoder for unsupervised feature learning in intrusion detection system [6]. This system is built based on random forest method and stacked nonsymmetric deep autoencoders. This system utilized NSL-KDD and KDD Cup '99 as input data sources for the performance evaluation. The implementation of this system is done with TensorFlow on graphics processing unit (GPU). This system got the promising accuracy for intrusion detection system. The system implementation evaluated that this approach achieved the higher precision, recall, and accuracy and reduced the time of training. The comparison of mainstream DBN method and

stacked nonsymmetric deep autoencoder was performed. The comparison results showed that this system improved the accuracy by 5% and reduced the training time by 98.81%.

The authors proposed the combination of convolutional neural network with the TensorFlow like cognitive computing method in intrusion detection system [4]. NSL_KDD dataset is used as input data source for this system. The presentation of network traffic according to the connections of TCP/IP is performed and the training of this technique is done with the signatures of known attack. The performance evaluation showed that this system achieved the promising precision by 99.82%, the F1-score by 96.34%, the accuracy by 98.92%, and the recall by 92.34%. This approach performed the integration of existing system as big data and TensorFlow by providing scalability for huge amount of data.

This paper presented an approach for intrusion detection system according to temporal convolutional neural networks, establishing the best detection that has the ability to solve huge amounts and high-dimensional data [5]. Moreover, this system can be applied for not only host-based intrusion detection systems but also network-based intrusion detection systems. The experimental results proved that the proposed approach performed better than other detection techniques with a lower false positive rate and higher accuracy. This proposed approach promised 90.5% accuracy for NSL-KDD dataset while requiring 543KB storage. Furthermore, the proposed system provided data preprocessing decisions for the systems that require complexity reduction and efficiency improvement.

3. Background Theory

An intrusion detection system is a network traffic monitoring system for detecting mistrustful activities and it gives an alarm if the mistrustful activities are found. It is a software application for analysis and monitoring the network system for the detection of harmful activity and brokerage of policy prior to the serious damage in network and the corruption of data assets. The intrusion detection system was developed in 1986 by Dorothy E. Denning. The aim of the intrusion detection system is in order to discover various types of security violations outside the system brokerage and inside the hateful system features and prevalence of data misuse. The rule-based feature matching strategy including normal action features in safe library is utilized. The comparison of this strategy with audited usage features for alerting any abnormal actions. The intrusion detection system can perform the detection various intrusions such as trojan horses, misuse of legitimate users, impersonating attempts, and viruses.

3.1. Random Sampling

Random sampling is one kind of probability sampling in that every instance possesses the equality in probability for the selection. A random selected instance is an unbiased description for the total population. If the instance does not represent the population, the variation becomes a sampling error. Random sampling is a method for choosing each participant or a subset of the population for providing the statistical inferences from them and estimation the characteristics of the whole population. It can be utilized as a data reduction method as it allows a large data set to be represented by a much smaller random data instance (or subset).

3.2. Undersampling

Undersampling is a balancing approach for asymmetric datasets with maintaining all of the data in the minority class and reducing the size of the majority class. In another words, the deletion of samples from the majority class are performed and this sampling can occur to the invaluable information loss in the model. Majority classes are classes which provides the larger proportion of the dataset [3]. Minority classes are classes which provides the smaller proportion of the dataset. Assume that D is a large data set, and that contains the number of instances, N. Simple random sample without replacement (SRSWOR) of size s: This creation is done by deleting s from the N instances at D ($s < N$), in where the deletion probability of any instance in D is 1=N, i.e., all instances are equally likely to be sampled. This sampling is appropriate in such conditions as there is plenty of data for an accurate analysis. All rare instances are utilized however the number of abundant instances is reduced for the creation of two equally sized classes.

3.3. Oversampling

Over-sampling is a balancing approach for asymmetric datasets with maintaining all of the data in the majority class and increasing the size of the minority class by adding the instances to it. In another words, the duplication of samples to the minority class in the training dataset and this sampling can occur overfitting to some models as learning algorithms focus on replication of minority instances. Simple random sample with replacement (SRSWR) of size s: This is similar to SRSWOR, except that each time an instance is chosen from D, it is recorded and then replaced, i.e., after an instance is drawn, the back placement is performed at D so that the choosing may be done again. This sampling is appropriate in such conditions as there is no enough information. One class is the majority, or abundant, and the other class is the minority, or rare. The number of rare instances is increased in this sampling.

3.4. Convolutional Neural Network

A convolutional neural network (CNN) is a type of deep, feedforward neural networks. It applies the multilayer perceptron and it has been developed for the reduction of processing needs [9]. It can be applied for the collection of spatial data or sequential data and it is widely utilized in speech recognition, time series analysis and image processing, etc. The architecture is shown in Figure 1.

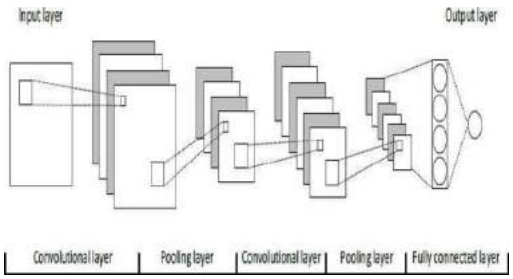


Figure 1. Convolutional Neural Network

It consists of an input layer, an output layer and a hidden layer which contains many pooling layers, convolutional layers, normalization layers, and fully connected layers.

Input layer: Text documents or images are kept by the representation of vector.

Convolutional layer: The output is decided with the calculation of dot products between set of weights at input layer. The aim of this layer is the observation of features. Every convolutional layer is specified by various parameters containing kernel size, zero padding, stride, input size, and the map stack. Then, the calculation of input signal is performed with an activation function, Rectified Linear Units (ReLUs) using equation 1.

$$f(x) = \max(0, x) \quad (1)$$

This activation function is used on the input data; therefore, the dimensions of the input and the output are same.

Pooling Layer: This layer acts as a mediator between many convolutional layers. It performs the reduction of the spatial dimensions of the input data. This is similar to the previous convolutional layer as this layer sweeps the filtering among all input data however this filtering does not possess any weights. Two kinds of pooling operations are maximum pooling, and average pooling. Maximum pooling takes the selection of the pixel by the maximum value for sending to the output array when the movement of filter among the input. Average pooling performs the computation of the average value in the receptive field by sending to the output array when the movement of filter among the input. This layer provides many benefits to convolutional neural network whereas there is the information loss at this layer. The reduction of noise features, the efficiency improvement, and the overfitting prevention are provided by this layer.

Fully-Connected Layer: This layer is same with the output layer of multilayer perceptron (MLP). The aggregation of information from the final feature maps and the generation of final classification are performed. The fully connection of all neurons with all neurons in the previous layer is taken. The reduction of data dimension at pooling layer to a single dimension and the connection with every neuron are taken. The classification is performed with activation function, SoftMax using equation 2.

$$\sigma(\vec{z})i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \quad (2)$$

Where, $\vec{z}$ is the input vector, z_i is the elements of the input vector, e^{z_i} is the standard exponential function, and K is the number of classes in the multi-class classifier. The SoftMax function does the testing of result by using training data. Finally, this provides the result which the input image to which the relating class.

4. Proposed System

The aim of this system is to develop the intrusion detection system using convolutional neural network model on unbalanced dataset and balanced dataset. In this system, the dataset for network intrusion detection, NSL-KDD data set, is employed. The data set consists of 42 features, 41 features grouped into four categories, such as essential features, content features, time-based, and host-based features. The last feature is about all the data of other features. Figure 2 shows different features in dataset. The system architecture is described in Figure 3. Features under various categories is shown in Table 1.

F#	Feature name	F#	Feature name	F#	Feature name
F1	Duration	F15	Su attempted	F29	Same srv rate
F2	Protocol type	F16	Num root	F30	Diff srv rate
F3	Service	F17	Num file creations	F31	Srv diff host rate
F4	Flag	F18	Num shells	F32	Dst host count
F5	Source bytes	F19	Num access files	F33	Dst host srv count
F6	Destination bytes	F20	Num outbound cmds	F34	Dst host same srv rate
F7	Land	F21	Is host login	F35	Dst host diff srv rate
F8	Wrong fragment	F22	Is guest login	F36	Dst host same src port rate
F9	Urgent	F23	Count	F37	Dst host srv diff host rate
F10	Hot	F24	Srv count	F38	Dst host serror rate
F11	Number failed logins	F25	Serror rate	F39	Dst host srv serror rate
F12	Logged in	F26	Srv serror rate	F40	Dst host rerror rate
F13	Num compromised	F27	Rerror rate	F41	Dst host srv rerror rate
F14	Root shell	F28	Srv rerror rate	F42	Class label

Figure 2. Different Features in Dataset

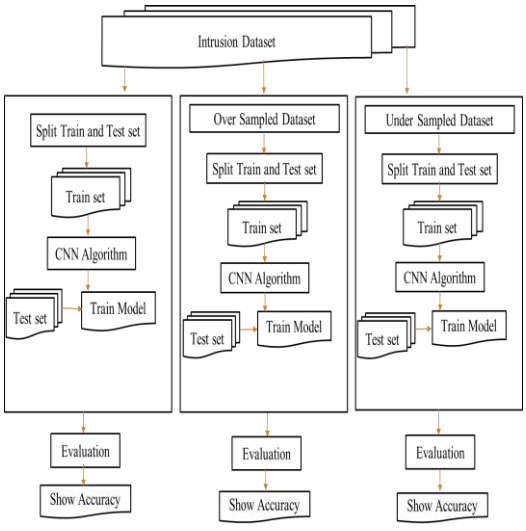


Figure 3. System Architecture

Table 1. Features under various categories

Category	Features
Basic Features	Feature 1 to Feature 10 all
Content Features	Feature 11 to Feature22 all
Time-based Features	Feature 23 to Feature 31 all
Host-based Features	Feature 32 to Feature 41all

The NSL-KDD dataset contains 1 training set and 2 testing sets:

- KDDTrain+: The full NSL-KDD train set including attack-type labels
- KDDTest+: The full NSL-KDD test set including attack-type labels
- KDDTest-21: A subset of the KDDTest+ which does not include records with difficulty level of 21 out of 21

This system applies 1 training set and 1 testing test: KDDTrain+ and KDDTest+. The transformation of NSL-KDD dataset to the 1-dimensional convolution architecture is performed. Moreover, this dataset includes non-numeric and numeric features. As the training input and testing input feeding to the convolutional neural network is in the form of numeric matrix, the conversion to numeric attribute must be performed. In addition, the one-hot encoder is utilized for the conversion of category features in the dataset into numeric matrix as the usage of one-hot encoder can handle the issue in the category conversion to integer. In this system, 32 kernels with 1*3 dimension and 5 convolutional layers are used. For each convolutional layer, maximum pooling and rectified linear unit (ReLU) are used and the pooling size is 4 for only first pooling layer and is 2 for other pooling layers. SoftMax function is used at fully-connected layer. Accuracy, precision, recall, and f-measure are the key metrics of performance evaluation of the proposed system.

5. Performance Evaluation

In this system, NSL-KDD dataset is used. From this dataset, 1 training set and 1 testing test: KDDTrain+ and KDDTest+ are utilized. KDDTrain+ dataset contains 125,973 network traffic samples and KDDTest+ has 22,554 network traffic samples. This system model is

trained in Keras which is based on Tensorflow. This system is firstly tested with unbalanced nature of original dataset. Then the system is tested with balanced nature by applying undersampling and oversampling. The key metrics of performance measures (accuracy, recall, f-measure, and precision) are evaluated for this proposed system analysis. The accuracy is computed using equation 3, the precision is calculated with equation 4, the computation of recall is done with equation 5, and the f-measure is computed according to equation 6.

Figure 3 shows the comparative results for the performance of proposed model for Abnormal type on unbalanced, balanced by oversampling and balanced by undersampling. Figure 4 describes the comparative results for the performance of proposed model for Normal type on unbalanced, balanced by oversampling and balanced by undersampling. The comparative results for the performance of proposed model on unbalanced, balanced by oversampling and balanced by undersampling are illustrated in Figure 5.

$$\text{Accuracy (\%)} = \frac{\text{Exactly predicted sample}}{\text{Total number of samples}} \times 100 \quad (3)$$

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive + False Positive}} \quad (4)$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive + False Negative}} \quad (5)$$

$$\text{F} - \text{measure} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision + Recall}} \quad (6)$$

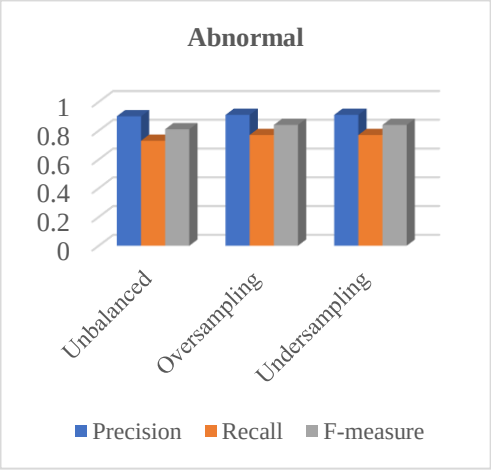


Figure 3. Performance Comparison of Unbalanced and Balanced on Abnormal

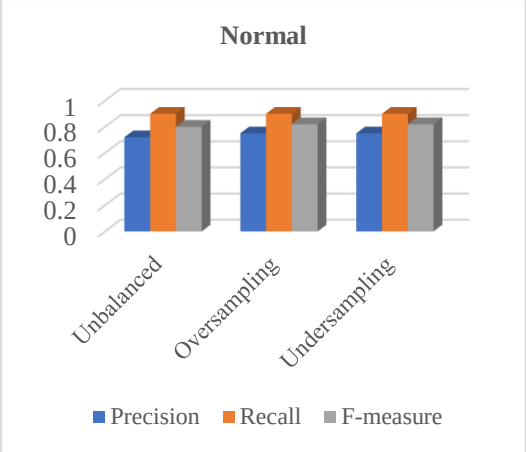


Figure 4. Performance Comparison of Unbalanced and Balanced on Normal

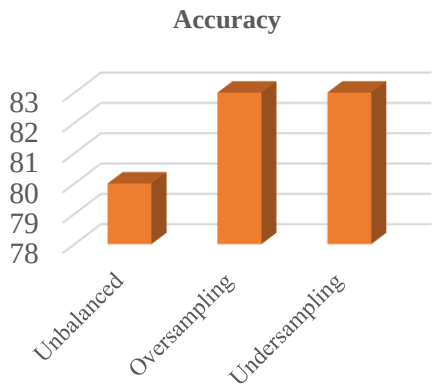


Figure 5. Performance comparisons of Unbalanced and Balanced

According to the evaluation results, the improvement of accuracy of CNN is achieved by applying random sampling techniques: oversampling and undersampling.

6. Conclusion

The proposed system is in order to achieve the improvement in intrusion detection effectiveness as the development of most existing intrusion detection system with the machine learning approaches did not support for the prevention by newly formed attacks using last data. Therefore, convolutional neural network deep learning model is used for developing the intrusion detection system. We implemented a deep learning model to train the model with NSL-KDD data set, namely convolutional neural network. This system achieves accuracy of 80% in unbalanced dataset and 83% in balanced dataset.

References

[1] B. A. Tama, M. Comuzzi, and K. H. Rhee, "TSE-IDS: A Two-Stage Classifier Ensemble for Intelligent Anomaly-based Intrusion Detection System", IEEE Access, (Volume: 7), IEEE, 11 July, 2019, pp. 94497 – 94507.

[2] I. Abrar, Z. Ayub, F. Masoodi, A. M. Bamhdi, "A Machine Learning Approach for Intrusion Detection System on NSL-KDD Dataset", Proceedings of the International Conference on Smart Electronics and Communication (ICOSEC 2020), IEEE, Trichy, India, 10-12 September, 2020, pp. 919-924.

[3] J. Han, M. Kamber, and J. Pei, "Data Mining: Concepts and Techniques (The Morgan Kaufmann Series in Data Management Systems) 3rd Edition", Elsevier Science Ltd, USA, 22 June, 2011, pp. 1-703.

[4] L. Heng, and T. Weise, "Intrusion Detection System Using Convolutional Neuronal Networks: A Cognitive Computing Approach for Anomaly Detection based on Deep Learning", 2019 IEEE 18th International Conference on Cognitive Informatics & Cognitive Computing (ICCI*CC), IEEE, Milan, Italy, 23-25 July, 2019, pp. 34-40.

[5] N. Fu, N. Kamili, Y. Huang, and J. Shi, "A Novel Deep Intrusion Detection Model Based on a Convolutional Neural Network", 26th International Conference, ICONIP 2019, Springer, Sydney, NSW, Australia, December 12–15, 2019, pp. 52-59.

[6] N. Shone, T. N. Ngoc, V. D. Phai, and Q. Shi, "A Deep Learning Approach to Network Intrusion Detection", IEEE Transactions on Emerging Topics in Computational Intelligence, Volume: 2, Issue: 1, IEEE, February, 2018, pp. 41-50.

[7] R. Primartha, and B. A. Tama, "Anomaly Detection using Random Forest: A Performance Revisited", 2017 International Conference on Data and Software Engineering (ICoDSE), IEEE, Palembang, Indonesia, 01-02 November, 2017, pp. 12-17.

[8] Y. Ding, and Y. Zhai, "Intrusion Detection System for NSL-KDD Dataset Using Convolutional Neural Networks", 2018 2nd International Conference on Computer Science and Artificial Intelligence (CSAI 2018), Association for Computing Machinery, New York, NY, United States, December, 2018, pp. 81-85.

[9] Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning", vol. 521, Nature, London, 27 May, 2015, pp. 436-444.

[10] (2018, Oct.) KDD Cup 1999. [Online]. Available: http://kdd.ics.uci.edu/databases/kddcup99/.

Strengthening Malaria Diagnosis and Treatment using CART and Rule-based Classification

Mya Myintzu, Dr Yu Mon Zaw
University of Computer Studies, Yangon
myamyintzu@ucsy.edu.mm, yumonzaw@ucsy.edu.mm

Abstract

Malaria is a life-threatening disease caused by parasites that are transmitted to people through the bites of infected mosquitoes. The most important malaria activity is the early diagnosis and prompt treatment of the disease. Malaria diagnosis and treatment system helps to identify the symptoms of each patient and treatment given. To strengthen Malaria diagnosis, Classification and Regression Tree (CART) algorithm is applied and monitoring treatment processes is carried out with Rule-based algorithm. In this proposed system, annual dataset of 2,910 malaria patients from Paletwa Township, Chin State of 2017 are used for diagnosis of malaria and national malaria treatment dataset will be used to give correct treatment. Dataset from Kaggle is learnt to understand the attributes selection. Classification with CART algorithm makes the accurate diagnosis and better to follow up.

Keywords: CART, malaria, classification, data mining, malaria treatment

1. Introduction

Nowadays, the systems for management of data, information and knowledge are offering new potential for improvement of different sectors. Data mining algorithms are mainly distinguished as descriptive or predictive. Classification is the data analysis task, where a model or classifier is constructed to classify the category class labels. Classification and regression trees (CART) algorithm is a model of binary partitioning decision tress which can evaluate relationships between different types of data values of malaria diagnosis. It is useful to find the independent variable that creates the best similar group when splitting the data element such as the symptoms of malaria.

In this paper, CART algorithm was choosing to demonstrate the malaria diagnosis cases. The system uses patient status as class label, and this class label has two classes: positive and negative. To classify for malaria infection, sixteen significant symptoms are chosen for every patient. The patient records will be stored in database. If the patient status results with positive, monitoring treatment case will be carried out. Rule-based classifier was using to determine the treatment which is based on the age group of each malaria patient. The treatment system follows the National Malaria Treatment Guideline which is co- published by Ministry of Health and World Health Organization.

The rest of this paper is organized with five sections. Related work of the system is presented at section 2, theoretical field of paper is described at section 3. Section 4 includes the proposed system and implementation of approach. The conclusion will be remark at section 5.

2. Related Work

The authors of [10] released a result of Decision Support System for Mosquito Borne Disease. They introduced decision support system and has advanced clinical support for Mosquito borne disease diagnosis. The authors of [16]**Error! Reference source not found.** r eleased Diagnosis of Malaria by Using Reduct Generation Algorithm. In their paper, they mentioned the data classification model from available training data set and classifying objects according to their attributes. They also developed the diagnosis of malaria patient's symptoms by using reduct generation algorithm under the set theory. Finally, the authors of [20] presented Diagnosis of TB Disease by Using Decision Tree Induction. In their system, decision tree inductive method was used to determine the appropriate TB disease classification according to decision tree

rules. This classifier classifies TB diseases based on the symptoms of each patient.

3. Background Theory

3.1. Data Mining

Data Mining is used to turn raw data into useful information. By using software to find patterns in large amounts of data, businesses and organization can develop more effective strategies [1]. Data mining involves exploring large amounts of information and studying meaningful patterns and trends. It can analyze the relationships and patterns of data, based on user requests.

Data mining is gaining popularity in different research fields due to its boundless applications and approaches to mine the data in an appropriate method. It is one of the approaches for approximating the nearby future consequences [11]. In advanced research. Data mining have a great potential to enable healthcare systems to use data more efficiently and effectively [17].

3.2. Classification

Classification is the process of classifying a record. It could be more than two classes to classify. Another process of data analysis is prediction [22]. It is used to find a numerical output. Same as in classification, the training dataset contains the inputs and matching numerical outputs. The algorithm builds the model or a forecaster according to the training dataset. This model should find a numerical output when the new data will be entered. Unlike in classification, this method has no class label. The model predicts a continuous-valued function or ordered value. Commonly, regression is used for prediction. In this paper, CART algorithm is used to develop decision tree.

3.2.1. Process of Classification

Classification is the process of classifying a record. A simple example of classification is to check whether it is raining or not. The answer can be yes or no. Sometimes there may be more than two classes to classify the options. Another data analysis process is forecasting. It is used to search for a digital output. As with classification, the training data set contains input data and corresponding numeric output values. The algorithm generates a model, or a predictor based on the training data set [21]. The model must find the digital output when new data is provided. Unlike sort, this method has no class tag. The model predicts a continuous function with values or an ordered value. Regression is commonly used for forecasting.

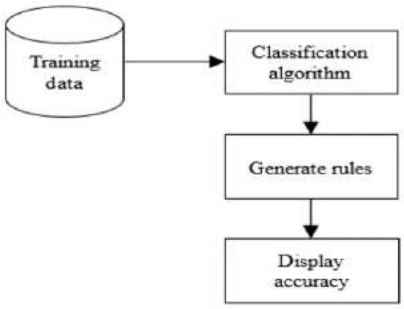


Figure 1. Process of Classification

3.3. Decision Tree

Decision tree algorithms were originally intended for classification. Decision tree induction constructs a flowchart like structure where each internal node denotes a test on an attribute, each branch corresponds to an outcome of the test, and each external node denotes a class prediction. At each node, the algorithm chooses the best attribute to partition the data into individual classes. When decision tree induction is used for attribute subset selection, a tree is constructed from the given data. All attributes that do not appear in the tree are assumed to be irrelevant. The set of attributes appearing in the tree form the reduced subset of attributes [11].

3.3.1. CART Algorithm

Decision tree can be used for classification or regression predictive modeling problems. CART is a machine learning method. It is a data mining procedure to present the results of a complex data set in the form of decision tree. Decision trees are then used to classify new data. The CART algorithm provides a foundation for important algorithms like bagged decision trees, random forest and boosted decision trees [8]. This algorithm can be used for both classification and regression. It starts with the training set as a root node.

CART can produce only binary tree and handle mission values automatically by using

surrogate/ substitute splits. CART uses Gini index impurity function for train data. Missing values in variables can be estimated by using surrogate variables so that practical data can be used whenever possible within the tree.

3.3.2. Selection of Attributes

The Gini index is used in CART algorithm. It considers a binary split for each attribute. The attribute that maximizes the reduction in impurity or the minimum Gini index is selected as the splitting attribute [12].

CART use Gini index to consider a binary split for each attribute. For discrete-valued attribute such as categorical variables, having *v* distinct values $\{a_1,a_2,\ldots, a_v\}$, there are 2^v possible subsets. For example, if anemia has two possible values, namely {yes,no}, the possible subsets are {yes,no}, {yes}, {no}, {}. Excluding the power set of {yes,no} and the empty set from the consideration do not represent a split. Therefore, the best split on *anemia* is 2^v-2. It was based on the binary split on that attribute.

For the next step of CART algorithm, Gini index is used for measuring attribute selection. The smallest impurity criterion or minimum Gini index value is chosen in this step and CART uses impurity function to select the best attribute. To measure the impurity of D, the data partition or the set of training tuples is as below.

$$Gini(D) = 1 - \sum_{i=1}^{n}(p_i)^2$$

In facts, p_i is the probability that a tuple in *D* belongs to class C_i. The sum is computed over *n* class.

A weighted sum of the impurity of each resulting partition was computed. For each attribute, if a binary split on *A* partitions, *D* will be divided into D_1 and D_2 and the Gini index of *D* given that partitioning is

$$Gini_A(D) = \frac{|D_1|}{|D|}Gini(D_1) + \frac{|D_2|}{|D|}Gini(D_2)$$

The attribute *A* maximizes the reduction in impurity is –

$$\Delta Gini(A) = Gini(D) - Gini_A(D)$$

The attribute that maximizes the reduction in impurity is selected as the splitting attribute.

When the sample data has 23 records and 14 attributes, searching all possible binary splits to each attribute according to its value. The sample dataset is presented at Table 1.

Let *sudden_fever* be a set concisting of *n* data samples. The expected information needed to classify will be given below.

Table 1. Sample dataset of patients' record

sudden fever	headache	bleeding	muscle pain	vomiting	diarrhea	weakness	jaundice	shock	urination loss	dyspnea	convulsion	anemia	hemoglobinuria	Diagnosis
1	1	0	0	1	0	1	0	0	1	1	1	1	0	Malaria
1	0	1	1	1	0	0	0	1	1	0	1	1	1	Malaria
0	1	0	0	1	0	0	0	0	0	0	0	0	0	Chikungunya
1	1	1	1	1	1	0	0	0	0	1	0	0	1	Dengue
1	1	0	1	1	1	0	1	0	0	1	1	1	1	Malaria
1	0	0	0	1	0	1	1	0	0	0	0	0	0	Rift Valley fever
1	0	1	1	0	1	1	0	1	1	1	1	1	1	Malaria
1	1	0	1	1	0	0	1	1	1	1	0	1	1	Malaria
0	0	1	0	1	0	1	0	0	1	1	0	0	0	Yellow Fever
0	0	0	1	0	1	1	1	1	0	1	1	1	1	Malaria
0	1	0	0	1	1	0	1	0	1	1	1	1	1	Malaria
1	1	0	1	1	1	1	0	1	0	0	1	0	1	Malaria
1	1	0	1	1	1	1	1	1	1	1	1	0	1	Malaria
1	0	0	1	1	0	0	0	0	1	1	0	0	0	Zika
1	0	1	0	0	0	1	0	1	1	0	0	1	0	Ill
1	1	0	1	0	0	1	1	1	1	1	1	0	1	Malaria
0	0	0	1	1	0	1	1	1	0	1	1	0	1	Plague
0	1	0	0	0	1	0	0	0	0	0	0	1	0	Tungiasis
0	1	1	0	0	1	1	0	1	0	1	1	1	1	Malaria
1	1	0	1	1	1	0	1	1	0	0	1	1	0	Malaria
1	0	1	1	0	1	1	1	1	1	1	1	1	1	Malaria
1	1	1	0	1	0	0	0	0	0	1	1	1	1	Malaria
1	1	1	1	1	1	0	0	0	0	1	0	0	1	Dengue

Use the equation of Gini(D) to this data and find the purity of given data.

Gini (D) = 1 – (14/23)2 – (9/23)2 = 0.4764

After applying $Gini_A(D)$, to search the Gini index values of each attribute are as follows:

$Gini_{sudden\_fever\{yes,no\}}$ (D) = 16/23 (1 - (11/16)2 – (5/16)2) + 7/23 (1 – (3/7)2 – (4/7)2) = 0.4479

$Gini_{sudden\_fever\{no,yes\}}$ (D) = 7/23 (1 - (3/7)2 – (4/7)2) + 16/23 (1 – (11/16)2 – (5/16)2) = 0.2400

3.3.3. Rule-based Classifier

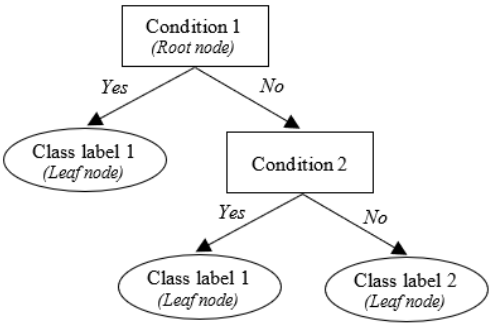


Figure 2. Rule-based data mining classifier

The Rule Based Data Mining Classifier is a technique used for data mining. Rules are a good method of representing information to understand. The effectiveness of a rule-based

classifier depends on features such as the quality of the rules, rule ordering, and properties of the set of rules. It is to discover consistencies and different scenarios in data stated in the IF-THEN rule [4].

4. System Implementation

4.1. Overview of the System

Malaria is a potentially life-threatening disease. At the same time, malaria is preventable and curable. Early diagnosing can benefit to give suitable treatment for decreasing mortality rate of malaria [13]. Users of this system can identify the malaria infected patient without receiving laboratory results of malaria. In this system, users can keep patient records well-organized.

In this system, the user needs to be collected training and testing data first. Each record of patients can be stored in the patient database. In the next step, the decision tree could be developed after calculating the impurity functions with CART algorithm. As a result, the rules will be generated that derived from the decision tree.

In the process of testing, those data are classified with rules to produce target class values are matched with identified values. The result of malaria positive or negative will be displayed to the user at the final stage.

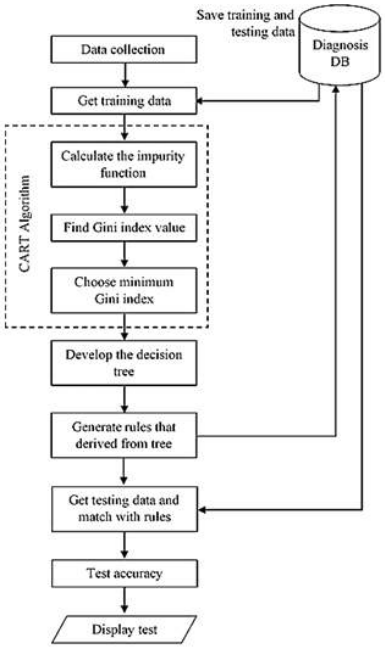


Figure 3: System flow of identifying malaria infected result with CART algorithm

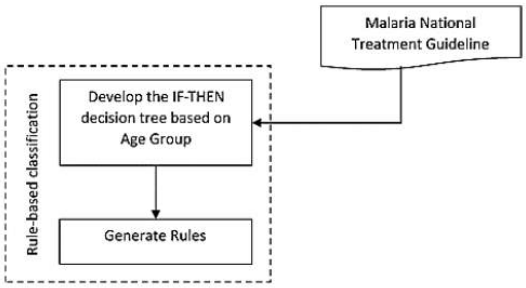


Figure 4: System flow of identifying malaria treatment with Rule-based classifier

According to the presented result, the treatment can be decided with the age group of each patient. This proposed system was also developed IF-THEN decision tree after identifying the malaria infection result of recorded patients. The rule-based classifier was used to generate the suitable treatment of malaria positive patients. The treatment is referencing from National Malaria Treatment Guideline which is co- published by Ministry of Health and World Health Organization and it's still active nowadays.

4.2. Design of the Proposed System

The epidemiology of malaria in Myanmar is highly complex. If the patient cannot get proper treatment, the infection can become severe and may cause kidney failure, seizures, mental confusion, coma, and death.

Table 2. Major Attributes of patients' record

No	*Symptoms*	*Datatype*	*Value*
1	Sudden_fever	Boolean	Yes, No
2	Vomitting	Boolean	Yes, No
3	Weakness	Boolean	cannot sit un-aid, cannot stand un-aid, cannot walk un-aid
4	Anaemia	Boolean	Uncomplicated, Severe
5	Jaundice	Boolean	Yes, No
6	Headache	Boolean	Yes, No
7	Renal_Failure	Boolean	Yes, No
8	Hemoglobinuria	Boolean	Yes, No
9	Dyspnea	Boolean	Yes, No
10	Convulsions	Boolean	Yes, No
11	Bleeding	Boolean	Yes, No
12	Shock	Boolean	Yes, No
13	Pain	Boolean	Yes, No
14	Cough	Boolean	Yes, No
15	Diarrhoea	Boolean	Yes, No
16	Oversweating	Boolean	Yes, No

In this system, the possible symptoms are used as the attributes of each patient. These are discrete values and are described as Table 2. In this implementation, 2,910 records of patient data are imported. The system was developed with PHP and the user interface was designed with HTML and CSS.

4.2.1. System Flow Diagram from User Side

There are two roles of users, administrator as well as user. When starting the application, the log in for will be displayed. Firstly, the user can make data entry to test malaria positive or not according to the significant symptoms. The data entry form is displayed in Figure 5.

Figure 5. Malaria Diagnosis Data Entry Form

After saving the completed entry record, the model box will be displayed to identify the patient was infected malaria diseases. Glucose-6-phosphate dehydrogenase (G6PD) deficiency is relatively common in populations exposed to malaria. This deficiency appears to provide some protection from this infection, but it can also cause the risk after administration of some antimalarial drugs. If the patient is suffering G6PD deficiency, the system alerts to referral the patient for hospitalization.

Every single entry will be recorded in dataset. The system administrators can view all patient. Administrator can also search with desire keywords in the list of patients as well as export with an excel sheet as well as random patients list.

4.3. In Depth Analysis of Generated Rules

4.3.1. Diagnosis Rules

Healthcare provider will survey and ask about the symptoms and malaria history. Depending on the type of parasite, symptoms can be mild in some conditions. Some patients don't feel the signs and symptoms of malaria even they infected. The system provides sixteen symptoms to distinguish malaria positivity results. In facts, the eleven attributes such as sudden fever, vomiting, anemia, jaundice, renal failure, hemoglobinuria, dyspnea, convulsion, shock, muscle pain and over sweating were selected to pre-defined outputs. Among those symptoms, if the patient suffers six symptoms out of those, the diagnosis of his or her could be assumed as positive.

4.3.2. Treatment Rules

The objective of the national antimalarial treatment policy is to provide safe and rapidly effective antimalarial treatment to all patients with malaria and to prevent the emergence and spread of drug resistance [13]. The treatment given is ruled with age group of patients. In the output view, the end user can easily identify the age group by extracting rules from a decision tree.

5. Analysis of System Performance

The performance of system needs to be evaluated by testing with one-third of patient records. Before confirming the accuracy of measuring, the rules of diagnosis will be check whether it is working or not in real time usage. The accuracy of classification is the total number of true predictions separated with the total number of predictions thru dataset. For system performance, accuracy is inappropriate for imbalanced classification problems.

In this system, hold out provides for estimating accuracy. The given data are randomly partitioned into two dependent sets, a training dataset and testing dataset. Commonly, two-third of the data allocated to the training dataset and the remaining one-third is allocated to the test set.

The confusion matrix provides not only the performance of a predictive model, but also which classes are being predicted correctly and which are incorrect as well as which type of errors are being made. The simplest confusion matrix is for a two-class classification problem, with class 0 (negative) and class 1 (positive). In fact, each cell in the table has a specific and well-understood name, summarized as follows:

		Predicted class	
		C_1	C_2
Actual class	C_1	True positives	False negatives
	C_2	False positives	True negatives

Figure 6. Confusion matrix for positive and negative tuples

The confusion matrix is a useful tool for analyzing to recognize tuples of different classes. A confusion matrix for two classes (C1 and C2) is shown in figure 6. Given two classes, in terms of positive tuples, C1 and negative tuples, C2. The calculation of sensitivity and specificity measures is used in accuracy measure.

$$\text{sensitivity} = \frac{\text{t-positive}}{\text{positive}}$$

$$\text{specificity} = \frac{\text{t-negative}}{\text{negative}}$$

where, sensitivity is true positive rate, specificity is true negative rate, t-pos is true Positive, -neg is true negative, pos is positive and neg is negative. The accuracy is a function of sensitivity and specificity and describes as below:

$$\text{accuracy} = \text{sensitivity} * \frac{\text{positive}}{(\text{positive}+\text{negative})} + \text{specificity} * \frac{\text{negative}}{(\text{positive}+\text{negative})}$$

When 917 records are corrected out of 1,000 test records and total positive case will be 798, then

Sensitivity = (917-798)/798 = 0.14
Specificity = (917-119)/119) = 6.70
Accuracy=0.14(798/(798+119))+6.70*(119/(798+119)) = 0.99*

When 689 records are corrected out of 750 test records and total positive case will be 344, then

Sensitivity = (689-344)/344 = 1.00
Specificity = (689-345)/345 = 0.99
Accuracy=1.00(344/(344+345))+0.99*(345/(344+345)) = 0.99*

When 459 records are corrected out of 500 test records and total positive case will be 258, then

Sensitivity = (459-258)/258 = 0.77
Specificity = (459-201)/201) = 1.28
Accuracy=0.77(258/(258+201))+1.28*(201/(258+201)) = 0.99*

When 230 records are corrected out of 250 test records and total positive case will be 172, then

Sensitivity = (230-172)/172 = 0.33
Specificity = (230-58)/58 = 2.96
Accuracy=0.33(172/(172+58))+2.96*(58/(172+58)) = 0.99*

When 91 records are corrected out of 100 test records and total positive case will be 64, then

Sensitivity = (91-64)/64 = 0.42
Specificity = (91-27)/27 = 2.37
Accuracy=0.42(64/(64+27))+2.37*(27/(64+27)) = 0.99*

To clarify the performance of Malaria Diagnosis System, it was tested with different amount of sample data. The accuracy of system classifies increased with the number of sample data amount. The result of the test accuracy is described in Table 3.

Table 3. Result of Classifier Accuracy with Different Amount of Sample Data

No of sample records	No of test data records	No of corrected records	No of failed records	Classifier's accuracy (%)
1000	1000	917	83	99%
1000	750	689	61	99%
1000	500	459	41	99%
1000	250	230	20	99%
1000	100	91	9	99%

6. Conclusion

This system is using the historical records of malaria patients and presenting the prediction of diagnosis by using data mining technique especially CART classification algorithm. This system is not intended to replace the medical experts but to help the experts to identify the diagnosis of malaria patient and to give correct treatments. This system can support some of the basic heath staffs from hard-to-reach area to provide early diagnosis and giving treatment correctly.

This system cannot be trusted completed to diagnosis of malaria. Computerized systems are well-adjusted to do repetitive tasks. These never get tired, bored or fatigued. Still, there can be failures of a computer system due to internal and external reasons. By the reference from Library of Medicine, malaria can hide in people with no symptoms, called “asymptomatic malaria”. This system is not useful when some patients who suffer malaria have no significant symptom. This is one of the major limitations of this systems, but it will be useful most of the patients for diagnosing in common.

References

[1] Akanksha A Kherdikar Kurlekar and Anusuya S. 2011. "A Study on Role of Data Mining in Research Methodology." Indian Journal of Commerce & Management Studies. March 3. www.sscholarshub.net.

[2] Akash Ramaswamy, Chakrapani Mahabala, Sridevi Hanaganahalli Basavaiah, Animesh Jain, and Ravi Raj Singh Chouhan. 2020. "Asymptomatic malaria carriers and their characterization in hotpops of malaria at Mangalore." National Center for Biotechnology Information, June.

[3] 2022. Associative Classification in Data Mining. June 22. https://www.geeksforgeeks.org/ associ ative-classification-in-data-mining/.

[4] B, Nidhi. 2022. Rule Based Data Mining Classifier: A Comprehensive Guide 101. May 30. https://hevodata.com/learn/rule-based-data-mining/.

[5] Biao Qin, Yin Xia, Sunil Prabhakar and Yicheng Tu. 2009. A Rule-Based Classification Algorithm for Uncertain Data. IEEE International Conference on Data Engineering.

[6] Brownlee, Jason. 2020. "Machine Learining Mastery." Machine Learining Mastery. April 8. https://machinelearningmastery.com/classification-and-regression-trees-for-machine-learning/.

[7] n.d. "Data Mining-Decision Tree Induction." httrps://www.tutorialpoint.com/data_mining/dm_dt .htm.

[8] Deepankar. 2021. "Decision Tree with CART Algorithm." April 19. https://medium.com.

[9] Hari, Vijaya. 2009. Empirical Investigation of CART and Decision Tree Extraction form Neural Networks. Ohio University.

[10] Hnin Su Wai and Daw Myint Myint Maw. 2009. "Decision Support System For Mosquito Borne Disease." In Decision Support System For Mosquito Borne Disease, by Hnin Su Wai and Daw Myint Myint Maw. Yangon: University of Computer Studies, Yangon.

[11] Jiawei Han, Micheline Kamber and Jian Pei. 2012. Data Mining Concepts and Techniques. Morgan Kaufmann Publishers is an imprint of Elsevier.

[12] Jonathan Abeles and David J Conway. 2020. "The Gini Coefficient As a Useful Measure of Malaria Inequality Among Populations." https://doi.org/10.11.86/s12936-020-03489-x.

[13] Ministry of Health, Myanmar. 2002. Guidelines for Diagnosis and Treatment of Malaria in Myanmar. Malaria Manual Development Committee.

[14] Mohd Mahmood Ali and Lakshmi Rajamani. n.d. "Decision Tree Inductiion: Data Classification using Height-Balanced Tree." International Conference: Information and Knowledge Engineering.

[15] Njoku, Obinna Chilezie. 2019. Decision tree and Their Application for Classification and Regression Problems. Missouri State University.

[16] Nyi Nyi Latt and Khin Moe Sann. 2011. "Diagnosis of Malarias by Using Reduct Generation Algorithm." In Diagnosis of Malarias by Using Reduct Generation Algorithm, by Nyi Nyi Latt and Khin Moe Sann. Hinthada: Computer University (Hinthada).

[17] Ogundele I.O, Popoola O.L, Oyesola O.O and Orija K.T. 2018. "A Review on Data Mining in Healthcare." International Journal of Advanced Research in Computer Engineering & Technology (IJARCET).

[18] Parvez Ahmad, Saqib Qamar and Syed Qasim Afser Rizvi. 2015. "Techniques of Data Mining In Healthcare." International Journal of Computer Applications.

[19] Seif, George. 2018. "A Guide to Decision Trees for Machine Learining and Data Science." November 30. https://towardsdatascience.com/a-guide-to-decision-trees-for-machine-learning-and-data-science.

[20] Soe Kalayar Naing and Nyein Nyein Myo. 2011. "Diagnosis of TB Disease by Using Decision Tree Induction." In Diagnosis of TB Disease by Using Decision Tree Induction, by Soe Kalayar Naing and Nyein Nyein Myo. Taungoo: Computer University (Taungoo).

[21] Supajittree Boonnamnuay, Nittaya Kerdprasop and Kittisak Kerdprasop. 2018. "Classification and Regression Tree with Resampling for Classifying Imbalanced Data." International Journal of Machine Learning and Computing.

[22] T.C.Olayinka and S.C.Chiemeke. 2019. "Predicting Paediatric Malaria Occurrence Using Classification Algorithm in Data Mining." Journal of Advances in Mathematics adn Computer Science.

[23] Win Min Thit, Jaramit Raewkungwal, Ngamphol Soonthornworasin, Nawanan Theera Amppmpunt, Boonchai Kijsanayotin, Saranath Lawpoolsri, Sid

Naing and Wirichada Pan ngum. 2016. Electronic Medical Record in Myanmar User Perceptions at Marie Stopes International Clinics in Myanmar. Bangkok: Mahidol University.

[24] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh and Q. Yang. 2007. Top 10 Algorithms in Data Mining. Survey, Knowl Inf Syst.

Classification of Bank Depositor using ID3 and Naive Bayesian Classifiers

Moe San Phyu, Zaw Tun
Faculty of Computer Science, University of Computer Studies (Sittway)
sanm58833@gmail.com, zawtun78@gmail.com

Abstract

Nowadays, banks are financial institutions whose activities are to collect funds from the public in the form of deposits (saving deposit and time deposit). Deposits are an alternative for customers because the interest offered on deposits is higher than regular savings. So, this system is proposed as the bank depositor classification system by using data mining (DM) methods. Among many DM methods, this system uses the ID3 and Naive Bayesian classifiers to classify bank customer's data. This system predicts which customer will subscribe to a long-term deposit proposed by a bank. Moreover, this system analyses the sensitivity, specificity and accuracy of ID3 and Naive Bayesian classifiers. This system can help the bank for identifying customers who will potentially open a time deposit so that it can be used to assist the performance and operations of the bank.

1. Introduction

One of the economic improvements in the world can be visible through the emergence of financial institutions, particularly in the banking area. Banks are financial institutions whose fundamental activities are gathering funds from people (funding) and directing these funds back to the community (lending) and giving other bank services. Generally, banks themselves benefit from customers that can be utilized as a source of funds in the form of checking accounts, savings and time deposits. Furthermore, the form of source of funds that became one of the bank's backbones is deposits.

Deposits can be an option for customers because the interest presented on deposits is higher than ordinary savings. In managing customer data, the amount of information is extremely huge. In this way, a bank customer data classification system is required that can classify between customers who have the opportunity to open a deposit or not. This system also helps the operation of the bank.

For bank deposit classification, this system utilizes the ID3 decision tree and Naive Bayesian classifiers. This system also shows the effectiveness of these classifiers. By analyzing the sensitivity, specificity and accuracy, this system compares these two classifiers to know which classifier is more precise than other.

This paper is organized into seven sections. In the first section, the introduction of the system is presented. Then, related work is descried in the second section. ID3 and NB classifier are presented in the third and fourth sections. In the fifth section, holdout method is described. In the sixth section, this paper explains the proposed system architecture with detail explanation and experimental results. In the final section, this paper describes the conclusion of the system.

2. Related Work

In 2021, M. H. Effendy and D. Anggraeni [1] used Naive Bayes and K-Nearest Neighbor classifiers to classify the bank customer data. This system used the bank customer data that consists of 4521 records and 17 variables. Results of this study indicate that the KNN method is better than the NBC method.

In 2021, F. Safarkhani and S. Moro [2] used J48 decision tree classifier to predict a customer will be a long-term deposit or not. Combination of resampling and feature selection has been applied to 4119 instances of a Portuguese bank from the UCI repository. This framework was to understand the effectiveness of the J48 model at predicting the success of telemarketing calls for selling bank long-term deposits.

3. ID3 Decision Tree Classifier

A record enters the tree at the root node of the ID3 (Iterative Dichotomiser) classifier, which then performs a test to determine which child node the record will encounter next. The classification of each record that lands at a specific leaf of the tree is the same. Each leaf follows a different route from the root. That route is an expression of the record classification rule [3, 4].

3.1. ID3 Algorithm

Algorithm: Generate_decision_tree.

- create a node *N*;
- if *samples* are all of the same class, *C* then return *N* as a leaf node labeled with class *C*;
- if *attribute-list* is empty then return *N* as a leaf node labeled with common class in *samples*;
- select *test-attribute*, the attribute among *attribute-list* with the highest information gain;
- label node *N* with *test-attribute*;
- for each known value a_i of *test-attribute* grow a branch from node *N* for the condition *test-attribute*=a_i;
- let s_i be the set of samples in *samples* for which *test-attribute*=a_i;
- if s_i is empty then attach a leaf labeled with the most common class in *samples*;
- else attach the node returned by Generate _decision _tree [5];

3.2. Information Gain

Information gain is an attribute selection measure [6]. This method is as follows:

$$\text{Info}(D) = -\sum_{i=1}^{m} P_i \text{Log}_2(P_i) \quad (1)$$

$$\text{Info}_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \text{Info}(D_j) \quad (2)$$

$$\text{Gain}(A) = \text{Info}(D) - \text{Info}_A(D) \quad (3)$$

where, P_i is the probability that an arbitrary tuple in partition D. Info(D) is the average amount of information needed to identify the class label. It is also known as Entropy of D. The $|D_j|/|D|$ acts as the weight of the j^{th} partition. $\text{Info}_A(D)$ is the expected information required to classify a tuple from D based on the partitioning by A. Attribute A with the highest information gain, Gain(A), is chosen as the splitting attribute at Node N [6].

4. Naive Bayesian (NB) Classifier

The Bayes theorem, which offers a method of determining the posterior probability, P(c|x), form P(c), P(x), and P(x|c), is the foundation of NB classifiers. This NB classifier makes the assumption that the impact of a predictor's value on a particular class's (c) value is unrelated to the values of other predictors [7].

4.1. NB Classification Algorithm

NB classification algorithm is as follows:

1. Each data sample is represented by n-dimensional feature vector (n-attributes), $X = (x_1, x_2 \ldots x_n)$.
2. There are m classes, $C_1, C_2, \ldots, C_m$. Given an unknown data sample X, NB classifier assigns this sample to class C_i if and only if
   $P(C_i \backslash X) > P(C_j \backslash X)$ for $1 \leq j \leq m, j \neq I$ (4)
   The class C_i for which $P(C_i \backslash X)$ that is maximized, called maximum posteriori hypothesis. By Bayes theorem,
   $P(C_i \backslash X) = P(X \backslash C_i)\, P(C_i)/ P(X)$ (5)
3. As P(X) is constant for all classes, only $P(X \backslash C_i)\, P(C_i)$ need to be maximized.
4. Given data sets with many attributes, the Naive assumption of class conditional independence is made. Thus,
   $P(X \backslash C_i) = \prod_{k=1}^{n} P(x_k \backslash C_i)$ (6)
   The probability $P(x_1 \backslash C_i)$, $P(x_2 \backslash C_i)$,… $P(x_n \backslash C_i)$ can be estimated from the data samples.
5. To classify an unknown sample X, $P(X \backslash C_i)\, P(C_i)$ is evaluated for each class $C_{i,}$. Sample X is then assigned to the class C_i if and only if
   $P(X \backslash C_i)\, P(C_i) > P(X \backslash C_j)\, P(C_j)$ for $1 \leq j \leq m, j \neq i$
   In other words, it is assigned to the class C_i for which $P(X \backslash C_i)\, P(C_i)$ is the maximum [6].

5. Holdout Method

For assessing classifier accuracy, holdout method randomly partitions the given data into two independent sets, a training set (two thirds of the data) and a test set (one third of the data).

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP+FN}} \quad (7)$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN+FP}} \quad (8)$$

$$\text{Accuracy} = \frac{\text{TN+TP}}{\text{TN+TP+FN+FP}} \quad (9)$$

where TP is true positive, TN is true negative, FN is false negative and FP is false positive. Sensitivity refers to the probability of a positive test, conditioned on truly being positive. Specificity refers to the probability of a negative test, conditioned on truly being negative.

6. Proposed System Architecture

This system is proposed as the bank depositor classification system by using ID3 and NB classifiers. At first of the system, the user can choose the desired classifiers. If the user chooses the ID3 classifier, this system classifies the user inputted unknown sample according to the decision rules. For ID3 classification, this system first calculates each information gain for each attribute from the training bank data. If the attribute has highest information gain, this system chooses these attributes as the root node. After choosing root node, this system continues to choose leaf node by calculating the information gain of each attribute. By using root node and leaf nodes, this system generates the decision tree. Then, this system generates the decision rules from the decision tree. By using generated decision rules, this system classifies the user who can be bank's depositor or not.

Figure 1 shows the system flow diagram of the system. If the user chooses the NB classifier, the user must first input the unknown sample X (unknown bank information). Because of NB classifier performs the classification process based on unknown sample X. After accepting unknown sample, this system calculates the probability of each class ($P(C_i)$) by using training bank's data. Then, this system also calculates the probability of each attribute of unknown sample X ($P(X\backslash C_i)$). To choose highest probability, this system calculates the multiplication $P(X\backslash C_i)$ and $P(C_i)$. To determine the user who is bank's depositor or not, this system produces the class that has highest probability result.

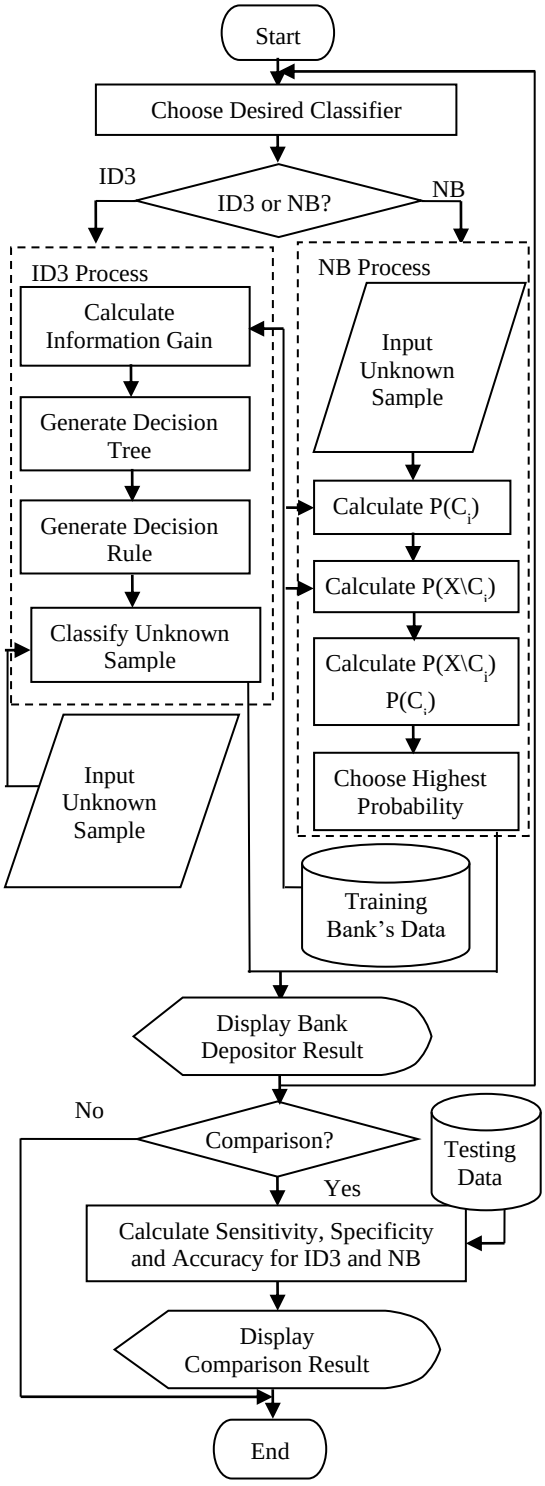


Figure 1. System Flow Diagram

By calculating sensitivity, specificity and accuracy of ID3 and NB classifiers as the performance, this system allows the user to know which classifier is more than another classifier.

To measure the performance of each classifier, this system uses the testing data. After measuring the performance of each classifier, this system displays the comparison result to the user.

6.1. Bank's Attribute Information

This system extracts the bank customer's dataset from the https://archive.ics.uci.edu/ml/datasets.php website. The size of this dataset is "4503" KB. Dataset includes the "4516" records, 16 attributes and one class. Bank customer's attribute information are shown in Table 1.

Table 1. Bank's Attribute Information

Attribute	Information
Age	Age of Bank Customer
Job	Type of job (Admin, Unknown, Unemployed, Management, Housemaid, Entrepreneur, Student, Blue-collar, Self-employed, Retired, Technician, Services)
Marital	Marital status (Married, Divorced, Single)
Education	Unknown, Secondary, Primary, Tertiary
Default	Has credit in default? (Yes, No)
Balance	Average yearly balance in euros
Housing	Has housing loan? (Yes, No)
Loan	Has personal loan? (Yes, No)
Contact	Unknown, Telephone, Cellular
Day	Last contact day of the month
Month	Last contact month of year (Jan, Feb,…, Dec)
Duration	Last contact duration, in seconds
Campaign	Number of contacts performed during this campaign and for this client
Pdays	Number of days that passed by after the client was last contacted from a previous campaign
Previous	Number of contacts performed before this campaign and for this client
Poutcome	Outcome of the previous marketing campaign (Unknown, Other, Failure, Success)
Class	Yes No

6.2. Explanation of the System

Proposed bank depositor classification system is explained by using ID3 and NB classifiers. This system is tested by using 16 patient records. Sample bank training data is shown in Table 2.

Table 2. Sample Bank Training Data

Age	Job	Martial	…	Poutcome	Class
29	admin	single	…	Failure	no
35	management	single	…	Failure	no
33	management	married	…	Failure	yes
61	admin	married	…	Success	yes
34	technician	married	…	Other	yes
32	blue-collar	single	…	Failure	no
25	admin	single	…	Failure	yes
32	technician	single	…	Success	no
59	technician	married	…	Success	yes
26	blue-collar	single	…	Failure	no
37	management	married	…	Success	yes
51	blue-collar	married	…	Failure	no
73	retired	married	…	Failure	yes
83	retired	married	…	Failure	yes
35	blue-collar	single	…	Other	yes
34	management	married	…	Other	no

According to the ID3 classifier, this system calculates information gain for each iteration. First iteration gain results are shown in Table 3.

Table 3. First Iteration Gain Results

Attribute Name	Information Gain Results
Age	0.014
Job	0.007
Marital	0.182
Education	0.023
Default	0
Balance	0.989
Loan	0.001
Contact	0
Day	0.036
Month	0.661
Housing	0.239
Duration	0.989
Campaign	0.055
Pdays	0.989
Previous	0.462
Poutcome	0.057

In the first iteration, the "Balance" attribute is the root node because it has highest information gain. After finishing first iteration, this system continues to calculate the information gain for next iterations. In this sample, this system generates decision tree after finishing the fourth iteration. Decision tree is shown in Figure 2.

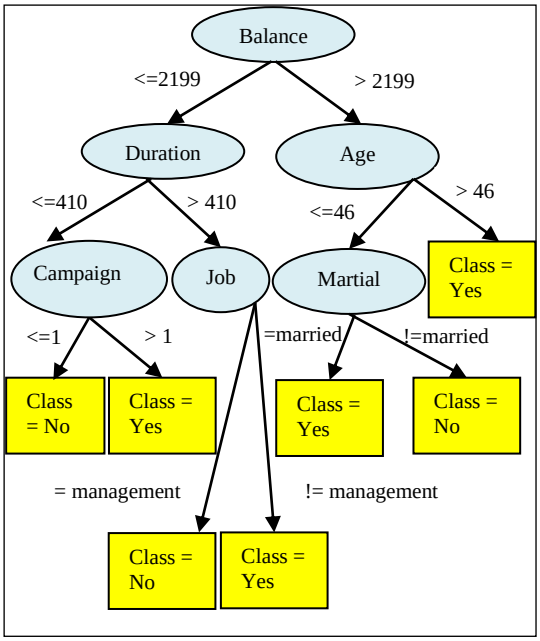


Figure 2. Decision Tree

According to the decision tree, this system generates the decision rules for unknown sample classification. **Rule 1** is {IF "Balance ≤ 2199" AND "Duration ≤ 410" AND "Campaign ≤ 1" THEN Class = No}. **Rule 2** is {IF "Balance ≤ 2199" AND "Duration ≤ 410" AND "Campaign > 1" THEN Class = Yes}. **Rule 3** is {IF "Balance ≤ 2199" AND "Duration > 410" AND "Job = Management" THEN Class = No}. **Rule 4** is {IF "Balance ≤ 2199" AND "Duration > 410" AND "Job != Management" THEN Class = Yes}. **Rule 5** is {IF "Balance > 2199" AND "Age ≤ 46" AND "Martial = married" THEN Class = Yes}. **Rule 6** is {IF "Balance > 2199" AND "Age ≤ 46" AND "Martial != married" THEN Class = No}. **Rule 7** is {IF "Balance > 2199" AND "Age > 46" THEN Class = Yes}.

By using the decision rules, this system classifies the user inputted unknown sample. The unknown sample (unknown bank information) is "Age = 29", "Job = admin", "Marital = single", "Education = secondary", "Default =no", "Day= 16", "Balance = 1350", "Housing = yes", "Loan = no", "Contact= cellular", "Month= apr", "Duration= 185", "Campaign= 1", "Pdays=330", "Previous= 1" and "Poutcome= failure". According to the **Rule 1**, this system produces that the "**Bank Customer's Deposit Result is No**" for user inputted unknown sample.

According to the Naive Bayesian classifier, this system calculates the probability for each attribute. Based on probability results, this system classifies the unknown sample. The probability results of NB classifier are shown in Table 4.

Table 4. Probability Results of NB Classifier

Attribute	Probability Results	
	Class (No)	Class (Yes)
Age: 29	0.142857	0
Job: admin	0.142857	0.222222
Marital: single	0.714286	0.222222
Education: secondary	0.428571	0.222222
Default: no	1	1
Balance: 1350	0.142857	0
Housing: yes	0.571429	0.555556
Loan: no	1	1
Contact: cellular	0.714286	0.888889
Day: 16	0.142857	0.333333
Month: apr	0.142857	0.111111
Duration: 185	0.142857	0
Campaign: 1	0.857143	0.777778
Pdays: 330	0.142857	0
Previous: 1	0.285714	0.333333
Poutcome: failure	0.714286	0.444444

After calculating each attribute probabilities, this system obtains the "Class = No" probability that is "0.000000011611" and the "Class=Yes" probability that is "0". So, this system produces that the "**Bank Customer's Deposit Result is No**" for user inputted unknown sample.

6.3. Experimental Result of the System

To compare the performance of ID3 and NB classifiers, this system calculates the sensitivity, specificity and accuracy of these two classifiers. For experimental result, this system uses "4516" bank data records. According to holdout method, this system splits these bank data records into "3008" training and "1508" testing records.

Table 5. Sensitivity, Specificity and Accuracy Results of ID3 and NB Classifier

Performance Measurement	ID3 Classifier	NB Classifier
Sensitivity	80 %	74 %
Specificity	53 %	73 %
Accuracy	79 %	76 %

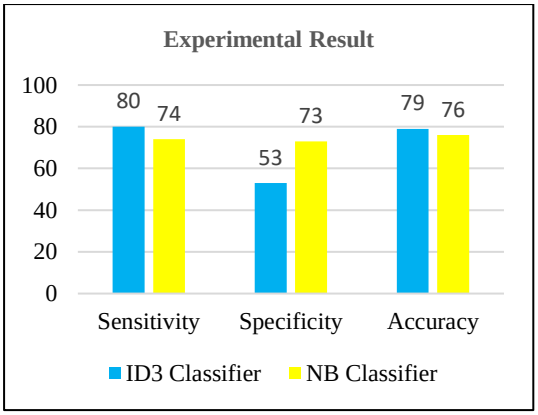


Figure 3. Experimental Result of the System

By using equation 7, 8 and 9, this system calculates the sensitivity, specificity and accuracy. These results are shown in Table 5. Experimental results of the system are shown in Figure 3. According to the performance measurement results, the accuracy of ID3 classifier is more precise than the NB classifier.

7. Conclusion

For bank depositor classification, this system used the ID3 decision tree and Naive Bayesian (NB) classifiers. This system resulted in each classifier to predict whether a customer will subscribe to a long-term deposit or not. This system is also implemented to show which classifier is more than another. To compare the performance of ID3 and NB, this system calculates the sensitivity, specificity and accuracy of each classifier. Finally, the system is helpful for bank manager during bank depositor classification.

References

[1] M. H. Effendy, D. Anggraeni, “Classification of Bank Deposit Using Naive Bayes Classifier (NBC) and K-Nearest Neighbor (KNN)”, *Proceedings of the International Conference on Mathematics, Geometry, Statistics and Computing*, 2021.

[2] F. Safarkhani and S. Moro, “Improving the Accuracy of Predicting Bank Depositor's Behavior Using a Decision Tree”, *Applied Sciences*, 2021.

[3] Umadevi, S. and Marseline, J. "A Survey on Data Mining Classification Algorithms", *International Conference on Signal Processing and Communication*, pp. 264-268, 2017.

[4] Sharma, S. and Agrawal, J. "Machine Learning Techniques for Data Mining: A Survey", *IEEE*, 2013.

[5] A. E. Permanasari and A. d. W. Lkram, "A Web-based Decision Support System of Patient Time Prediction using Iterative Dichotomiser 3 Algorithm", *IEEE*, 2019.

[6] H. Jiawei and K. Micheline, *Data Mining Concepts and Techniques*, USA, 2001.

[7] K. Vembandasamy, R. Sasipriya and E. Deepa, "Heart Diseases Detection using Naive Bayes Algorithm", *International Journal of Innovative Science, Engineering & Technology*, vol. 2, no. 9, pp. 441-444, 2015.

Water Demand Prediction in Irrigation System using KNN Algorithm

Wint Wah Loon, Thin Lai Lai Thein
University of Computer Studies, Yangon
Wintwahloon2@ucsy.edu.mm, tllthein@ucsy.edu.mm

Abstract

Over the coming decades, the global agriculture sector will face growing challenges due to rising food demand. Increasing crop output while avoiding a significant increase in farmland area is a sustainable answer to this problem. Best management techniques should be found and adopted in order to accomplish this. To do so, a deeper comprehension of how climate change and growing-season weather variability affect crop productivity is necessary. The amount of water needed for irrigation in a field of agricultural depends on many factors. These elements include the age of the plant, humidity, temperature, and soil moisture/need for soil water. Despite the several solutions suggested, there is still a lot of water under the floors and over flood in the agricultural field. Given the need for irrigation, artificial influence should be regarded as a significant effect factor. Technology can assist in conserving a significant amount of water in agriculture. This system uses K-nearest neighbor to anticipate the equilibrium between water supply and demand, which calls for an effective water supply system (KNN).

This system is implemented using C# programming language on Microsoft Visual Studio ID and Microsoft SQL Server is also used for Database Engine.

Keywords: irrigation, humidity, temperature, soil moisture, KNN

1. Introduction

Weather forecasting and predicting the amount of soil water that will happen in a space can be exceptionally dreary. It would include fastidious perception of the air conditions and cloud development alongside the formation of models to mimic barometrical circumstances and cumulus cloud connection, which likewise prompts a serious level of intricacy [2]. How much water utilized for water system is assessed without thinking about the successful measure of precipitation that will be knowledgeable about a given region. At the point when unreasonable measure of water is utilized it might prompt over water system, Water logging and may likewise bring about saltiness in this way lessening crop yield. Anyway, utilizing less water may likewise bring about under water system and diminishes yield effectiveness.

Consequently, assessing the perfect proportion of water that will be provided for inundating the harvests is of central significance. The strategy includes dissecting different geological factors, for example, land geography, slant, channel elements, for example, soil surface, construction, and profundity, alongside meteorological boundaries, for example, temperature, radiation, relative stickiness, and wind speed, to anticipate how much compelling precipitation that will be gotten over indicated geographic locales [1]. With this, the harvest is not entirely set in stone consistently.

The objective of the strategy is to help ranchers with utilizing the water required or not really for water systems and assist them with picking the right water system situation that ought to be carried out for the ideal development of yields. Information mining methods, for example, characterization are utilized to acquire information about how much water expected for water system which would go far in reinforcing the agrarian area. This also allows ranchers to plan ahead of time how much water should be set aside for the water system in the event of a monsoon rainstorm. This system will predict the water supply needed or not for specific crops by using K-nearest neighbor (KNN).

The abstract of the proposed paper describes a brief explanation of the system, while the introduction section discusses weather forecasting, prediction, and water needed

prediction. Section 2 discussed the related work of the system, section 3 discussed motivation, and section 4 discussed background theory. Section 5 is used to present the system implementation, and Section 6 will describe the conclusion.

2. Related Work

Rainfall and weather expectation include refined PC displaying and amusement for precise forecast [3]. Displaying of such non-direct frameworks have been achieved by utilizing fake brain networks which have been utilized in the framework proposed in papers [8] and [9] An Counterfeit Brain Organization here is utilized to predict the way of behaving of such nonlinear frameworks. Demonstrating of such frameworks basically includes the utilization of delicate figuring.

Delicate figuring has three fundamental parts, to be specific, Artificial Neural Network (ANN), Fluffy rationale and Hereditary Algorithm. [3] Delicate processing is a model which manages surmised models where an estimate reply or result is accomplished. Measurable signs picked are fit for separating the patterns, which can be viewed as components for making the models. Fake brain networks have been utilized by many individuals to display the cumulus cloud communication in different places, for example, Thailand [10] and in a lot more places to gauge the got precipitation. They have likewise been utilized in the approaches proposed in [8] and [9]. Their endeavors however are gathered in the making a proficient displaying framework to foresee got precipitation and relatively few propose ideas to utilize the precipitation got in a viable manner.

In paper [11], a framework has been depicted with a contextual investigation of the Bijapur region of Karnataka to give data on precipitation qualities and its expectation from verifiable informational collections, the determination of harvests in light of the gauge on taluka premise. It gives a definite examination of the precipitation received throughout recent years, yet information mining methods have not been applied in an effective manner in order to mine sufficient information to take care of a portion of the issues faced by ranchers, for example, assessing how much water is needed for a water system. In [4], the assessment of powerful precipitation depends principally on 3 factors: specific dampness, Temperature and precipitation are calculated using a neuro fuzzy framework that is divided into two sections: fluffy rationale and brain organization. Different factors, for example, the land incline, soil surface, soil design, and wind speed have not been thought of. This might permit little errors to creep in while computing how much compelling precipitation there is. This may not be alluring if the edge for blunder is small, particularly in a nation like India where ideal yield is of fundamental significance.

3. Motivations

When the climate varies, then automatically field parameters also suddenly change. Whenever there is heavy rainfall or temperature variation, this may become very hard to analyze and it causes a major problem. Given this issue, designing a water demand prediction system for irrigation systems is critical. Therefore, choosing a time series model that respects these traits is necessary in order to produce the most appropriate representation. The short-term water demand time series was forecasted by this system using the k-nearest neighbor (kNN) methodology. When using the kNN technique, a pattern recognition algorithm, the anticipated values are directly based on the most comparable prior observations.

4. Background Theory

To anticipate whether water is needed or not, a machine learning approach is used, which makes use of numerous categorization algorithms. A classifier uses training data to map input variables to target classes. Classifiers used in the paper to forecast whether water would be needed versus not are briefly outlined. Generally speaking, these classifier-based predictions can be divided into two types: single classifier-based predictions and ensemble classifier-based predictions.

4.1. Single Classifier based Prediction

Classifiers are trained to predict unknown test cases. The following classifiers are used while detecting fake job posts:

a) Naive Bayes
b) Multi-Layer Perceptron Classifier
c) K-nearest Neighbor
d) Decision Tree Classifier

4.2. Ensemble Approach based Classifiers

The ensemble approach enables numerous machine learning algorithms to work together to improve the overall system's accuracy.

5. System Implementation

There are steps in this detection framework such as Load dataset, Data Preprocessing, Prediction and Accuracy Results.

I. Load Dataset

The experiment dataset is loaded into the system's machine learning repository during this stage.

II. Pre-processing

The raw dataset will go through data cleaning processes like tokenization, stop-word removal, and stemming during the preprocessing phase. The clean dataset will be utilized for the subsequent feature extraction and selection phase.

i. **Tokenization** is the process of breaking down the text corpus into individual elements.

ii. **Removing Stop Words**: Stop words are extraneous words that frequently appear in texts. For instance, phrases like "so," "and," "or," "the," etc. First, all stop words are eliminated. The stop words in the illustration below are: you, are, that, have, and the, which are eliminated by this method.

iii. **Prediction**: During this phase, there are training and testing procedures. 40% will be allocated to testing, and 60% to training. After finishing step iii, there should be features that are regarded as spam. Consequently, the dataset must be trained using a machine learning technique (KNN).

III. Prediction and Accuracy Result

The prediction is the main phase of system and the KNN classifier is used to predict and then accuracy is determined and discussed in section 6.

5.1. K-nearest Neighbor Classifier (KNN)

It is a supervised machine learning approach that employs proximity to classify or predict how a single data point will be grouped. In order to learn by analogy, nearest-neighbor classifiers compare a given test tuple with training tuples that are similar to it. The letter "K" stands for the number of closest neighbors to a new unknown variable that needs to be forecasted or categorized. In order to determine what class a new unknown data point belongs to, it seeks to locate all of its nearest neighbors. It is also known as a lazy learner algorithm since it saves the training dataset rather than learning from it immediately. Instead, it uses the dataset to perform an action when classifying data. The item is just put in the class of the object's one nearest neighbor if k=1. The advantages of KNN are: Easy to implement, adapts easily ad few hyper-parameters.

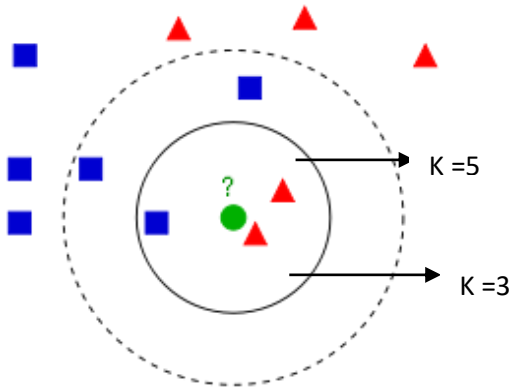


Figure 1. KNN

KNN works as follows:

Step-1: Select the number K of the neighbors
Step-2: Calculate the Euclidean distance of K number of neighbors
Step-3: Take the K nearest neighbors as per the calculated **Euclidean distance**.

The training tuples are described by n attributes. Each tuple represents a point in an n-dimensional

space. When given an unknown tuple, a k-nearest-neighbor classifier searches the pattern space for the k training tuples that are closest to the unknown tuple. "Closeness" is defined in terms of a distance metric, such as Euclidean distance. The Euclidean distance between two points or tuples, say,

X1 = (x11, x12,..., x1n) and X2 = (x21, x22,…....., x2n) , is

$$dist(X_1, X_2) = \sqrt{\sum_{i=1}^{n} (x_{1i} - x_{2i})^2}. \quad (1)$$

Where, x1, x2 = two points in Euclidean n-space
X1i - X2i = Euclidean vectors, starting from the origin of the space (initial point)
n = n-space

Step-4: Among these k neighbors, count the number of the data points in each category.
Step-5: Assign the new data points to that category for which the number of the neighbor is maximum.

5.2. The Proposed System

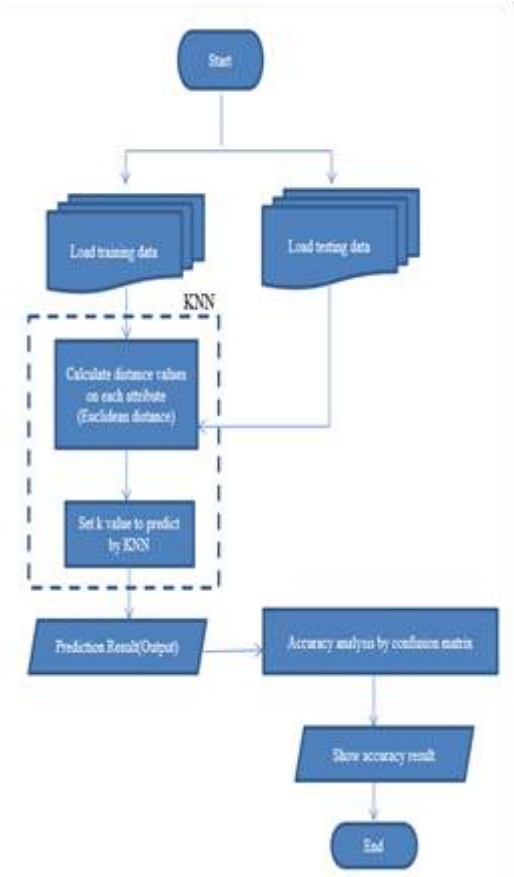


Figure 2: The Process Flow

The proposed water demand prediction on irrigation will be emphasized for two sample training datasets (Plants A dataset and Plants B dataset). Each attribute of the training dataset contents and the testing dataset contents distance are calculated by Euclidean distance. Then, the resultant distance values are fed to KNN for prediction of water demand by suitable selected K values. For the accuracy evaluation, the confusion matrix will be used.

Table 1. Sample Dataset of Plant A (Watermelon)

Temp	Humid	SM	Plant Type	Remark
26	67	27.5	Watermelon	Water Need
25	56	51	Watermelon	No Water Need
27	94	32	Watermelon	Water Need
22	78	50.113	Watermelon	No Water Need
31	65	24.2	Watermelon	Water Need

Table 2. Sample Dataset of Plant B (Banana)

Temp	Humid	SM	Wind Speed	Plant Type	Remark
36	53	31	3	Banana	Water Need
20	69	37.4	5	Banana	No Water Need
31	75	28	5	Banana	Water Need
20	78	36.7	7	Banana	No Water Need
32	67	25.6	6	Banana	Water Need

6. Evaluation Metrics

Five evaluation metrics, which are precision, recall, F-measure, accuracy and failure-ratio, are used to evaluate the effectiveness of the system. These are calculated by using Eq. (4.1) -(4.5) respectively.

$$\text{Accuracy} = \frac{TP+TN}{TP+FN+FP+TN} \quad (1)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (3)$$

$$\text{F- measure} = \frac{Precision * Recall}{Precision+Recall} \quad (4)$$

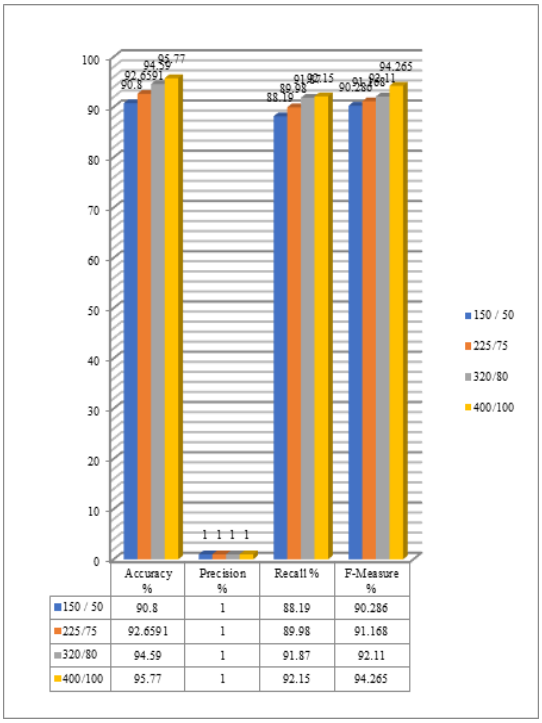
Where:

- TP refers to the number of true positive reviews.
- TN refers to the number of true negative reviews.

- FP refers to the number of false positive reviews.
- FN refers to the number of false negative reviews.
- Number of Misclassified Reviews refers to the reviews labelled to the class label which was not included in the actual class labels.
- Total Number of Reviews refers to the number of all reviews.

6.1. Experimental Result

In each analysis of 3 different training dataset and testing dataset pairs are used (Test1: Training Data 150 records and Testing Data 50 records; Test 2: Training Data 225 records and Testing Data 75 records; Test 3: Training Data 300 records and Testing Data 100 records; Test 4: Training Data 400 records and Testing Data 100 records. This system made the experiment result's performance evaluation based on Accuracy, Precision, Recall and F-measure of each analysis. The analysis results of 4 different dataset are shown in Figure 3.



	Accuracy %	Precision %	Recall %	F-Measure %
150 / 50	90.8	1	88.19	90.286
225/75	92.6591	1	89.98	91.168
320/80	94.59	1	91.87	92.11
400/100	95.77	1	92.15	94.265

Figure 3: Experimental Results of Analysis

Based on the analysis, the system can give better detection result if the more trained data can feed to this system.

7. Conclusion

The need for irrigation water is mostly influenced by human factors and climate change. The primary climatic factor is rainfall, and human factors such as irrigation area and technology level are also important. The impact of human factors is becoming increasingly noticeable with the advancement of water-saving irrigation equipment. The water demand forecasting model incorporates the idea of the water-saving improvement coefficient based on the dual feature of "artificial-natural". The approach has a better simulation effect than time series analysis and conventional regression, and it can more accurately reflect the impact of water-saving technologies and adjusting planting structure on irrigation water.

References

[1] Paulo José A. Oliveira and Dominic L. Boccelli,"k-earnest Neighbor for Short Term Water Demand Forecasting", World Environmental and Water Resources Congress 2017.

[2] A.M.ASCE, "Water Demand Prediction Using Machine Learning", https://www.mifratech.com, 2022.

[3] Bhuvana, Shashikala, "Water Demand Prediction using KNN Algorthim", Department of Computer Science and Engineering, Dayananda Sagar Academy of Technology, 2022.

[4] P. Cunningham and S. J. Delany,K -Nearest Neighbor Classifiers,‖ Mult. Classif. Syst., no. May, pp. 1–17, 2007, doi: 10.1016/S0031-3203(00)00099-6.

[5] E. G. Dada, J. S. Bassi, H. Chiroma, S. M. Abdulhamid, A. O. Adetunmbi, and O. E. Ajibuwa, "Machine learning for Prediction Outcomes: review, approaches and open research problems, Heliyon, vol. 5, no. 6, 2019, doi: 10.1016/ j.heliyon. 2019.e01802.

Performance Comparison of Supervised Machine Learning Algorithms for Credit Card Fraud Detection

Tin Zar Lin, Khin Lay Myint
University of Computer Studies (Monywa)
tinzarlin.cu@gmail.com, khinlaymyint.cu@gmail.com

Abstract

Credit cards are widely used due to the development of e-commerce and mobile intelligent devices. Credit card fraud events take place frequently and the results of which are many financial problems. Fraudsters use some technology to steal credit card owner's information. Therefore, effective fraud detection methods are very important. Credit card fraud detection is based on the fraudulent transactions. Generally, credit card fraud activities can occur both online and offline. A variety of methods can be used to find fraudulent transactions. In this system, Support Vector Machine (SVM) and Random Forest Algorithm (RFA) are used to find fraudulent transactions and to evaluate the performance of those transactions. All of these algorithms are based on supervised machine learning algorithms. In this system, two different credit card datasets are used. The two credit datasets are highly unbalanced data and the number of non-fraudulent transactions greatly exceeds the number of fraudulent transactions. In this system, the two credit datasets are resampled using random under-sampling. After resampling, a balance dataset is obtained. A balance dataset is modeled with SVM and RFA to classify fraud and non-fraud. After modeling, a confusion matrix is obtained. A confusion matrix is partitioned into four blocks TP, TN, FP and FN. The performance of SVM and RFA are evaluated based on confusion matrix. Finally, displays the comparison of performance results and execution time of two algorithms on two different credit datasets.

Keywords: Credit card fraud detection, random under-sampling, splitting, confusion matrix, Support Vector Machine (SVM), Random Forest Algorithm (RFA)

1. Introduction

Credit cards are widely used due to the development of e-commerce and mobile intelligent devices. Moreover, credit cards have made online transactions easier and more convenient. In our every life, various transactions are made through credit card payments, card less transactions like Google Pay and PayPal, etc. Physical cards are required in face-to-face transfers for legal activities, while virtual cards are required when you conduct illegal or fraudulent activities.

It is now simpler to commit fraud because the physical card is not required for online transactions and the card's information is sufficient to complete a payment [8]. Credit card fraud is increasing day by day in the real world. Credit card fraud can occur in both online and offline transactions. Thus, without the cardholder's knowledge, these fraudulent credit card operations may result in several fraudulent transactions. In order to conduct transactions, fraudsters are monitoring for sensitive data including credit card numbers, bank accounts, and other user details.

Fraudsters steal users' credit cards when conducting transactions locally, and when conducting transactions online, they steal the user's identity and login credentials. In today's technologically advanced world, credit card fraud has developed to be a significant problem in bank transactions. There are many fraudulent transactions that cannot be facilely identified by the utilizer and by the banking ascendancy which leads to the loss of sensitive data [2].

Therefore, fraud detection is necessary and essential. Fraud detection is basically dividing the transaction between fraudulent and non-fraudulent can enjoy their shopping or any other transactions easily without any delay. There are various models in order to detect fraud transactions based on the behavior of the

transactions and these methods can be used as supervised machine learning algorithms. This system utilizes the supervised machine learning algorithms which are implemented on two Credit Card Datasets from Kaggle. In this system, Credit Card Fraud Detection System is developed based on transaction features by applying Support Vector Machine (SVM) and Random Forest Algorithm (RFA) classifiers.

This paper discusses the Performance Comparison of Supervised Machine Learning Algorithms for Credit Card Fraud Detection Systems as follows. Section II describes the Related Work. Section III describes the credit card fraud detection system, the Datasets Description, the methodology of the system and the related theory of the system. Section IV describes the experimental results of the system and conclusion respectively.

2. Related Work

Machine Learning approaches are vital is the detection of card fraud throughout many efficient areas for data processing one of them is the identification of card fraud.

The accuracy, specificity, and sensitivity performance metrics are used by the authors of [1] to evaluate and compare the performance of Support Vector Machine (SVM) Kernel methods, which are trained on transactional data. The model is evaluated existing classifiers such as Naive Bayes, Decision Trees, KNN, Logistic Regression, and SVM. To create a balanced dataset from the severely skewed data, the positive class is down-sampled and the negative class is up-sampled. The results suggest that SVM Kernel methods show conventional approaches in terms of all three-performance metrics, including sensitivity, accuracy, and specificity over traditional techniques. Analysis and observation show that the RBF Kernel function outperforms other methods and provides 96% accuracy and 96% sensitivity. The highest sensitivity, when compared to other techniques, is provided by the linear kernel function, which achieves 90%.

Authors in [2] examined the performance of two various of Random Forest models. Our experiment provides use of a real-world B2C dataset of credit card transaction. Even though Random Forest performs well on small sets of data, there are still several problems, including as imbalanced data. This performance analysis gives an accuracy of approximately 90% in this credit card fraud detection system.

By deploying a genetic algorithm and a multivariate normal distribution on an unbalanced credit card transaction dataset, the authors of [3] defined the general concept of an outlier. After being trained and evaluated on the same dataset, the model's prediction accuracy was compared to that of an artificial neural network, support vector machine, and decision tree. The model's results revealed an impressive F score of 93.5%, while the F scores for the artificial neural network, support vector machine, and decision tree were 68.5%, 84.2%, and 80.0%, respectively.

In applications including the detection and prevention of financial frauds like money laundering, credit card theft, tax evasion, check theft, and embezzlement, authors in [4] have shown to be of great help. The Gradient Boosting algorithm, one of the several Machine Learning algorithms applied, has an edge over the others. Gradient Boosting outperformed the methods proposed by 95.9%.

The Random Forest classifier performs best with 96.7% accuracy, according to the authors of [5], who investigated the decision tree, Random Forest, logistics regression, and naive bayes classification machine learning algorithms. The differences in the performances of each algorithm were minimal.

The performance of classification algorithms was evaluated by the authors in [6] based on financial loss. The records of credit card transactions in a bank in Europe over a two-day period were used to develop a weighted SVM model using random under-sampling methodology, using the number of transactions as a weight for fraudulent data points. The results demonstrate that credit card fraud detection performance can be significantly improved by utilizing new criteria and novel weighting schemes. Most importantly, this strategy will reduce the financial losses caused by banks due to credit card fraud.

Authors in [7] use the Random Forest Algorithm to accurately identify credit card fraudulent transactions. This algorithm is based on a supervised learning algorithm where it utilizes a decision tree for the classification of the dataset. A confusion matrix is obtained once the

dataset has been classified. Based on the confusion matrix, the Random Forest Algorithm's performance is evaluated. The accuracy of the results from processing the dataset is around 90%.

Authors in [8] evaluated the performance of two different of random forest models. Credit card transaction data from real-life B2C transactions is used. Despite producing strong results on small datasets, the random forest algorithm still has massive problems, such as imbalanced data. The accuracy of the results from processing the dataset is around 98.67%.

To verify the effectiveness of these algorithms, authors in [9] utilized classifier ensembles (Bagging, AdaBoost, Random Forest, and Gradient Boosting) in fraudulent credit card transactions. Various performance measures were used to evaluate and compare the results. There remain a number of issues.

3. Proposed System

Credit Card Fraud Detection System uses SVM and RFA for the classification of Datasets. This system uses two credit card datasets. Import the credit card datasets from publically available Kaggle. The dataset is. CSV (Comma Separated Values) file. Firstly, credit card datasets are analyzed to understand credit card use in e-commerce and the usage of various models of credit card types. Resampling the datasets is done after analyzing the datasets. The main issue with this system is the problem of class imbalance.

Therefore, a resampling procedure is necessary. Then, in order to analyze the datasets, the datasets are divided into two categories: the train dataset and the testing dataset. After dividing the datasets, SVM and RFA are applied, which improve better performance to identify credit card fraud.

The datasets will be divided into four categories using SVM and RFA, and the results will be presented in the form of a confusion matrix. The dataset will be divided into four blocks for the confusion matrix: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). The obtained confusion matrix is used for the performance analysis. The performance of credit card fraud transactions may be obtained in this analysis, and the results of the performance tests of two algorithms on two credit card datasets can then be compared.

3.1. Datasets Description

The Credit Card Datasets are skewed and highly imbalanced the first task to scale and sample the dataset into equal fraud and non-frauds. The format of the two Credit Card Datasets is .CSV (Comma Separated Values) file. In this system, the datasets have the two portions. The first portion is dataset 1 and the second portion is dataset 2.

The dataset 1 is obtained from Kaggle and it contains transactions of European credit card holders of September 2013 for two days which has 284,807 transactions of which 492 are fraud one's class as 0 and 284,315 non-fraud class as 1. The dataset is highly unbalanced; the number of frauds is 0.172% of all transactions.

Table 1. Dataset 1 from European Credit Card Holders Transactions

Time	V1	V2	V3	V4	V5	V6	V7	V8
0	-1.35981	-0.07278	2.536347	1.378155	-0.33832	0.462388	0.239599	0.098698
0	1.191857	0.266151	0.16648	0.448154	0.060018	-0.08236	-0.0788	0.085102
1	-1.35835	-1.34016	1.773209	0.37978	-0.5032	1.800499	0.791461	0.247676
1	-0.96627	-0.18523	1.792993	-0.86329	-0.01031	1.247203	0.237609	0.377436
2	-1.15823	0.877737	1.548718	0.403034	-0.40719	0.095921	0.592941	-0.27053
2	-0.42597	0.960523	1.141109	-0.16825	0.420987	-0.02973	0.476201	0.260314
4	1.229658	0.141004	0.045371	1.202613	0.191881	0.272708	-0.00516	0.081213
7	-0.64427	1.417964	1.07438	-0.4922	0.948934	0.428118	1.120631	-3.80786
7	-0.89429	0.286157	-0.11319	-0.27153	2.669599	3.721818	0.370145	0.851084
9	-0.33826	1.119593	1.044367	-0.22219	0.499361	-0.24676	0.651583	0.069539
10	1.449044	-1.17634	0.91386	-1.37567	-1.97138	-0.62915	-1.42324	0.048456

It contains only numerical input variables which are the result of a Principle Component Analysis (PCA) transformation. The original features are not provided because PCA transformation is performed on them and the attributes are represented v1 to v28 that are PCA features because of confidentiality issues and dimensionality reduction to protect user identities and sensitive features. The known features of Amount and Time is the number of second elapsed between these transactions and the first transactions in the dataset.

The dataset 2 of credit card transactions is collected from Kaggle and it contains transactions of German credit card holders. It contains a total of 1000 credit card transactions.

It considers fraud transactions as “class 0” and nonfraud transactions as “class 1”. The dataset 2 contains the number of nonfraud 700 and the number of frauds 300 and it contains 20 features. These features are Account Balance, Duration of Credit (Month), Payment Status of Precious

Credit, Purpose, Credit Amount, Value Savings/Stocks, Length of Current Employment, Installment Percent, Sex & Marital Status, Guarantors, Duration in Current Address, Most Valuable Available Asset, Age (Year), Concurrent Credit, Type of Apartment, Number of Credit at this Bank, Occupation, Number of Dependents, Telephone and Foreign Worker.

Table 2. Dataset 2 from German Credit Card Holders Transactions

Account B	Duration c	Payment S	Purpose	Credit Am	Value Sav	Length of	Instalmen	Sex & Mar
1	18	4	2	1049	1	2	4	2
1	9	4	0	2799	1	3	2	3
2	12	2	9	841	2	4	2	2
1	12	4	0	2122	1	3	3	3
1	12	4	0	2171	1	3	4	3
1	10	4	0	2241	1	2	1	3
1	8	4	0	3398	1	4	1	3
1	6	4	0	1361	1	2	2	3
4	18	4	3	1098	1	1	4	2

3.1.1. Before Resampling Dataset 1

Unbalanced data typically represents unequal class distributions within a dataset. Unbalanced data are those kinds of datasets in which the target class has an unequal distribution of observations, i.e., one class label has a very high number of observations while the other has a very low number. Most credit card transactions in credit card datasets are non-fraudulent, and very few classes are transactions that are fraudulent.

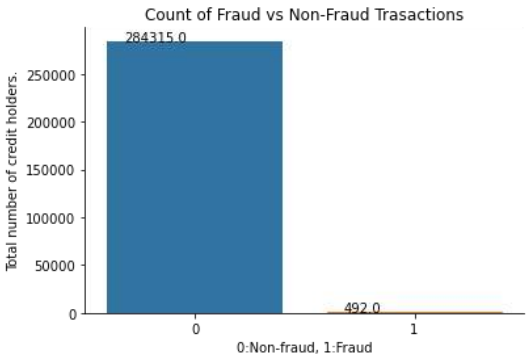


Figure 1. Count of Fraud and Non-Fraud Transactions before resampling

The number of non-fraudulent has 284315transactions and the number of fraudulent has 492 transactions in datasets.

Figure 2. Counts of Fraud and Non-Fraud Transactions before resampling

The number of non-fraudulent has 700transactions and the number of fraudulent has 300 transactions in datasets.

3.1.2. After Resampling

Sampling is a technique for dealing with unbalanced data, and it is frequently used to transform unbalanced datasets into balanced datasets by using random under-sampling.

Figure 3. Counts of Fraud and Non-Fraud Transactions after resampling

After random under-sampling, the number of non-fraudulent has 492transactions and the number of fraudulent has 492 transactions in datasets. The total number of transactions in resampled dataset is 984.

After random under-sampling, the number of non-fraudulent has 300transactions and the number of fraudulent has 300 transactions in datasets. The total number of transactions in resampled dataset is 600.

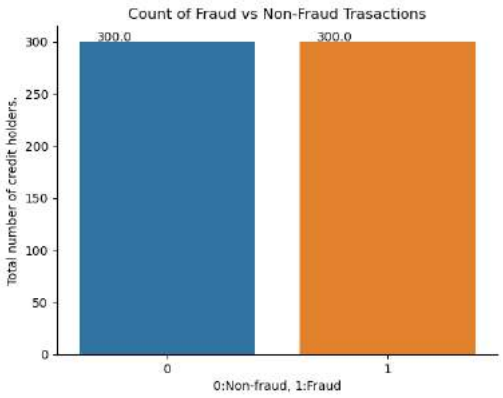


Figure 4. Counts of Fraud and Non-Fraud Transactions after resampling

3.2. Methodology of the System

The proposed system contains six main phases. Figure 5 shows the Conceptual Architecture for Credit Card Fraud Detection with SVM and RFA.

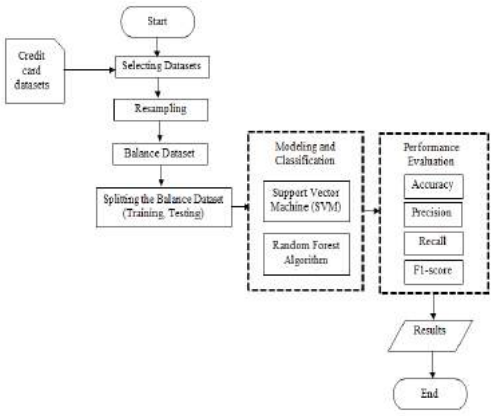


Figure 5. The Conceptual Architecture for Credit Card Fraud Detection with SVM and RFA

In a credit card fraud detection system, first of all, there must be credit card datasets. The datasets were used from Kaggle. Both of those two datasets are unbalanced datasets. The features of unbalanced datasets do not perform well during classification. Therefore, it is necessary to resampling the unbalanced dataset to become a balanced dataset. In resampling, there are oversampling and under-sampling. In this system, random under-sampling is used. In random under-sampling, rows are randomly selected to delete the majority class from the dataset. This has the effect of reducing the numbers of rows in the majority class. This process can be repeated until the desired class distribution is achieved, such as an equal number of rows for each class. This approach may be more suitable for those datasets where there is a class imbalance although a sufficient number of rows in the majority class.

After under-sampling, a balance dataset is obtained. In dataset 1 and 2, the total number of transactions are 948 and 600 respectively. And then, the balance dataset is split into two categories as training and testing for comparing and analyzing the dataset. After splitting the dataset, support vector machine (SVM) and random forest algorithm (RFA) are modeled and classified. After modeling the SVM and RFA, a confusion matrix is obtained. The performance of SVM and RFA are evaluated based on the confusion matrix. Finally, the performance results and execution time of two algorithms compare on two different credit card datasets.

3.3. Support Vector Machine (SVM)

Support Vector Machine (SVM) is a supervised learning algorithm that is used for both classifications and regression problems. It is used for classification problems in Machine learning. SVM classifies the two classes using a hyperplane. This hyperplane has the largest margin in a high dimensional space to separate given data into classes. The margins between the two classes represent the longest distance between the closest data points of those classes.

3.4. Random Forest Algorithm

It is a machine-learning technique that constructs multiple decision trees. The final decision is made based on the outcome of the majority of the decision trees. It can be used for both regression and classification purposes. But this algorithm is mainly used for classification purposes. This algorithm creates decision trees on the sample data and gets the prediction from each sample data. It is called as ensemble method. This algorithm is better than the single decision tree because it reduces over-fitting by averaging the results.

3.5. Resampling (Oversampling and Undersampling)

This technique is used to upsample or downsample the minority or majority class. When

we are using an imbalanced dataset, we can oversample the minority class using replacement. This technique is called oversampling. Similarly, this technique can randomly delete rows from the majority class to match them with the minority class which is called undersampling. After sampling the data, we can get a balanced dataset for both majority and minority classes. So, when both classes have a similar number of records present in the dataset, we can assume that the classifier will give equal importance to both classes.

3.6. Confusion Matrix

It is a specific table that is used to measure the performance of the algorithm. It is used to summarize the performance of the classification algorithm. It shows the error in the performance of the algorithm in the form of a matrix hence it's called an error matrix. The matrix is based on actual and predicted parameters.

Table 3. Confusion Matrix

	Non-fraud (Predict)	Fraud (Predict)
Non-fraud (Actual)	TN	FP
Fraud (Actual)	FN	TP

True Positive (TP): The Fraud cases that the model predicted as "Fraud"
False Positive (FP): The Non-fraud cases that the model predicted as "Fraud"
True Negative (TN): The Non-fraud cases that the model predicted as "Non-fraud"
False Negative (FN): The Fraud cases that the model predicted as "Non-fraud"

Accuracy: Accuracy is one way to measure how well an algorithm can classify a data point. Based on all the data points, accuracy refers to the number of correctly predicted points.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

Precision: The precision measures the number of observations correctly predicted out of those that are all positive.

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2)$$

Recall: The recall is the proportion of correctly predicted positive observations to all observations in a class.

$$\text{Recall} = \frac{TP}{TP+FN} \quad (3)$$

F1-score: It is a weighted average of Precision and Recall forms the F1 score. As a result, this score accounts for both false positives and false negatives.

$$\text{F1-score} = \frac{2*recall*precision}{recall+precision} \quad (4)$$

4. Experimental Results

This section shows the details and results of the experiments. The system performs the accuracy results on the training dataset and testing dataset according to split by 0.2, 0.3, and 0.4 respectively. The table 4 shows the performance results of dataset 1. For splitting amount 0.2, 0.3, and 0.4, RF is better than SVM in terms of Accuracy, Recall, F1-score and Execution Time. SVM is better than RF if only one Precision.

Table 4. Performance Comparison for dataset1

Algorithm	Splitting Amount	Performance Comparison for dataset 1				
		Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Execution Time (second)
SVM	0.2	0.904	0.991	0.813	0.893	44.293
RF		0.941	0.972	0.91	0.938	0.485
SMV	0.3	0.897	0.981	0.81	0.886	47.767
RF		0.941	0.975	0.905	0.937	0.424
SVM	0.4	0.893	0.981	0.807	0.885	31.64
RF		0.93	0.959	0.896	0.931	0.437

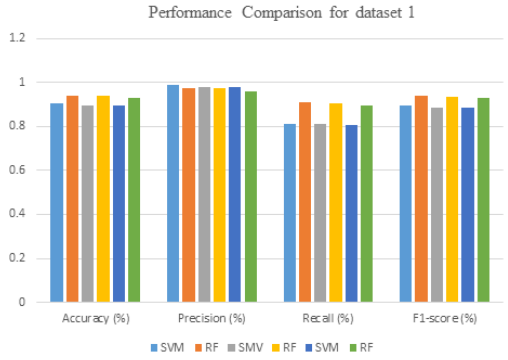


Figure 6. Performance Comparison for dataset 1

The figure 6 describes the performance comparison for dataset 1.

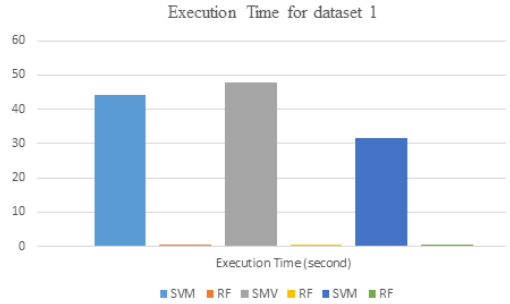


Figure 7. Execution Time for dataset 1

The figure 7 describes the execution time for dataset 1. The execution time of RF is faster than SVM.

Table 5. Performance Comparison for dataset2

Algorithm	Splitting Amount	Performance Comparison for dataset 2				
		Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Execution Time (second)
SVM	0.2	0.677	0.685	0.64	0.659	122.328
RF		0.764	0.754	0.776	0.764	0.394
SMV	0.3	0.662	0.697	0.619	0.652	142.794
RF		0.744	0.758	0.737	0.747	0.462
SVM	0.4	0.681	0.687	0.636	0.66	104.192
RF		0.743	0.746	0.721	0.731	0.383

The table 5 shows the performance results of dataset 2. For splitting amount 0.2, 0.3, and 0.4, RF is better than SVM in terms of Accuracy, Recall, F1-score and Execution Time. SVM is better than RF if only one Precision.

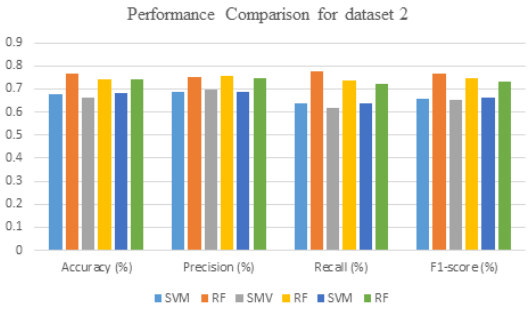


Figure 8. Performance Comparison for dataset 2

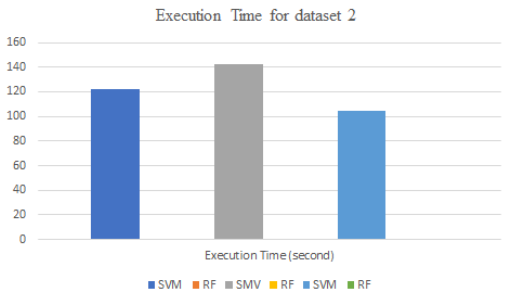


Figure 9. Execution Time for dataset 2

The figure 8 describes the performance comparison for dataset 2. The figure 9 describes the execution time for dataset 2. The execution time of RF is faster than SVM.

5. Conclusion

In this paper, the machine learning classification techniques Support Vector Machine (SVM) and Random Forest Algorithm (RFA) are used. The analysis results in the conclusion that RFA works better and takes less time to execute than SVM. In terms of Accuracy, Recall, and F1-score, RFA outperforms SVM algorithms. The precision results of SVM are better than RFA in datasets 1 and 2. By comparing two credit card datasets, it is found that dataset 1 has better than the performance of dataset 2. To conclude, RFA is more effective than SVM at detecting credit card fraud.

References

[1] Shoe Kumar, Vinit Kumar Gunnjan, Mohd Dilshad Ansari and Rashmi Pathak (2022) “Credit Card Fraud Detection Using Support Vector Machine”.

[2] B.N. V Madhubabu, T. Vyshnavi, K. Ashok (2021) “Credit Card Fraud Detection Algorithm using Decision Tree based Random Forest Classifier”.

[3] Angela Makolo and Tayo Adeboye (2021) “Credit Card Fraud Detection System using Machine Learning”.

[4] Kartik Madkailar, Manthan Nagvekar, Preity Parab, Riya Raikar, Supriya Pathi (2021) “Credit Card Fraud Detection System”.

[5] D. Tanouz, R Raja Subramanian, D. Eswar, G V Paarmeswawa Reddy, A. Ranjith Kumar, CH V N M (2021) “Credit Card Fraud Detection Using Machine Learning.

[6] Dongfang Zhang, Basu Bhandari, Dennis Black (2020) “Credit Card Fraud Detection Using Weighted Support Vector Machine”.

[7] M. Suresh Kumar, V. Soundarya, S. Kavitha, E.S. Keerthika, E. Aswini (2019) “Credit Card Fraud Detection using Random Forest Algorithm”.

[8] Shiyan Xuan, Guanjun Liu, Zhenchuan Li, Lutao Zheng, Shuo Wang, Changjun Jiang (2018) “Random Forest for Credit Card Fraud Detection”.

[9] Jasmina Novakovic, Suzana Markvic (2020) "Classifier Enesmbles for Credit Card Farud Detction".

Comparison of Classification Algorithms for Breast Cancer Prediction

Khin Swe Win Latt, Yi Mar Myint
University of Computer Studies, (Monywa)
khinswewinlatt@gmail.com, yimarmyint.ucsmonywa@gmail.com

Abstract

Data mining methods are being used more frequently on medical data to find trends and patterns that can help with diagnosis and decision-making. Data classification is an essential data mining technique. It is the process of identifying a set of models that describe and differentiate between data classes or concepts. Then, these models are used to predict the class of objects with unknown class labels. In this paper, the K-Nearest Neighbor (KNN), Support Vector Machine (SVM), and Decision Tree algorithms are used on the dataset taken from the UCI repository. In this work, KNN, SVM, and decision tree algorithms are compared for the diagnosis of Wisconsin breast cancer. The performance evaluation of the proposed methods has been compared with accuracy, precision, and recall.

Keywords: Classification; K-Nearest Neighbor, Decision Tree, Support Vector Machine, Breast Cancer Prediction

1. Introduction

The WHO reported that breast cancer is the most common cancer among women globally. It is the second most dangerous cancer after lung cancer [1]. Breast cancer is the abnormal growth of breast cells in women and rarely in men. A number of variables, including genetics, family history, obesity, and lifestyle, can affect the chances of developing breast cancer. Early diagnosis is critical for breast cancer [13]. If chances of cancer are predicted at an early stage, then the survival chances of a patient may increase. Using machine learning algorithms for the prediction of abnormal tumors is one way to describe breast cancer. The system collects a breast cancer dataset from the UCI Machine Learning Repository. In the dataset, there are 699 instances and 10 attributes, with two classes: 2-benign and 4-malignant. Benign for not cancer, malignant for cancer. These numbers were assigned by the pathologist based on their characteristics. A high value indicates a greater chance of malignancy. The study lists breast cancer as one of the main issues (benign and malignant tumors). One of the most cancerous tumor types, a benign tumor, does not invade nearby tissue or spread throughout the host. Malignant tumors are more dangerous and can be fatal to people. It is referred to as breast cancer. When breast tissue cells grow abnormal cells, a malignant tumor is created. It is critical to detect breast cancer in its early stages with advanced techniques and equipment.

Classification is the most fundamental and crucial task, and it is a prerequisite for data mining and machine learning. Classification methods are crucial for research and practical application. Researchers in a wide range of disciplines, such as pattern recognition, statistics, vision analysis, disease diagnosis, and other areas, have proposed a number of classification methods. Several classification algorithms, such as K-Nearest Neighbor, Decision Tree, and Support Vector Machine are used in this paper. The classification of the breast cancer dataset and the effect of these methods were estimated using confusion metrics like accuracy, recall, and precision.

The rest of this paper is arranged as follows: Section 2 reviews the related work. Section 3 describes the proposed system methodology. In Section 4, a system flow diagram is explained. In Section 5, a dataset description is presented. Section 6 presents the analysis and experimental findings. The study's conclusion is the last.

2. Related Work

Classification is one of the most important and fundamental tasks in machine learning and

data mining. Using data mining and machine learning on various medical datasets, numerous studies have been done to categorize breast cancer.

Ch. Shravya et al. [2] researched the prediction of breast cancer using supervised machine learning techniques. The SVM, KNN, and Logistic Regression algorithms were used by the author, and the output of various classifiers was compared and evaluated.

Nurul Amirah Mashudi et al. [3]: "Comparison of some machine learning techniques in breast cancer classification." In their research, the authors found that, when compared to the other methods, the ensemble classification method had the highest accuracy.

N. Yusoff and F. Kabir Ahmad [4] used a random forest classifier to classify breast cancer types based on fine needle aspiration biopsy data. The algorithm has been tested on secondary FNA data with a variety of features.

In the study by Dr. S.N. Singh and Shivani Thakral [5], the performances of supervised learning classifiers such as Naïve Bayes, Bayesian logistic regression, Simple Cart, and J48 are compared to find the best classifier in breast cancer datasets. The experimental result showed the highest accuracy and lowest error rate.

Bharat, A., et al. [6] experimented on the Wisconsin Breast Cancer (original) dataset. They used four different machine learning algorithms: Support Vector Machine, Naïve Bayes, K-nearest neighbor, and C4.5. To identify the best classification accuracy, they compared the efficiency and efficacy of various algorithms in terms of accuracy, precision, sensitivity, and specificity. SVM was able to illustrate its strength in terms of efficiency and effectiveness based on recall and accuracy.

Asri et al., machine learning algorithms for the risk prediction and diagnosis of breast cancer have been compared. The Wisconsin Breast Cancer dataset is used to predict and diagnose by using Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Naive Bayes (NB), and Decision Tree (C4.5). When the analysis results are compared, the SVM classification method gives the results with the highest accuracy and lowest error rate [7].

Ahmed et al. conducted a study on a similar study researching Machine Learning models to diagnose breast cancer [12]. They had access to a bigger dataset with 1189 instances. The SVM classification model predicts breast cancer recurrence with the lowest error rate and maximum accuracy.

3. Classification Method

In machine learning, there are two steps: the learning step and the prediction step. In the learning step, the provided training data is processed and a model is developed. In the prediction step, a model is developed to predict the feedback from the given data. In data analysis, data samples are classified by a group of attributes. The process of classifying includes building a classifier model that is trained on a set of training data where each data point has already been assigned to the associated class. This develops a model of the distribution of class labels. Then, whether the values of the features are known or unknown, they are used to categorize new data.

3.1. K-Nearest Neighbor

K-nearest neighbor (KNN) is a supervised machine learning technique used for classification and regression. The KNN algorithm is used to predict a data set's class or property [9]. Based on learning by analogy, the nearest neighbor classifier. The instances are described by n-dimensional numeric attributes. Every tuple in an n-dimensional space corresponds to a point. Euclidean distance is used to define the closest neighbors, dist (X, Y) [10]. The Euclidean distance between two points or tuples, X = ($x1$, $x2$,.., xn) and Y= (y1, y2,.., yn), is

$$d(X,Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (1)$$

Algorithm

- The dataset is entered and split into a training set and a testing set.
- Calculate the separation between a chosen instance and the training set using the testing sets.
- List distances in ascending order.
- The samples are the most common class of the three first training samples (k = 3).

3.2. Decision Tree

A decision tree is the most powerful and widely used machine learning tool for classification and prediction. A decision tree is like a tree's structure; for detecting breast cancer, its leaf nodes are classified as benign and malignant. And then certain rules are established to check if the tumor is benign or malignant. For this reason, the C4.5 decision tree can be used for classification and is often referred to as a statistical classifier. The C4.5 algorithm, which also chose attributes from a dataset with the highest information gain for splitting, improved the ID3 algorithm [10]. The gain ratio is used by C4.5 to divide the dataset according to the goodness function. Information (entropy) is required to classify any arbitrary tuple in D.

$$Info\ (D) = \ -\sum_{i=1}^{m} p_i\ \log_2(p_i) \qquad (2)$$

Where p_i is the probability that an arbitrary tuple in D belongs to class C_i {i=1, 2, 3,.., m} and is estimated by p_i = |C_i , D|/|D|.

$$Info_A(D) = \ \sum_{i=1}^{v} \frac{|D_j|}{|D|} * \ Info(D_j) \qquad (3)$$

Where $Info_A(D)$the expected information of each attribute in data D and v is the type of data in that attribute. Information gained by branching on attribute A.

$$\text{Gain (A) = Info (D) - } Info_A(D) \qquad (4)$$

In other works, Gain (A) tells us how much would be gained by branching on A. Knowing the values of A will reduce the amount of information that is needed, as expected. The information gain measurement seeks to test several possible values.

$$SplitInfo_A(D) = -\sum_{j=1}^{V} \frac{|D_j|}{|D|} * \log_2 \frac{|D_j|}{|D|} \quad (5)$$

Where SplitInfo (A) is the expected split information required to classify a tuple from D based on the partitioning by A. The attribute with the maximum gain ratio is selected as the splitting attribute.

Equation 6 may be used to compute the gain ratio for each attribute.

$$GainRatio\ (A) = \ \frac{Gain(A)}{SplitInfo_A(D)} \qquad (6)$$

3.3. Support Vector Machine

Support Vector Machine (SVM) is a supervised and linear machine learning method that is commonly used to solve classification problems. Support vector machines also support a kernel method called the kernel SVM, which enables us to handle non-linearity. An SVM creates a hyperplane (the following figure [11]) in multidimensional space to separate different classes. To reduce classification errors, the best hyperplane is generated iteratively. The formula for a separating hyperplane is

$$W.X + b = 0 \qquad (7)$$

Where W is a weight vector, $W = \{w_1, w_2, \dots, w_n\}$; n is the number of attributes; and b is a scalar.

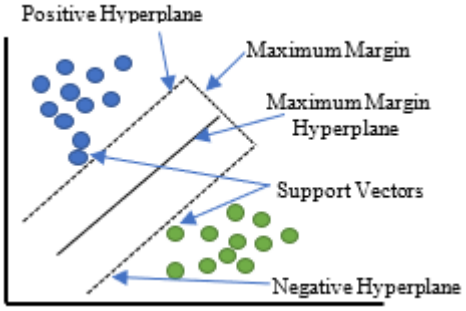


Figure 1. Support Vector Machine

Figure 1 explains that the purpose of a support vector machine is to find a maximum marginal hyperplane (MMH) that divides a dataset into classes as evenly as possible. The MMA is the training pair of either class that is closest to it in terms of distance [10]. On the lines from both classes, the SVM algorithm determines the closest point. Support vectors are the names for these points. It helps find the best line or decision boundary; this best boundary or region is called a hyperplane. The distance between the vectors and the hyperplane is called the "margin," and the purpose of SVM is to maximize this margin. The hyperplane with the maximum margin is called the "better hyperplane." Algorithms for support vector machines employ a set of mathematical functions known as the kernel. The kernel's task is to receive data as input and change it into the necessary form. Various SVM algorithms use various kinds of kernel functions, such as linear, radial basis functions (RBF), and sigmoid.

4. System Flow Diagram

Figure 2 shows flow the diagram for the comparison of classification algorithms to predict breast cancer. The system takes the breast cancer dataset as input and produces the predicted results as output. In the preprocessing step, there are 16 instances that have a single missing attribute value, which are denoted by a question mark ('?'). These values are removed by the system because it can't consider missing values. The system employs the three classification algorithms named KNN, decision tree, and SVM to compare their accuracies. For comparison, the system uses three evaluation metrics, including accuracy, recall, and precision.

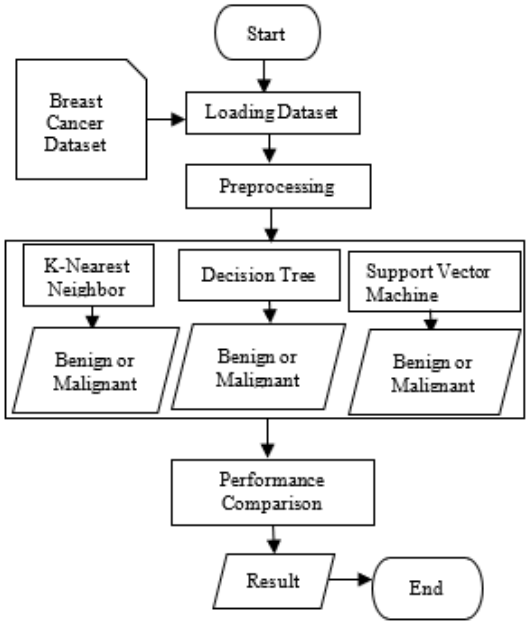


Figure 2. System Flow Diagram

4.1. Data Preprocessing

Data preprocessing refers to the process of converting unstructured data into structured data as well as removing null and missing values from a dataset. In this dataset, there are 699 instances including 16 missing values, which are denoted by a question mark ('?'). In the preprocessing stage, missing values cannot be considered, so these values are removed from the system. The system is working on the remaining 683 instances.

5. Dataset Description

In this study, the Wisconsin Breast Cancer (WBC) dataset was collected from the UCI Machine Learning Repository [8]. The breast cancer dataset has 699 instances and 10 attributes. The pathologist rates each type of behavior on a scale of 1 to 10 for each one. A large value represents a greater chance of malignancy. The attributes in the breast cancer dataset are as follows:

Table 1. Dataset description

ID	Attribute Name	Value
1	Clump thickness	1-10
2	Uniformity of cell size	1-10
3	Uniformity of cell shape	1-10
4	Marginal adhesion	1-10
5	Single epithelial cell size	1-10
6	Bare nuclei	1-10
7	Bland chromatin	1-10
8	Normal nucleoli	1-10
9	Mitoses	1-10
10	Class	2 for Benign or 4 for Malignant

The information is divided into two categories: benign (doesn't spread to other bodies or encroach on nearby tissue); and malignant (invades surrounding tissue).

6. Experimental Results and Comparison

The number of test records that are correctly and incorrectly predicted by the classification model serves as a measure of the model's performance [3]. A confusion matrix is used to calculate the test record counts, as shown in Table 2. The confusion matrix is used to calculate the measurements of accuracy, recall, and precision.

True Positive (TP): The model correctly predicted that it was malignant.
False Positive (FP): The model wrongly predicts that the benign is malignant.
True Negative (TN): The model correctly predicts that it is benign.
False Negative (FN): The model wrongly predicts that the malignancy is benign.

Table 2. Confusion matrix

		Predicted Class	
		Benign	Malignant
Actual Class	Benign	TN	FP
	Malignant	FN	TP

The following is the equation for precision, recall, and accuracy.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (8)$$

$$Recall = \frac{TP}{TP+FN} \quad (9)$$

$$Precision = \frac{TP}{TP+FP} \quad (10)$$

Table 3. Performance comparison of three classifiers (training 70%, testing 30%)

Method	Accuracy (%)	Recall (%)	Precision (%)
Decision Tree	0.96	0.93	0.95
KNN (k=3)	0.99	0.97	0.97
SVM	0.99	1.00	0.95

Table 3 shows the performance comparison of the three classifiers, such as Decision Tree, KNN, and SVM with training set at 70% and testing at 30%. In KNN, k = 3 is set, and in SVM, a linear kernel function is used in this experiment. The system is analyzed with different k values for KNN, different training sizes for the three classifiers, and linear kernel functions for SVM. For each model, the confusion matrix was calculated in order to obtain a significantly better class prediction. The confusion matrices for all three machine learning techniques are displayed in Table 4. Out of the three machine learning techniques, SVM can correctly predict that malignancy is malignancy.

Table 4. Confusion matrix of the three classifiers (70%, testing 30%)

Methods	Predicted Class		Actual Class
	Benign (B)	Malignant (M)	
Decision Tree	157	2	B
	3	39	M
KNN (K=3)	159	1	B
	1	43	M
SVM	158	2	B
	0	44	M

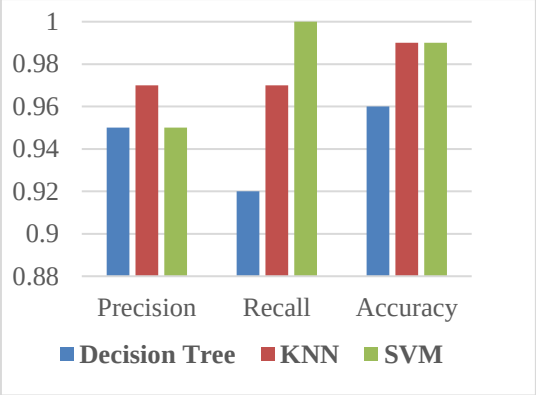


Figure 3. Comparison of performance evaluation for classification methods

Figure 3 shows the performance comparison of the three classifiers, named decision tree, KNN, and SVM. It is experimented on in a 70% training and 30% testing set. By comparing classification results, the system found that SVM with a linear kernel is better than other classification algorithms for the breast cancer classification dataset, which was split 70%/30% into train and test.

Table 5. The performance for even k values of KNN using the confusion matrix

K	Actual Class	Predicted Class	
		B	M
2	B	160	0
	M	4	40
4	B	160	0
	M	1	43
6	B	160	0
	M	1	43
8	B	160	0
	M	0	44
10	B	160	0
	M	0	44
12	B	159	1
	M	0	44
14	B	160	0
	M	0	44

Table 6. Performance for even k values of KNN

K	Accuracy (%)	Recall (%)	Precision (%)
2	0.98	0.90	1.00
4	0.99	0.97	1.00
6	0.99	0.97	1.00
8	1.00	1.00	1.00
10	1.00	1.00	1.00
12	0.99	1.00	0.97
14	1.00	1.00	1.00
Average	0.99	0.97	**0.99**

Table 7. The performance for odd k values of KNN using the confusion matrix

K	Actual Class	Predicted Class	
		B	M
3	B	159	1
	M	1	43
5	B	159	1
	M	1	43
7	B	159	1
	M	0	44
9	B	159	1
	M	0	44
11	B	159	1
	M	0	44
13	B	159	1
	M	0	44
15	B	159	1
	M	0	44

Table 8. Performance for odd k values of KNN

K	Accuracy (%)	Recall (%)	Precision (%)
3	0.99	0.97	0.97
5	0.99	0.97	0.97
7	0.99	1.00	0.97
9	0.99	1.00	0.97
11	0.99	1.00	0.97
13	0.99	1.00	0.97
15	0.99	1.00	0.97
Average	0.99	**0.99**	0.97

Table 6 and 7 shows performance for even and odd values of KNN. The system did the analysis on different k values of the k-NN classifier. The system is analyzed for even and odd k values within a range of 2 to 15. It is made by splitting a 70% training and 30% testing dataset. When a comparison is made between the averages of even and odd K values, the accuracies of both are the same. The recall rate of odd values is higher than that of even values, but in the precision rate, even values are more common than odd values.

7. Conclusion

In this paper, the three classification algorithms' comparative analysis on problems with breast cancer prediction is presented. The algorithms can assist in proper treatment methods for a patient diagnosed with breast cancer. The system analyzed even and odd K values within a range from 2 to 15. When a comparison is made between the averages of even and odd K values, the accuracies of both are the same. The recall rate of odd values is higher than that of even values, but in the precision rate, even values are greater than odd values. By comparing classification results, the system found that SVM with a linear kernel function is better than that of other classification algorithms for the breast cancer classification dataset, which was split 70%/30% into train and test. If a person with the disease is wrongly thought to be absent, the recall rate will increase. Simultaneously, the accuracy of the system decreases. It has been found that the recall rate is more important in terms of recall and precision. The recall rate of SVM is better than that of other classification methods, so SVM is found to be better. The analysis results showed

that the SVM algorithm is more effective for the prediction of breast cancer datasets.

Acknowledgement

I would like to express my gratitude to the University of Computer Studies (Monywa) for allowing me to do this research work. Thanks to Daw Yi Mar Myint for her valuable comment, feedback, support, and encouragement.

References

[1] Ghebreyesus, T. A.: *Breast Cancer*, World Health Organization, 26 March 2021, https://www.who.int/news-room/fact-sheets/detail/breast-cancer

[2] Ch. Shravya, K. Pravalika, and Shaik Subhani, "Prediction of Breast Cancer Using Supervised Machine Learning Techniques," ISSN: 2278-3075, Volume-8, Issue-6, April 2019, pp. 1106–1110.

[3] Nurul Amirah Mashudi et al., "Comparison on Some Machine Learning Techniques in Breast Cancer Classification," IEEE EMBS Conference on Biomedical Engineering and Sciences (IECBES), 2021, pp. 499.

[4] F. Kabir Ahmad and N. Yusoff, "Classifying breast cancer types based on fine needle aspiration biopsy data using a random forest classifier," December 2013, DOI: 10.1109/ISDA.2013.6920720.

[5] S.N. Singh and S. Thakral, "Using Data Mining Tools for Breast Cancer Prediction and Analysis," 4th International Conference on Computing, Communication, and Automation (ICCCA), 2018, pp. 1-4.

[6] A. Bharat: Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis, (2018), pp.4.

[7] H. Asri, H. Mousannif, H. Al Moatassime, and T. Noel, "Using machine learning algorithms for breast cancer risk prediction and diagnosis," *Procedia Computer Science, 83*, pp. 1064-1069.

[8] "UCI Machine Learning Repository: Breast Cancer Wisconsin Dataset," [Online] Available: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29. [Accessed: 29-Dec-2015].

[9] M. Amrane et al., "Breast Cancer Classification Using Machine Learning," 78-1-5386-5135-3/18/$31.00, 2018 IEEE, p. 4.

[10] Jiawei Han, Micheline Kamber, and Jian Pei, "Data Mining Concepts and Techniques", 3rd Edition, Morgan-Kaufmann-2011, ISBN 978-0-12-381479-1, pp. 327-436.

[11] Bashir Alam, "Implementation of Support Vector Machine (SVM) Using Python", https://hands-on.cloud/implementation-of-support-vector-machine-svm-using-python/#h-overview-of-svm-algorithm

[12] Mohamed M., Cesa G., Cohen T. S., and Welling M.: "A Data and Compute Efficient Design for Limited-Resources Deep Learning", arXiv, July (2020), Available: http://arxiv.org/abs/2004.09691

[13] M. Jadhav et al., "Breast Cancer Prediction using Supervised Machine Learning Algorithms", Oct 2019, e-ISSN: 2395-0056, p-ISSN: 2395-0072, [Online]. Available: http://www.irjet.net

Renewable Vehicles Registration System using Information Retrieval Process

Khin Moe Wai, Hsu Mon Kyi
University of Computer Studies, Taunggyi
khinmoewai@ucstgi.edu.mm, hsumonkyi@ucstgi.edu.mm

Abstract

Nowadays, vehicle registration has been existence in all countries. Vehicle registration, management and the control of the vehicle license is a composite process. The current process is manually and numerous problems have been encountered. Therefore, there is need to improve the current situation. According to their existing problems, the Renewable Vehicles Registration System (RVRS), which uses an online system, has been implemented for the Directorate of Road Transport Administration Department. The (RVRS) system uses Model View Controller (MVC) of Laravel framework for collecting, storing and searching vehicle data. By using this (RVRS), the office staff of the Road Transport Administration Department could not waste spend their time when they make the list of registration vehicles, extending or renewing and on the part of the people, there is no need to come to the office, and they can easily renew their vehicle. The RVRS can easily view vehicle that are due to be renewed, and also check if they are prohibit of the police. This system applies information retrieval process to find out the relevant information. The Renewable Vehicles Registration System to accurate search the vehicle information by using Boolean and Wildcard retrieval, the primary retrieval method in this system that procedures of searching, renewal vehicles of notification, prohibit vehicles from police case list and retrieving recorded data and information from database.

Keywords: Information Retrieval, MVC framework, vehicle registration, renewal vehicles

1. Introduction

Vehicle registration is the main aim to see that road safety. As the number of vehicles increases, so does the rate of vehicle collisions and then unregistered vehicles increase. Therefore, vehicle registration becomes an important field in order to easily find the owner of the vehicle. The vehicle registration RVRS system provides ease and accurate access of vehicle owner's information by applying Information Retrieval (IR). This system is to provide the details of the list of the vehicles. The RVRS has two sections, admin and user. The main function of the admin is to vehicle renewal notice is issued and vehicles applying for renewal are taxed and the vehicles are renewed. The main function of the user is to if you apply for a vehicle renewal and got permission to renew it, you can renew it by paying tax.

In addition, a renewed vehicle document includes a Quick Response (QR) code containing about the vehicle and its owner information with embeds. When the vehicle travels, the relevant departments check the document containing the QR code, registered vehicle information will find by scanning the vehicle and the QR code. The RVRS system is an online system to access the internet requires government assistance in registered vehicles. Moreover, public service is importance. RVRS system will be able to download or export excel & CSV files directly from the database in Laravel framework. In order to find out the vehicles that have booked and taxed on a daily basis, the files can be downloaded and stored, and the data can be exported and stored as a CSV file.

This system refers to the relevant information retrieval process, searching procedures and notification of renewed vehicles. It includes prohibit vehicles from police registers and retrieving recorded information and data from a file or database. Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers) [2]. The RVRS system has tested for accurate information using

valid input data from the Directorate of Road Transport (Pyin Oo Lwin) office. RVRS system uses Boolean operator and Wildcard Searching in the main document database to help RVRS system users find information easily.

2. Related Works

In this section, we will review the research paper on the information retrieval technique applied.

Abass O. Adisa [1] introduced an improved vehicle registration and licensing system tested using correctly chosen input data to ensure adequate reliability, accuracy and evaluated its compliance with the specified requirement the system.

Arash Habibi Lashkari [2] proposed to explain IR methods and evaluate them from a Boolean simple method with two viewpoints. They proposed a simple method for ranking terms and documents on information retrieval and implement the method and checking the result. In this system, Visual Basic on Oracle database and analyze the result is implement Query based or non-Query based algorithm, and Ad-Hoc filtering or browsing method.

Muhammad Bello Aliyu [3] introduced the technique of Boolean search string is proposed in detail along with the evaluation of the effectiveness of the technique. The information retrieval for the search on several relevant documents was searched using the Boolean search strings technique. However, the free text query and the Boolean search strings results analyze the search for the documents. It shows that the Boolean search strings return more relevant documents than the free-text query.

To the best of our knowledge, the RVRS system is searching for information that can find the exactly described data with Boolean and Wildcard method are search techniques, in which searching is conducted from different search processes. These techniques are most widely used in an information retrieval system.

3. Background Theory

Information is transmitted or received knowledge concerning a special. Retrieval refers to searching through keep information to find information relevant to the task at hand. Because of this, information retrieval (IR) deals with the representation, storage, organization, and approach to information items. Here, documents, Web pages, online catalogs, structured records, and multimedia objects include types of information items. Text and searching are indexed chief goal of the IR is indexing for useful documents in a collection. Libraries were among the first institutions to adopt IR systems for retrieving information [4].

Information retrieval is to provide the user with the “best possible” information from a database. That meets the user’s needs to identify information by information retrieval system. The IR system assists the users in finding the information they require but it does not explicitly return the answers to the question. It notifies concerning the existence and location of documents that might contain of the needed information. Information in a database can only be searched and retrieved when search algorithms and procedures provide a corresponding search mechanism [5]. The types of searches are Basic Search and Advanced Search. Basic search has Boolean operators, Phrase searching, Truncation (or) Wildcard searching, Proximity searching and Focusing (or) Limiting a search. Advanced search has Fuzzy searching, Weighted searching, Query expansion and Multiple databases searching Boolean search and Wildcard search are used in RVRS system.

Boolean Searches are simple words (AND, OR and NOT) used as conjunctions to combine or admit keywords in a search. Connect and define the relationship between the search terms use these. Combine two or more terms use the AND operator. The searcher to specify alternatives among the search terms are allowed the OR operator. Exclude the term from a set of resources use the NOT operator [4].

Wildcard search, is a search technique, in which, the search is showed for different forms of a word having the same common root. It is one of the most widely ratified methods in information retrieval system. In this technique, root word is taken with truncation mark and search is showed [4]. Wildcards are used to search for alternate spellings and differences on a root word. Wildcard characters cannot be used in place of the first letter of a word or within an exact phrase search, and word roots must contain at least three letters preceding a wildcard.

3.1. System Overview

Renewal is the online booking system in the RVRS system. RVRS has two sections, admin and user. The work of Admin functions is prohibited, including notification list, Check Booking list, Calculate Bills, vehicle renewal, transition, and vehicle search. The vehicle information output is stored in the database, which retrieves for prohibition list, notification list, and vehicle renewal. The database was implemented in Mysql database, and 15844 data of the Directorate of Road Transport (Pyin Oo Lwin) office are used in the system.

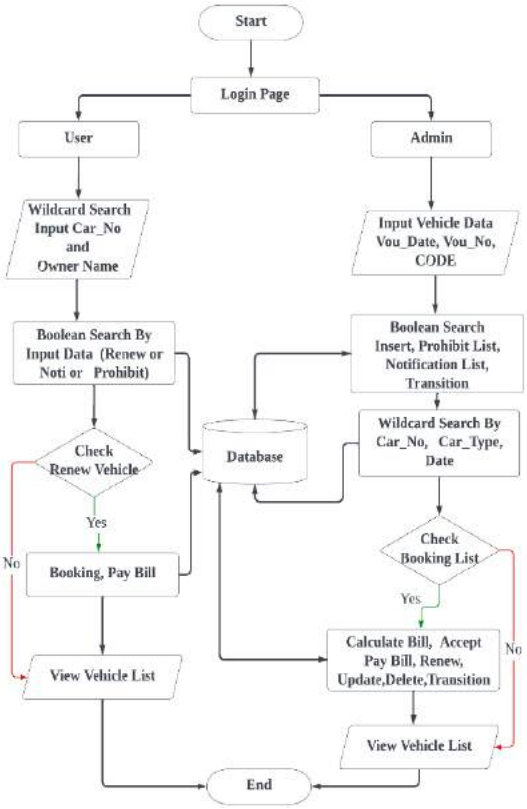


Figure 1. Flowchart of Renewable Vehicles Registration System

Admin will advertise banned vehicles that cannot be renewed. In addition, expired vehicles, also advertise vehicles that need to be renewed and vehicles that can renew three (3) months in advance. Vehicles to be renewed and police vehicles are notified using the Boolean Search method. When the admin calculates the tax to renew the vehicle, the vehicle inspection fee, vehicle renewal fee, vehicle types fee, vehicle certificate fee and holo fees must be calculated. In order to ensure security, when issuing a certificate for renewed vehicles, a certificate will be issued that includes the vehicle number, vehicle type, owner's name, registration number and vehicle expiration date in the QR code. The vehicle search permits the administrator to search for a vehicle number. The renewable vehicle registration system (RVRS) flow is shown in Figure 1.

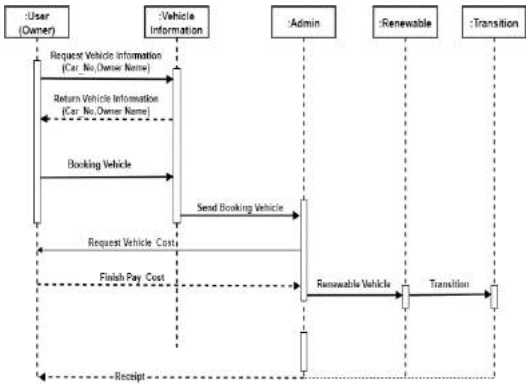


Figure 2. Sequence Diagram for Renewable Vehicle Process

Figure 2 illustrates renewable vehicle registration start the process from the sequence diagram user renew request to the end of renew process. In this sequence, the user will first search by vehicle number or vehicle registrant's name to renew the vehicle. Once the vehicle information is received and then start taking the booking to renew the vehicle. The admin will check if the vehicle in the booking list is a prohibited vehicle or a vehicle that needs to be renewed. After that, the tax will be calculated. If the user is notified of the right to renew the vehicle tax, the tax will be paid. The vehicle renewal and transition will be processed as soon as the admin receives proof of tax payment. After that, the owner of the vehicle will be notified of the extension. The system database design incorporates a relational data model. The following figure 3 illustrates database schema of the RVRS system.

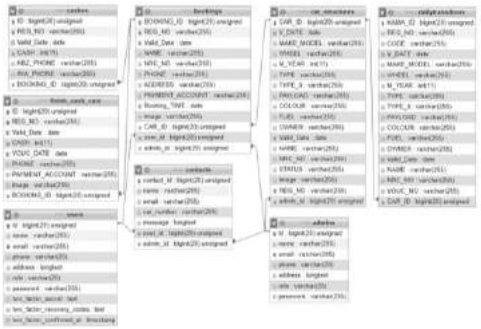


Figure 3. Database Schema of the RVRS System

3.2. Boolean and Wildcard Retrieval Method with RVRS

Information retrieval process is an important system. It is a process that facilitates when the user searches for relevant information in the document. Therefore, information retrieval processes such as Boolean search and Wildcard search are used in the RVRS system for notified vehicle information and prohibit vehicle lists. Before presenting the algorithm flow, the notations and symbols are described in Table 1.

Table 1. Notations and Symbols

Symbols	Descriptions
C_{no}	Vehicle Number
R_{VL}	Renewable vehicle list
C_{VD}	Valid date of car
CT_p	Personal car
CT_{lt}	Light truck car
C_{info}	Vehicle information
O_{Name}	Owner Name of car
P_L	Prohibited car lists

3.2.1. Notification Vehicle List with Boolean Search

Algorithm 1: Boolean Search Method for Notification Vehicle List

Parameters: C_{no}, R_{VL}, C_{VD}, CT_p and CT_{lt}
Begin
check car registration number is prohibited car or not
if (C_{no}) is Prohibited List) then
return no_renew( )
else
begin
$R_{VL} = C_{VD} \wedge (CTp \vee CT_{lt})$
//search renewable vehicle List information R_{VL} according to valid date C_{VD} and car types CTp, CT_{lt} using Boolean search and (∧), or(∨)
$C_{VD} = f(T_m)$

// valid date function f(Tm) within current date to three months

end

End

Boolean search is used in the Notification Vehicle List. In the notification vehicles list, the expiration date of the vehicle that is younger than the current date or more than 90 days is pulled out with the "AND" operator of the Boolean search. In addition, the types of vehicles were also checked by extracting the information using the "OR" Boolean search operator.

3.2.2. Prohibit List Retrieval Process with Wildcard Search

A wildcard search is used in the prohibited vehicle list. Wildcard Search is used when searching for a vehicle number or vehicle owner from the prohibit vehicle list. If we search owner name is 'myint*' then all the records will be retrieved where term 'myint' come out in full or part of the word.

Algorithm 2: Wildcard Search Method for Prohibit List

Parameters: C_{info} , C_{no}, O_{Name} , P_L
Begin
check car registration number is prohibited car or not
if (Cno_P_L) =true) then
begin
$C_{info} = C_{no} + O_{Name}$
//Prohibited car information C_{info} according to car number C_{no} and Owner name O_{Name}
O_{Name} = f(*Search)
// Wildcard * search function
end
End

4. Implementation of the RVRS system

RVRS system used Laravel framework of a php programming language. Laravel framework for development as it has a lot of superiority and gives a proper MVC structure to efficiently manage Databases, Backend and Frontend separately. The system uses a portal based on the MVC Framework for easy and fast data collection. These features made developing websites easier and most. The RVRS system uses information retrieval especially methods such as Boolean Search and Wildcard Search when searching and retrieving data.

4.1. Search Query Results

In this section, the two-search query result outputs were collected to analyze the query result.

In this RVRS system, 15844 vehicle list records are stored in the database. Figure 4 is data retrieved 8679 notified vehicle list from all

vehicle lists. When retrieving a notification vehicle list, this system is used the 'AND' and 'OR' operators from Boolean Search to list only notified vehicles.

Figure 4. Notified vehicle lists using Boolean Search Query

The next query is the wild card search method for a prohibited vehicle. The wildcard search is used to make it easier to search by vehicle number or owner name. Figure 5 was searched to find out how many vehicles there are with vehicle number 1R.

Figure 5. Prohibited vehicle lists using Wild Card Search Query by Car Number(1R)

The below table can be seen the relevant documents from 84 prohibited vehicle lists. From all records of 84, 5 relevant records by owner name (%Myint%) and 4 records by car number(1R) can be retrieved.

Table 2. Example Wild Card Search Query Results

Number of Prohibited List	Wild Card Query Result	Remark
84	4	Car No. (1R)
	5	OwnerName (%Myint%)

From the above table, it can be seen that irrelevant records from the total number of prohibited lists can be removed by Wild card search strings with relevant query results. However, different results can be seen that the different search query strings.

5. Conclusion

The information retrieval system will ensure reliable and robust services for the renewable vehicle's registration. RVRS system can also easily search for vehicles that are due for renewal, and you can also find out the list of prohibited vehicles. When requesting information from the relevant departments and when looking for the desired information, the information can be obtained quickly, saving time. There is no need to visit the office in person to renew the vehicle for the public, and renewal of the vehicle can be done at any time and place.

In the future, this system can extend the driver's license renewals, initial vehicle registrations, and mobile application services through the RVRS system.

References

[1] A.O.Adisa and S.I.Eludiora, "*An Improved Vehicle Registration and Licensing System*", FUOYE Journal of Engineering and Technology, vol.6, no.1, March. 2021.

[2] A.H.Lashkari, F.Mahdavi and F.Mahdavi, "*A Boolean Model in Information Retrieval For Search Engines*", International Conference on Information Management and Engineering (IEEE), 2019.

[3] M. B Aliyu, "*Efficiency of Boolean Search strings for Information Retrieval*", American Journal of Engineering Research (AJER), vol.6, no.11, pp.216-222, Nov. 2017.

[4] "*Information Retrieval System*" NIOS Class 12th Library and Information Science English Medium Solve Assignment TMA 2021.

[5] "*What is information retrieval used for?*" https://www.engati.com/glossary/information-retrieval

Information Management System for Middle School Level

Nway Htwe Aung, Khin Zezawar Aung, Cherry Phyo Wai
University of Computer Studies (Taunggyi), Myanmar
nwayhtweaung@ucstgi.edu.mm, khinzezawaraung@ucstgi.edu.mm,
cherryphyowai@ucstgi.edu.mm

Abstract

The Information Management System for Middle School Level offers a simple application for the management of students' information. Nowadays, technology is used to make jobs easier. Higher education and offices have started to use information technology. This system reduces teachers' and staffs' workloads, reduces time-consuming tasks. This system improves communication between the school, students, and parents. This system helps teachers and staff control the students' information data, exam results, and performance assessments. It uses a system to know parents about their children's Monthly Report Card (MRC) and Performance Assessment Record. This system converts children's marks from numbers to grades. In student information searching, it implements exchange sort algorithm and binary search algorithm.

Keywords: Information Management System, Education, Binary Search Algorithm

1. Introduction

Today's use of technology has greatly improved in all aspects and fields. In the same way, management has started using technology in the field of education. It is very important to properly manage and store information when using technology. Many educational institutions try to set up computer systems to manage students' data. Some universities and colleges are already using computerized system. However, at the Department of Basic Education, the use of computerized system in the office is rare and cannot be used in basic education schools. Organizations related to education are trying to manage students' information data using computers. This system is intended for use in basic education. In this system, parents can easily check their children's grades and performance assessments at any time. This system allows teachers and staff to easily (insert, update, and delete) students' information data, marks, and performance assessments. Marks are also automatically calculated as grades. Students' information data, grades, and performance assessment will be stored in the database. The information in the database will be retrieved and used as needed. When retrieving the data from the database, the first thing to do is sort it using exchange sort. In this system, binary search techniques are used to search a student's information, scores, and performance assessment by entering their ID number.

2. Related Work

S. Liu did the research on the application of the binary search algorithm using an RFID system, that is to say, to identify supermarket shopping information [4]. This RFID system was used in the supermarket shopping system, which networks information for identification. According to the research, it is very convenient to identify the articles by using the binary search algorithm in the RFID supermarket shopping systems. In this research, the application of RFID technology is evaluated for supermarket shopping system and utilized the anti-collision algorithm to classify upon the basis of the unique sequence code. This system focuses on networking, databases, and RFID.

Modified Binary Search Algorithm in 2014 by A.Chadha is explosive search [1]. This paper intends to modify the traditional binary search algorithm, where it inspects the existing input element with the middle element of the given element sets for each iteration. The linear search and binary search algorithms are mostly used in many search applications. Binary search with a complexity of O(log2 n) have a high time complexity. The binary search algorithm is

effective in reducing execution time and increasing efficiency. This paper proposes modifications to the binary search algorithm that improve on the worst cases of traditional binary search. The algorithm is more standardized compared to others, as it removes unnecessary relation at the fundamental stage itself. It can even be expanded to add the string domain.

Another search technique that was presented by G. Bagogun in 2019 is quadratic search [2]. This paper focuses on searching for some of the findings within a limited time in the linear search algorithm. In the linear search algorithm, a search will continue through an entire section to obtain the predetermined result. Sorting and searching are the two essential processes in computer science. The binary search algorithm is one of the most powerful search methods for functional purposes. On the other hand, a linear search is one of the most convenient search methods that can be utilized in the current situation.

3. Methodology and Dataset

This system provides details of students' information; scores, performance assessments can be supported to easily search and store. This system provides to manage middle schoolers information from basic education.

A data structure is specially designed to organize, process, retrieve, and store data. Many types of data structures are designed to organize data for a specific purpose. Data structures are simple for users to access and retrieve the data they need in many ways. Data structures are fundamental for the organization's ability to collect information. So, machines and humans can easily understand it.

Sorting is the process of arranging data in a particular order. Data structure makes it more convenient to reorganize the data or element in ascending or descending order lexicographically, numerically, or according to what the user defines. The importance of sorting is due to the fact that data searching can be greatly improved if data is stored in a sorted manner. Sorting can also be used to make data more readable. In this system, student ID numbers are sorted using exchange sort.

"Searching in data structure" means finding the predetermined information from the data set stored in computer memory. There are many types of data sets: graphs, trees, arrays, and linked lists. Moreover, searching in the data structure is to find the specific element from collected items. Searching in the data structure could be implemented by using algorithms to control the elements in the collected data structure. In this system, a binary search technique is used to search a student's information, scores, and performance assessments by entering their ID number.

3.1. Exchange Sort

The exchange sort is mostly the same as the bubble sort. But exchange sort is different from bubble sort. In exchange sort, the first element must be compared with each following element of the array and make changes if necessary. If the first change in the array is finished, take and compare the second element, which can be out of order, and make changes if necessary. The exchange sort algorithm takes and compares every element in the array and changes those elements that are in the wrong position, just like the bubble sort algorithm.

The exchange sort algorithm is:

```
Inputs:          n, SID[1],……SID[n]
Output:          Sorted keys in the array SID.
Algorithm:       Exchange Sort
Begin
        for (i=1; i<=n-1; i++)
                for (j=i+1; j<=n; j++)
                if(SID[j] < SID[i])
                        temp = SID [j]
                        SID[j]   = SID [i]
                        SID[i] = temp
                endif
End
```

The explanation of symbols in this algorithm are described in Table 1.

Table 1. The Symbols in exchange sort

No	Symbol	Description
1	SID	Student ID
2	n	Size of Student Information Array

3.2. Binary Search

Binary search is a well-equipped search algorithm that can retrieve data from sorted lists. So, the list must be sorted when the user searches for an element in the array by using the binary search algorithm. In a binary search, find the middle element and compare it with the desired element. If the match is correct, return and exist from the array. If the match is not correct and the desired element is greater than the middle element, continue to the right-hand side. And then, search the middle element in the right-hand side. If the desired element is smaller than the middle element, continue to the left-hand side and search for the middle element in the array. This process must be continued until the match is found.

The binary search algorithm is:

```
binarySearch (arr, x, low, high)
if low > high
    return False
else
    mid = (low + high) / 2
    if x == arr[mid]
        return mid
    else if x > arr[mid]
        return binarySearch (arr, x, mid + 1, high)
    else
        return binarySearch (arr, x, low, mid - 1)
```

The explanation of symbols in this algorithm are described in Table 2.

Table 2. The Symbols in Binary Search

No	Symbol	Description
1	arr	Sorted Student Information Array
2	x	Value to be searched
3	low	0
4	high	Size of Student Information Array

3.3. Dataset

The dataset used in this system is the 2022 academic year's middle schoolers' information from No. 1 Basic Education High School, Hopone Township. The proposed system uses 380 student information records.

3.3.1. Database Design

The proposed system uses three tables in database. These are student, mark, and event tables. The student table connects to the mark table and event table with one-to-many connections respectively.

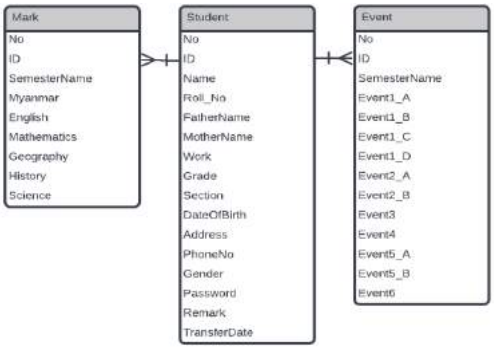


Figure 1. Database Flow Diagram

4. System Design and Implementation

This system is implemented for two roles. One role is admin, which includes both staff and teachers. Another role is user, which includes parents. The system administrator must enter their username and password when logging in. If admin login is successful, admin can manage (insert, update, and delete) students' information data, grades, and performance assessments. Then, admin can search student information by the student's ID. The proposed system uses binary search and exchange sort algorithms as searching techniques. The binary search algorithm can only search for sorted data. Before searching, the system needs to sort the data using the exchange sort algorithm. Now, schools report grades to parents as a mark of an exam. Admin can calculate a student's marks for grading from information searched by the student's ID. The user must enter their username and password when logging in. If user login is successful, user can view profile, grade and events. The system flow diagram is described below.

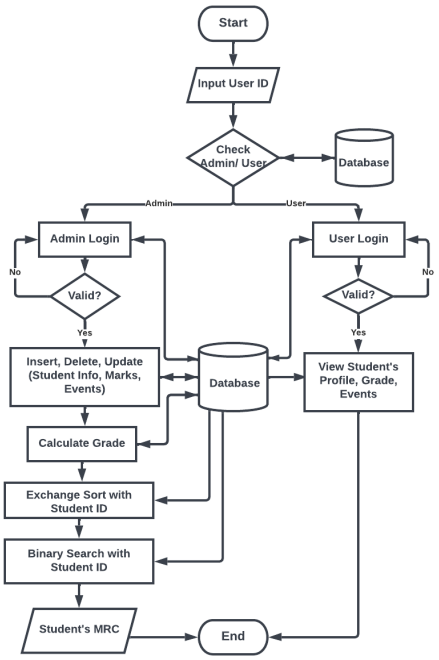


Figure 2. System Flow Diagram

Figure 3. All Student Form

In the all-student form, the admin will see the detailed information of all middle schoolers in the school. Admin can easily view student information and the number of students by grade, section and who have transfers and drop out.

Figure 4. Student Register Form

Figure 3 Admin can easily register a new student in the registration form. Admins can easily select grade, section, and date of birth when registering. This system will generate ID and Roll_No automatically. The student's information will be stored in the database when admin fills in the student's information correctly and clicks the Register button.

Figure 5. Search Student Form

In the search student form, admin can search by entering the student ID and clicking the search button. On the left side of the form, the admin can view student information details, and on the other side, the admin can view the mark by grade. The administrator can also view the exam results based on each subject's mark.

Figure 6. Edit Student Form

In this form, the admin can enter the ID and click the search button to edit student data. Admin can edit the student's status (transfer or dead) and add the transfer date. Edit correctly the student's information on the left side and click the "update student info" button. Edit correctly the student's mark on the right side and click on the "update mark" button. Click on the "view grade" button below, and the admin can view the student's result.

Figure 7. Add Mark Form

Admin enters the student's ID and semester in the add mark form, then fills out the student's mark and clicks the add mark button. Admin can show some information about the student on the left and mark information on the right. The student's mark will be stored in the database when admin fills in the student's mark correctly and clicks the "add mark" button. By clicking on the clear button, the admin can fill in the students' marks again according to the above steps.

Figure 8. Manage Events Form

Admins enter the student's ID and semester in the Manage Events form, then easily edit and delete the student's performance assessment by clicking the Update and Delete buttons.

Figure 9. User Home Page

The user home page form includes a grade form, personal assessment record form, profile form, and password change form for parents and students to view. Parents and students will only be allowed to view and not edit or delete.

5. Conclusion

This system is intended for middle school students. This system saves time, money, and space by reducing the amount of paperwork and manpower required. Collected information could be saved, updated, and accessed at any time.

The system performance of searching algorithm is measured by comparing the number of comparisons needed in the best and worst case as time complexity. When the target item is found at index one, the best case complexity of binary search and linear search is O(1). The worst-case time complexity for binary search is O(logN) comparisons, and for linear search, it is O(N) comparisons before the target item is found. Therefore, binary search algorithm is faster than linear search algorithm. Compared to binary search, linear search requires more time. Thus, the proposed system uses a binary search algorithm for searching student information by student's ID.

References

[1] A.Chadha, 2014. Modified Binary Search Algorithm. International Journal of Applied Information Systems, Volume 7-No2, Aprial 2014-www.ijais.org.

[2] G. Balogun,2019. A Modified Linear Search Algorithm.Vol.12, No.2, June 2019, pp.43-54.

[3] S.Bharamagoudar, 2013. Web Based Student Information Management System. International Journal of Advanced Research in Computer and Communication Engineering Vol.2, Issue 6, June 2013.

[4] S.Liu, 2019. The research of the application of the binary search algorithm of RFID system in the supermarket shopping information identification. Wu et al, EURASIP Journal on Wireless Communications and Networking (2019) 2019:27.

[5] S.Reddy, 2019. Student Management System. International Research Journal of Computer Science (IRJCS) Issue 06, Volume 6 (June 2019).

Library Management System for University of Medicine (Taunggyi)

Myat Thu Aung, War War Khaing
University of Computer Studies (Taunggyi)
myatthuaung@ucstgi.edu.mm, warwarkhaing@ucstgi.edu.mm

Abstract

The Library Management System (LMS) is a system that can retrieve the information of digital library. The librarians have to work within the allotted time for arranging, sorting books in the book shelf. LMS can assist the librarians to work easily. The users need not stand in a queue for a long period to return or borrow a medical book from the library. The LMS is designed with the basic featured such as librarian can insert, view, update and delete books and student's details in it. The authorized persons can only access the LMS system and login with user id and password. The librarian can smoothly activate the system without expert advice. The registration member can view book list and borrow the book on the LMS. Boolean search method is used in this system for book information retrieval. The system is implemented using web-based technologies which include C# and MSSQL and runs on Windows operating system.

Keywords: Boolean Search, LMS

1. Introduction

Library Management System (LMS) is the Web based System and all data everything is associated with Library can be managed from any device and any place because it is web-based Library Management System.

Library management system is a place where a collection of books and resources are available which can be accessible by the users. It enhances the dissemination of medical knowledge and spiritual civilization among the medical students. This knowledge optimizes the medical student to achieve a better result in academic as well as personal skill development. Numerous tedious processes reduce the efficiency of the library. If the medical books are missed, it can make the librarians confused. It can cause a monotonous among the staff. Therefore, to solve the issues of traditional library management system, web-based library management is proposed.

Library Management System contains the major library operations such as registration of book and student and borrower and return of a book. The first section is the registration of the user for the system and then can be modified by authorized users. The second section is the registration of new books and new users for the library information system. The third section is borrowing and returning books which is the major area needed by the users. There are two end users who are the admin (librarian) and students for the library management system.

2. Literature Review

This section presents the literature review of library management information system concerning with the Boolean search technique.

Shanmugam et al implemented the library management system that acts as a tool to transform traditional libraries into digital libraries. Their system supports the librarians to encounter all the issues concurrently. Their model is developed with dot net technology. C# language is used to build the front-end application and MSSQL server is exploiting as database [9].

Naik et al launched information search retrieval system in libraries. This paper discussed about the information search and retrieval system used in digital libraries. Digital Libraries present still another environment for information retrieval, presenting new and different challenges and expanded research agenda. Some of these challenges arise from the nature of the content in digital libraries, others from the nature of the tasks performed and the characteristics of the users of digital libraries. Information Retrieval system which presents the basic layer that is applied in conceptualization processes and discussed the models of the IR system. Their model is developed with Boolean model, Vector

model, generalized vector space model and Probabilistic model [8].

Muhammad presented the efficiency of Boolean search strings for information retrieval system. In this paper, the technique of Boolean search string is explored in details along with the analysis/evaluation of the effectiveness of the technique. The technique was evaluated based on three criteria: Number of documents retrieved, the time taken to retrieve them and relevance of the documents to the query or research question. The analysis of this technique shows that Boolean search strings technique returns more relevant articles compared to the free text query by at least 77% and in shorter time frame. According to Muhammad, Boolean search strings are very useful for information retrieval [1].

3. System Design and Methodology

3.1. Material and Method

Library Management System uses Boolean search method. This method is a method of keyword search that is to find candidates in database. Boolean searches allow to combine keywords and phrases using the words AND, OR, NOT (known as Boolean operators) to limit, broaden, or define search term. AND Operator is used to retrieve information from the database that ALL search terms must be present in the resulting records. For example, Subject AND Published year. OR operator is used to retrieve the information from the database that ANY of search terms can be present in the resulting records. For example, Author OR Published year. NOT Operator is used to retrieve the information from the database to ignore concepts that may be implied by search terms. For example, Subject Anatomy NOT "Published Year 2015".

3.1.1. AND operator

The operator 'AND' is used when the user needs to retrieve all information about two or more keywords. For example, a search term containing the words, subject 'Biochemistry' and published year '2012' will bring results containing both Biochemistry and 2012. The use of the 'AND' operator allows the user to combine more than one search term in a single query.

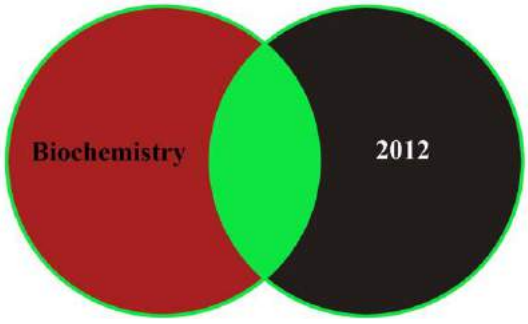


Figure 1. Boolean Search for AND operator

3.1.2. OR operator

The operator 'OR' is used to retrieve search results containing any of the search terms. This allows the user to combine search terms in a single query and retrieve results on any of the search terms. The operator 'OR' allows the user to search broadly. Thus, the user can get more results.

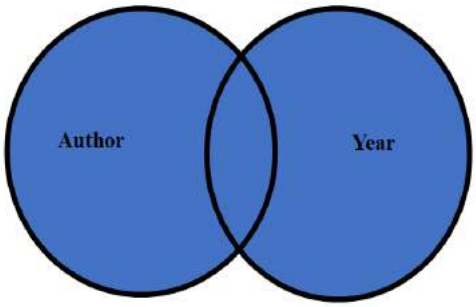


Figure 2. Boolean Search for OR operator

3.1.3. NOT operator

The 'NOT' operator is used to get less the results for a single query. The use of 'NOT' helps the user to eliminate or exclude (irrelevant) terms or records in a search result.

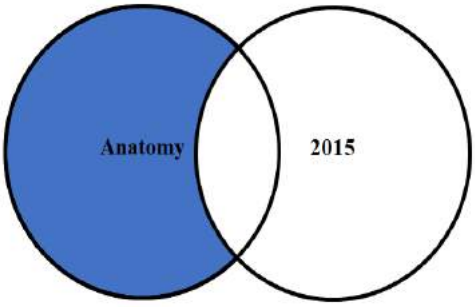


Figure 3. Boolean Search for NOT operator

3.2. Boolean Search Algorithm for Library Management System

Step1-Input Subject.Text, Author.Text, Year.Text
Step2- Output Book Information
Algorithm: Boolean Search
Step3- Begin
3.1 IF (Subject.Text == Subject && Year.Text == Year)
3.2 Show Book Information From View Book Table
3.3 Else IF (Author.Text == Author || Year.Text == Year)
3.4 Show Book Information From View Book Table
3.5 Else (Subject.Text == Subject && Year.Text != Year)
3.6 Show Book Information From View Book Table
End
Step4- Log out
Step5– End

4. System Architecture

In this section, the system flow diagram and system implementation will be explained.

4.1. System Flow Diagram

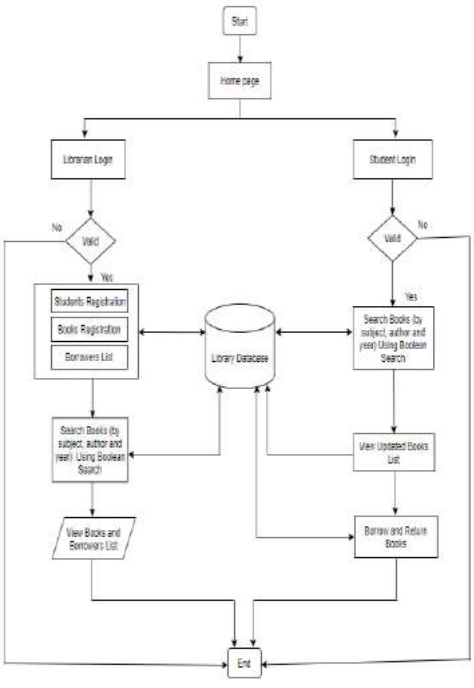


Figure 4. System Flow Diagram

In this system, two parts administration and user (student) are shown in figure 4.1. The administrator needs user name and password to enter the system.

The administrator must register the book and student information list for this system. Then the administrator checks daily for overdueing and borrowing book list. If member information or book information change, only the administrator can insert, update and delete the member entry and book entry.

User must have a username and password to enter the system if a registration member. And then user can search and borrow available books. Each student can rent two books at a time and must return not later than one week after renting date.

Finally, administrators can view the rent list information and overdue list information (weekly, monthly, yearly) by rent member name, book title, rent date and return date.

4.2. Administrator (Librarian)

Administrator needs login name and password to enter the system. The administrator will be able to insert, update, delete and view the book's data and student's information. The administrator is the person who plans to modify the information of the library.

4.3. User (Student)

User (Student) also needs user name and password to enter the system. The students can search (by year, by author, by subject), borrow and return the book.

4.4. System Implementation

This system is a web-based library management system using the ASP.Net MVC framework. There are two sections in this system: Admin and User. Firstly, Admin Section will be explained.

4.4.1. Admin Login Page of Library Management System

In Admin login page, the administrator must enter the user name and password correctly to access the system. If the user's name and

password are incorrect, this page will be logout. If the user's name and password are correct and then the administrator can do another process for the next page as shown in Figure 4.2.

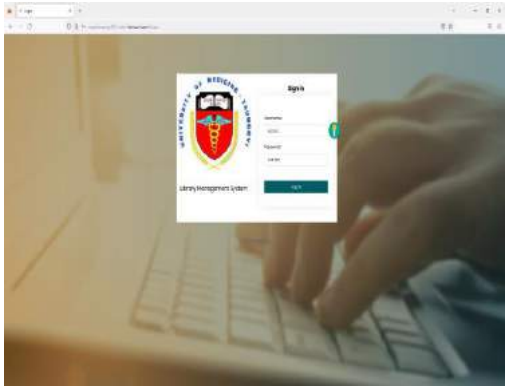


Figure 5. Admin login page

4.4.2. Student Registration Page (Admin Page)

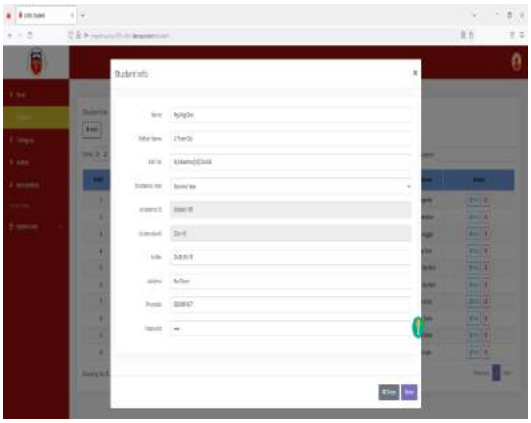


Figure 6. Student Registration Page

In this page shows the registration of the student page. Administration can be registered students' information.

4.4.3. Book Registration Page (Admin Page)

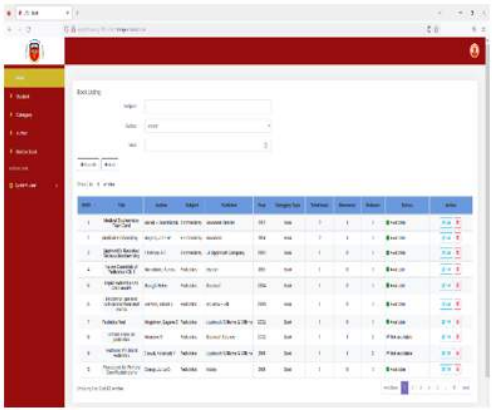


Figure 7. Book Registration Page

In this page, admin can view registered books list. If admin wants to insert, update and delete, he /she can do process, in this page that are shown in figure 7.

4.4.4. Borrowed Book page for Student (Student Page)

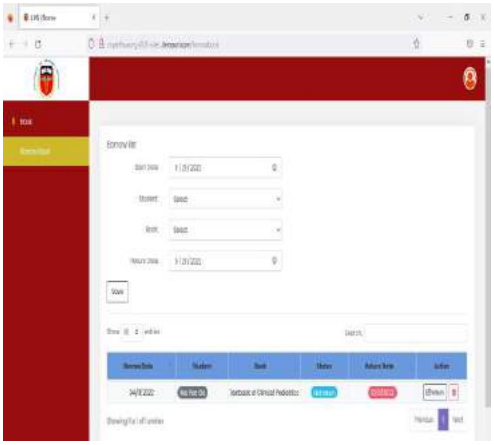


Figure 8. Borrowed Book Page on Library Management System for Student (User)

In this page, the registration book can be borrowed by student. If a book would be borrowed by a student and need to select information such as the book, enter start date and return date from the dropdown list box. And then student can also click the save button of this page and borrower process of the book that has been finished. The student clicks the return button for submission when student wants to return the book. After successful submission, the book could be automatically accessed by student using the Library Management system as shown in figure 8.

4.4.5. Available and not Available Book List Page for Student

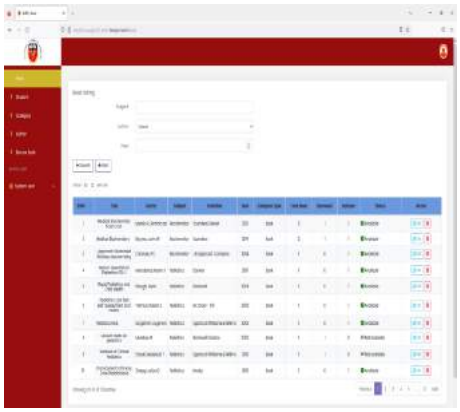


Figure 9. Available and not available book list page

In this page, the book list shows available and not available books of the library management system. The person who rent a book must look a book list page first, if there is a book they want to rent a book as show in figure 9.

4.4.6. Searching Book List Page (Student)

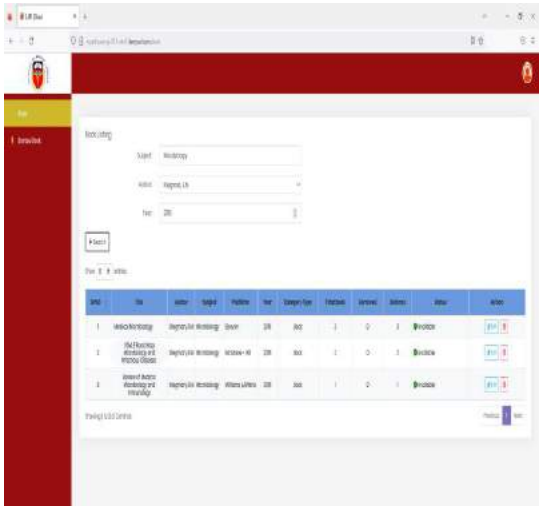


Figure 10. Searching Book list of Library Management System (by AND operator)

In this page, the book can be searched by student using field of the book such as subject, author and year for book list of library management system.

5. Results and Discussion

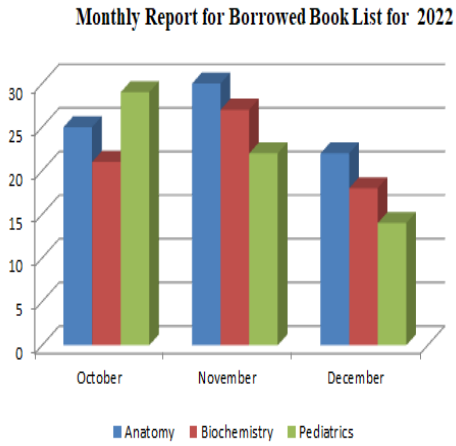


Figure 11. Monthly Report for Borrowed Book List

By using this system, borrowed book list for students can be done as shown in figure 11. In this figure shows how many books are rented for every month. A few categories of booklist (Anatomy, Biochemistry and Pediatrics) are compared as clustered column for monthly report. This figure shows the most interesting books by students in three months.

6. Conclusion and Further Extension

6.1. Conclusion

The issues of the traditional medical library are identified and promote it to easy access for the library information retrieval system. In the Library Management System, the librarian can insert, update and delete the student's information and book details into the database. The students have a Unique ID for accessing any book from the library. Through the ID, the librarian can check the user's details and book's details. The LMS reduces library staff work and makes the system efficient. Boolean search algorithm is used to search the book information and implemented using web-based technologies which include C# and MSSQL. The system serves as improvement in books and students data management.

6.2. Limitation Further Extension

There are some limitations in this system. This system intends to University of Medicine's students in Taunggyi only, and staff who work in University of Medicine, Taunggyi cannot borrow books. The student can borrow up to two books within one week at a time. If the student does not return the books after one week, student has to pay fine. When the book is lost or damaged, the fine amount has to be calculated by librarian. Librarian can modify and access this system without expert advice. Maximum numbers of records are set up in this system. Member detail information has to be changed by librarian only. Students can change their password only.

In the future, if someone modifies this system, he/she can scan every book's pages and upload to the system. Then, the students can easily read, download and print the books by online system. If this system is upgraded, they will be fined if the students borrow books and lose the books. More information can be inserting into book table such as international serial book number (ISBN) to enhance the system to be more perfect. In addition, the system should be enhanced so that one book can have more than one author.

Currently the system only supports one author per book. As some book is having more than one author, user may miss to search the book.

References

[1] A. B. Muhammad, "Efficiency of Boolean Search strings for Information Retrieval", American Journal of Engineering Research (AJER), Vol-6, Issue-11, 2017.

[2] A. Biswas, D.J. Borgohain, "Global Research in Library Management from 2010 to 2020: A Bibliometric Investigation based on Scopus", University of Nebraska-Lincoln, December 2021.

[3] Alade, O.A., Adewale O.S, Ojokoh B.A., "Algorithmic Relational Design Model for Online Library Information Retrieval System", International Journal of Scientific & Technology Research, volume 3, issue 8, august 2014.

[4] A. Samuel, A. Godfred, X. He, "Design and Implementation of Library Management System", International Journal of Computer Applications, September 2018.

[5] A. T. Mary, S. Ramya, Mr. S. K. Murthy, Dr. A. Valarmathi, "Enhanced Library Management System", International Journal of Creative Research Thought, Volume 5, Issue 4 October 2017.

[6] B. R. Adebayo, "Library Management System with topic modeling and its adaptability to open and distance learning libraries", National Open University of Nigeria, 2019.

[7] GUTL, CH, GARC'I A-BARRIOS, V.M, "The Application of Concepts for Learning and Teaching", Proceeding of 8th International Conference on Interactive Computer Aided Learning, Villach, Australia, ICL 2005.

[8] N.R. Naik, A.M. Rao, "Information Search and Retrieval System in Libraries", 8th International CALIBER – 2011, Gao University, Gao, March 02-04, 2011.

[9] Shanmugam A.P, Ramalakshmi, A, Sasthri, G Baalachandran, S, "Library Management System", Journal of Xi'an University of Architecture & Technology, 2020.

[10] T.C. Chun, Ms. C.K. Mui, "Library Management System", Cambell University, U.S.A, 2010-2011.

[11] T.W. Araya, Ass. Pro. A. Mengsteab, "Designing Web-based Library Management System", International Journal of Engineering Research & Technology, Vol.9 Issue 10, October 2020.

Staff Information and Leave Management System for University of Medicine (Taunggyi)

Swe Zin Than, Soe Soe Lwin
University of Computer Studies (Taunggyi)
swezinthan-m@ucstgi.edu.mm, swezinthan122053@gmail.com

Abstract

Most universities operate staff information on paper. Therefore, it is time-consuming, inaccuracy, non-security, and has no information reliability. Therefore, the web-based staff management system is proposed for staff at the University of Medicine (Taunggyi), among the five universities of Medicine throughout Myanmar. In many universities, the staff is entitled to different kinds of leaves. In this system, staff can only request both causal leave and earned leave. By using a web-based staff information and leave management system, staff can request leave online without going to the office. And then, staff can also view whether their requested leave is approved or rejected and know the remaining leaves. Staff can also view their own personal information. Authorities of the university can insert, delete, and update staff's detailed personal data and manage leaves of staff. An administrator can approve or reject the leaves of all staff. Staff leave can be calculated by using a rule-based approach. Staff data can be easily searched by using the B+ tree algorithm.

Keywords: Rule-based approach, B+ tree method

1. Introduction

Staff information and leave management system is a web-based staff management system that automates every step of the staff management process without compromising functionality. An online staff management system gives staff the freedom to request, approve, reject and manage vacation requests from anywhere, at any time, and from any device.

Staff information and leave management system can be applied easily by the staff and management of an organization or institution properly distributed; A platform that enables tracking and granting leave. In many institutions, the staff is entitled to different kinds of leaves; Casual Leave, Earned Leave or Annual Leave, Medical Leave, Maternity Leave or Paternity Leave, Quarantine Leave, Extraordinary Leave without Pay, Special Disability Leave, Hospital Leave, Seamen's Sick Leave, Apprentice Leave, and Vacation Leave.

These leaves are recorded according to the organization's policy. The administrative department is commonly considered to be one of the most important assets in any organization. All the records of staff are kept as a part of the administrative department function. Every organization's administrative department drives information and administrative staff drives day-to-day operations. Most organizations use a traditional approach to requesting and managing leave.

According to the traditional method, academic staff needs to submit a leave application themselves to the administrative department through the Head of Department (HOD). This method is time-consuming and error-prone; It requires more paperwork and is tedious to maintain. In this web-based system, staff can request leave online. Therefore, this system is proposed for retrieving information faster and easier. By using this system, staff of the University of Medicine can save time and reduce manpower.

2. Literature Review

A. Adamu [1] proposed the successful management of staff leave in higher education institutions. That system was created using a three-tier software architecture and implemented with web-based technologies like CSS, JS, HTML, and MySQL. The technology enables academic institution employees to request and monitor their leave requests in a timely manner.

C. Yu [2] proposed a college student management system (basic student information,

attendance information, and scholarship management) based on the K-mean clustering algorithm. Cluster analysis is performed on the student campus activity data in the system, dividing students into several characteristic categories and pointing out the main characteristics of each category.

M. U. Singh [3] proposed an employee information system to maintain the data of employees, to make it easy to control employees, to divide jobs and access control of employees, and use technology for accuracy. The Head of the department can add, update and delete employee details, and accept or reject employee leaves. This system uses HTML, CSS, and PHP at Front End and Microsoft SQL Server at Back End.

M. S. P Dalke [4] proposed a web-based application for a staff management system with two users 1. Staff and 2. Head of Department. The data of the entire system will be handled by the MySQL database. The staff can apply for leave on this system by filling out the leave form online and then sending it to the HOD for approval. The performance analysis of staff is implemented using K Means Clustering Algorithm.

M. Ramanan [5] proposed a web-based leave management system that has been developed to overcome applying for leave manually, which is time-consuming. The leave requested through an online platform gets approval from higher authorities. The system is an automated one that aids in reducing time to leave requests and eliminates paperwork.

N. Choudhary [6] proposed for the college to maintain the leave records of the staff and maintains the leave application of the staff. This system intends to reduce paperwork and maintain leave records with a particular website for leave maintenance and reduce the formalities and time delay faced by faculty members for the approval of leaves. Their system deals with the record of leaves taken by faculties within the institute where higher authorities like HODs and directors will approve/reject the leave applications requested by the employees.

Y. M. Win [7] proposed university student information systems and problems, registered, examinee, pass, fail, honors/qualify, Missing, drop-out, one academic year. The system provides student information about registration, classroom management, facility management, extra curriculum activities, teachers, scholarship, hostel facilities, and transportation. Microsoft visual studio 2015, SQL server database, and Power BI is used for the system.

3. Background Theory

3.1. Material and Method

Databases should provide a quick and easy means to insert, read, and update data. Inserting and reading data may be done quickly and effectively using B+ trees. In actual Database implementation, the database stores data using a combination of B-tree and B+ tree. The B+ tree is used to store the actual records, while B-tree is used for indexing. The B+ tree offers sequential search capabilities, giving the database more control to find non-index values in a database.

A B+ tree is a more complex type of self-balancing tree in which all values are presented at the leaf level. Multilevel indexing is used in B+ trees. The index of indices is produced in multilevel indexing. It enables faster and simpler easier data access.

On a B+ tree, the data pointers are only presented at the leaf nodes, as opposed to a B+ tree where the leaves are connected and the data pointers are presented internally. A B+ tree allows for faster operations.

3.1.1. Properties of a B+ Tree

- All leaves are at the same level.
- The root has at least two children.
- Each node except the root can have a maximum of m children and at least $m/2$ children.
- Each node can contain a maximum of $m - 1$ key and a minimum of $\lceil m/2 \rceil - 1$ key.

3.2. Searching on a B+ Tree

The following steps are followed to search for data in a B+ Tree of order m. The data to be searched staff ID (SID).

- Start from the root node. Compare SID with the keys at the root node [SID 1, SID 2, SID 3, ... SID $m - 1$].
- If SID < SID 1, go to the left child of the root node.

- Else if SID == SID 1, compare SID 2. If SID < SID 2, SID lies between SID 1 and SID 2. So, search in the left child of SID 2.
- If SID > SID 2, go for SID 3, SID 4, ...SID m-1 as in steps 2 and 3.
- Repeat the above steps until a leaf node is reached.
- If SID exists in the leaf node, returns true else returns false.

Search Staff ID = 45 on the following B+ tree.

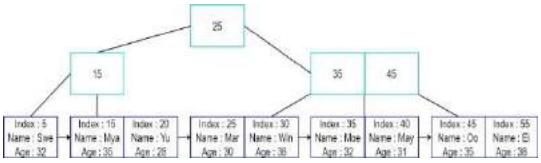


Figure 3.1. B+ tree

Compare Staff ID with the root node.

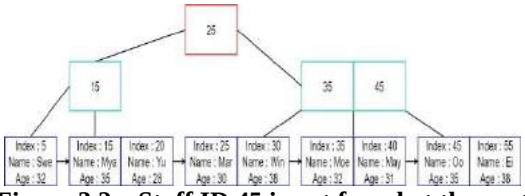


Figure 3.2. Staff ID 45 is not found at the root

Since Staff ID > 25, go to the right child.

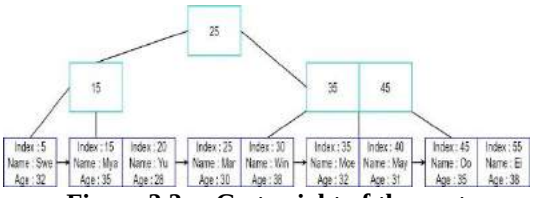


Figure 3.3. Go to right of the root

Compare Staff ID with 35. Since Staff ID > 30, compare Staff ID with 45.

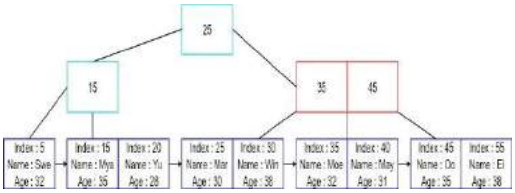


Figure 3.4. Staff ID 45 not found

Since Staff ID ≥ 45, so go to the right child.

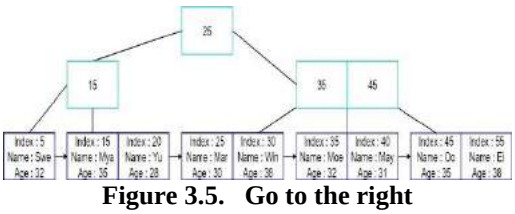


Figure 3.5. Go to the right

Staff ID = 45 is found

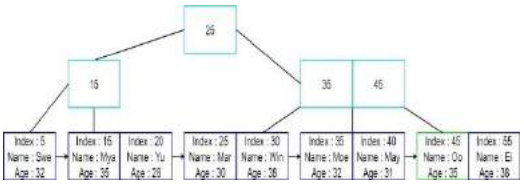


Figure 3.6. Staff ID 45 is found

3.3. System Requirement

C# programming language is used in this system. ASP.NET MVC is used for a C# Web application framework. The server is used with an IIS server. MSSQL is used for the database. Browser such as Microsoft Edge, Internet Explorer, Chrome, and Firefox is used to access the web application in this system. Editor (Visual Studio) is used for the development of the system.

Processor 11th Gen Intel (R) Core (TM) i5-113567@ 2.40 GHz 2.24 GHz, RAM 8.00 GB, System Type 64-bit operating system, x64-based processor, and SSD (256 GB) or HDD (2 TB) is used in this system.

3.4. Rule-based Approach

Rule-based classifiers make use of a set of IF-THEN rules for classification. A rule can be represented in the following:

IF the condition THEN the conclusion

The IF part of the rule is called the rule antecedent or precondition. The THEN part of the rule is called the rule consequent. For example, these rules are used for giving leaves to staff.

Rule 1:
IF Casual Leave Then
give the10 Days per year
Rule 2:
IF Earned Leave Then
give the 1 Month per year
Rule 3:
IF Medical Leave Then
give the 12 Month
Rule 4:
IF Maternity Leave or Paternity Leave Then
give the 6 months
Rule 5:
IF Quarantine Leave Then
give the 21 Days

Rule 6:
IF Extraordinary Leave without Pay Then
give them any time
Rule 7:
IF Extraordinary Leave without Pay Then
give the pension for late
Rule 8:
IF Special Disability Leave Then
give the 24 months
Rule 9:
IF Hospital Leave Then
give the 3 Months per 3 years
Rule 10:
If Seamen's Sick Leave Then
give the 3 Months per 3 years
Rule 11:
IF Apprentice Leave Then
give the Study period
Rule 12:
IF Vacation Leave Then
give the 1 month per year

In this system, staff of the University of Medicine (Taunggyi) can only apply for casual leave and earned leave.

4. System Architecture

In this section, the system flow diagram, the process flow of the system, and system implementation will be explained.

4.1. System Flow Diagram

In this system two parts administrator and user (staff) as shown in figure 4.1. The administrator needs a login name and password to enter the system. The administrator will be able to insert, update, and delete data when needed in this system. The administrator is the person who plans to search and view the information of the staff. The administrator can approve or reject staff leaves.

The user (staff) also needs a login name and password to enter the system. Staff can see the profile information, apply for leaves and view leave history.

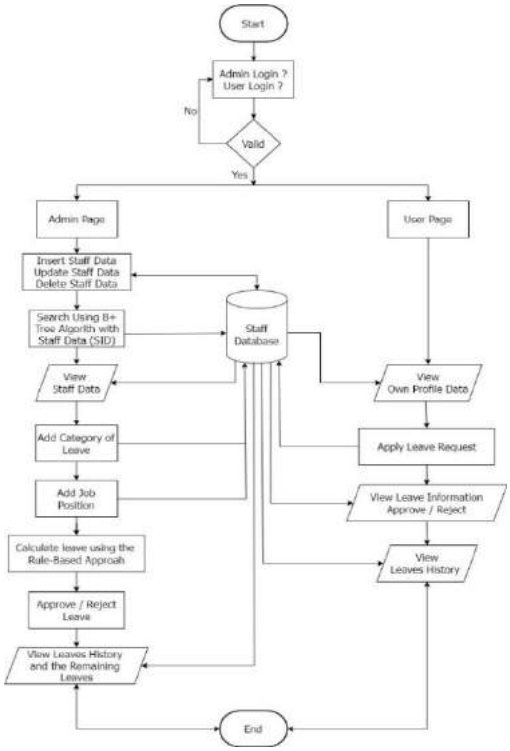


Figure 4.1. System Flow Diagram

4.2. Process Flow of System

Begin
1 Check the user roles (Admin or User) in the system
2 If Admin then
Begin
2.1 Admin has the right to see all staff's data in the system.
2.2 Admin has the right to view staff's information in the system.
2.3 Admin can add leave category and the number of leave days.
2.3 Admin can calculate the leaves in the system.
2.4 Admin can accept or reject the leaves of the staff in the system.
2.5 Admin can see the remaining leaves of the staff and leave history.
2.6 Admin can insert, update, and delete, all staff data.
End
3 else if user then
Begin
3.1 User can view the personal profile
3.2 User can apply the leaves
3.3 User can see the leaves history
3.4 User can see the remaining leaves
End
End

4.3. System Implementation

This system is a web-based staff information management system using the ASP.Net MVC framework. There are two sections in this system: Admin and User. Firstly, Admin Section will be explained.

4.3.1. Admin Login Form

Figure 4.2. Admin Login Form

The system administrator must enter a username and password. Then click Log in button. If the filled information is correct, the administrator could enter the system.

4.3.2 All Staff Information Form

Figure 4.3. Staffs Information

Admin can view all staff information in the University of Medicine, staff's detailed personal profile, and edit staff information by clicking the edit button as shown in figure 4.3. A personal detail data page is shown in figure 4.4.

Figure 4.4. Staff Information Detail

4.3.3 Category of Leave Types Form

Figure 4.5. Category of Leave Types and number of days

Admin can add a category of leave types and number of days and can also delete leaves category. There is Casual Leave, Earned Leave or Annual Leave, Medical Leave, Maternity Leave or Paternity Leave, Quarantine Leave, Extraordinary Leave without Pay, Special Disability Leave, Hospital Leave, Seamen's Sick Leave, Apprentice Leave, and Vacation Leave.

4.3.4 Searching Staff ID 45 Personal Data by using B+ Tree Form

Figure 4.6. Searching Staff ID 45

By using B+ tree method, staff's personal data can be searched by staff ID. Admin can search staff's personal profile by entering the staff ID (SID). Figure 4.6 shows staff's data of SID 45 and staff information can be edited by clicking edit button.

In B+ tree, the data pointers are present only at the leaf nodes whereas the data pointers are present in the internal, leaf or root nodes on a B tree. Search operations on a B+ tree is faster than on a B tree.

4.3.5 Approve or Reject Leave Form

Figure 4.7. Approve or Reject Leave

Admin can approve or reject leave when the staff apply for the leave. If the staff has the remaining leave, the admin can approve the leave. If the staff has no remaining leave, the admin can reject the leave. Approve or reject leave page is shown in figure 4.7.

4.3.6 Leave History Form

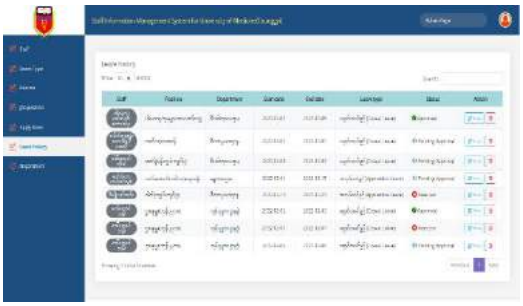


Figure 4.8. Leave History of Staff

Admin can view all staff's leave information (Staff Name, Position, Department, Start Date, End Date, Leave Type, Approved or Rejected). If the admin has not decided to approve or reject leave, the status shows pending approval.

4.3.7 User Login Form

Figure 4.9. User Login

The System users must enter a username and password as shown in figure 4.9. And then the click Log in button. If the filled information is correct the user could enter the system.
The user can view their own personal profile and cannot edit profile information.

4.3.8 Leave Request Form

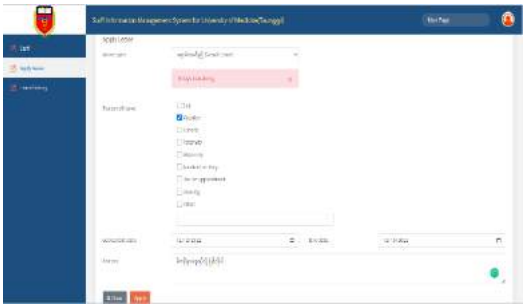


Figure 4.10. Leave Request

In figure 4.10., Staff can fill leave request form and request leave from the authority of the university.

4.3.9 Leave History Form with Leave Status (Pending approval or Approved or Rejected)

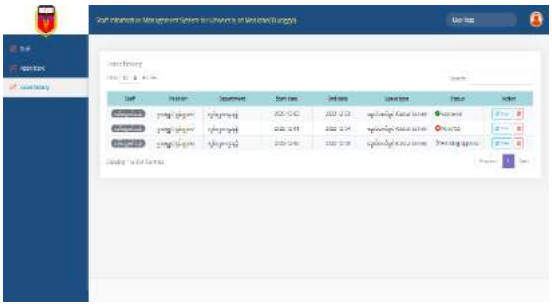


Figure 4.11. Leave History with Leave Status (Pending approval or Approved or Rejected)

Before the admin has not decided whether the leave is approved or not, the leave history form with leave status (pending approval), after the admin has decided that the leave is approved, the leave history form with leave status (approved), and after the admin has decided that leave is rejected, leave history form with leave status (rejected) will appear as shown in figure 4.11.

4.3.11 Leave Request Form with Remaining Leave

If the staff does not have enough remaining leave, the system will show a leave request form with remaining leave that he/she cannot apply for the leaves as shown in figure 4.12.

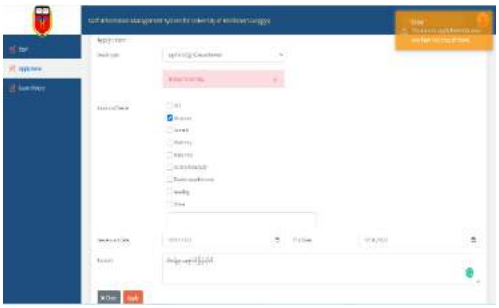


Figure 4.12. Leave Request Form with Remaining Leave

5. Staff's Leave Analysis

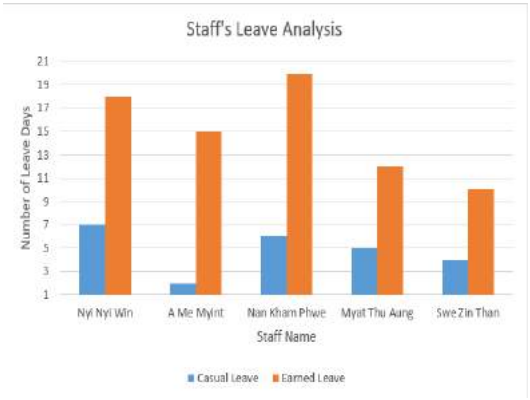


Figure 5.1. Staff's Leave Analysis

By using this system, staff's leave analysis can be done as shown in figure 5.1. Figure 5.1 shows the number of leave days taken by five staffs of University of Medicine (Taunggyi) for two categories of leave types (causal leave and earned leave). X axis represents the staff and Y axis denotes the leave duration. The leave type is highlighted in different colors. Staff's leave analysis can be easily done by using this system.

6. Conclusion and Further Extensions

6.1 Conclusion

This system can be used by the University of Medicine (Taunggyi). This system reduces paperwork and manpower. Staff's personal information can be easily searched by using the B+ tree algorithm. Staff can easily apply for leaves and view remaining leaves by using this system. By using this system, an authority of the university can also manage staff's data and make decisions whether the leave is approved or rejected.

6.2 Limitations and Further Extensions

In this system, staff can only apply for two leaves: causal leave and earned leave. In this system, staff data of the University of Medicine (Taunggyi) is used. Other staff data from the University of Medicine will not be used in this system.

In the future, the remaining leaves: medical leave, maternity leave or paternity leave, quarantine leave, extraordinary leave without pay, special disability leave, hospital leave, seamen's sick leave, apprentice leave and vacation leave will be computed as a further extension. The staff of various payrolls will be computed in the future. The client-Server architecture will be a further extension.

References

[1] A. Adamu, "Employee Leave Management System", Department of Computer Science, Ibrahim Badamasi Babangida University, Lapai, Niger state, Nigeria, FUDMA Journal of Science (FJS), Vol. 4 No.2, June, 2020, pp 86-91.

[2] C. Yu, Y. Wang, "College Student Management System Based on K-means Clustering Algorithm" Steven Holzner etal, Black book- Java 2, JDK 5 Edition.

[3] M. S. P. Dalke, M. S. A. Deshmukh, M. J. G. Dalav, M. V. N. Sasane. "WEB BASED STAFF MANAGEMENT SYSTEM", International Journal of Science Technology & Engineering. Volume: 3, Issue:09, March, 2017, pp 271-276.

[4] M. Ramana, "WEB BASED LEAVE MANAGEMENT SYSTEM FOR UNIVERSITY COLLEGE OF JAFFNA", Journal of Research Technology and Engineering. Eng. 2(3) 2021, pp 106-113.

[5] M. U. Singh, H.S. Fartyal, K.A.A. Zubair, A. Laddhasystem, "Employee Management System", International Research Journal of Engineering and Technology (Irjet), Volume: 06 Issue: 05 | May 2019

[6] N. Choudhary, A. Khalfe, Y. Khan, M. Ansari, "Leave Management System or Aiktc", International Research Journal of Engineering and Technology (IRJET), Anjuman-I-Islam's Kalsekar Technical Campus, Maharashtra, India, Volume: 07, Issue:03, March, 2020, pp 1715-1717.

[7] Y. M. Win, N. N. Win, "Information Retrieval System for University's Student Data", Dagon University Research Journal 2020, Vol. 11.

Spam Detection in Twitter By Using K-nearest Neighbor (KNN)

Shwe Tha Zin, Zin Thu Thu Myint
University of Computer Studies, Yangon
shwethazin@ucsy.edu.mm, zinthuthumyint@uccsy.edu.mm

Abstract

Twitter is the popular social networking site with approximately 300 million monthly users and 500 million tweets every day. This is the primary reason spammers use Twitter to disseminate malicious software that takes the user's personal information and tweets containing bogus or malfunctioning URLs, assertively follow or unfollow users and trending fake tweets to capture users' attention, and spread pornography adverts. Twitter has apparently collected and analyzed the data of active users in recent years; the research plainly demonstrates that over 32 million users have engaged with the server for casual information on a daily basis. As a result, recognizing and filtering malicious tweets or trends that are damaging or unwelcome for users is critical in today's social world. This system proposes to analyze the tweets and classify them into spam and ham based on the words involved in tweets. Although there are various machine learning and deep learning methods to classify and detect spam tweets, this system will implement the clustering and binary detection model that is used KNN.

Keywords: tweets, spam, ham, KNN

1. Introduction

Twitter has become the fastest news dissemination application and is widely used by people of all ages. Twitter is used for information, job searching, education, and marketing strategy implementation. People may now find out what's going on in different countries around the world with a single swipe. There is also a risk of distributing incorrect or irrelevant information via tweets. Many people are being duped by spammers with malicious content that claims to deliver free accessories or other evading strategies that collect the user's personal information and utilize it unfairly. In general ‘spam’ is a term for bulky irrelevant material, whereas 'ham' refers to desirable or needy bulk messages. Many digital marketers utilize keywords to manipulate and profit from Twitter trends. This is an online marketing method that disseminates product data or other information to potential users worldwide. However, the unwelcome information given by marketing strategists causes a stir among consumers and is labeled as spam. Many people who want to focus on current developments are frequently diverted by unpleasant or malicious content comments posted by the spammers.

2. Related Work

Spam is a major issue in electronic communication, particularly large-scale email systems. Spam email classification is an important activity in an email system. Many factors influence the effectiveness of this process, including the quantity of features, feature selection algorithms, symbol representation, and classifier. This work focuses on email classifiers that use the Multilayer Perceptron (MLP) technique to classify spam and ham emails. During preprocessing, the system employs term frequency and inverse document frequency (tf-idf) and fisher score feature selection methods. These methods allow for the selection of relevant features as well as the addition of benefits such as improved accuracy and reduced time complexity email classification system [1].

Twitter is currently among the top social networks in the world for monthly active members, right after Facebook and Instagram. Twitter is mostly used by users to learn more about breaking news or to stay up with news in general by following hot topics. Twitter has evolved into a platform for sharing the newest concepts as news breaks in the form of comments and replies. As a result, a number of mobile applications that make use of the Twitter API have been created to give their users a quick way to access trending topics. From a marketing perspective, Twitter trending

topics provide an efficient way for web marketers to promote their marketing materials. Spam contents in Twitter were discovered to be obtrusive and irritating for certain users, necessitating the development of a mobile application to deliver spam-free Twitter trending topics contents. This study creates an Android application framework that enables programmers to create custom spam detection classifiers for Twitter content as application libraries. This study uses two categorization techniques. i.e. Naïve Bayes and K-Nearest Neighbor, to identify spam in Twitter trending topics. The Naïve Bayes and K-Nearest Neighbor classification methods are able to detect spam and ham contents with 82% and 71% accuracy respectively [2].

3. Motivations

Twitter is much more likely to have spam information than other social networking sites because so many people use it every day. Twitter users have the option to "follow" other accounts that interest them. The connection between users is bi-directional as opposed to simplex links on other social media sites, which could result in one person not following one of his followers. As a result, spammers have a chance to spread their spam. Spamming typically refers to sending the user undesired information or data. This project's primary goal is to find tweets that are perceived as spam.

On Twitter, a user is only identified by their username and, optionally, by their real name. It can be problematic to accept an Associate in nursing erroneous friend request from a stranger. Whether or not the victim is aware of the aggressor in the actual world, user will still click any link contained in communications. In the unlikely event that the attacker fabricates their phishing mails using information stolen from their friends in informal organizations. Consequently, our approach can assist in locating those spam tweets and provides better results than the current system with reasonable accuracy.

4. Background Theory

Machine learning is a type of computational learning that uses algorithms to understand and predict outcomes from provided data. Making accurate predictions and understanding the system better are the two basic objectives of machine learning.

Artificial intelligence is divided into a subfield called machine learning (AI). Every machine learning algorithm primarily consists of three tasks: representation, evaluation, and optimization. The suggested system will be implemented KNN It is a potent predictive modeling technique. Multi-class classification issues are solved with KNN classifiers. KNN models employ the class probabilities and conditional probabilities. When compared to other algorithms, the primary benefit of employing KNN classifier is that learning occurs quickly.

A. Single Classifier based Prediction

Classifiers are trained for predicting the unknown test cases. The following classifiers are used while detecting spam posts:

(a) Naive Bayes
(b) Multi-Layer Perceptron Classifier
(c) K-nearest Neighbor
(d) Decision Tree Classifier

B. Ensemble Approach based Classifiers

Ensemble technique enables numerous machine learning algorithms to work together to improve the system's accuracy.

5. System Implementation

These are steps in this detection framework such as Data Collection, Data Preprocessing, Feature Extraction, Classification and Accuracy Results.

i. Data Collection

The experiment dataset is to download in this step from the Kaggle.com machine learning repository.

ii. Pre – processing

For Pre-processing phase, data cleaning operations like tokenization, stop-word removal, and stemming will be carried out on the raw dataset. The following step of feature selection and extraction will employ the clean dataset.

i. **Tokenization** is the process of separating the text corpus into its component parts.
ii. **Removing Stop Words:** Stop words are words that frequently appeared in the text

yet were unneeded. For instance, phrases like "so," "and," "or," "the," etc. First, all stop words are eliminated. The stop words in the illustration below are: you, are, that, have, and the.

iii. **Feature Selection and Extraction:** Feature selection is a step before class classification. The dataset will be used to identify the appropriate features.

iv. **Classification:** During this phase, there are training and testing procedures. 40% will be allocated to testing, and 60% to training. After finishing step iii, there should be features that are regarded as spam. Consequently, the dataset must be trained using a machine learning technique (KNN).

5.1. Term Weighting Schemes (TF-IDF method)

Term Frequency (TF) is term weighting determined by how often a word appears in a manuscript. The impact of a word on a document increases as a word's TF value increases. The weighting approach known as Inverse Document Frequency (IDF) bases its calculations on the number of words that appear in each document. One of the most straightforward and effective weighting methods for the data is TF-IDF. Due to its straightforward formulation and effective operation on a variety of different data sets, TF-IDF and its algorithm version are the default choice. This method's formulation is as follows:

$$W(d, t) = tf (t, d) * \log (N / n_t)$$

Where:

w(t,d) = term weight in document d

tf (t,d) = term frequency in document

N = the total number of document

n_t = number of documents that have term t

5.2. K-nearest Neighbor Classifier (KNN)

KNN is an algorithm for classification. There are principally two processes in categorization:

1. Learning Step: Using the training data a classifier is constructed.

2. Assessment of the classifier.

As the new unlabeled information is sorted by determining which classes its neighbors belong to using the training data a classifier is constrained by using the closest neighbor approach. This concept is incorporated into the KNN computation. When using KNN, a certain value of K is fixed, this aids in organizing the cryptic tuple. KNN does two things when a new unlabeled tuple appears in the dataset:

The K closest neighbors acted, or the K focuses closest to the new data of interest, are broken down first.

Second, KNN determines which class the new information should be placed in by using the neighbors' classes.

Given the training dataset: { (x(1), y(1)) , (x(2), y(2)), ... , (x(m), y(m)) }

Step1: Store the training set

Step2: For each new unlabeled data,

A. Calculate Euclidean distance with all training data points using the formula

$$dist(X_1, X_2) = \sqrt{\sum_{i=1}^{n} (x_{1i} - x_{2i})^2}.$$

The formula defines dist (X_1,X_2) is the separation of two documents. The number of distinctive words in the document collection is n. The term "r" in document "x1" is given the weight "x1," and the term "r" in document "x2" is given "x2".

B. Find the k- nearest neighbors

C. Assign class containing the maximum number of nearest neighbors.

5.3. The Process Flow

In the accompanying figure, the rapid process flow of the suggested system is shown. Training phase and testing phase are the two categories into which the system's processing phases can be divided. Data pre-processing (Tokenization and Removing Stop Words) shall be carried out before weight or classification in any phase. The pre-processing steps are briefly explained in the section above, and the last step is feature selection using KNN and TF.IDF as seen in figure 1.

In this system, TF-IDF and KNN is employed to categorize and find spam and ham in a group of tweets. The goal of this technique is to stop twitters from sending spam. Furthermore, it is implied that a crucial stage in this suggested method is data cleaning. The confusion matrix determines how accurate the algorithm is. Due to the combination

of these factors, our technique will provide an accurate result in recognizing spam and ham TF-IDF (feature extraction) and KNN algorithm.

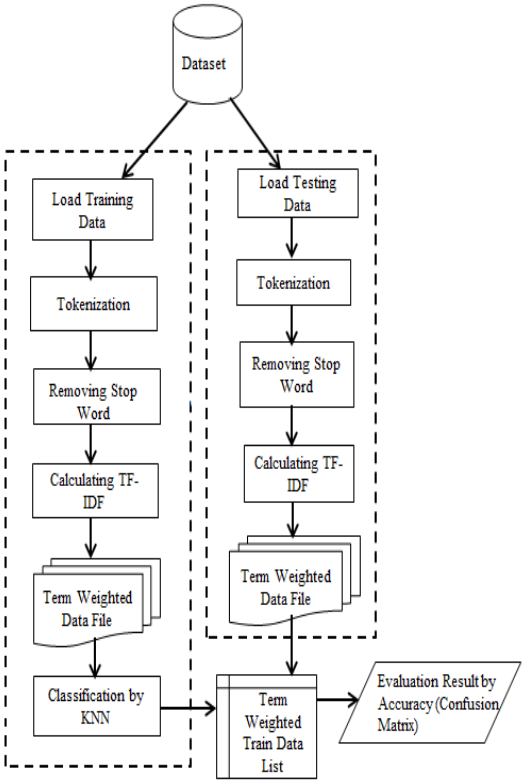


Figure1. The Process Flow

6. Evaluation Metrics

6.1. Twitter Dataset

To evaluate our proposed Twitter spam detection technique, we apply twitter dataset from Kaggle.com, which contains 6153 ham tweets and 5815 spam tweets. The sample datasets are shown in table 1.

Table1: Twitter Sample Dataset

N0	Tweet Comment	Class
1	@misslyndaleigh good afternoon lynda hope your well xxxx	ham
2	hopefully the super characters will be in it but I doubt it	ham
3	And I'm sick now, I just want to go home and die tbh	ham
4	Now Playing: Young MA - Hot Sauce -> https://t.co/eMX9JgTv3v	Spam
5	#TopVideo Gay couples officially tie the knot after ruling http://t.co/pL2hilBtHa	Spam
6	All the action! It's action packed from start to finish! https://t.co/izPzgq2iXW	Spam

6.2. Confusion Matrix

A confusion matrix is a table that is often used to describe the performance of a classification model on a set of test data for which the true values are known. Four evaluation metrics, which are precision, recall, F-measure and accuracy are used to evaluate the effectiveness of the system.

Table 2: Confusion Matrix

	Ham	Spam
Ham	TP	FN
Spam	FP	TN

$$\text{Accuracy} = \frac{TP+TN}{TP+FN+FP+TN} \quad (1)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (3)$$

$$\text{F-measure} = \frac{2*Precision*Recall}{Precision+Recall} \quad (4)$$

Where:

- TP is the quality of ham tweets which are correctly classified as ham.
- TN refers to the number of spam which are exactly considered as spam.
- FP represents the amount of spam which are wrongly labeled as ham.
- FN indicates the data for ham tweets which are treated as ham by mistake.

7. Experimental Result

In this tweets' Spam and Ham three experiments are conducted to test the detecting system. To analysis, the dataset is split into 70% training data and 30% of testing data. This system used Accuracy, Precision, Recall, and F-measure of each analysis to evaluate the performance of the experiment results. As in shown in Figure 2, almost performance values are higher than 90%. About 90% accuracy is maintained in this proposed work.

Based on the analysis, the system can give better detection result if the more trained data can feed to this system.

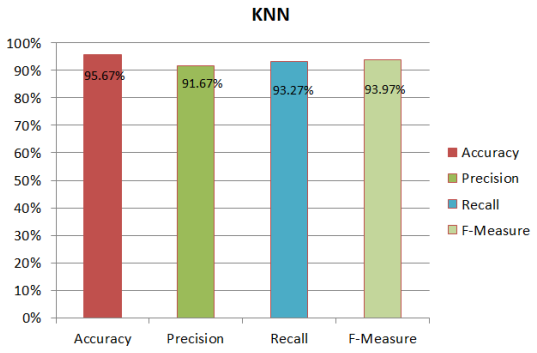


Figure 2: Experimental Results of Analysis

8. Conclusion

Twitter is a great starting point for studying social media. The system preprocesses the gathered spam tweets using machine learning techniques. The KNN classifier is used to classify the spam tweets and choose the spam features. The user would find it simple to acquire the summarized report on the opinion through the proposed system spam tweets.

References

[1] KamalanathanKandasamy, PreethiKoroth: An Integrated Approach to Spam Classification on Twitter Using URL Analysis, Natural Language Processing, and Machine Learning Techniques, 2018 IEEE Student s' Conference on Electrical, Electronics and Computer Science

[2] Aryo Pinandito, Rizal Setya Perdana, Mochamad Chandra Saputra, "Spam Detection Framework for Android Twitter Application Using Naïve Bayes and K-Nearest Neighbor Classifiers", Information System Department, Computer Science Faculty, Universities Brawijaya, 2017.

[3] Bratko, A., Filipiˇc, B., Cormack, G., Lynam, T. Zupan, B.: Spam filtering using statistical data compression models. The Journal of Machine Learning Research 7 (2016) 2673–2698

[4] F. Murtagh, ―Multilayer perceptrons for classification and regression, Neurocomputing, vol. 2, no. 5–6, pp. 183–197, 1991, doi: 10.1016/0925-2312(91)90023-5.

[5] P. Cunningham and S. J. Delany, ―K -Nearest Neighbour Classifiers, Mult.Classif. Syst., no. May, pp. 1–17, 2007, doi: 10.1016/S0031-3203(00)00099-6.

[6] H. Sharma and S. Kumar, ―A Survey on Decision Tree Algorithms of Classification in Data Mining, Int. J. Sci. Res., vol. 5, no. 4, pp. 2094–2097, 2016, doi: 10.21275/v5i4.nov162954.

[7] E. G. Dada, J. S. Bassi, H. Chiroma, S. M. Abdulhamid, A. O. Adetunmbi, and O. E. Ajibuwa, "Machine learning for email spam filtering: review, approaches and open research problems, Heliyon, vol.5, no.6, 2019, doi: 10.1016/j.heliyon.2019.e01802.

[8] L. Breiman, ―ST4_Method_Random_Forest, Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1017/CBO9781107415324.004.

[9] B. Biggio, I. Corona, G. Fumera, G. Giacinto, and F. Roli, ―Bagging classifiers for fighting poisoning attacks in adversarial classification tasks," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 6713 LNCS, pp. 350–359, 2011, doi: 10.1007/978-3-642-21557-5_37.

Healthcare Question and Answer System Based on Sequence to Sequence Model

Ei Zin Phyo, Tin Zar Thaw
University of Computer Studies, Yangon
eizinphyo@ucsy.edu.mm, tinzarthaw@ucsy.edu.mm

Abstract

The healthcare sector is one of the most important domains that impacts the entire global population and is closely linked to the development of any country. There are millions and billion pieces of healthcare information available, but making the right information accessible when needed is very important. The advent of question and answer system has been applied as a promising solution and an efficient approach for retrieving significant healthcare information easily and time saving. The deep learning algorithms are used to train the data and bring output within a specific range by using statistical analysis. Recurrent neural network-based sequence-to-sequence model is one of the most commonly researched models to implement artificial intelligence question answering system. This system has been implemented using neural network where it uses bidirectional recurrent neural network as encoder and Luong attention recurrent neural network as decoder. The system presented a healthcare question and answer system to provide healthcare information based on three long attention alignment functions: general, dot and concat with three datasets: NCI, NIHSeniorHealth and NHLBI in MedQuAD.

Keywords: Healthcare dataset, recurrent neural network, sequence to sequence model, bidirectional recurrent neural network

1. Introduction

The Question and Answer (QA) system is an automated approach given a context to retrieve correct answers to the questions asked by human in natural language. Today, people are busy with work receptions, work at work, and additional addictions to the Internet. Particularly in providing healthcare services for rural and distant areas, is important because the healthcare situation and focusing on this realistic issue in developing countries. Therefore, healthcare QA systems are advancing rapidly around the world. Conversational user interface is a software that runs simple and structurally.

With the help of question answering QA systems, patients can connect with one another quickly. And when appropriately applied, they can assist healthcare professionals in exceeding patient expectations and enhancing patient outcomes. Critics of these platforms have repeatedly questioned the effectiveness because of the lack of face-to-face connectivity and empathy between the patient and the healthcare service provider. Because of technologies such as AI, ML, and NLP, QA system is said to have attained a level where they can gauge human feelings. The uniqueness in individual behavior can confound the QA system.

The utilization of QA framework has spread from customer client benefit to things of life and passing. QA systems are entering the healthcare industry and can help solve many of its problems. So, healthcare facilities in general are a crucial resource for poor nations, but they are frequently challenging to set up due to a lack of awareness and infrastructure development. Many internet users rely on it to find answers to the questions about healthcare. The user can also seek medical advice in an easier way and get exposure to various diseases and diagnostic available for it. To make communication more efficient, QA system has developed for answering healthcare information. This paper applied the healthcare QA system using the concepts of NLP and deep learning algorithms. The prediction is carried out using the sequence-to-sequence algorithm.

This paper is organized as follows: Section 2 discusses related work and Section 3 explains background theory. Healthcare dataset is discussed in Section 4 and Implementation of the proposed system and experimental result are described in Section 5 and Section 6. Finally,

Section 7 presents conclusion and Section 8 presents limitation and future works.

2. Related Work

A Question Answering (QA) system gives a direct answer to a frequently asked question in natural language. There are two types of QA systems: open domain and closed domain. Closed-domain systems are the task of answering questions from a particular narrow domain and offer answers on specific topics, while open-domain systems are based on general ontologies and broad unrestricted knowledge. In this paper, the closed domain QA system for healthcare development is presented. Many researchers have proposed different approaches to deal with the problem of question and answer systems for healthcare development.

The authors [6] have proposed an attention-Based Recurrent Neural Networks for short text classification for Twitter mining for public health monitoring. They found that the proposed system was to automatically filter Tweets which are relevant to the syndrome of asthma/difficulty breathing based on bi-directional Recurrent Neural Network architecture with an attention layer (termed ABRNN). Konar.K [4] have studied that the detailed comparison between lexicon-based approach and machine learning based approach. They found that Lexicon based approach was easy to implement, easy to understand, less complex when compared with machine learning approach and Machine Learning based approach provided accuracy rate that was very much higher when compared with Lexicon approach, very good performance.

The authors [2] had developed a chatbot for Medical Purpose using Deep Learning (Neural Network). They implemented an Artificially Intelligent Chat-bot using applications of Deep Learning to fight COVID-19 including to solve COVID-19 problem with the help of the symptoms provided by patient itself and help them to give the correct antibiotics/ medicines and precautions. The authors [7] presented a chat bot for the student to know about the admission process of the college from anywhere with internet connection and receive fast replies. They had proposed this system using chatterbot algorithm that was a python library that makes it easy to generate automated responses to a user's input. The paper was used pattern-matching, natural language processing and data mining to answer to the queries for the college administrators. This paper presented the question answering system with the three-attention based sequence to sequence model to apply the healthcare application development.

They discussed that accurate generative chatbots were usually trained on large datasets of question–answer pairs [3]. But companies usually own small domain-specific datasets about products, services, or used technologies. They found that effective solutions to create generative seq2seq-based chatbots from very small data. Encoder–decoder LSTM-based approaches was suitable for English language than other languages for a morphologically complex language. In this paper, the sequence-to-sequence based question answering system to answer healthcare questions based on the MedQuA dataset.

3. Background Theory

Sequence to Sequence models depend on RNN design and comprises of two RNNs: an encoder and a decoder. The encoder-decoder model is a way to use recurring neural networks (RNNs) for sequential prediction problems. It was originally developed for machine translation problems, although it was found to be effective at sequence-to-sequence prediction problems such as text summarization and question answering. RNNs involves two recurrent neural networks, one for encoding the input sequence, called the encoder, and a second for decoding the encoded input sequence into the target sequence called the decoder. This system presented a healthcare question-answering system to provide healthcare information based on three attention sequence to sequence models: general, dot and concat with three sub datasets: NCI, NIHSeniorHealth and NHLBI in MedQuAD. The presented system has used Tensorflow and Keras to build our Seq2Seq model. Both are an open-source end-to-end machine learning platforms and Keras contains Python library to provide an interface for modeling artificial neural networks.

3.1. Recurrent Neural Network Units

Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) are basic cells of RNN. An encoder takes the input sequence and encapsulates

them as the internal state vectors. Yields of the encoder are rejected and as it were inside states are utilized. GRU couples forget as well as input gates. GRU use less training parameters and so utilize less memory, execute faster and train faster than LSTM's whereas LSTM is more accurate on dataset utilizing longer sequence [5].

3.2. Encoder

This system is used Bidirectional RNN as encoder and describes the algorithm process of Bidirectional RNN as follows in [8].

Inputs:

- Input sequence: batch of input sentences;
- Input length: list of sentence lengths corresponding to each sentence in the batch;
- Hidden state: hidden state: n_layers, num_directions, batch_size and hidden_size;

Computation Graph:

- Convert word indexes to embedding.
- Pack padded batch of sequences for RNN module.
- Forward pass through GRU.
- Unpack padding.
- Sum bidirectional GRU outputs.
- Return output and final hidden state.

Outputs:

- Outputs: output features from the last hidden layer of the GRU;
- Hidden state: updated hidden state from GRU;

3.3. Attention

The attention mechanism was introduced to improve the performance of the encoder-decoder model for machine translation, Question Answering and Natural Language Inference. There are two popular attention mechanisms: Lounge and Bahdanau attentions. Luong attention used top hidden layer states in both of encoder and decoder. Luong attention gets the decoder hidden state at time t. At that point calculate attention scores and from that get the context vector which will be concatenated with hidden state of the decoder and then predict. Bahdanau attention take concatenation of forward and backward source hidden state of top hidden layer. Bahdanau has only concat score alignment model. Luong has three diffferent types of alignments and shown the equations in (1) to (3).

Bahdanau recommend uni-directional encoder and bi-directional decoder. Luong has both as uni-directional. Luong also recommends taking just the top layer outputs; in general, the model is simpler. The three types of Loung attention are dot, general and concat. Dot is the simplest function to produce the alignment score by multiplying the hidden states of the encoder and the hidden state of the decoder. General is similar to the dot function, with the exception that a weight matrix is added to the equation as well. Concat is that the hidden state of the decoder and the hidden states of the encoder are added up before going through a linear layer. This means that the hidden status of the decoder and the hidden status of the encoder will not have individual weight matrix, but a shared matrix instead, unlike Bahdanau Attention. After being passed through the linear layer, a tanh activation function will be applied on the output before being multiplied by a weight matrix to produce the alignment score.

$$Score(h_t, h_{\bar{s}}) for\ dot = h_s^T * h_s^T \quad (1)$$

$$Score(h_t, h_s^-) for\ general = h_s^T * W_a * h_s^T \quad (2)$$

$$Score(h_t, h_s^-) for\ concat = h_a^T \tanh\ (W_a[h_t + h_s^T]) \quad (3)$$

When calculating the context vector (Wa), the global attentional model considers all the hidden states of the encoder. A variable-length alignment vector (a_t) equals to the size of the number of time steps in the source sequence is inferred by comparing the current target hidden state (h_t) with each of the source hidden state (h_s). The alignment score is referred to as a content-based function for which three different alternatives are being considered.

3.4. Decoder

This system is used Lounge attention RNN as decoder and describes the algorithm process of decoder as follows:

Inputs:

- Input step: one time step (one word) of input sequence batch;
- Last_hidden: final hidden layer of GRU;

- Encoder_outputs: encoder model's output: max_length, batch_size and hidden_size;

Computation Graph:

- Get installing of current info word.
- Forward through unidirectional GRU.
- Calculate consideration loads from the current GRU yield.
- Multiply consideration loads to encoder yields to get new "weighted entirety" setting vector.
- Concatenate weighted setting vector and GRU yield.
- Predict next word without softmax.
- Return output and last shrouded state.

Outputs:

- Output: softmax normalized tensor giving probabilities of each word being the correct next word in the decoded sequence;
- Hidden state: final hidden state of GRU: n_layers, num_directions, batch_size and hidden_size;

4. Dataset Description

This system is used three sub dataset of MedQuAD that is a collection of 47k question and answer pairs. MedQuAD includes 47,457 medical question-answer pairs created from 12 NIH websites :National Cancer Institute, Genetic and Rare Diseases Information Center , Genetics Home Reference, MedlinePlus Health Topics, National Institute of Diabetes and Digestive and Kidney Diseases, National Institute of Neurological Disorders and Stroke, NIHSeniorHealth, National Heart, Lung, and Blood Institute, Centers for Disease Control and Prevention, MedlinePlus A.D.A.M. Medical Encyclopedia, MedlinePlus Drugs and MedlinePlus Herbs and Supplements [1]. This system used three websites to construct a collection of question-answer pairs. They are provided below:

- National Cancer Institute (NCI): It contains free text from 116 articles on various cancer types and 729 QA pairs.
- NIHSeniorHealth: This website contains health and wellness information for older adults and it contains 71 articles and 769 question and answer pairs.
- National Heart, Lung, and Blood Institute (NHLBI): It contains text from 135 articles on diseases, tests, procedures, and other relevant topics on disorders of heart, lung, blood, and sleep. It contains 559 question and answer pairs.

5. Implementation of the Proposed System

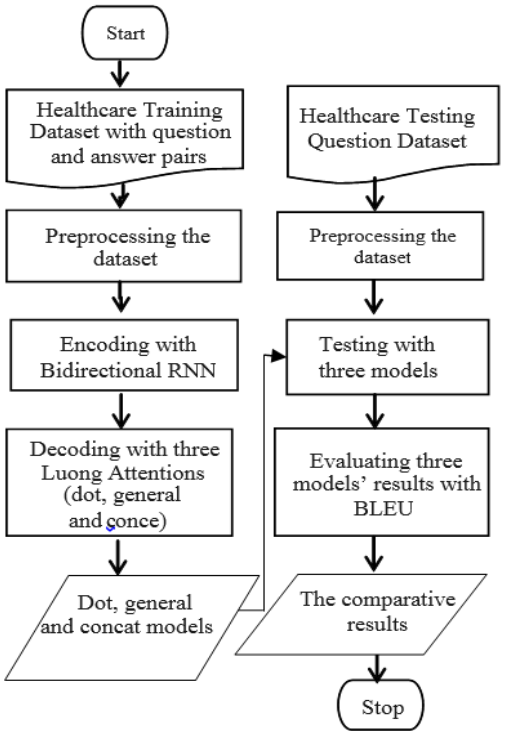


Figure 1. The system flow of the healthcare QA System

The Figure 1 shows the system flow of the proposed system. The eighty percent of training and twenty percent of testing are used in the presented system. The system takes healthcare training dataset with question and answer pairs as input. The input pairs are preprocessed to change over all letters to lowercase and trim all non-letter characters with the exception of essential accentuation. After that all sentence pairs of given dataset are trimmed to count Unique Words. To accommodate sentences of various sizes in a similar clump, sentences are being padding with zero cushion to lounge information tensor of max_length and batch_size. TensorFlow [9], an open-source machine learning library, was used to implement the models. The proposed methodology was evaluated on three healthcare datasets, which contains healthcare question and answer pairs. Most of the answer sentences contain more than 200 words. Although increasing

the training sentences improves the accuracy of the model, the proposed system trains the sentences that number of words are less than or equal to 200 because of execution time. The default hyper parameters used to train nine models based on three datasets and three loung attention alignment functions are as follows.

After preprocessing step, the system encodes the dataset with Bidirectional RNN and decode lounge attention decoder according the three lounge attention alignment functions: dot, general and concat. The system trains these three sequence to sequence models on each three dataset: NCI, NIHSeniorHealth and National Hearth. For three datasets, the system has produced nine models. After creating nine models, the system is asked healthcare information and answers the appropriate answer. To evaluate three models based on three datasets, Bilingual Evaluation Understudy (BLEU) metric is used to compute the quality of system generated answers. This metric is understandable, language independent and adaptation for evaluating a generated sentence to a reference sentence. So, this metric is used to measure the proposed model. The output of BLEU score range is always between 0 and 1, value close to 1 show that the system generated answer is more analogous to the expert generated answer and 0 is no match at all.

BLEU=min(1,hypothesis length/reference length) *(maximum number of words occurs in reference/ total number of words in hypothesis) (4)

Where hypothesis means dataset's expert answer and reference is QA system's answer. Then, the system shows the comparative results of the models.

6. Experimental Results of the System

The proposed system is compared with the nine sequence to sequence models based on three datasets.

Table 1. Hyper parameters for sequence-to-sequence model

Hyperparameter	Size
Encoder Layer	2
Decoder Layer	2
Hidden Size	500
Batch Size	64
Sentence Maximum Length	200
Sentence Minimum Length	3
dropout	0.1
Checkpoint Iteration	1200

In this system, the sentence length of the pairs is set with 200. The default hyper parameters that are shown in table 1 are used to train nine models based on three datasets and three loung attention alignment functions. Figure 2 shows the comparison results of the NCI dataset. According to the comparison results, the system answers 1, 8 and 7 sentences based on concat, dot and general models respectively with 100% accuracy. The bot answers 20, 25 and 28 sentences based on concat, dot and general models respectively with greater than or equal to 75% accuracy. The bot answers 43, 58 and 40 based on concat, dot and general models respectively with greater than or equal to 60% accuracy.

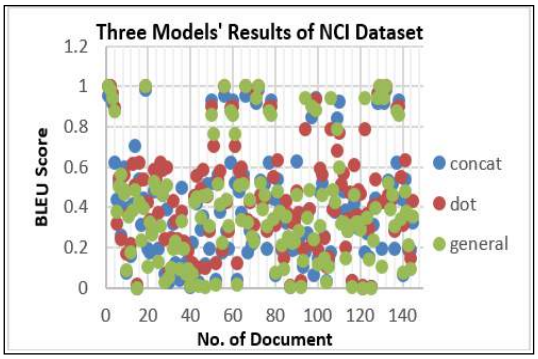


Figure 2. Three Models' Results of NCI Dataset

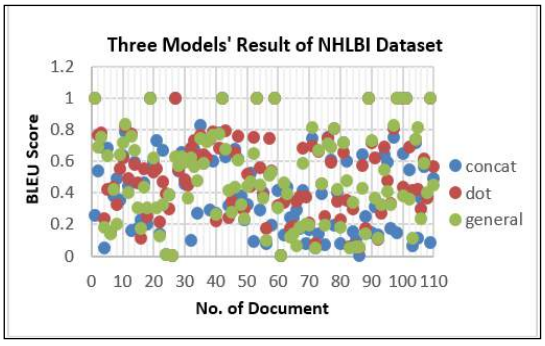


Figure 3. Three Models' Results of NHLBI Dataset

The comparison results of the NHLBI dataset are shown in Figure 3. According to the comparison results, the system answers 8, 11 and 10 sentences based on concat, dot and general models respectively with 100% accuracy. The bot answers 14, 25 and 21 sentences based on concat,

dot and general models respectively with greater than or equal to 75% accuracy. The bot answers 46, 57 and 49 based on concat, dot and general models respectively with greater than or equal to 50% accuracy.

For NHLBI dataset, three models have been evaluated and the results are shown in figure 4. The system answers 0, 2 and 1 sentences based on concat, dot and general models respectively with 100%. The bot answers 2, 10 and 3 sentences based on concat, dot and general models respectively with greater than or equal to 75%. The bot answers 33, 61 and 84 based on concat, dot and general models respectively with greater than or equal to 50%.

Table 2. BLEU scores of nine different Seq2Seq models for three datasets

Dataset Name	Unique Words	Dot's BLEU Score	General's BLEU Score	Concat 's BLEU Score
NIHSenior Health	6249	0.42	0.48	0.31
NCI	6272	0.46	0.41	0.39
NHLBI	5901	0.52	0.48	0.42

Table 2 shows the result of BLEU scores of nine different Seq2Seq models for three datasets. According to the experimental results, the QA model on dot function is the best than other models. The QA model on concat produces less BLEU score than other models. So the healthcare QA system should use Dot and general functions to get more accurate answers.

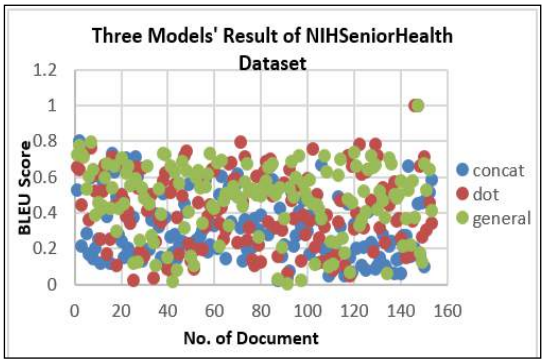


Figure 4. Three Models' Results of NHLBI Dataset

According to the experimental results, if dataset contains same questions and difference answers, the system BLEU score decreases. Although the healthcare topic is difference, the questions have same structure and most of words are same and then the system BLEU score can decrease. For example, the two question and answer pairs are displayed with BLEU score.

Q1: What is (are) Psoriatic arthritis?
A1: Psoriatic arthritis is a joint problem (arthritis) that often occurs with a skin condition called psoriasis.
System Answer: psoriatic arthritis is a joint problem arthritis that often occurs with a skin condition called psoriasis.
BLEU Score: 1.0
Q2: What is (are) Septic arthritis?
A2: Septic arthritis is inflammation of a joint due to the bacteria that cause gonorrhea has different symptoms and is also called gonococcal arthritis.
System Answer: psoriatic arthritis is a joint problem arthritis that often occurs with with a condition called psoriasis.
BLEU Score: 0.4294447297159816

For above sample question and answer pairs, two questions have same structure and contain only two different words: "Psoriatic" and "Septic" but the answers and healthcare topics are different. But the system answers the similar results for the two questions. The system can answer the first question accurately. But the system answers the first psoriatic information for the second septic question and it tends to decrease BLEU score.

7. Conclusion

The healthcare question answering will be implemented seq2seq model of deep learning to answer right information and adapt self-learning. The system will learn using neural networks where bidirectional RNN one is used as encoder and Luong Attention RNN is used as decoder. According to three Loung Attention Alignment functions: dot, general and concat, this system created nine healthcare question answering models based on three sub datasets: NIHSeniorHealth, NCI and NHLBI in the MedQuAD dataset. This system evaluates three models' performance using BLEU score. According to the experimental result, the accuracy of concat models are lower than the other accuracy of other models. The accuracy of Dot models is higher than the other models. This system uses the same no. of iterations but the

models need to be changed no. of iterations according to the size of datasets to make it more accurate. Generally, the model can be trained with increasing no. of hidden layers to make it more accurate and no. of iterations in model training.

8. Limitation and Future Work

The proposed system uses default setting of the hyperparameters and the different parameters will be used to compare the performance of the models in the future. Moreover, the other datasets in the MedQuAD will be applied in the healthcare question and answer system based on sequence-to-sequence model.

References

[1] Ben Abacha and Demner-Fushman, Dataset Link: MedQuaD Collection of 47k QA pairs (2019): ttps://paperswithcode.com/dataset/ medquad.

[2] Himanshu Gadge, Vaibhav Tode, Sudarshan Madane, Prateek Kachare and Anuradha Deokar, “A Chatbot for Medical Purpose using Deep Learning”, International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181, Vol. 10 Issue 05, May-2021.

[3] Jurgita Kapočiūtė-Dzikienė, “A Domain-Specific Generative Chatbot Trained from Little Data”, March 2020, online Applied Sciences.

[4] Konar.K,“A Comparative Study on Chatbot Based on MachineLearning and Lexicon Based Technique”, International Journal of Innovative Science and Research Technology, Volume 5, May-2020.

[5] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural net-works. Signal Processing, IEEE Transactions on, 45(11):2673–2681, 1997.

[6] Rojas, I., Joya, G., Catala, A. (eds), “Attention-Based Recurrent Neural Networks (RNNs) for Short Text Classification: An Application in Public Health Monitoring”, Advances in Computational Intelligence. IWANN 2019.

[7] Shingte, Kshitija and Chaudhari, Anuja and Patil, Aditee and Chaudhari, Anushree and Desai, Sharmishta, Chatbot Development for Educational Institute (June 6, 2021). Available at SSRN: https://ssrn.com /abstract=3861241 or http://dx.doi.org/ 10.2139/ssrn. 3861241 Mayank.

[8] Salim Akhtar Sheikh, Vineeta Tiwari, Sunita Singhal. "Generative model chatbot for Human Resource using Deep Learning", 2019 International Conference on Data Science and Engineering (ICDSE), 2019.

[9] “TensorFlow,” Available at https://www. Tensorflow. org/.

Customer Churn Prediction using Logistic Regression and Decision Tree (CART) Techniques

Nan Ei Phyo Htet, Sandi Winn Aye
University of Computer Studies, Yangon
naneiphyohtet@ucsy.edu.mm, sandiwinnaye@ucsy.edu.mm

Abstract

The term "customer churn" is used to indicate those customers who are about to leave for a competitor or end their subscription. Customer churn or customer attrition has become an important issue for organizations particularly in subscription-based businesses, where customers have a contractual relationship which must be ended. According to numerous studies, acquiring new clients is significantly more expensive than keeping the ones you already have. As a result, businesses are concentrating on creating precise and trustworthy predictive models to pinpoint potential clients who will churn in the near future. The proposed model uses data from telecom companies on a range of aspects, including customers who left within the last month, services that each client has signed up for, demographic data about customers, and customer account information. This model is presented using machine learning techniques, particularly Logistic Regression (LR) and Decision Tree (CART), to forecast churn for telecoms companies. Comparisons are made to determine the algorithm's efficacy using the provided dataset. The results from a strategy based on Logistic Regression (LR) can predict the telecom market better than Decision Tree (CART) techniques.

Keywords: machine learning techniques, customer churn, customer attrition, Logistic Regression, Decision Tree (CART)

1. Introduction

Customers typically have a wide range of options when selecting a supplier of telecommunications services. They are free to select any service provider they want and to leave their existing one. Customers who choose to choose a different provider than their present one cause the existing provider to lose business and money. Churn is the term for the percentage of customers who leave and stop using the service. Any firm must have a consistent consumer base in order to succeed. Businesses strive to keep customers happy and keep them around for a long time. However, in the actual world, the telecom sector's customer churn rate might reach 25% yearly. Additionally, getting a new customer is 10 times more expensive than keeping an existing one. For business owners, this presents a significant difficulty.

Customers with high levels of loyalty can help businesses increase their core competitiveness and perform better. As a result, many businesses invest a significant amount of money to attract new clients because losing clients is expensive for any company. According to research by Reichheld et al., a company's profits from its current clients increase with the length of their commercial relationship. Customers' net present value will rise by 25% to 95% for every 5% improvement in customer retention rates in the business environment. According to Jones and Sasser's [4] research, an organization's average profit rate will climb by 25% to 85% when its customer turnover rate drops by 5%.

Moreover, recruiting new customers is five to six times more expensive than maintaining existing ones, this is a costly issue. Because of this, the practical significance of the customer churn prediction is that it will help businesses financially. Therefore, identifying the churners can aid businesses in keeping customers, and maintaining relationships with current clients is more crucial [1]. Customer Churn generally means that the customers who are about to move their usage of service to a competing service provider. Different Churn prediction methods gives the prediction about customers who likely to churn in the near future and churn management help to identify such churners and give them some positive offers in order to reduce churn effect.

These customers can be identified using their behavior, customer account information, services that the customer sing up for and demographic details etc. Data can be processed and analyzed using data mining techniques to spot trends and behavioral patterns as well as to improve and optimize business operations to cut down on high-churn consumers. Numerous studies have demonstrated the effectiveness of machine learning algorithms in predicting churning and non-churning events by learning from historical corporate data [6]. All consumer data collected over time is included in the data used in this.

In the experiment of proposed system, Logistic Regression and Decision Tree (CART) algorithms are mainly concentrated to be an efficient prediction model for customer turnover. This experiment focuses on regression-based and tree-based machine learning methods and algorithms for prediction of churn in telecom industries.

2. Related Work

This section gives an analysis on the various works that have been proposed in the area of churn prediction, stating both their merits and demerits.

The study, "Customer Churn Prediction System using Machine Learning," was proposed by Vrushabh Jinde and Prof. Amit Savyanavar [8]. In this study, they suggested a Random Forest algorithm-based approach for predicting client attrition. They made predictions using data from South Asian telecom companies, and the results were very accurate. However, because they employed the Random Forest algorithm, it takes a lot of time. In addition, a lot of features have been used.

The second study, "Churn Prediction in Telecommunication using Logistic Regression with Logit Boost," was written by Hemlata Jain, Ajay Khunteta, and Sumit Srivastava [4]. In this study, Logistic Regression and Logit Boost were utilized as two machine-learning algorithms for forecasting customer attrition. An actual database from the American company Orange and the WEKA Machine-learning program were used in the experiment. The outcomes were displayed using various evaluation metrics.

The title "Customer Churn Prediction Model and Identifying Features to Increase Customer Retention Based on User Generated Content" was written by Essam Abou el Kassem, Shereen Ali Hussein, Alaa Mostafa Abdelrahman, and Fahad Kamal Alsheref [3]. It is a useful tool for predicting customer churn. This study adopted two main strategies: the first was to determine the key elements that influence customer churn, and the second was to identify consumers with a high propensity to leave by examining social media.

In the first method, a dataset is created using real-world surveys, and it is then examined utilizing matched learning techniques like Deep Learning, Logistic Regression, and Naive Bayes algorithms. The second strategy involves predicting customer attrition by looking at user-generated content (UGC), such as comments, postings, messages, and product or service evaluations. They used sentiment analysis to determine the text's polarity (positive/negative) when assessing the user-generated content. The findings demonstrate that while the algorithms were equally accurate, they differed in how they arranged the qualities according to their relative importance in the conclusion.

3. Background Theory

The related background theory with the research work is presented in this section.

3.1. Customer Churn

Customers who are about to switch to a rival or cancel their subscription are referred to as "churners." Attrition of customers, often known as "customer churn," has become a significant problem for businesses, particularly those that rely on subscription models and with whom consumers have binding contracts that must be terminated [1]. Because of the saturated market, the difficulty in attracting new customers, and the ease with which it is possible to switch to another provider, customer churn is a significant problem in the telecommunications industry. It is widely accepted that acquiring a new customer costs six to seven times more than retaining an existing one [7,9].

Instead of trying to win over new consumers with lower retention rates, telecom operators would be better off investing in their current base and earning their continued trust. Unhappiness with the service's quality, excessive expenses, uncompetitive pricing, a lack of benefits for client loyalty, poor support, a long wait for a solution to a problem, customer dissatisfaction, the amount of

service consumption, privacy issues etc. are all potential reasons for churn [5].

3.2. Machine Learning Techniques

Customer turnover can be predicted using a variety of different machine learning algorithms. Techniques like Logistic Regression and Decision Tree (CART) are applied in this proposed system.

3.2.1. Logistic Regression

A type of statistical analysis called logistic regression is used in predictive analytics. It is an effective method for figuring out and forecasting client satisfaction. When the result of the dependent variable is discrete, such as 0 or 1, yes or no, and A, B, or not, logistic regression is used. A binary prediction of a categorical variable, such as customer churn, that depends on one or more predictor variables, such as customer features, is also produced using this method [2].

$$\text{Logit } Y = \left[\frac{Y}{1-Y}\right] \quad (1)$$

$$Y = \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_n x_n}}{1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_n x_n}} \quad (2)$$

where,

- β_0 = Constant Coefficient
- β_n= Coefficient of x_n
- x_n = independent variable (where n= 1,2,…,n)
- P(Y)= Probability that Y equals 1

3.2.2. Decision Tree (CART)

One of the most often used categorization techniques in data mining and machine learning is the decision tree. Because it has a straightforward hierarchical structure that facilitates user comprehension and decision-making [5]. As splitting criteria, entropy, information gain, gain ratio, and the Gini index are employed. Leo Breiman, Jerome Friedman, Richard Olshen, and Charles Stone, a group of statisticians, proposed Classification and Regression Trees (CART) in 1984.

CART is a method for learning binary decision trees that can result in either classification or regression trees. The term "binary" suggests that a decision tree node can only be divided into two groups. CART and C4.5 are quite similar, although CART supports numerical target variables and does not construct rule sets, whereas C4.5 does. CART's splitting criteria use the GINI Index (diversity index) to determine which attribute the branch should be formed in. CART covers missing attribute values as well as data with numerical or categorical values.

$$\text{Gini}=1-\sum_{i=1}^{c}(P_i)^2 \quad (3)$$

```
1  d=0, endtree=0
2  Note(0)=1, Node(1)=0, Node(2)=0
3  while endtree<1
4        if
   Node(2^d − 1) + Node(2^d) + … + Node(2^(d+1) − 2) = 2 − 2^(d+1)
5              endtree = 1
6        else
7              do i = 2^d − 1, 2^d, …, 2^(d+1) − 2
8                    if Node(i) > -1
9                          Split tree
10                   else
11                         Node(2i + 1) = -1
12                         Node(2i + 2) = -1
13                   end if
14             end do
15       end if
16 d = d + 1
17 end while
```

Figure 1. CART Pseudocode

3.3. EDA Analysis

An essential part of a data analyst or scientist's daily practice is exploratory data analysis (EDA). It allows for a thorough examination of the dataset, the formulation or disapproval of hypotheses, and the development of prediction models with a solid basis. It uses a number of statistical tools and data manipulation techniques to describe and appreciate the relationship between elements and how they can affect a company.

Exploratory data analysis, or EDA, should be the first step in any project involving data analysis or data science. Exploratory data analysis is an essential method for performing early studies on data in order to detect patterns, spot anomalies, test hypotheses, and triple-check presumptions with the help of summary statistics and graphical representations.

Exploratory data analysis (EDA) is the process of examining a dataset to look for patterns and abnormalities (outliers) and creating hypotheses based on our understanding of the dataset studies on data in order to detect patterns, spot anomalies, test hypotheses, and triple-check presumptions with the help of summary statistics and graphical representations.

Exploratory data analysis (EDA) is the process of examining a dataset to look for patterns and abnormalities (outliers) and creating hypotheses

based on our understanding of the dataset. EDA comprises creating summary statistics for the numerical data in dataset and creating different graphical representations to facilitate the understanding of the data.

4. Dataset Description

The model compares the effectiveness of two machine learning algorithms: Logistic Regression (LR) and Decision Tree (CART) using data gathered from the IBM Watson Analytics Community. The dataset comprises of 21 columns and 7043 rows of different customer-related data. Using the train-test split technique, 75% of these data will be used for training and 25% for testing.

5. Implementation of the Proposed System

The system flow of the proposed system is depicted in Figure 2. The system uses 75% of dataset as training data and 25% of dataset as testing data. The main goal is to assess how well two data mining methods—Logistic Regression and Decision Tree—perform in predicting telecom customer churn.

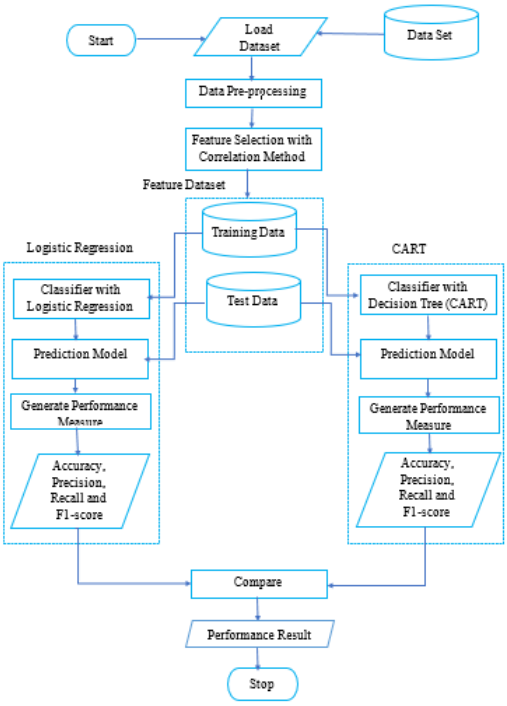


Figure 2. The system flow of the Proposed System

Getting the data from the IBM Watson Analytics Community is the first and most important stage. The dataset typically includes a wide variety of errors and noisy data. The pre-processing phase has been completed, and the data have been cleansed to make them usable. The output of the data pre-processing is noise-free data that may be used for further processing stages.

Next, the feature selection process employs the correlation approach. The feature selection approach is used to eliminate irrelevant, unnecessary, and redundant features that don't significantly improve prediction accuracy. Additionally, it assists in enhancing the classifier's overall efficiency in terms of processing and memory usage. The information is placed in the Selected Feature Dataset, which is split into a training and a testing dataset.

The classifiers: Logistic Regression and Decision Tree (CART), are trained using a training data set. After that, the prediction model is created and prepared for testing. Testing data are given to the prediction model as input, and this system's effectiveness is assessed. The accuracy, precision, recall, and the F1-score of each model are calculated in comparing performance of these two models.

6. Experimental Results

The Telecom Customer Churn dataset undergoes an experiment to determine the most effective supervised machine learning model for forecasting customer churn. The data used in this experiment are the dataset without feature selection, the dataset with feature selection, and the dataset with the proposed system, which includes screening for multicollinearity.

The confusion matrix was used to determine accuracy, precision, recall, and F-measure in order to assess the effectiveness of classifiers in churn prediction. The percentage of total predictions that were accurate is known as accuracy.

$$\textbf{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \quad (4)$$

The percentage of correctly predicted affirmative cases is known as precision.

$$\textbf{Precision} = \frac{TP}{TP + FP} \quad (5)$$

The percentage of affirmative cases that were correctly identified is known as the "recall rate".

$$\text{Recall} = \frac{TP}{TP + FN} \quad (6)$$

The harmonic mean of recall and precision is known as the F-measure, commonly known as the F1_score.

$$\text{F-measure} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (7)$$

In term of accuracy, precision, recall, and F1-score as evaluation metrics, it has been found that the supervised machine learning algorithms: Logistic Regression and Decision Tree (CART) perform better when feature selection and multicollinearity are processed. The Logistic Regression machine learning algorithm outperforms the Decision Tree (CART) algorithm in all studies conducted in terms of all assessment measures.

6.1. Experimental Results of Without Feature Selection

The following results are obtained from the initial dataset. The results of Logistic Regression and Decision Tree (CART) are shown in Figure 3 and Figure 4.

```
-----------------------------------------------------------
Accuracy of Logistic Regression:  0.7940841865756542
Execution time: 0.13391685 seconds
-----------------------------------------------------------
              precision    recall  f1-score   support

           0       0.84      0.89      0.86      1291
           1       0.64      0.53      0.58       467

    accuracy                           0.79      1758
   macro avg       0.74      0.71      0.72      1758
weighted avg       0.78      0.79      0.79      1758
```

Figure 3. Logistic Regression Result of Without Feature Selection Model

```
-----------------------------------------------------------
Accuracy of Decision Tree:  0.7201365187713311
Execution time: 0.07295632 seconds
-----------------------------------------------------------
              precision    recall  f1-score   support

           0       0.81      0.81      0.81      1291
           1       0.47      0.48      0.48       467

    accuracy                           0.72      1758
   macro avg       0.64      0.64      0.64      1758
weighted avg       0.72      0.72      0.72      1758
```

Figure 4. Decision Tree (CART) Result of Without Feature Selection Model

The confusion matrix for both models is shown in the following Figure 5.

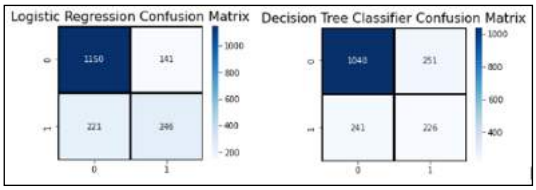


Figure 5. Confusion Matrix for Without Feature Selection Model

The performance evaluation of Logistic Regression and Decision Tree (CART) model (without feature selection) as bar chart is shown in the following Figure 6.

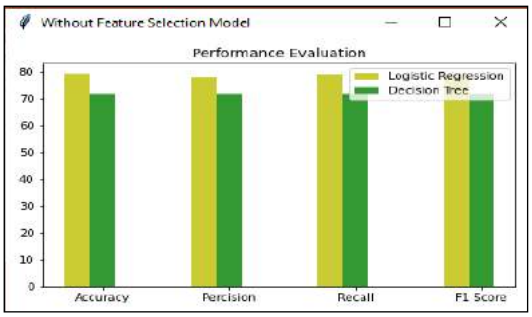


Figure 6. Performance Evaluation of Without Feature Selection Model

6.2. Experimental Results with Feature Selection

In this experiment, the correlation matrix is used for feature selection. Therefore, among 21 features, "Senior Citizen", "Partner", "Dependents", "tenure", "Phone Service", "Internet Service", "Contract", "Paperless Billing", "Payment Methods", "Monthly Charges" and "Total Charges" (11 features) are used for this system. The Logistic Regression and Decision Tree (CART) models' accuracy has increased slightly from 79% to 81% and 72% to 73%, respectively. The results are shown in the following Figure 7 and Figure 8 respectively.

```
-----------------------------------------------------------
Accuracy of Logistic Regression:  0.8083048919226393
Execution time: 0.26317286 seconds
-----------------------------------------------------------
              precision    recall  f1-score   support

           0       0.85      0.90      0.87      1291
           1       0.67      0.55      0.61       467

    accuracy                           0.81      1758
   macro avg       0.76      0.73      0.74      1758
weighted avg       0.80      0.81      0.80      1758
```

Figure 7. Logistic Regression Result of With Feature Selection Model

```
Accuracy of Decision Tree:  0.7286689419795221
Execution time: 0.07132268 seconds
              precision    recall  f1-score   support

           0       0.82      0.80      0.81      1291
           1       0.49      0.52      0.51       467

    accuracy                           0.73      1758
   macro avg       0.66      0.66      0.66      1758
weighted avg       0.73      0.73      0.73      1758
```

Figure 8. Decision Tree (CART) Result of With Feature Selection Model

The confusion matrix for both models is shown in the following Figure 9.

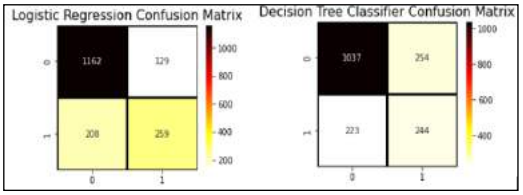


Figure 9. Confusion Matrix of With Feature Selection Model

The performance evaluation of Logistic Regression and Decision Tree (CART) model (with feature selection) as bar chart is shown in the following Figure 10.

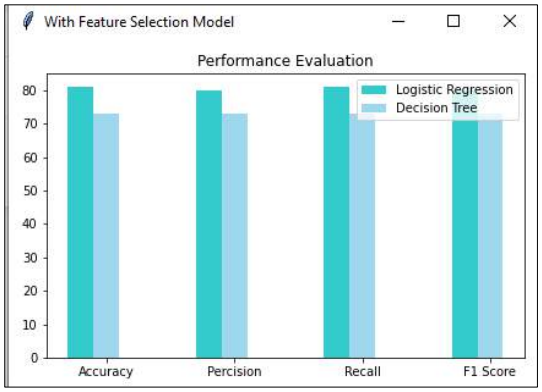


Figure 10. Performance Evaluation of With Feature Selection Model

6.3. Experimental Results of the Proposed System

The proposed system is done by using "Feature Selection and Checking Multicollinearity". In this system, "Senior Citizen", "Partner", "Dependents", "tenure", "Phone Service", "Internet Service", "Contract", "Paperless Billing", "Payment Methods" and "Monthly Charges" are the best 10 select feature for building two models (LR and CART).

The obtained accuracy of Logistic Regression is 83% while CART get 74% of accuracy. This is the best accuracy among three experiments: Without Feature Selection, With Feature Selection and Proposed System (With Feature Selection and Multicollinearity).

```
Accuracy of Logistic Regression:  0.8270762229806599
Execution time: 0.12798548 seconds
              precision    recall  f1-score   support

           0       0.86      0.92      0.89      1291
           1       0.72      0.57      0.64       467

    accuracy                           0.83      1758
   macro avg       0.79      0.75      0.76      1758
weighted avg       0.82      0.83      0.82      1758
```

Figure 11. Logistic Regression Result of Proposed System

```
Accuracy of Decision Tree:  0.7394766780432309
Execution time: 0.03999376 seconds
              precision    recall  f1-score   support

           0       0.83      0.81      0.82      1291
           1       0.51      0.54      0.52       467

    accuracy                           0.74      1758
   macro avg       0.67      0.68      0.67      1758
weighted avg       0.74      0.74      0.74      1758
```

Figure 12. Decision Tree (CART) Result of Proposed System

In Logistic Regression, the likelihood of a client leaving is determined by the data, which showed that 1187 and 267 are, respectively, true positives and false positives, or correctly identified instances. These instances amount to 1454, or 83% of the total 1758 occurrences that make up the total. Decision Tree Model is correctly classified the data 1043 and 254 respectively and get 74% of the accuracy.

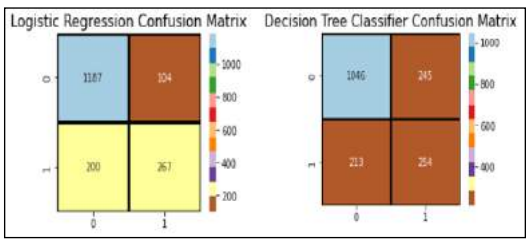


Figure 13. Confusion Matrix of Proposed System

The performance evaluation of Logistic Regression and Decision Tree (CART) model

(proposed system) as bar chart is shown in the following Figure 14.

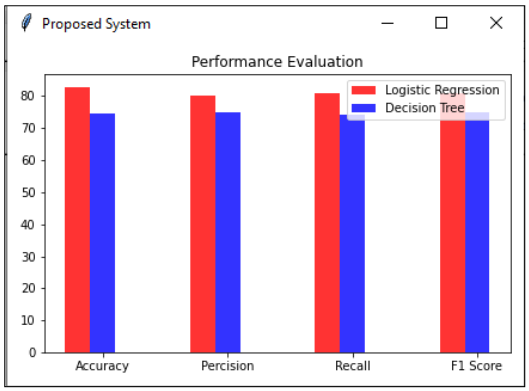


Figure 14. Performance Evaluation of Proposed System

6.4. Model Comparison

When it comes to the overall performance of how well they forecast the outcome, the initial model, the initial model with feature selection, and the proposed models with multicollinearity are not particularly different from one another.

Nevertheless, the simplified model is leaner, quicker, and less resource-intensive because there is a significant difference in the number of variables. The performances of the initial model, the initial model with feature selection, and the final models are contrasted in the following Figure 15.

Comparison	Without Feature Selection Model		Feature Selection Model		Proposed Model	
	LR	CART	LR	CART	LR	CART
Accuracy	79	72	81	73	83	74
Execution time	0.1339s	0.0729s	0.2631s	0.0713s	0.1279s	0.0399s
Precision	78	72	80	73	82	74
Recall	79	72	81	73	83	74
F1-score	79	72	80	73	82	74

Figure 15. Models Comparison

7. Conclusion

Customer churn is one of the most significant sources of revenue for the telecom industry, therefore it helps businesses generate more money. In this proposed system, the performance of two machine learning techniques: Logistic Regression and Decision Tree (CART) are analyzed. The Telecom dataset from the IBM Watson Analytics Community is used in this proposed system. In this system, three experiments: without feature selection, with feature selection and proposed system (with feature selection and multicollinearity) are carried out by using these two modelling techniques. In the first experiment, Logistic Regression model get 79% accuracy score while Decision Tree (CART) has 72%. In the second experiment, the accuracy of Logistic Regression and Decision Tree (CART) improves from 79% to 81% and from 72% to 73% respectively. The last experiment, the proposed system gives the highest accuracy score: improving from 81% to 83% in Logistic Regression and from 73% to 74% in Decision Tree (CART) respectively. This proposed system suggests that Logistic Regression model outperforms than the Decision Tree (CART) model.

References

[1] Abdelrahim Kasem Ahmad, Assef Jafar and Kadan Aljoumaa, “Customer churn prediction in telecom using machine learning in big data platform”, Journal of Big Data, 2019

[2] Chinnu P Johny, Paul P. Mathai, “Customer Churn Prediction: A Survey”, International Journal of Advanced Research in Computer Science, Volume 8, No. 5, May – June 2017

[3] Essam Abou el Kassem, Essam Abou el Kassem, Shereen Ali Hussein, Fahad Kamal Alsheref, “Customer Churn Prediction Model and Identifying Features to Increase Customer Retention based on User Generated Content”, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 11, No. 5, 2020

[4] Hemlata Jain, Ajay Khunteta, Sumit Srivastava, “Churn Prediction in Telecommunication using Logistic Regression and Logit Boost”, International Conference on Computational Intelligence and Data Science (ICCIDS 2019)

[5] Mohamad Firman Maulana, Meriska Defriani, “Logistic Model Tree and Decision Tree J48 Algorithms for Predicting the Length of Study Period”, Journal Penelitian Ilmu Komputer, System Embedded & Logic, Vol. 8 (1): 39 – 48 (March 2020)

[6] Pretam Jayaswal, Bakshi Rohit Prasad, Divya Tomar, and Sonali Agarwal, “An Ensemble Approach for Efficient Churn Prediction in Telecom Industry”, International Journal of

Database Theory and Application Vol.9, No.8 (2016), pp.211-232.

[7] Shin-Yuan Hung, David C. Yen, and Hsiu-Yu Wang, "Applying data mining to telecom churn management," Expert Syst. Appl., vol. 31, pp. 515–524, 2006.

[8] Vrushabh Jinde and Prof. Amit Savyanava," Customer Churn Prediction System using Machine Learning", International Journal of Advanced Science and Technology Vol. 29, No. 5, (2020), pp. 7957-7964

[9] Yasser Khan, Shahryar Shafiq, Abid Naeem3, Sabir Hussain, Sheeraz Ahmed and Nadeem Safwan, "Customers Churn Prediction using Artificial Neural Networks (ANN) in Telecom Industry", International Journal of Advanced Computer Science and Applications, Vol. 10, No. 9, 2019.

Mobile App Recommendation System using K-Means and Item-Based Collaborative Filtering

Khin Aye San, Thu Thu Zan
University of Computer Studies, Yangon, Myanmar
*khinayesan1@ucsy.edu.mm, thuthuzan@*ucssittwayy.edu.mm

Abstract

Recommender Systems are being commonly used in many web applications such as online shopping and e-commerce websites to give valuable suggestions about their products, and items. The main purpose of these systems is to support better recommendations about the items to the users for their interested items. There is a variety of approaches that is used in recommendation such as item-based collaborative filtering, user-based collaborative filtering, model-based collaborative filtering, and content-based recommendation. In this paper, a novel mobile app recommendation system is proposed and the main goal of this system is to provide useful recommended applications to the mobile phone users that are relevant to their interests or needs. This system uses K-means clustering algorithm to cluster the users based on their age and rating values and item-based collaborative filtering based on rating values given by the users. By using this proposed system, mobile users can get very effective recommendations about applications without waste of time and effort.

Keywords: Recommendation, K-Means, Item-Based Collaborative Filtering, Mobile Applications

1. Introduction

Nowadays, recommender systems are an important role in applications because many online users are interested in online applications such as online shopping, e-commerce sites. Therefore, many websites intend to update their systems by using recommendation techniques and recommend applications. Many websites were used to handle the user inputs and some systems used collaborative filtering approach.

Then, the recommendation system has been a recent focus of researchers and practitioners. These systems aim to filter the information and data analysis to help the users in finding the interested items by recommended applications. It learns from a user and recommends applications that they will find most valuable among the available applications.

The recommender system allows e-commerce websites to recommend products to customers by providing relevant information to help them in online shopping tasks. The collaborative filtering is becoming the most popular method to decrease information conflicts. The works in collaborative filtering are working like database creation of users' preferences. These systems have significant success on the Internet and many big companies use collaborative filtering. The purpose is to avoid the bottleneck by exploring the relationships between items, rather than the relationships between users.

Item-based recommendation computes to find similar items to other items that the user interested. The quality of item-based algorithms as same as the user-based algorithms with less computation because the relationships between items are relatively static. This system considers K-means clustering and item-based collaborative filtering method although there are different types of methods used in developing recommendation system. The purpose of this system is to recommend the mobile applications for the users that are relative to their wanted applications. In this system, the K-means clustering algorithm is used to cluster the registered users based on their age and rating values of applications and item-based recommendation is presented to determine the similarities between the items. Then the system uses computed similarity values to give recommended applications.

The remaining sections of this paper is presented as follows. Section 2 will present the related works about the proposed system. Section 3 explains the background theories that are

applied in the proposed system and Section 4 presents the design of the proposed system. Section 5 represents the evaluation metrics, and experimental results and finally Section 6 ends with conclusion and future work.

2. Related Work

The main goal of a recommender system is to provide useful recommendations to users for items that might interest them. The real-world examples are Amazon book suggestions, or Netflix movies recommendations. The design of recommender systems depends characteristics of data available and on the domains. For example, Netflix movie watchers give rating style from numerical value from 1 to 5. While profile attributes are focused in content-based filtering systems, the analysis of historical interactions are based in collaborative filtering systems; and hybrid techniques attempt to combine both designs. There is an active research area in the architecture and their evaluation on real-world problems [9].

There has been many researches and approaches about recommender systems by using collaborative filtering. In the study [1], the merging of user-item based and content based collaborative filtering approach was proposed that provides a few numbers of recommendations. The main contribution of this study is to help in placing better items in relatively smaller list of recommended items and curtail the size of the recommendation list.

The research work [2] proposed an algorithm to balance three current similarity measurements such as: Adjusted cosine similarity, Cosine-based similarity, and Pearson correlation similarity. In this study, there is a comparison between the improved algorithm of traditional measurement metrics and the existing algorithm of the traditional metrics.

A recommender system give useful mobile game applications for the mobile phone users was proposed in the study [3]. This study emphasizes only on item-based collaborative filtering method to recommend game applications.

The researchers in the study [4] proposed a hybrid model to achieve high-quality e-commerce recommendations. The proposed model based on the effective combination of collaborative filtering techniques. The model consists of the following components: item-based collaborative filtering and user-based collaborative filtering. This model also computed the similarity values between the predicted objects. To introduce the recommender system for the medium-scale e-commerce platforms this study can become a methodological basis.

In this paper, a mobile app recommendation system is proposed to recommend mobile applications for mobile phone users with effective and meaningful recommendations. By using the combination of K-Means clustering and item-based collaborative filtering, the proposed system will be more useful and effective than the existing systems.

3. Background Theory

As the recommender systems support personalization for users in buying products, they are becoming popular in commercial tasks. They not only give ecommerce sites for personalization by changing product recommendations according to customers' preferences but also give benefits for ecommerce sites.

The largest ecommerce websites applies recommender systems to provide their customers in buying products. For example, Amazon provides different types of recommendations to help the customers in downloading decisions. This site gives suggestions about the products to users that they have already rated. Users receive recommendation for a list of applications based on the applications that have already rated.

3.1. K-Means Clustering

K-Means Clustering, unsupervised learning algorithm can solve the clustering problems in data science or machine learning. It groups the unlabeled dataset into different clusters. Here, K is defined as the number of pre-defined clusters that need to be created in the process. It groups unlabeled data into different clusters. It is a convenient way to discover the categories of groups in the unlabeled dataset. As each cluster is associated with a centroid, it is known as a centroid-based algorithm.

This algorithm intends to minimize the

distance values between center points and data points to put into the corresponding clusters. The input of the algorithm is unlabeled dataset, divides the dataset into k-number of clusters, and repeats the process until doesn't change the best centroid values. The value of k is predefined in this approach [6].

The main two tasks of k-means clustering algorithm are as follows:

(1) Determines the best k center points values by using and iterative process

(2) Assigns each data point to its closest k-center. Those data points which are near to the particular k-center, create a cluster.

Thus, each cluster has datapoints with some commonalities. The step-by-step procedure of K-means algorithm is described as follows.

Step-1: Select the number K to decide the number of clusters.

Step-2: Select random K points or centroids

Step-3: Assign each data point to their closest centroid, which will form the predefined K clusters.

Step-4: Calculate the variance and place a new centroid of each cluster.

Step-5: Repeat the third steps, which means reassign each datapoint to the new closest centroid of each cluster.

Step-6: If any reassignment occurs, then go to step-4 else go to FINISH.

Step-7: The model is ready.

Figure 1. Steps of K-Means Clustering

The age and rating value are used in this as two dimensions in this proposed system. The system uses K-Means clustering to group the user clusters.

3.2. Collaborative Filtering

The collaborative filtering is a technique used by some recommender systems. Applications of collaborative filtering typically involves very large datasets. Collaborative filtering methods have been applied to many different kinds of data including sensing and monitoring data, such as in mineral exploration, environmental sensing over large areas or multiple sensors financial data, such as financial service institutions that integrate many financial sources or in electronic commerce and web applications where the focus is on user data, etc. Collaborative filtering methods have user-based nearest neighbor algorithms and item-based nearest neighbor algorithms [7].

Collaborative filtering is probably the most widely implemented and best understood technique. The collaborative filtering is a process of filtering or evaluating items using the opinions of others. The collaborative filtering is to filter data based on the similarity of the characteristics of the consumer so it can provide new information to the consumer group that is almost the same. The difference in interest in some members of the group creates new source information that may be useful to other group members.

In general, there are three steps in recommendations: the similar user discovering, finding neighbors, and neighbors' prediction. Collaborative filtering generates predictions or recommendations to users or subscribers intended for one or more items. Items can consist of anything that can be provided by someone such as books, movies, art, article, or tourisms destination. Unavailability rating indicates there is no information linking users with an item. Most systems collect rating values explicitly or implicitly or both. Explicit rating that is obtained opinions on certain items. Implicit rating that is obtained through the action of the customer [5].

3.2.1. Item-Based Collaborative Filtering

The steps and procedure of item-based collaborative filtering are described as follows:

- The system request users to give rating values for their interested items.
- In order to determine similar items, the system correlates rating values.
- The system computes predicted rating values of items that are not rated by users
- If these new items seem to be preferred, the system recommends them to the user.
- Then, the user knows as predicted rating.

The rating style is also important in item-based collaborative filtering and the rating style

that is used in this proposed system is described in Table 1. The rating value 1 is defined as most disliked and 5 is defined as the most liked the items.

Table 1. Rating Style

Rating Values	Description
1	Most Disliked
2	Disliked
3	Medium
4	Liked
5	Most Liked

3.2.2. Adjusted Cosine Similarity

The Adjusted Cosine Similarity algorithm can modify the value of similarity between items. In addition, the algorithm also can estimate the frequent change of items and user relationship. It predicted similarities by forming an offline similarity model that automatically saves time and memory for counting when a user accesses a list of items. The popular similarity model which implemented in recommender systems is given in equation [3].

There are multiple options related to choosing the similarity measure. Adjusted-cosine similarity, Pearson correlation and cosine vector similarity are well-known similarity measures used to compute the similarity.

There is different in similarity calculation between user-based and item-based. While user-based computed with matrix rows, item-based computed with matrix columns. Computing similarity by using basic cosine measure in item-based case has one important drawback the difference in rating scale between different users are not taken into account. The adjusted cosine similarity eliminates this drawback by subtracting the corresponding user average from each co-rate pair. The formulation for adjusted-cosine similarity [8] is described below:

$$sim(i, j) = \frac{\sum_{u \in U} (R_{u,i} - \bar{R}_u)(R_{u,j} - \bar{R}_u)}{\sqrt{\sum_{u \in U} (R_{u,i} - \bar{R}_u)^2} \sqrt{\sum_{u \in U} (R_{u,j} - \bar{R}_u)^2}} \quad (1)$$

Where $R_{u,i}$ is rating value of user u for item i,
$R_{\hat{u}}$ is average rating value of user u for all items,
R_{u}, j is rating of user u for item j.

4. Design of the Proposed System

The recommender system of the mobile phone applications needs to collect data. These applications are downloaded from apkpure.com, Google play store and kaggle.com. The information of application is collected by name, version, size, and source. The collected data and rating values are obtained by surveying. There are two main components of this recommender system: admin and user. The system flow diagrams of these two components are presented as shown in Figure 2 and 3.

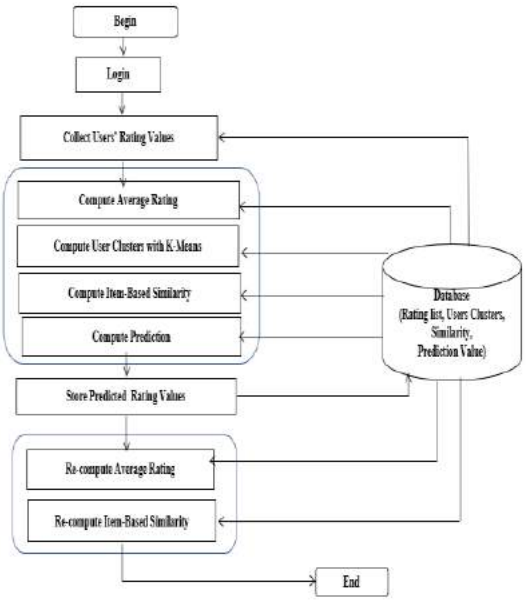


Figure 2. System Flow Diagram for Admin

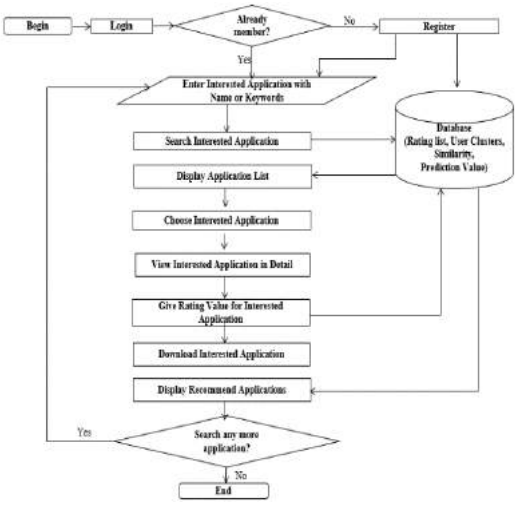


Figure 3. System Flow Diagram for User

4.1. Steps of the Proposed System

In this section, the steps of the proposed system will be presented and the steps of the proposed system are described as follows:

1. Calculate Average Rating
2. Compute User Clusters
3. Calculate the similarity values between items according to rating list
4. Build the similarity matrix
5. Calculate normalization values
6. Calculate the predicted rating values
7. Find denormalization values
8. Recompute the average rating with predicted rating values
9. Recalculate the similarity values with new predicted values.

Firstly, the proposed system uses the following formula to compute the average rating by each user. The ratio of total rating of the user and total number of applications is defined as average rating.

$$\text{Average Rating } \bar{R}_u = \frac{Total\ rating\ of\ user}{Total\ Number\ of\ Applications}$$

Then, the proposed system computes the user clusters by using K-Means clustering algorithm based on their behaviors (age and average rating values). Figure 4 shows the sample clustering results of the registered users by the proposed system.

No	User Name	Age	Avg Rating	Cluster1 [17.1] k1	Cluster2 [20.2] k2	Cluster3 [19.2] k3	Minimum Distance Value	Cluster Number
				Formula D=sqrt((x2-x1)^2+(y2-y1)^2)				
1	Mg Aung Myat	18	2.24	1.59	2.01	1.03	1.03	3
2	Mg Naing Lin	20	1.76	3.09	0.26	1.03	0.26	2
3	Ma Phoo Pwint Thazin	19	2.52	2.51	1.13	0.52	0.52	3
4	Ma Yoon Nadi Zaw	20	2.02	3.17	0.02	1.00	0.02	2
5	Mg Myo Iin	17	1.90	0.90	3.00	2.00	0.90	1
6	Mg Aung Bo Bo Zaw	20	2.45	3.33	0.45	1.10	0.45	2
7	Ma Khin Htet Htet Kyaw	18	2.07	1.47	2.00	1.00	1.00	3

Figure 4. Cluster Registered Users

As the further steps, the proposed system computes the similarity matrix by using adjusted cosine similarity, normalized rating values, predicted rating values by using weighted sum and de-normalization. The formulas that used in this proposed system are described below.

$$NR_{u,N} = \frac{2(R_{u,N} - Min_R) - (Max_R - Min_R)}{(Max_R - Min_R)}$$

Where, $R_{u,N}$ current rating user u gave item N
$NR_{u,N}$ is normalized rating values
Let, Max_R be the maximum rating =5

$$P_{u,i} = \frac{\sum_{N \in similarTo\,(i)} (s_{i,N} * NR_{u,N})}{\sum_{N \in similarTo\,(i)} (|\, s_{i,N}\,|)}$$

Where, $P_{u,i}$ denotes the predicted rating for item i by users u,
$S_{i,N}$ is similarity between item i and N items(from the similarity matrix)
$NR_{u,N}$ is the normalized rating
Once the predicted rating values have been calculated, the similar items will be sorted and recommended for the users.

$$R_{u,N} = \frac{1}{2}((NR_{u,N} + 1) * (Max_R - Min_R)) + Min_R$$

Where, $R_{u,N}$ is the current rating user u gave item N
$NR_{u,N}$ is normalized rating for predict rating value
Let, Max_R be the maximum rating = 5

The user needs to login into this system. If he/she is already a member, he/she can search the interested applications using with criteria: name, keyword, and source. Moreover, they can view the application list and detail respectively. The user can give or not the rating value between (1 to 5). After the user has given the rating value, he/she can see the recommended applications. The users can make searching anymore by their interested application if they want. The system will recommend the similar mobile applications with the user's interested searching application as shown in Figure 5.

Figure 5. Recommend Mobile Applications

5. Evaluation Metrics

The section describes the evaluation metrics that are widely used in recommender systems. The performance of the proposed system will be shown in terms of precision and recall.

Precision and recall that are the most popular metrics used for evaluating information retrieval systems.

$$Precision = \frac{No.of\ Relevant\ Applications\ Retrieved}{No.of\ All\ Retrieved\ Applications}$$

$$Recall = \frac{No.of\ Relevant\ Applications\ Retrieved}{No.of\ All\ Relevant\ Applications}$$

5.1. Experimental Results

This section describes the experimental results of the proposed system in terms of precision and recall. These values were calculated the similarity values above 0.5 and depending on 64 different applications that have rating values by 26 users. The experimental results are described in Table 2.

Table 2. Experimental Results

No	Application Name	Relevant App	All Recommend	All Relevant	Precision	Recall
1	Facebook	17	21	17	80.95	100
2	Viber	17	23	17	73.91	100
3	Telegram	20	24	20	83.33	100
4	Youtube	15	21	15	71.43	100
5	Messenger	14	18	14	77.78	100
6	Block Puzzle Conquer	17	23	17	73.91	100
7	Blossom Blast Saga	21	24	21	87.50	100
8	Bomb Squad	15	21	15	71.43	100
9	Mini World Block Art	16	18	16	88.89	100
10	Fruits Legend	21	22	21	95.45	100

According to the experimental results that are shown in Table 2, the system shows that the precision values from 70 % above 90% and recall values are 100 %.

6. Conclusion

The proposed system helps the users to search their interested mobile applications that they want to download or would like to view. This system applies the K-Means clustering to cluster the users based on their age and rating styles and item-based collaborative filtering to find similar applications. By using the clustering step, this system will provide more useful and effective recommendation results about mobile applications for mobile phone users than the traditional item-based collaborative filtering. As future work, there is necessary to develop a novel recommendation system in large dataset with more applications and users rating values. The proposed system will also apply in other applications such as e-commerce sites, etc.

References

[1] Anand, Shanker, Tewari, “Generating Items Recommendations by Fusing Content and User-Item based Collaborative Filtering”, ICCIDS 2019

[2] Chigozirim, Ajaegbu, “An optimized item-based collaborative filtering algorithm”, JAIHC 2020

[3] Khin Mar Cho, Mya Sandar Kyin, “Mobile Game Applications Recommendation System with Item-Based Collaborative Filtering”, IJARND 2020

[4] Galyna, Chornous, “A hybrid user-item-based collaborative filtering model for e-commerce recommendations”, JIS 2021

[5] Arissa, Aprillia, Nurcahyani, “Recommendation Systems for Mobile Phone Device Application Downloading Using Item Collaborative Filtering Method Implementation", 2009.

[6] Nadia, Fadhil, AL-Bakri, “Collaborative Filtering Recommendation Model Based on K-Means Clustering”, A1-NJS 2019

[7] Mya Thu Zar, "Book Recommender System Using Collaborative Filltering and Content-Based Techniques", 2015.

[8] Rohan Katariya, V Krishna Mohan, “Implementing Collaborative Filtering Approach in Online Event Recommendation System", 2018

[9] Zaw Lin Htay, " Collaborative Filtering-Based Agoregated Query (CFAg Query) in Online Car Recommendation”,2016.

Skin Cancer Diagnosis using Support Vector Machine based on Gray Level Co-occurrence Matrix

Myint Myint Wai, Win Lelt Lelt Phyu
University of Computer Studies, Yangon
myintmyintwai@ucsy.edu.mm, wllphyu@ucsy.edu.mm

Abstract

*As today's world is more developed than ever, the way to diagnose the diseases is better and more precise. It is found that cancer is one of the most common causes of death. In this paper Local Binary Pattern (LBP)*and *Gray level co-occurrence matrix (*GLCM) methods were used to extract features from *dermoscopy images and then Support Vector Machine was used to classify melanoma (the most serious type of skin cancer) or benign (non-cancerous). Five features, contrast, dissimilarity, homogeneity, energy and correlation were extracted by GLCM. Radial Basis Function (RBF) Kernel of SVM was trained with the features and, then tested the images and classified whether the cancer or not. Performance was evaluated with the confusion matrix by testing accuracy, specificity, sensitivity and precision.*

Keywords—Texture feature extraction, Gray level co-occurrence matrix (GLCM), Support vector machine (SVM), Texture features.

1. Introduction

Skin is the largest area of the human body and skin cancer is one of the most dreaded cancers now. Nearly all skin cancers can be cured if found and treated early. If the disease is not known and left untreated, it can easily spread to the whole body. It can be very difficult to treat, and it can affect the life expectancy. Although it has been treated, it can be found that the hard leaves remain on the body unfortunately. Therefore, early detection and prompt treatment of skin cancer are crucial. By accurately diagnosing the disease and assisting clinical decision-making, early classification of skin lesions may improve the likelihood of curing cancer before it spreads.

Software systems can predict outcomes more correctly with the use of machine learning (ML), a type of artificial intelligence (AI), without needing to be explicitly told to do so. Machine learning algorithms use historical data as input to forecast new output values. These factors serve as the driving force for this study, which examines how well Support Vector Machine, a type of machine learning, performs in classifying skin lesions from dermoscopic pictures. In this paper, the model is trained and tested upon the dataset made available by International Skin Imaging Collaboration (ISIC).

The structure of the paper is as follows. In the next section, related work is presented. Section 3 describes the image pre-processing and segmentation. Section 4 represents Feature extraction. Section 5 presented Support Vector Machine. Section 6 reports experimental results. Finally, Section 7 presents conclusion.

2. Related Work

Image classification uses machine learning. There are a lot of machine learning algorithms are also used K-nearest neighbor algorithm, naïve bayes and Random Forest used to classify of image.

The physicians can properly classy skin lesions with the assistance of the authors of [1] Automated systems based on machine learning. Four major phases: are in the proposed system. Two basic steps of Preprocessing the images are hair removal and image enhancement. The thresholding technique was used to partition the skin lesion, converting all pixels with intensities higher than the threshold to foreground values. The background values were applied to the remaining pixels. A user-friendly Graphical User Interface (GUI) is then employed to show the system, allowing for step-by-step classification and the visualization of the statistical features that are utilized for classification.

The authors of [2] are tested with two groups: with and without pre-processing to detect melanoma or not. Without pre-processing system contains only two steps: image gaining and classification. The researcher used (Naïve Bayes) as a tool to classify the skin cancer and aims to use more meaningful data to improve skin cancer detection.

The authors of [3] Dermoscopy pictures are obtained with care by utilizing various contraptions and light conditions, rendering clashing covering data. And so, joining a covering correction step in the pre-dealing with organize might be essential. Seven different types of skin lesions are classified by using step by step algorithm and Random Forest Algorithm applied to make decision the type of diagnosis.

3. Image Pre-Processing and Segmentation

In this section, the image preprocessing steps and segmentation of the system are described. Proposed framework is taken input image, process it and segment the process image using simple thresholding method. Then texture features get is extracted from the image segmentation using local binary pattern and gray level co-occurrence matrix.

3.1. Image Pre-Processing

There are three steps in preprocessing stage as follow:

(i) Converts the image from rgb to grayscale
(ii) Image enhancement and
(iii) Noise Removal

3.1.1. RGB to Grayscale

The brightness of the 8-bit image ranges from 0 to 255, with 0 being black and 255 denoting white. Pure red is encoded as (255,0,0), pure green as (0,255,0), and pure blue as (0,255,0). (0,0,255). The first number in every RGB encoding represents the amount of red, the second value the amount of green, and the final value the quantity of blue. The three numbers' ranges are 0 to 255. Black, white, and all the shades of gray in between are how grayscale images are represented. Any gray value can be represented as an RGB value by a group of three equal numbers. Black is (0,0,0), white is (255,255,255), and medium gray is (127,127,127). The higher the numbers, the lighter the gray. The purpose of converting color images to grayscale is due to the use of grayscale in the GLCM algorithm.

Grayscale intensity = 0.299r + 0.587g + 0.114b.
A shade of dark purple has an RGB value of (100, 0, 150). The weighted average is
Grayscale intensity = 0.299(100) + 0.587(0) + 0.114(150),
Grayscale intensity = 47.

3.1.2. Image Enhancement

Digital images can be improved and changed using filtering techniques. Image filters are used for edge recognition, noise reduction, and blurring. Image filters are mostly used to reduce high and low frequencies (using smoothing techniques) (image enhancement, edge detection). A 2-D convolutional operator is used in this filter. Images used to be blurry. It also eliminates noises and details.

The two-dimensional Gaussian kernel is used in this system. The Gaussian function is used to set a weight value for each pixel in a linear type filter that also includes the gaussian filter. It's linear procedure involves multiplying each adjacent neighbor pixel and adding the results to produce the result for a particular coordinate point denoted by (x, y). The mechanism of it is to move the center of a filter mask from one point to another. The two-dimensional gaussian form is defined as follows:

$$G(x) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(\frac{-x^2-y^2}{2\sigma^2}\right) \quad (1)$$

Where, σ represents as the standard deviation of the same distribution, x and y are expressed as coordinate points (rows and columns) in image pixels.

3.1.3. Noise Removal

The goal of noise removal is to identify and eliminate undesirable noise from images. Pixel values vary as a result of noise. In this system unwanted noise is eliminated using a median filter. The kernel can have a dimension of n x n and be designed to convolve or glide over an image that is m × m in size. The value of a specific pixel is replaced with the median value

of the n x n kernel after this procedure obtains the median value of the n x n kernel on the image. 2D Median filtering example using a 3 x 3 sampling window:

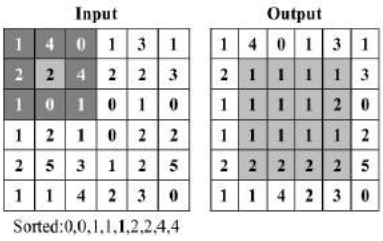
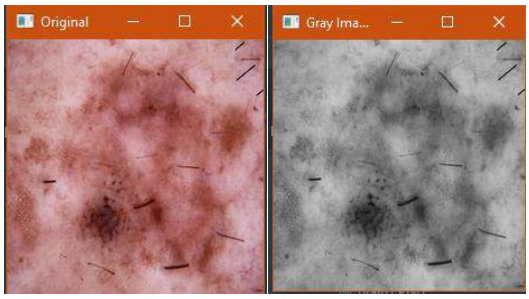
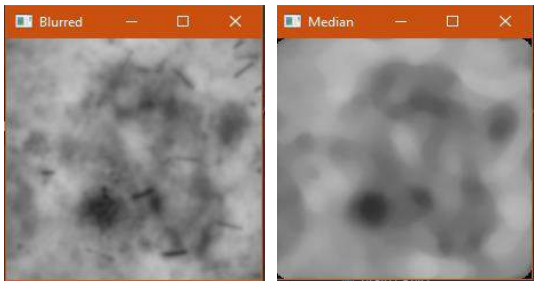


Figure 1. Noise Removal Sample Calculation



(a) Original Image (b) RGB to Grayscale image



(c)Enhancement image (d) After noise removal image

Figure 2. Preprocessing images

3.2. Segmentation

In digital image processing, thresholding is the simplest and popular method of segmentation. Simple thresholding is used in this system. The same threshold value is used for each pixel. The pixel value is set to 0 if it is below the threshold; otherwise, it is set to the maximum value.

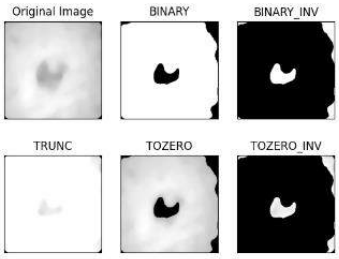


Figure 3. Segmented Images

4. Feature Extraction

The technique of turning raw data into numerical features that can be handled while keeping the information in the original data set is known as feature extraction. Both basic features, such the extraction of color, texture, and shape, and domain-specific features can be found in image features. In this system, Local Binary Pattern (LBP) and Gray Level Co-occurrence Matrix (GLCM) were used for extracting texture features.

4.1. Local Binary Pattern (LBP)

The LBP texture operator thresholds the area around each pixel to label the pixels in the image, and it considers the result to be a binary integer using the following equation:

$$\mathrm{LBP}(x_c, y_c) = \sum_{i=0}^{7} s(g_i - g_c)2^i$$

$$s(x) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases} \qquad (2)$$

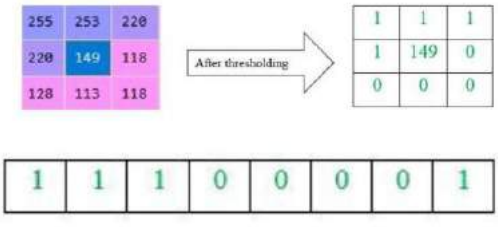


Figure 4. Sample Calculation of LBP

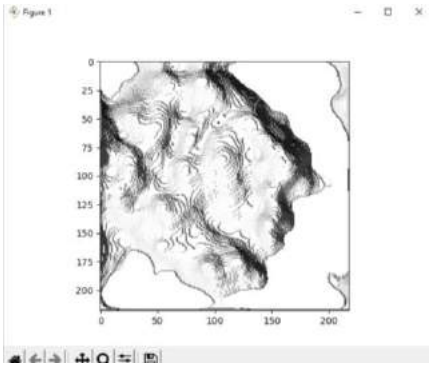


Figure 5. Sample Output Image of LBP

4.2. Gray Level Co-occurrence Matrix (GLCM)

GLCM is used to calculate the dependence of gray level in an image. In GLCM, the number of rows and columns are exactly equal to the number of gray levels in the image. In propose system are used contrast, dissimilarity, homogeneity, energy and correlation.

Figure 6. Extracted feature values from the sample images

5. Support Vector Machine (SVM)

As one of the most effective machine learning algorithms, Support Vector Machine (SVM) is particularly well-liked. It is a supervised machine learning algorithm that may be applied to both regression and classification on linear and non-linear data. It separates the data into different categories by finding the best hyperplane and maximizing the distance between points. In the SVM classifier, it is easy to have a linear hyper-plane between these two classes. SVM works best when the dataset is small and complex. SVM without any kernel have similar performance but depending on features, one may be more efficient than the other.

The Radial Basis Function (RBF) kernel is one of the most powerful, useful, and popular kernels in the Support Vector Machine (SVM) family of classifiers. RBF kernels place a radial basis function centered at each point, then perform linear manipulations to map points to higher-dimensional spaces that are easier to separate. RBF kernels are among the most often used types of kernelization and are also used in this system because of their similarity to the Gaussian distribution. SVM Classifier Algorithm is presented as follow:

1.Training vectors: xi, i = 1. . . L
2. Consider a simple case with two classes:
3. Define a vector y:

$$yi = \begin{cases} 1 \; if \; xi \; in \; class \; 1 \\ -1 \; if \; xi \; in \; class \; 2, \end{cases}$$

4. A hyperplane which separates all data and separating hyperplane with:

$$w^T x + b = 0$$
$$(w^T x_i) + b > 0 \; if \; y_i = 1$$
$$(w^T x_i) + b < 0 \; if \; y_i = -1$$

5. Decision function $f(x) = sign(w^T x + b)$,
x:test data
Variables: w and b are coefficients of a plane

6. Select w, b with the maximal margin.
Maximal distance between $w^T x + b = \pm 1$

6. Proposed System

The architecture of Skin Cancer Diagnosis using Support Vector Machine based on Gray Level Co-occurrence Matrix is presented in Figure 6.

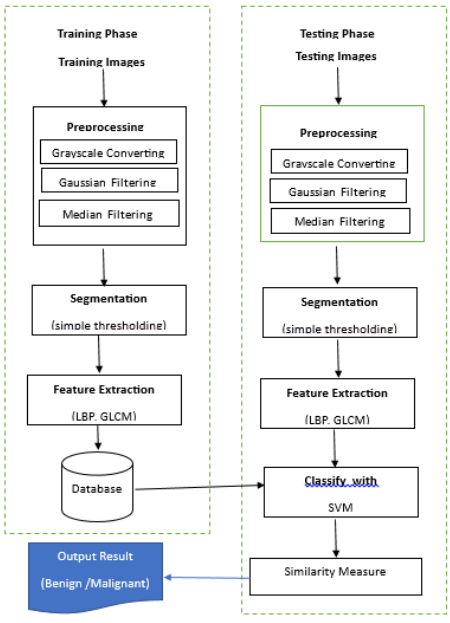


Figure 7. Proposed System Architecture

7. Experimental Result

The following table contains the information of the images that are used to train and test for the proposed system:

Table 1. Dataset Description

Image	Train Data Set	Test Data Set
Size	224 * 224	224 * 224
Quantity	2536	375
Kind	Benign, Melanoma	
Source	International Skin Imaging Collaboration (ISIC).	

Performance of the proposed skin cancer diagnosis system is analyzed with three evaluation parameters: Accuracy, Sensitivity, and Specifcity.

Accuracy = (TP + TN) / (TP + TN + FP + FN)

Specificity = (TN) / (TN + FP)

Sensitivity = (TP) / (TP + FN)

Precision = (TP) / (TP+FP)

The performance evaluation of the system is printed as the following figure 7.

Table 2. Evaluation result of the system

	Precision	Recall	F1-score	support
Benign	0.89	0.91	0.90	269
Malignant	0.76	0.73	0.75	107
Accuracy			0.86	376
Macro avg	0.83	0.82	0.82	376
Weighted avg	0.86	0.86	0.86	376

According to this experiment, the accuracy of the benign is better than malignant. The system accuracy is 0.86.

8. Conclusion

This paper proposed the diagnosis of skin cancer using Support Vector Machine and the two feature extraction methods, LBP and GLCM. By using this system, the experimental result gets the accuracy of 86%. In this approach, features from LBP are first extracted, and only five features from GLCM are then employed, yet the testing showed that this achieved the right level of accuracy.

References

[1] POORNIMA M S1, Dr. SHAILAJA K2 “Detection of Skin Cancer Using SVM” International Research Journal of Engineering and Technology (IRJET) in Volume: 04 Issue: 07 | July -2017.

[2] S. Sasikala1*, S. Arun Kumar2, S.N. Shivappriya3 and Priyadharshini T4 “Towards Improving Skin Cancer Detection Using Transfer Learning” in 2020.

[3] Mustafa Qays Hatem, “Skin lesion classification system using a K-nearest neighbor algorithm”, March 2022.

[4] Ohood Fahdil Alwan College of Al Muqdad, “SKIN CANCER IMAGES CLASSIFICATION USING NAİVE BAYES ALGORITHM”, April, 2022 by University of Diyala, Diyala, Iraq oh85ood@gmail.com

[5] S. Nandhini, Mohammed Abdul Sofiyan, Sushant Kumar, Adnan Afridi, “Skin Cancer Classification using Random Forest”, November 2019.

[6] Uzma Bano Ansari1 “Skin Cancer Detection Using Image Processing” International Research Journal of Engineering and Technology (IRJET) Volume: 04 Issue: 04 | Apr -2017.

Prediction of Employee Attrition using Bayes Risk Post-Pruning in Decision Tree

Win Pa Pa May Phyo Aung, Nilar Aye
University of Computer Studies, Yangon
winpapamayphyoaung@ucsy.edu.mm, nilaraye@ucsy.edu.mm

Abstract

Employee attrition is the departure of employees from the organization for any reason (voluntary or involuntary), including resignation, termination, death, or retirement. Attrition is widely understood to be one of the major problems affecting organizations today. Losing employees has many direct and indirect impacts across a company. It occurs when an employee leaves and is not replaced at all or for a significant amount of time, resulting in a reduction of the workforce. In this system, Decision Tree classifier is used to analyze the causes of employee attrition. And then Bayes Risk Post-Pruning (PBMR) technique is applied to reduce the condition of overfitting on decision tree. The proposed system performance is evaluated various evaluation standards such as precision, sensitivity and F1 score values based on IBM Human Resource Analytic Employee Attrition and Performance dataset from Kaggle site. The proposed system compares the accuracy between before applying post-pruning and after applying Bayes Risk post pruning is applied. This system findings help organizations overcome employee attrition by improving the factors that cause attrition.

Keywords: Data Mining, Machine Learning, Decision Tree, Post-Pruning, Bayes minimum risk

1. Introduction

Artificial intelligence, in general terms, is the ability of a machine to replicate intelligent human behavior. One subfield of artificial intelligence is machine learning. Data—numbers, photographs, or text is the foundation of machine learning. Examples of data include bank transactions, pictures of individuals or even specific bakery goods, repair records, time series data from sensors, or sales reports. The information is collected and made ready to be served as training data, or the material on which a machine learning model will be trained. The application works best when there is more data. Machine learning (ML) has been developed and effectively applied to a wide number of real-world areas, making it one of the fastest-growing disciplines of study. This study presents a comparative analysis of before pruning tree and after pruning tree, to predict employee attrition.

Employee attrition in a company refers to the loss of workers by customary methods like retirement and resignation, clients passing away from old age, or layoffs brought on by a shift in the organization's target demographics. An organization's high rate of staff attrition is a serious problem because it has a big influence on them. Employees depart a business with priceless tacit knowledge that is frequently the source of the company's competitive advantage [10]. The cost of hiring and training new workers, as well as business disruption, is borne by the company as a result of employee turnover. On the other side, higher retention results in lower hiring and training costs as well as the gradual addition of more seasoned employees to the workforce. In order to prevent employee turnover, organizations nowadays have a strong business interest in understanding the causes of staff attrition. As a result, predicting employee attrition and figuring out the main causes of attrition become crucial goals for an organization to achieve in order to improve its human resource strategy [8].

If an employee is fired by their employer for any reason, there could be a number of reasons why they leave, such as a lower salary, a lack of job satisfaction, personal reasons, or environmental concerns. The term for it is involuntary attrition (Kaur & Vijay, 2016). On the other hand, voluntary attrition is sometimes referred to as the employee who leaves an employer. If the person is talented, the organization loses out from this type of attrition. Everyone wants a bigger pay and job security in

the current environment. Because of this, workers quit their jobs right away if they have better opportunities elsewhere.

Machine learning methods are becoming a significant part of predicting employee attrition in the modern field of computer science. Based on the employee's past performance data, such as age, experience, education, previous promotion, and so on, these approaches offer forecasts. The HR department is aware of employee attrition in advance based on the forecast results. As a backup plan for the worker who intends to quit in the near future, the HR department has already begun recruiting new workers.

The IBM Human Resource Analytic Employee Attrition and Performance dataset used in this study is a freely accessible dataset from the Kaggle Dataset Repository. Employee satisfaction, salary, seniority, and demographic information are the four main components of the dataset. The dataset includes a number of attributes that affect the predicted variable "Attrition," which indicates whether an employee left the organization or not based on 1,470 cases and 35 attributes. Attrition is the identified class, and there are 237 instances of "Yes" and 1233 instances of "No" in it.

This study compares employee attrition predictions made before and after decision tree pruning in order to determine whether method is more accurate and efficient at predicting employee attrition.

2. Related Work

The numerous studies that have been offered in the field of employee attrition prediction are analyzed in this section, along with their benefits and drawbacks, in detail.

Norsuhada Mansor, Nor Samsiah Sani, and Mohd Aliff proposed the study, titled "Machine Learning for forecasting Employee Attrition." In this study, the performance of three different classifiers—the artificial neural network classifier and the decision tree classifier—was compared. The authors of this study employed the Employee Attrition and Performance dataset from IBM Human Resource Analytic. For optimization, they also used techniques like regularization and parameter adjustment. The objective of this study was to forecast employee attrition using machine learning classification models.

Ahmed Mohamed Ahmed, Ahmet Rizaner, and Ali Hakan Ulusoy wrote the second article, "A novel decision tree classification based on post-pruning with Bayes minimum risk" [4]. The post-pruning method discussed in this study takes into account a number of evaluation criteria, including attribute choice, accuracy, tree complexity, time spent pruning the tree, precision/recall scores, TP/FN rates, and area under the ROC. The Zoo, Iris, Diabetes, Labor, and Blogger datasets were among the five used by this system. This study demonstrated that the suggested approach generated classification accuracy that was superior to Reduced-error Pruning (REP) and Minimum-error Pruning (MEP). The results of the experiments demonstrated that, across all test datasets, the suggested technique generated classification accuracy that was superior to REP and MEP.

The authors of the article "Predicting Employee Attrition Using Machine Learning Approaches" are Ali Raza, Kashif Munir, Mubarak Almutairi, Faizan Younas, and Mian Muhammad Sadiq Fareed [3]. In order to predict employee attrition, the four cutting-edge machine learning techniques Extra Trees Classifier (ETC), Support vector machine (SVM), Logistic Regression (LR), and Decision Tree Classifier (DTC) were used in this work. In order to identify the causes causing employee attrition, the Employee Exploratory Data Analysis (EEDA) was performed. The primary elements that contribute to employee attrition, according to their analysis, are monthly income, hourly wage, job level, and age. Results of the study were based on the IBM HR employee attrition dataset. Compared to other machine learning algorithms, the Extra Trees Classifier (ETC) that was proposed was more accurate.

Krishna Kumar Mohbey published the article titled "Employee Attrition prediction using Machine Learning Approaches." This study compares the effectiveness of Naive Bayes, SVM, decision trees, random forests, and logistic regression as machine learning techniques. The result that has been supplied will assist us in determining the conduct of employees that can be taken into account the following time. In comparison to other machine learning techniques, experimental data show that the logistic regression strategy can achieve up to 86% accuracy.

3. Background Theory

In this section, the background theory of the proposed system is presented.

3.1. Employee Attrition

The term "employee attrition" refers to the common process through which workers depart from a company for various causes, such as resignation. Attrition among employees can be caused by a variety of circumstances [1]. Employee turnover is greater than that of new hire turnover. When an employee leaves the company, there is a loss to the company since the positions go unfilled. Understanding an organization's progress level is aided by its attrition rate. The high attrition rate indicates that employees are regularly quitting their jobs. Benefits to the organization are lost as a result of the high attrition rate [2]. The attrition rate needs to be kept under control to keep the organization growing.

Many different forms of employee attrition aid in our comprehension of the process. The type of attrition is determined by whether a worker leaves the organization voluntarily or not. When a company terminates an employee's employment, this is an instance of involuntary attrition. An employee leaving one company to work for another is referred to as the external attrition type. When a worker receives a promotion to another position within the same company, internal attrition takes place. The number of persons quitting a company is measured by the attrition rate. By calculating the attrition rate, the problems can be pinpointed and can be issued that must be resolved in order to stop employee attrition. The number of departing workers is divided by the average number of workers during a period of time to determine the attrition rate. The company can track their development over time by looking at the attrition rate.

3.2. Machine Learning Techniques

Several different machine learning techniques can be used to forecast employee attrition. This suggested system employs methods like Decision Tree (ID3) and Bayes Minimum Risk.

3.2.1 Decision Tree (ID3)

In the decision tree method, the appropriate property for each node of a created decision tree is often determined using the information gain methodology. As a result, the characteristic of the current node can be chosen that has the biggest information gain (entropy reduction at the maximum level). This will result in the least amount of data being required to classify the training sample subset produced via later partitioning. Therefore, the degree of mixture of various kinds for all generated sample subsets will be minimized to a minimum when this property is used to divide the sample set included in the present node. Therefore, using an information theory technique will successfully lower the necessary number of object classification divisions.

For the purpose of building a decision tree, the information gain for each and every attribute is calculated, and the attribute with the largest information gain is designated as the root node. Arcs are used to indicate the remaining possible values. Following that, it is determined whether or not each of the potential outcome occurrences falls within the same class. A single name class is used to indicate instances of the same class; otherwise, splitting attributes are used to categorize the instances.

The Concept Learning System (CLS) algorithm, which is a recursive top-down divide-and-conquer algorithm, is the foundation of ID3. Information theory is used by the ID3 family of decision tree induction algorithms to choose the attribute shared by a group of instances to split the data on next. In this manner, attributes are selected repeatedly until a comprehensive decision tree that categorizes each input is obtained. Some of the initial occurrences can be incorrectly classified if the data is noisy. In the case of noisy data, it might be possible to prune the decision tree to lower classification errors. This learning algorithm learns rather quickly, and the decision tree classification system it produces learns fairly quickly as well.

The ID3 algorithm can handle continuous attributes by discretizing them or by examining their values directly to identify the appropriate split point by applying a threshold to the attribute values.

```
ID3(D,X) =
  Let T be a new tree
  If all instances in D have same class c
    Label(T) = c; Return T
  If X = ∅ or no attribute has positive information gain
    Label(T) = most common class in D; return T
  X ← attribute with highest information gain
  Label(T) = X
  For each value x of X
    D_x ← instances in D with X = x
    If D_x is empty
      Let T_x be a new tree
      Label(T_x) = most common class in D
    Else
      T_x = ID3(D_x, X − {X})
    Add a branch from T to T_x labeled by x
  Return T
```

Figure 1. Pseudocode of ID3 algorithm

3.2.2 Bayes Minimum Risk

The Bayes Risk decision rule, which employs Bayes to reduce expectations of loss or risk, is crucial because costs associated with its implementation correspond to misclassification errors.

The risk corresponding to the cost of classifying the data with attributes x into class Ci, where the correct class is Cj, j=1,2,…m . Conditional risk is defined when classifying x into the class Ci as in equation (1).

$$\boldsymbol{l^i(x) = \sum_{j=1}^{2} \lambda_{j,i} \Pr(C_j|x)} \quad (1)$$

where;

$\lambda_{j,i}$ = cost of classifying the data into class C_i ,

C_j = true class

$Pr(C_j|\boldsymbol{x})$ = probability of a subject with attribute $\boldsymbol{x}$ predicted in class C_j

Pr(Cj |x) is calculated using Bayes' Theorem given in equation (2).

$$\Pr(Cj|\boldsymbol{x}) = \frac{\Pr(x \cap Cj)}{\Pr(x)} = \frac{\Pr(x|Cj).\Pr(Cj)}{\Pr(x)} = \frac{\Pr(x|Cj).\Pr(Cj)}{\sum_{l=1}^{m} \Pr(x|C_l).\Pr(C_l)} \quad (2)$$

One special case of risk matrix is zero-one-loss which has the same cost when misclassifying (classifying a subject with the i class as a j class or vice versa) as in equation (3).

$$\boldsymbol{\lambda_{j,i}} = \begin{cases} \mathbf{1}, \boldsymbol{if\ i \neq j} \\ \mathbf{0}, \boldsymbol{if\ i = j} \end{cases} \quad (3)$$

A post-pruning algorithm is run from the bottom (leaf node) – up (root node) by evaluating risk each subtree based on Bayes Risk. Based on zero-one loss risk associated with each parent node t is shown in equation (4).

$$\boldsymbol{R_t^i(x) = \sum_{j=1, j\neq i}^{2} \lambda_{j,i} \Pr(C_j|x)} \quad (4)$$

where:

$R_t{}^i$ $(\boldsymbol{x})$ = risk associated with node t when classifying subject with attribute $\boldsymbol{x}$ into class C_i

$Pr(C_j|\boldsymbol{x})$ = probability of a subject with attribute $\boldsymbol{x}$ predicted in class C_j

The risk associated with the leaf node of its parent node t is shown in equation (5).

$$\boldsymbol{R_l = \sum_{l=1}^{tl} R_l^i(x)} \quad (5)$$

where: $R_l{}^i$ $(\boldsymbol{x})$ = risk associated with leaf node l when classifying subject with attribute $\boldsymbol{x}$ into class C_i

tl = total leaf nodes in the subtree

3.3. EDA Analysis

Data scientists utilize exploratory data analysis (EDA) to examine and analyze data sets and describe their key properties, frequently using data visualization techniques. It makes it simpler for data scientists to find patterns, identify anomalies, test hypotheses, or verify assumptions by determining how to best modify data sources to obtain the answers they need.

EDA is generally used to investigate what data can disclose beyond the formal modeling or hypothesis testing work and offers a better understanding of the variables in the data set and their relationships. Additionally, it may assist you decide whether the statistical methods you are contemplating for data analysis are appropriate.

EDA's major goal is to assist in looking at data before making any assumptions. It can aid in spotting obvious mistakes, as well as better understanding data patterns, spotting outliers or unusual events, and spotting intriguing relationships between the variables.

Exploratory data analysis (EDA) is the process of examining a dataset to look for patterns and abnormalities (outliers) and creating hypotheses based on our understanding of the dataset. EDA comprises creating summary statistics for the numerical data in the dataset and creating different graphical representations to facilitate understanding of the data.

4. Dataset Description

The IBM Human Resource Analytic Employee Attrition and Performance dataset is used in the model to compare the performance of two machine learning algorithms: Decision Tree (ID3) and Bayes Minimum Risk pruning. Employee satisfaction, salary, seniority, and demographic information are the four main components of the dataset. 35 attributes and 1470 cases are included in the dataset. With 237 instances of "Yes" and 1233 cases of "No," the identified class is designated as "Attrition." The dataset is divided into two portions, referred to as the training dataset 70% and testing dataset 30% respectively.

5. Implementation of the Proposed System

Figure 2 shows the system flow of the suggested system. 30% of the dataset is used for testing, and the remaining 70% is used for training. The primary objective is to evaluate how successfully the data mining techniques Decision Tree and Bayes Minimum Risk predict employee attrition.

Pre-preprocessing the study's data is the first phase in its execution. Additionally, the data is comprehensive and there are no existing missing values. The dataset has 35 properties, making it high dimensional. Remove any extraneous characteristics that do not further the study's goals.

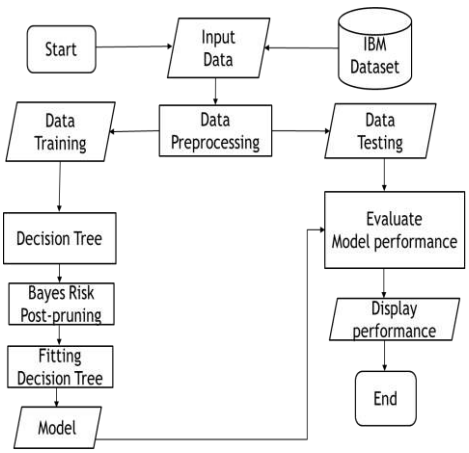


Figure 2. The system flow of the Proposed System

Feature scaling or normalization is used when transforming data during the preprocessing phase. The range of independent variables or data attributes are standardized using the normalization procedure. The discretization technique and changing the attribute type from numerical to nominal were part of the data cleaning and reduction operations. In light of the findings above, four (4) traits were eliminated, leaving the remaining thirty (30) attributes.

A training data set is used to train the Decision Tree (ID3) classifier. The prediction model is then developed and ready to be tested. The effectiveness of this method is evaluated using testing data as input to the prediction model. After creating the decision tree, the Bayes Risk Post-Pruning method was used to produce a decision tree that was suitable. In comparing the performance of the decision tree before and after Bayes Risk Pruning was used, the accuracy, precision, recall, and the F1-score of the model are calculated.

6. Experimental Results

In order to evaluate the performance of classifiers in attrition prediction, the confusion matrix was utilized to calculate accuracy, precision, recall, and F-measure. Accuracy is defined as the proportion of all predictions that were right.

$$\textbf{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \quad (4)$$

Calculating precision, also known as positive anticipated values, involves dividing the total number of predicted positive observations by the number of genuine predicted positive observations.

$$\textbf{Precision} = \frac{TP}{TP + FP} \quad (5)$$

The ratio of accurately predicted positive observations to all actual positive observations is known as sensitivity, recall.

$$\textbf{Recall} = \frac{TP}{TP + FN} \quad (6)$$

The harmonic mean of sensitivity and precision is known as the F1 score, F score, or F measure.

$$\textbf{F-measure} = \frac{2 * Precision * Recall}{Precision + Recall} \quad (7)$$

The supervised machine learning algorithm, Decision Tree, has been determined to be the most

effective in terms of accuracy, precision, recall, and F1-score as assessment metrics (ID3). The training dataset's model performance outperformed the testing dataset's performance in terms of post-pruning. Bayes Risk Post-Pruning is therefore required to solve the overfitting issues. The model's performance in the training dataset declined after Bayes Risk Post-Pruning was used, whereas it improved in the testing dataset. After applying Bayes Risk Post-Pruning, the decision tree's accuracy, precision, and recall values were higher than they would have been without post-pruning.

6.1 Experimental Results of the Proposed System

The proposed system is constructed utilizing the ID3 classification method. After obtaining the decision tree, some of its branches can contain noise or outliers. In order to eliminate unneeded branches or nodes, the system employs the Bayes Risk Post-Pruning approach. Decision Tree's accuracy after using Bayes Risk Post-Pruning was 86%; the accuracy before was 83%. Between the two experiments—without pruning and after pruning system—this one has the highest accuracy.

0.8344671201814059
[[342 22]
[51 26]]

	precision	recall	f1-score	support
0	0.87	0.94	0.90	364
1	0.54	0.34	0.42	77
accuracy			0.83	441
macro avg	0.71	0.64	0.66	441
weighted avg	0.81	0.83	0.82	441

Figure 3. Before Pruning Result of the Proposed System

0.8619909502262444
[[364 20]
[41 17]]

	precision	recall	f1-score	support
1	0.90	0.95	0.92	384
2	0.46	0.29	0.36	58
accuracy			0.86	442
macro avg	0.68	0.62	0.64	442
weighted avg	0.84	0.86	0.85	442

Figure 4. After pruning Result of Proposed System

The training data in Decision Tree, which revealed that 869 and 0 are, respectively, true positives and false positives, or correctly detected cases, predicted the likelihood of a client departing. Of the total 1029 occurrences, these cases make up 869, or 84% of the total. Testing data Model properly identified 369 and 0 data points, respectively, and achieved 82% accuracy.

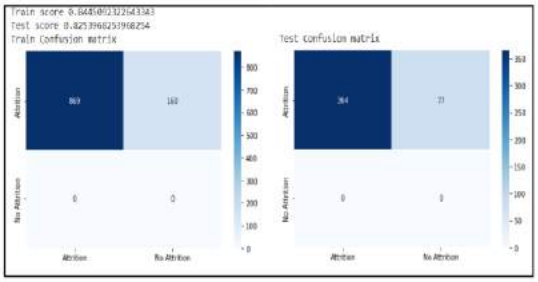


Figure 5. Confusion Matrix of Proposed System

The training data in after pruning Tree, which revealed that 856 and 38 are, respectively, true positives and false positives, or correctly detected cases, predicted the likelihood of a client departing. Of the total 1029 occurrences, these cases make up 894, or 86% of the total. Testing data Model properly identified 356 and 11 data points, respectively, and achieved 83% accuracy.

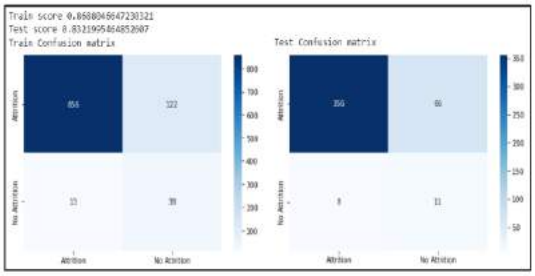


Figure 6. Confusion Matrix of After Pruning Decision Tree of Proposed System

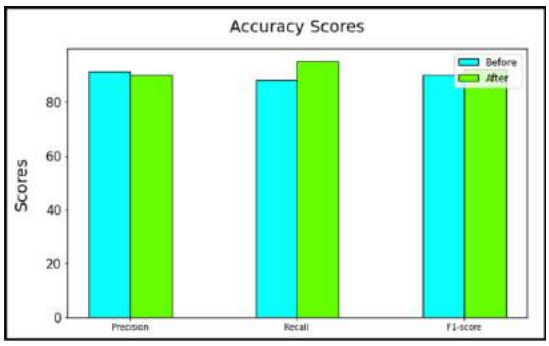


Figure 7. Performance Evaluation of Proposed System

6.2 Model Comparison

The original model, the initial model with feature selection, and the proposed models with

multicollinearity do not differ significantly in terms of how well they predict the outcome overall. However, due to the large difference in the number of variables, the simpler model is leaner, faster, and less resource-intensive. In the following Figure 8, the performances of the original model and the final model are compared.

Comparison	Proposed Model	
	Without Post-Pruning	After Post-Pruning
Accuracy	83	86
Precision	81	84
Recall	83	86
F1-score	82	85

Figure 8. Comparison the Performance of the Model

7. Conclusion

Any business must expect to experience attrition. Customers or workers may leave an organization due to attrition. The decision tree classification technique was used in this study's comparison to predict employee attrition. The IBM Human Resource Analytic Employee Attrition and Performance dataset serves as the basis for the suggested model's operation. For the purpose of overcoming overfitting issues, this study uses Bayes Risk Post-Pruning to condense a decision tree. When Bayes Risk Post-Pruning is used, the system compares the model's performance to that of the testing dataset. When the post-pruning procedure wasn't used, the proposed system had accuracy scores of 83%. The proposed approach acquired a score of 86% accuracy after applying Bayes Risk post-pruning. The system's precision, recall, and f1-score accuracy scores were 82% before the pruning stage. The system then scored with 85% accuracy following the pruning stage. Compared to the decision tree without post-pruning, Bayes Risk Post-Pruning enhanced the decision tree model's capacity to forecast fresh data. The greater accuracy, precision, and recall numbers when Bayes Risk Post-Pruning was used served as evidence of this. The system will improve the dataset feature space in the following study to produce more accurate findings utilizing various machine learning methods.

References

[1] Norsuhada Mansor, Nor Samsiah Sani and Mohd Aliff, “Machine Learning for predicting Employee Attrition”, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 12, No. 11, 2021

[2] Ali Raza, Kashif Munir, Mubarak Almutairi, Faizan Younas, and Mian Muhammad Sadiq Fareed, “Predicting Employee Attrition Using Machine Learning Approaches”, Appl. Sci. 2022, Published: 24 June 2022

[3] Ahmed Mohamed Ahmed, Ahmet Rizaner, Ali Hakan Ulusoy, “A Novel Decision tree classification based on Post-Pruning with Bayes Minimum Risk”, April 4,2018, PLoS ONE 13(4): e0194168

[4] Devina Christianti, Sarini Abdullah, Siti Nurrohmah, “Bayes Risk Post-Pruning in Decision Tree to overcome overfitting problem on Customer churn classification”, Conference paper. January 2020, DOI: 10.4108/eai.2-8-2019.2290487, ICSA 2019, August 02-03, Bogor, Indonesia

[5] Krishna Kumar Mohbey, “Employee Attrition prediction using Machine Learning Approaches”, 06 January 2020, DOI: 10.4018/978-1-7998-3095-5.ch005

[6] Rahul Yedida, Rahul Reddy, Rakshit Vahi, Rahul J, Abhilash, Deepti Kulkarni, “Employee Attrition Prediction”

[7] F.M. Javed Mehedi Shamart, Sovon Chakraborty, Md. Masum Billah, Protiva Das, “A comprehensive study on pre-pruning and post-pruning methods of decision tree classification algorithm”, 5th International Conference on Trends in Electronics and Informatics (ICOEI 2021) Tirunelveli, India, 3-5, June 2021, DOI: 10.1109/ICOEI51242.2021.9452898

[8] Nesserullah, “The Pros and Cons of pruning in classification”, Proceedings of Academics era 32nd International Conference, London, United Kingdom, 18th -19th October 2018

[9] Himani Sharma and Sunil Kumar, “A Survey on Decision Tree Algorithms of Classification in Data Mining”, International Journal of Science and Research (IJSR), April 2016.

[10] O.O Adeyemo & T.O Adeyemo and D. Ogunbiyi, “Comparative Study of ID3/C4.5 Decision Tree and Multilayer Perception Algorithms for the Prediction of Typhoid Fever”, African Journal of Computing & ICT, Vol 8 No.1, March 2015.