Classification of Publication Papers by Using K-nearest Neighbor Algorithm

May Zin Tun
University of Computer Studies, Pathein, Myanmar
Mayzintun87@gmail.com

Abstract

Classification is a data mining or machine learning technique used to predict group membership for data instances. Several major kinds of classification method including decision tree induction method, Bayesian networks method, k-nearest neighbor classification method, case-based reasoning, genetic algorithm and fuzzy logic techniques. Classification is the task of deciding whether a paper belongs to a set of pre-specified classes of papers. Automatic classification schemes can greatly facilitate the process of categorization. Categorization of documents is challenging, as the number of discriminating words can be very large. In this paper, we presented categorization of publication papers by applying k-nearest neighbor classification using the Euclidean Distance measure.K-nearest neighbor method is the simplest and most straightforward method among all classification methods. Hence, k-nearest neighbor method is used to classify different number of nearest neighbors for different categories, rather than a fixed number across all categories in this system.This system is intended to classify different categories from different papers in data sets and to save time for searching papers.

Keywords: Text Categorization, Data Mining, Classification

1. Introduction

Data Mining involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data set. These tools can include statistical models, mathematical algorithm and machine learning methods, Consequently, data mining consists of more than collection and managing data, it also includes analysis and prediction. Classification technique is capable of processing a wider variety of data than regression and is growing in popularity [6].

There are several applications for Machine Learning (ML), the most significant of which is data mining. People are often prone to making mistakes during analyses or, possibly, when trying to establish relationships between multiple features. This makes it difficult for them to find solutions to certain problems. Machine learning can often be successfully applied to these problems, improving the efficiency of systems and the designs of machines.

Today's organizations face a vast volume of knowledge and information. Most of the explicit knowledge is stored in different types of documents but only a few people (often only the authors of the documents) know where to locate them. There are plenty of ways to approach the problem of organizing knowledge in a company [2]. The objective of document classification is to reduce the detail and diversity of data and the resulting information overload by grouping similar documents together. The notion “document classification” is often used to subsume two types of analyses: document categorization and document clustering.

Text categorization is the automated assigning of natural language texts to predefined categories based on their content, is a task of increasing importance. A primary application of text categorization systems is to assign subject categories to documents to support information retrieval or to aid human indexers in assigning such categories[1]. Categorization may be used to filter out documents or parts of documents that are unlikely to contain extractable data.Document categorization may be viewed as assigning documents or parts of documents in a predefined set of categories. Usually this set is created once and for all with so called training documents and, remains unchanged over time. For applying document classification in general, some preprocessing tasks have to be executed [2].In this paper, we presented Euclidean Distance of k-nearest neighbor algorithm for different categories, rather than a fixed number across all categories. More sample (nearest neighbors) will be used for deciding whether a test document should be classified to a category, which has more samples in the training set. To classify a new document, the system finds the k-nearest neighbors among the training documents.

The rest of the paper is organized as follows. In Section 2, we reviews related work. We show the Section 3 gives a description of k-nearest neighbor approach in classification.General architecture of the proposed system in Section 4.. Section 5 describes experimental results of this system and section 6 draws some conclusion and future work.

2. Related Works and Background

There are several applications for Machine Learning (ML), the most significant of which is data mining. People are often prone to making mistakes during analyses or, possibly, when trying to establish relationships between multiple features. This makes it difficult for them to find solutions to certain problems. Machine learning can often be successfully applied to these problems, improving the efficiency of systems and the designs of machines. [3]

Every instance in any dataset used by machine learning algorithms is represented using the same set of features. The features may be continuous, categorical or binary. If instances are given with known labels (the corresponding correct outputs) then the learning is called supervised, in contrast to unsupervised learning, where instances are unlabeled. By applying these unsupervised (clustering) algorithms, researchers hope to discover unknown, but useful, classes of items.

In [5] Murthy (1998) provided an overview of work in decision trees and a sample of their usefulness to newcomers as well as practitioners in the field of machine learning. Decision trees are trees that classify instances by sorting them based on feature values. Each node in a decision tree represents a feature in an instance to be classified, and each branch represents a value that the node can assume. Instances are classified starting at the root node and sorted based on their feature values.

A Bayesian Network (BN) is a graphical model for probability relationships among a set of variables features. The Bayesian network structure S is a directed acyclic graph (DAG) and the nodes in S are in one-to-one correspondence with the features X. The arcs represent casual influences among the features while the lack of possible arcs in S encodes conditional independencies [6].

3. K-nearest Neighbor Classification

K-nearest neighbor (KNN) algorithm is the simplest of all machine learning algorithms. KNN classifier is an instance-based learning algorithm. An object is classified by a majority vote of its neighbors. With the object being assigned to the class most common among its k nearest neighbors. k is a small positive integer. If k=1, then the object is simply assigned to the class of its nearest neighbor. KNN classifier usually applies Euclidean distances as the distance metric. The training samples are described by n-dimensional numeric attributes. Each sample represents a point in an n-dimensional space. When given an unknown sample, a k-nearest neighbor classifier searches the pattern space for the k-training sample that are closest to the unknown sample. These k training samples are the k nearest neighbors of the unknown sample. Closeness is defined "Euclidean distance", where the Euclidean distance between two points $X = (x_1, x_2, ..., x_n)$ and $Y = (y_1, y_2, ..., y_n)$

$$d(X,Y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

According to table 2, we calculate Euclidean Distance of each class:

Natural Language Processing

$$\text{class}\,1 = \sqrt{(\text{-}1)^2 + (-1)^2 + (-1)^2 + (-1)^2 + 1^2 + 1^2.}$$
$$= \sqrt{6} = 2.4494$$

Data Mining

$$\text{class}2 = \sqrt{1^2 + (-1)^2 + (-1)^2 + (-1)^2.}$$
$$= \sqrt{4} = 2.0$$

Web Engineering, XML and Database

$$\text{class}\,3 = \sqrt{1^2 + 1^2 + (-1)^2 + (-1)^2 + (-1)^2 + (-1)^2.}$$
$$= \sqrt{6} = 2.4494$$

Signal Processing

$$\text{class}\,4 = \sqrt{1^2 + {}_1{}^2 + 1^2 + (-1)^2 + (-1)^2 + (-1)^2 + (-1)^2 + (-1^2).}$$
$$= \sqrt{8} = 2.8284$$

Finally, we compare the categories (classes) according to Euclidean distance values and display result this paper in Data Mining Category is minimum value.

Natural Language Processing	=2.4494
Data Mining	= 2.0
Web Engineering, XML and Database	= 2.4494
Signal Processing	=2.8284

4. Proposed System Design

The idea behind the k-Nearest Neighbor algorithm is to build a classification method using no assumptions about the form of the function, $y = f(x_1, x_2, ….. x_p)$ that relates the dependent variable, y, to the independent variables $x_1, x_2, ….. x_p$. In this syetem, k-nearest neighbor method is used to dynamically identify k- observations in the training data set that are similar to a new observation.

Data classification is a two-step process. In the first step, a model is built describing a predetermined set of data classes or concepts constructed by analyzing database tuples described by attributes and each tuple is assumed to belong to a predefined class. In the second step, the model is used for classification and the predictive accuracy of the model (or classifier) is estimated. The architecture of system is shown in Figure 1.

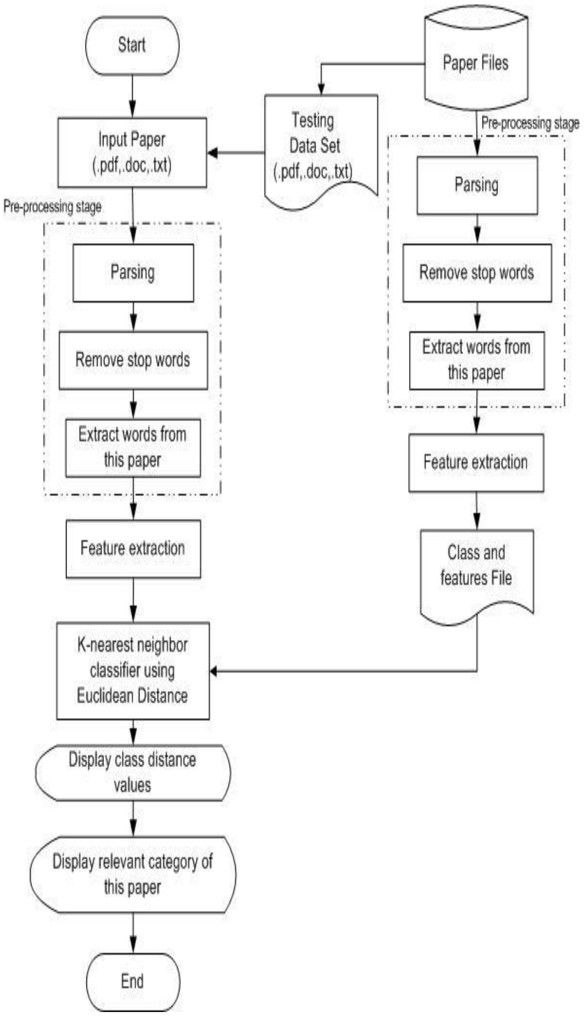


Figure.1. Proposed System Design

4.1. Preprocessing Stage

In the preprocessing stage, consisting of three parts, there are Parsing, Remove stop words and Extract words from the paper.Computers cannot automatically recognize words and sentences. A document is only a sequence of bytes. Computer does not "know: that a space character separates words in a document. Parsing each paper from testing paper dataset has two phase (1) parse individual word and (2) change uppercase to lowercase. Words in a document that are frequently occurring but meaningless in terms of information retrieval are called stop words. Stop list contain stop words, not to be used as index: prepositions, articles, pronouns, some adverbs and adjectives and some frequent words. Words stemming consists of reducing English word to their root word forms.

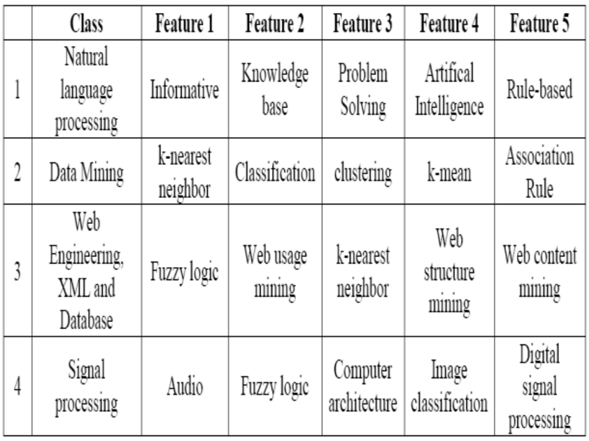
4.2. Feature Extraction

Feature extraction is the second step in document preprocessing. Therefore the training documents are parsed to determine a list of all words (features) contained in the documents. Afterwards feature reducing techniques are applied to reduce the dimension of the list created by the parsing process. Feature extraction is followed by feature selection. The main objective of this phase is to eliminate those features that provide only few or less important information.

4.3. Data Description

In this paper, the system is composed of seven predefined classes or categories among the k-nearest training samples such as Natural language processing, Signal processing, Web engineering, XML and database, Parallel and distributed computing, Image processing, Networking and Security and Data Mining. All instances correspond to points in an n-dimensional Euclidean space. Classification is done by comparing feature of the different points. In arbitrary instance is represented by X(x_1,x_2,..,x_n) denoted features. Target function may be discrete or real-valued (mean value of the k-nearest training). The following table 1 is shown sample 4 class (categories) and some sample features (attributes).

Table 1: Example of Classes and features

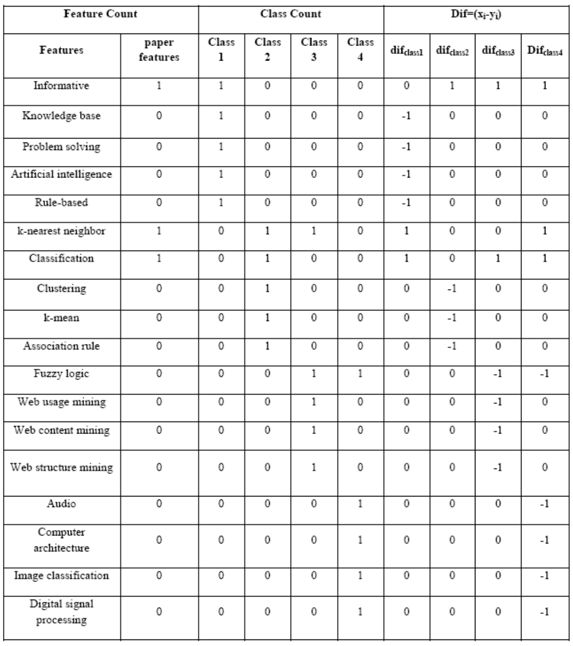


	Class	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5
1	Natural language processing	Informative	Knowledge base	Problem Solving	Artifical Intelligence	Rule-based
2	Data Mining	k-nearest neighbor	Classification	clustering	k-mean	Association Rule
3	Web Engineering, XML and Database	Fuzzy logic	Web usage mining	k-nearest neighbor	Web structure mining	Web content mining
4	Signal processing	Audio	Fuzzy logic	Computer architecture	Image classification	Digital signal processing

4.4. Different Class and Feature Matrix

By calculating Euclidean distance, the first column has features and paper features, features column is that extracted features during feature extraction in preprocessing step. Paper features column is that one feature include in paper is denoted "1" and otherwise "0". The second column is sample four classes (categories). Feature include in each class is assumed "1" and otherwise "0". The third column, we compute the different of paper feature and each class. The different class and feature matrix is shown in table 2.

Table 2: Different Class and Feature Matrix



Feature Count		Class Count				Dif=(x_i-y_i)			
Features	paper features	Class 1	Class 2	Class 3	Class 4	dif_{class1}	dif_{class2}	dif_{class3}	Dif_{class4}
Informative	1	1	0	0	0	0	1	1	1
Knowledge base	0	1	0	0	0	-1	0	0	0
Problem solving	0	1	0	0	0	-1	0	0	0
Artificial intelligence	0	1	0	0	0	-1	0	0	0
Rule-based	0	1	0	0	0	-1	0	0	0
k-nearest neighbor	1	0	1	1	0	1	0	0	1
Classification	1	0	1	0	0	1	0	1	1
Clustering	0	0	1	0	0	0	-1	0	0
k-mean	0	0	1	0	0	0	-1	0	0
Association rule	0	0	1	0	0	0	-1	0	0
Fuzzy logic	0	0	0	1	1	0	0	-1	-1
Web usage mining	0	0	0	1	0	0	0	-1	0
Web content mining	0	0	0	1	0	0	0	-1	0
Web structure mining	0	0	0	1	0	0	0	-1	0
Audio	0	0	0	0	1	0	0	0	-1
Computer architecture	0	0	0	0	1	0	0	0	-1
Image classification	0	0	0	0	1	0	0	0	-1
Digital signal processing	0	0	0	0	1	0	0	0	-1

Finally, we compare the categories (classes) according to Euclidean distance values and display result this paper in Data Mining Category is minimum value.

Natural Language Processing	=2.4494
Data Mining	= 2.0
Web Engineering, XML and Database	= 2.4494
Signal Processing	=2.8284

5. Measuring Performance

In this paper, we evaluated 50 conference papers in training dataset is the Sixth International Conference on Computer Application ICCA-2008. To evaluate the effectiveness of category assignments by classifiers to documents, the standard precision and recall are used in this paper. In equation (1), Precision is defined to be the ratio of correct assignments by the system divided by the total number of the system's assignment. In equation (2), Recall is the ratio of correct assignments by the system divided by the total number of correct assignments. The F1 measure combines precision (p) and recall(r) with an equal weight in the following form:

$$p = \frac{\mathrm{C}}{\sum N} \qquad \_\_\_\_\_\_\_\_\_\_\_\_ \ (1)$$

$$r = \frac{\mathrm{C}}{\mathrm{D}} \qquad \_\_\_\_\_\_\_\_\_\_\_\_ \ (2)$$

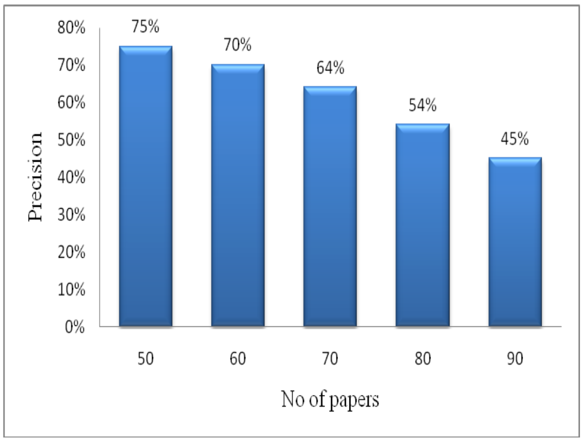


Figure 2: Performance evaluation for precision

Let p = Precision
r = Recall
C = The correct document by the system
D = The correct document in dataset
∑N = The total number of documents

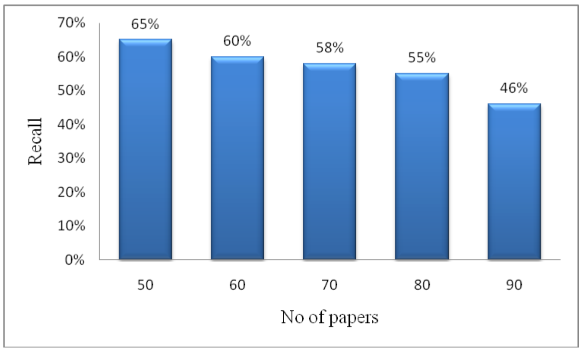


Figure 3: Performance evaluation for recall

6. Conclusion

K-nearest neighbor classification algorithm that learns importance of attributes and utilizes them in the similarity measure. Nearest neighbor will be used for deciding whether paper should be classified to a category. Classify features from paper belonging to small classes with parameter k. Closeness papers are to each other can be evaluated by calculating the Euclidean distance between the features. The users can classify conference papers by using this system. According to these categories, this system is proposed to classify conference papers. This system will recognize semantic analysis (verb, singular, plural, etc.) in the future. This system can classify all training papers in the test data set. This system will develop information retrieval technique for text search. By using this system to ease retrieval, the user can search paper in each category for retrieval and speeds up the retrieval time.

7. References

[1] David D.Lewis, Marc Ringuette, " A Comparison of Two learning Algorithms for Text Categorization", Proceedings of SDAIR-94.

[2] Hiede Brucher, Gerhard Knolmayer, Marc-Andre Mittermayer, "Document Classification Methods for Organizing Explicit Knowledge",

[3] Kotsiants. S.B, "Supervised Machine Learning: A Review of Classification Techniques", Informatica 31 (2007).

[4] Li Baoli, Yu Shiwen, and Lu Qin, "An Improved k-Nearest Neighbor Algorithm for Text Categorization", Proceedings of the 20th International Conference on Computer Processing of Oriental Languages, Syenyang China, 2003.

[5] Murthy, (1998), "Automatic Construction of Decision Trees from Data", A Multi-Disciplinary Survey, Data Mining and Knowledge Discovery 2: 345–389.

[6] Thair Nu Phyu, "Survey of Classification Techniques in Data Mining", IMECS 2009, March 18-20, Hong Kong.

[7] Pascal Soucy, Guy W . Mineau, "A Simple KNN Algorithm for Text Categoriztion",Department of Computer Science, Universite Laval,Quebec,Canada.

[8] Ciya Liao, Shaminopha, Paul Dixon,"Feature Preparation in Text Categorization", Oracle Corporation.

[9] Marcal Rusinol and Josep Llados, " Logo Spotting by a Bag-of-words Approach for Document Categorization", Computer Vision Center, Dept. Ciencies de la Computacio Edifici O, Universitat Autonoma de Barcelona 08193 Bellaterra (Barcelona), Spain.

[10] Esgardo Ferretti , Javier Lafuente and Paola Rosso, "Semantic Text Categorization using the K-nearest Neighbors Method" , LIDIC-Department of Computer Science, National University of San Luis, Argentina.

[11] Min-Ling Zhang and Zhi-Hua Zhou, "A K-nearest Neighbor Based Algorithm for Multi-label Classification" , National Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093 ,China.

[12] "Nearest Neighbor Learning by means of labeled and unlabelled datFranca Debole and Fabrizio Sebastiani, "Supervised Term Weighting for Automated Text Categorization".

[13] Dr.Subhash Ajmani, Vlife Sciences Technologies Pvt.Ltd, "Advantages of K-nearest Neighbor method for Developing QSAR models".

[14] Nearest Neighbor learning by means of labelled and unlabelled data.