Effective Analytics on Healthcare Big Data Using Ensemble Learning

Pau Suan Mung
University of Computer Studies, Yangon
Yangon, Myanmar
pausuanmung@ucsy.edu.mm

Sabai Phyu
University of Computer Studies, Yangon
Yangon, Myanmar
sabaiphyu@ucsy.edu.mm

Abstract

Healthcare big data is a collection of record of patient, hospital, doctors and medical treatment and it is so large, complex, distributed and growing so fast that this data is difficult to maintain and analyze using some traditional data analytics tools. To solve this difficulties, some machine learning tools are applied on such big amount of data using big data analytics framework. In recent years, many researchers have proposed some machine learning approaches on healthcare data to improve the accuracy of analytics. These techniques were applied individually and compared their results. To get better accuracy, this paper proposes one machine learning approach called ensemble learning, in which the results of three machine learning algorithms are combined. Soft voting method is used for combining accuracies. From these results, it is observed that ensemble learning can obtain maximum accuracy.

Keywords: *Ensemble learning, big data analytics, soft voting*

I. INTRODUCTION

Historically, healthcare industry is highly data intensive and large amount of data generated by healthcare industry is driven by record keeping and regulatory requirements for caring patient. Most of these healthcare data are in the form of hard copy and they are required to be digitized rapidly into digital format. These big amount of healthcare data can be used to improve the quality of healthcare and also to reduce the costs. These data can be used in wide area of medical fields and healthcare functions, health management and also disease surveillance.

One of the main purpose of healthcare services is to get the best cares or services to patients. Nowadays many organizations of healthcare services proposed many models of information system. Electronic health records (EHRs) and large amount of complex biomedical data are used to get personalized, predictive and preventive medicine. Genomics and post-genomics technologies can produce large size of raw data in the process of complex biochemical regulatory for the living creatures [4]. Since these EHRs data are heterogeneous, they must be stored in different data forms, different styles and different storage types. Such data can be unstructured, semi-structured or structured. They may also be discrete or continuous.

Healthcare sector needs to be modified and modernized with the big data analytics because big data techniques is new emerging technology. Careful analyzing of healthcare big data is required to be used big data techniques in healthcare inductor. Big data are difficult to analyze and manage with traditional computations. There are huge amount of data in healthcare such as list of patients, doctors and medicine, history of patient records. Big data analytics can be used for integration of heterogeneous data and data quality control. It can also be applied in analysis, modeling, integration and validation [5]. Comprehensive knowledge discovery from large amount of data can be provided in big data analytics application. Big data analytics will discover new knowledge and it can be provided for benefit to the patients, health workers and healthcare policy makers [7].

Big data analytics for healthcare and medical field can enable analysis of large amount of datasets of thousands of patients and correlation between these datasets. This analytics can also integrate the results of analysis of many scientific areas such as bioinformatics, medical imaging, health informatics and sensor informatics.

Using machine learning techniques on big data analytics tools can enhance the performance of the techniques. Many researchers have proved these techniques individually and then compared such results with other methods. This paper proposes an ensemble learning techniques for better accuracy on healthcare big data. A big data analytics tool, Spark, is used in this proposed system. With Spark MLlib, different machine learning tools such as Naïve Bayesian, Decision Tree and K-NN are used as base learners to get particular accuracy of each method and ensemble learning method is applied on the same dataset to get more accuracy than particular methods.

This paper is organized as follows: Section 2 describes the related research works. Characteristics of big data and some of big data analysis techniques are described in Section 3. Ensemble learning proposed in this paper is also included in this section. The next section, Section 4, includes the experimental results of this research work. Section 5 is the last section and it concludes the paper with conclusion and future works.

II. RELATED WORKS

Many machine learning algorithms have been applied in big data analysis and also healthcare big data. The authors in [7] proposed a framework for analysis of stock markets with machine learning algorithms. In this paper, forecasting on decision of stock trading was proposed using ANN and decision support model. Such decision was compared with other methods such as Naïve Bayes, SVM, K-Nearest Neighbor and Decision Tree Model.

In the paper [9], the authors proposed an algorithm called Ensemble Random Forest Algorithm to be analyzed on big data. It presented the difficulties of modelling the insurance business data with classification because of imbalanced of business data that was missing by user features and many other reasons. Heuristic bootstrap sampling approach was combined with the ensemble learning algorithm for mining on insurance business data with large-scale. Ensemble random forest algorithm was also proposed and it can be applied in the parallel computing process and Spark tool was used to optimize memory-cache mechanism. The performance of the proposed algorithm was evaluated by F-Measure and G-Mean. Its experimental results of this proposed system showed that it outperformed in both performance and accuracy with imbalanced data than other classification algorithms.

In the paper [2], the authors proposed efficiency and reliability classification approach for diabetes. The real data was collected from Sawanpracharak Regional Hospital, Thailand and this data was analyzed with gain-ratio feature selection. Naïve Bayesian, K-nearest neighbors and decision tree classification were used as base learners on the selected features. To apply the ensemble learning on these three algorithms, bagging and boosting were combined. Comparison of results of base learners and ensemble learnings were presented. Then the results of each ensemble learning with respective base learner were collected and compared to find the best method for its research work.

The EC3 ensemble learning was proposed in the paper [8]. In which, step by step processing of a novel algorithm named EC3, Combining Clustering and Classification for Ensemble Learning, was presented. Classification and clustering have been successful individually but they had their own advantages and limitations. The author proposed systematic utilization of both of these types of algorithms together to get better prediction results. Its proposed algorithm can also handle imbalanced datasets. 13 datasets from UCI machine learning repository were used and 60% was for training, 20% for testing and other 20% for validation. Six algorithms were used as base classifiers namely Decision Tree, Naïve Bayes, K-nearest neighbors, Logistic Regression, SVM and Stochastic Gradient Descent Classifier. Base clustering methods were DBSCAN, Hierarchical, Affinity, K-Means and MeanShift.

Big data tool applied on healthcare analytics was presented in the paper [5]. K-means clustering techniques was applied on healthcare big data and MongoDB was used as data storage. History data of patient's medical treatment were clustered according to their attribute values. This paper proved that using machine learning tools on big amount of healthcare data is efficient for patients, doctors and medical treatment.

III. ENSEMBLE MODEL FOR BIG DATA ANALYTICS ON HEALTHCARE DATA

Data mining has many specialized forms and machine learning is one of such specialized forms. In machine learning, models are learnt by supervised or unsupervised learning. A mathematical model is generated from a set of data in supervised learning and it includes both the inputs and the respective outputs. Classification and regression algorithms are the type of supervised learning. Classification algorithms can be used for the outputs that are restricted to a limited set of values. Regression can be used for the continuous outputs such as length and temperature. In unsupervised learning, a mathematical model is generated from a set of data. This model has only inputs and no desired output labels. It can be used to search structure in data, such as data points clustering. This type of learning can find patterns in data, and the input can be grouped into clusters or categories, as in feature learning. The process of reducing the number of features and dimensionality reduction can be applied on input data sets.

Ensemble learning method is used for finding better performance and it integrates multiple learning algorithms and produces more performance than the individual algorithm. This type of learning has two main categories namely serialization and parallelization. Serialization refers to the existence of strong dependency between some individual learners that generate result serially and including boosting. Parallelization has no dependence with other learners and therefore the learners can be trained concurrently, including random forest and bagging [6].

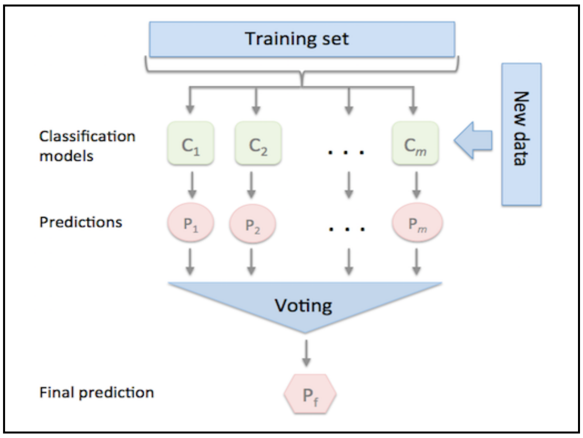

Figure 1. Voting for Accuracy

The basic idea of ensemble learning apply different learning models to get better classification or results [6]. The results from different classification models can be combined in two different ways, that are voting and averaging. Voting method is mainly used in classification and averaging method is commonly used in regression model. There are two common types of voting, hard and soft voting. Hard voting is also known as majority voting and in which each classifier votes individually and the majority of these votes is accepted. In soft voting, each classifier defines the probability values for a particular target class on each data point. By averaging these probabilities, the target label with the greatest average provides the vote [11].

Ensemble learning is also consolidation of some base models of machine learning methods to get one ideal model. Ensemble model can improve accuracy and robustness on single learning methods and also can overcome the constraints of a single method. Therefore it has some different learners called base learners. Base learners are known as powerless learners and their results are combined to get superior to strong learners [6]. Healthcare analytics is a supervised classification problem. Ensemble learning model proposed in this paper combines the results of three algorithms namely, Naïve Bayesian, Decision Tree and K-NN. The method for building ensemble learning and the algorithms used in ensemble learning are presented in the next subsection.

A. Ensemble Learning

Ensemble learning is a method that combines some machine learning algorithms to get better performance. The ensemble learning model is built in two steps. In the first step, all the base learners are used in parallel where the generation from a learner has an impact to the other learners. In the next step, the decisions or results of all base learners are combined in two different way namely, majority voting and weighted averaging. Result combination with majority voting is popular for classification and weighted averaging is popular for regression [6].

B. Machine Learning Algorithms Used

Ensemble learning is a method that combines some machine learning algorithms to get better performance. The ensemble learning model is built in two steps. In the first step, all the base learners are used in parallel where the generation from a learner has an impact to the other learners. In the next step, the decisions or results of all base learners are combined in two different way namely, majority voting and weighted averaging. Result combination with majority voting is popular for classification and weighted averaging is popular for regression [6].

i. Naïve Bayesian

Naïve Bayesian classifier is probabilistic classifier and that is based on Bayes theorem. This classifier is highly scalable and requires a number of parameters (features/predictors). Naïve Bayes is a simple method for building classification model. Class labels are assigned to problem instances and the class labels are drawn from some finite set. This classifiers can be trained in a supervised learning efficiently. It can be used in many complex situations in real-world environment. This method is outperformed by other approaches such as boosted trees or random forests. Naïve Bayesian classifier is used in application with automatic medical diagnosis. [9]. The advantage of Naïve Bayesian classifiers is that small number of training data is required to estimate for classification. The process of Bayes theorem is mathematical and to find the probability for a condition, that is mostly related with a condition already taken. Bayes' theorem is based on the following equation:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \qquad (1)$$

ii. Decision Tree

Decision tree is the form of tree structure and is used for classification or regression models. The data set is broken down into smaller and smaller subsets and therefore an associated decision tree is incrementally developed. Decision tree produces decision nodes and leaf nodes at its final result. Each decision node has two or more branches. The leaf node represents a decision or final result. The topmost decision node in a tree is root node. The root node is best predictor. A decision tree is also a top-down structure and the topmost is the root node. The data is partitioned into subsets that have similar values (homogenous). Entropy value is used in decision tree algorithm to get the homogeneity of a subset. The entropy is zero for the sample with completely homogeneous. If the sample is an equally divided, the entropy is one [5]. The entropy is calculated as:

$$H(x) = -\sum_{i=1}^{n} P(x_i) log_2 P(x_i) \qquad (2)$$

iii. K-NN

K-NN or K-Nearest Neighbors is a simple algorithm and it can be used for both classification and regression. In both cases, the input contains the k closest training examples for the feature space and the output depends on whether k-NN is used for classification or regression. All available cases are stored and new cases is classified using a similarity measure or distance function. The case is assigned to class of its nearest neighbor when K is 1. Weights can be assigned to the contributions for the neighbors and therefore the neighbors nearer can contribute more on average than the more distinct ones. Let d is the distinct to neighbor, a common weighting scheme contains each neighbor a weight of 1/d to each neighbor. The neighbors are in the set of objects. For the neighbors, the class is for K-NN classification and the object property value is for K-NN regression.

C. Proposed Model of Ensemble Learning

This paper proposed an ensemble learning model based on soft voting method. Compared to hard voting method, the class labels can be predicted based on the predicted probabilities p for classifier. Firstly, healthcare data are analyzed using base learners. Big data analytics framework, Spark, and its MLlib (Machine Learning Library) are used for these base learners: Naïve Bayesian, Decision Tree and K-NN. Then accuracies from these base learners on each data object are obtained and soft voting is applied on the average of these accuracies. The result of soft voting is used as prediction value of ensemble learning. Let A has the probabilities, 0.9 for positive and 0.1 for negative on class label, B has 0.8 for positive and 0.2 for negative, and C has 0.4 for positive and 0.6 for negative. Then the soft voting method produces positive as its prediction as it has the average value 0.7 greater than those of negative. The prediction value for the soft voting is calculated as:

$$y = max_i\left(\sum_{j=1}^{n} P_{ij}/n\right) \qquad (3)$$

IV. EXPERIMENTAL RESULTS AND ANALYSIS

In this experiment, three base learners or machine learning algorithms: Naïve Bayesian, Decision Tree and K-NN are used individually and their accuracy results are compared with those of proposed ensemble learning model. Experiments are done with Spark MLlib with Java. All techniques are implemented on a computer system with 8GB RAM, Intel Core i5 processor and Spark 2.4.4 framework on Mojave MacOS. The different datasets on healthcare are taken from UCI machine learning repository [12].

TABLE I. DATASET DESCRIPTION

Data set No.	Data sets Name	No. of Instances	No. of Attributes
1	Lung-cancer	598	57
2	Heart-disease	370	14
3	Diabetes-disease	100000	55
4	Cervical-cancer	858	36

TABLE II. ACCURACY COMPARISON

Data set No.	Classification Accuracy (%)			
	Naïve Bayes	Decision Tree	K-NN	Ensemble Learning
1	88.32	98.65	96.24	99.93
2	87.19	91.48	86.32	93.56
3	80.22	96.50	87.68	98.06
4	89.51	96.85	96.15	98.82

Each of base learners is trained with the training dataset and then testing datasets is applied to get the accuracy of each learner. For each of class label obtained by each data object with each base learner is recorded. According to soft voting method, these class label values are averaged to get the value for class label of ensemble learner. Then testing dataset is applied on each base learner and ensemble learner. The accuracy of each base learner is recorded and are shown in table 2.

In this experiment, four healthcare datasets from UCI machine learning repository, namely Lung Cancer, Heart Disease, Diabetes and Cervical Cancer, are used. As shown in table 2, the accuracy of Naïve Bayesian classifier is minimum for all datasets whereas those accuracies of Decision Tree and K-NN classifiers fluctuate for all datasets. Nevertheless, proposed ensemble learning has the highest accuracy rate. The most attractive reason of using ensemble learning is to get more accuracy. This experience also shows that the ensemble method gets more accuracy than any single method

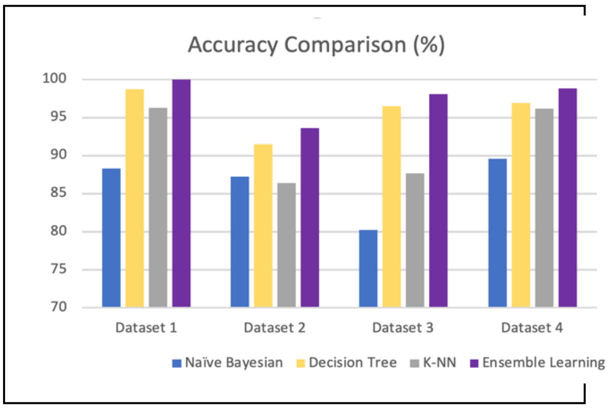

Figure 2. Accuracy Comparison

V. CONCLUSION AND FUTURE EXTENSION

This paper proposed an ensemble learning model that is based on the accuracy values of base learners: Naïve Bayesian, Decision Tree and K-NN classification algorithms. Soft voting method is used for combining accuracies of the base learners. By using this method, the proposed ensemble learning method has the highest accuracy than those of individual classifiers. There are many classification algorithms and many combining methods. Ensemble learning can also be applied with other classifiers and combining methods. Comparing the accuracies on these ensemble learnings are future works of the paper.

REFERENCES

[1] Junhai Zhai, Sufang Zhang and Chenxi Wang, "The Classification of Imbalanced Large Data Sets based on MapReduce and Ensemble of ELM Classifiers", Springer-Verlag Berlin Heidelberg, Springer 2015

[2] Nongyao Nai-arun and Punnee Sittidech, "Ensemble Learning Model for Diabetes Classificaion", Faculty of Science, Naresuan University, Phitsanulok, Thailand, Advanced Materials Research Vols. 931-932 pp. 1427-1431, Trans Tech Publications, Switzerland, 2014

[3] Ping Deng, Honghun Wang, Shi-Jinn Horng, Dexian Wang, Ji Zhang and Hengxue Zhou, "Softmax Regression by Using Unsupervised Ensemble Learning", 2018 9th International Symposium on Parallel Architectures, Algorithms and Programming (PAAP), IEEE 2018

[4] Prashant Dhotre, Sayali Shimpi, Pooja Suryawanshi, Maya Sanghati, "Health Care Analysis Using Hadoop", Department of Computer Engineering, SITS, Narhe, IJSTR 2015

[5] Priyanka Dhaka, Rahul Johari, "HCAB: HealthCare Analysis and Data Archival using Big Data Tool", Indraprastha University, New Delhi, India, IEEE 2016

[6] Shikha Mehta, Priyanka Rana, Shivam Singh, Ankita Aharma, Parul Agarwal, "Ensemble Learning Approach for Enhanced Stock Prediction", Department of Computer Science and Engineering, Jaypee Institute of Information Technology, Noida, India, IEEE 2019

[7] Shuaichao Gao, Jianhua Dai, Hong Shi, "Discernibility Matrix-Based Ensemble Learning", 2018 24th International Conference on Pattern Recognition (ICPR), Beijing, China, IEEE 2018

[8] Tanmoy Chakraborty, "EC3: Combining Clustering and Classification for Ensemble Learning", Dept of CSE, IIIT Delhi, India, 2017 IEEE International Conference on Data Mining, IEEE 2017

[9] Weiwei Lin, Ziming Wu, Longxin Lin, Angzhan Wen and Jin Li, "An Ensemble Random Forest Algorithm for Insurance Big Data Analysis", School of Computer Science, Guanzhou University, China, IEEE 2017

[10] Yong Liu, Qiangfu Zhao and Yan Pei, "Ensemble Learning with Correlation-Based Penalty", School of Computer Science and Engineering, The University of Aizu, Japan, 2014 IEEE 12th International Conference on Dependable, Autonomic and Secure Computing, IEEE 2014

[11] http://rasbt.github.io/mlxtend/user_guide/classifier/ Ensemble VoteClassifier/#methods, "Ensemble Vote Classifier", 2014-2019.

[12] https://archive.ics.uci.edu/ml/index.php, "UCI Machine Lerning Repository"