ISBN 978-981-14-4787-7
Proceedings of 2020 the 10th International Workshop on Computer Science and Engineering (WCSE 2020)
Yangon (Rangoon), Myanmar (Burma), February 26- February 28, 2020, pp. 141-145
doi: 10.18178/wcse.2020.02.024

Ensemble Learning Method for Enhancing Healthcare Classification

Pau Suan Mung [1+] and Sabai Phyu [2]

[1, 2] University of Computer Studies, Yangon, Myanmar

Abstract. Ensemble learning technique is proposed in this paper for better efficiency of healthcare classification and prediction. Healthcare industry is an ever-increasing rise in the number of doctors, patients, medicines and medical records. Medical history records are beneficial for not only individual but also human society. Three popular machine learning algorithms, namely Na ïve Bayes, Support Vector Machine and Decision Tree are applied on this history data as base learners. Two forms of ensemble learning namely bagging and boosting are applied with each base learner for better accuracy than using individually. Comparison results are presented and the experiments show that ensemble classifiers perform better than the base classifier alone. Cervical cancer dataset is used as case study.

Keywords: Ensemble learning, Base learners, Machine learning, Bagging and boosting

1. Introduction

Healthcare data is collection of data about patients, doctors, medicines, treatments and history record of patients and these are so large, distributed, complex and grow so fast. Therefore healthcare industry is highly data intensive. Maintaining and analyzing such large amount of data is a big problem. Traditional database management system are inadequate. Therefore healthcare researchers explore some novel approaches for data aggregation and analysis because of the increasing availability large amount of data of healthcare industry. Cervical cancer used as case study of this paper is the fourth most frequent cancer in women. In 2018, new 570,000 cases are infected and that is 6.6% of all female cancer rate. Approximately 90% of this type of cancer deaths are from the countries with income low or middle. The rate of this cancer can be reduced by prevention, early diagnosis. Effective screening and treatment schemes can be used to reduce this rate. Vaccines are also be provided to protect of human papilloma virus that are common cancer-causing types and therefore the risk of this cancer can be reduced [1].

Data mining or knowledge discovery in database is a computational process that can extract interesting patterns in large volume of history data. The main tasks of data mining include association, classification and clustering. Classification is a process of generating a model to predict appropriate classes of unknown data. Two basic classification processes are model construction and prediction [2]. Some classification techniques are emerged in business. Some are Na ïve Bayesian, decision tree, regression, k-nearest neighbor and neural network. Classification methods can be used in many areas of business such as medical, economy and industry applications.

Most machine learning techniques have their own specific results. Each of them has its own pros and cons. To use the concern of individual algorithms or to get benefits from many algorithms, ensemble learning become more important because they combine the predictions or results of several machine learning models to get the overall result of a system more efficient than single algorithms. Ensemble methods become popular for their prediction capacity than individual algorithms. They have already proven successful in both unsupervised and supervised learning. In this study, three machine learning algorithms are used as base learners and each of these are combined with boosting and bagging to improve their accuracies.

[+] Corresponding author. Tel.: +959428121862
E-mail address: pausuanmung@ucsy.edu.mm

This paper is presented as follows: related research works are presented in section 2 and theory background used in this system is presented in section 3, in which some machine learning approaches used as base learners are included, namely Naïve Bayes, Support Vector Machines and Decision Tree classification, and then ensemble learning methods, boosting and bagging, are presented. Section 4 is the experimental result and analysis. This paper is concluded at section 5 and further extension is discussed in this section.

2. Related Works

Many machine learning algorithms have been applied in real applications. In healthcare environment, many machine learning algorithms were used in classification of different diseases. To get more accurate prediction or classification, researchers proposed ensemble learning techniques in which more than one algorithm are used.

One of the ensemble learning techniques was proposed in paper [3]. It named EC3 – Combining Clustering and Classification for Ensemble Learning. In this paper, step by step processing of this novel algorithm 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 UCI datasets for 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, K-nearest neighbours, SVM, Naïve Bayes, Logistic Regression and Stochastic Gradient Descent Classifier. Base clustering methods were DBSCAN, Affinity, Hierarchical, K-Means and MeanShift.

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 analysed with gain-ratio feature selection. Naïve Bayesian, K-nearest neighbours and decision tree classification were used as base learners on the selected features. To apply the ensemble techniques, bagging and boosting were combined on each of these algorithms. 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.

Authors of the paper [4] proposed ensemble learning methods to enhance performance of network intrusion detection system and to reduce false positive rate using bagging, boosting and stacking. It proposed a prototype model using some base classifiers combining with ensemble learning methods. Four base classifiers: Naïve Bayes, decision tree, rule induction and nearest neighbours are used in this model. It proved the accuracy of 99% for detecting known intrusion and stacking can reduce the false positive rate with a high significantly amount of 46.84%.

Classification and regression tree (CART) with resampling techniques was used for classifying imbalanced datasets in the paper [5]. The authors introduced a simplified method for learning such techniques. Based on many metrices such as precision and classification on minority and majority data respectively, the proposed method was compared with other methods. Matthews Correlation Coefficient (MCC) was used because it is suitable with imbalanced data and classification metrices were true positive, true negative, false positive and false negative.

To find the hidden knowledge in medical field, a paper [6] proposed the graph based association rule mining and its proposed system was intended to use with large medical database. It included process of data mining such as data warehousing, data query and cleaning, and data analysis. 6549 obstetrical patients records were collected for exploratory factor analysis.

3. Ensemble Learning

Ensemble learning is one of machine learning techniques, in which multiple learning algorithms are used to get better predictive performance than any of machine learning algorithm alone. Ensemble methods are able to obtain more accuracy than the individual classifiers which make them up. The idea of ensemble learning model is that many weak learners are used together to build a strong learner because the combination of these weak learners generates increased accuracy than each weak learner. It is also known as committee-based learning because multiple classifiers learn to solve the same problem. An ensemble contains many hypothesis or learners which are usually derived from training data by using each of base learning algorithms. Ensemble methods are well known for their ability to boost weak learners [2].

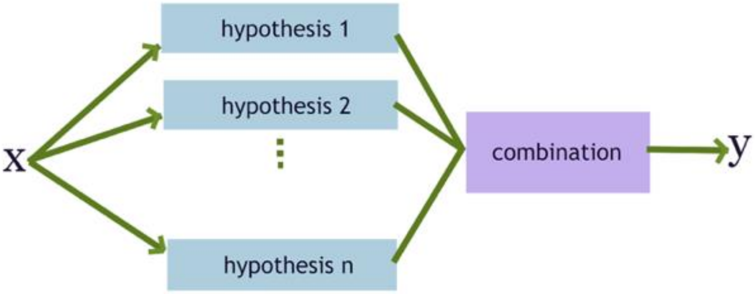

Fig. 1: Ensemble Learning

Two ensemble learning techniques: boosting and bagging, are applied in combination with the base learners. These ensemble learning methods are used to get the results more accurate than using single base classifier.

3.1. Base Classifiers

Three base classifiers: Naïve Bayes, Support Vector Machine and Decision Tree are used in this paper. With these base classifiers, ensemble learning will be applied.

3.1.1. Naïve Bayes

Naïve Bayes classifier based on Bayes theorem is a probabilistic classifier. This classifier is a simple method for building classification model. This classifier is highly scalable and requires a number of parameters (features/predictors). 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. Naïve Bayes classifier is used in application with automatic medical diagnosis [2]. The advantage of this 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.

3.1.2. Support Vector Machine

Support Vector Machine, SVM, can also be used for analysis of classification and regression and they are supervised learning models in machine learning. Each sample in training data is marked as belonging to one or the other of two categories. This algorithm generates a model to assign new example to one category or the other. This model is a representation of the example as points in space, mapped so that examples of the separate categories are divided by a clear gap that is as wide as possible. Next examples are mapped into the same space and predicted to belong to a category based on the side of the gap on which they fall [7].

3.1.3. Decision Tree

Decision tree can be used for classification and also regression, and they are in the form of tree structure. The data set is divided into smaller and smaller subsets until its leaves arrived and therefore an associated decision tree is an association that is developed incrementally. Decision tree produces decision nodes and leaf nodes at its final result. Each decision node has two or more child nodes or branches. The leaf node is a decision or final result. The topmost decision node 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 and entropy value equals one for the sample divided equally [8].

3.2. Boosting

Boosting is primarily used to reduce the variance and bias in a supervised learning technique. The idea of this method is to build the weak classifiers repeatedly to be correlated with the true classification. This technique can reduce the error caused by weak classifier significantly. It refers to algorithm family that converts weak learners (base learners) to a strong learner. Although the potential results are estimated by theoretical perspectives, the true value can only be obtained by applying the technique in the real world classification problems [9].

3.3. Bagging

Bagging or Bootstrap Aggregating can be used to improve the accuracy and makes the model more generalize by reducing the variance i.e. by avoiding overfitting. In this, multiple subsets of training dataset are taken and each subset is used to build the classifier. Then the outputs are combined by using voting to get the final decision with more accuracy. Variance can be reduced and overfitting can be avoided. Although this method is applied in decision tree normally, it can be used with other algorithm as well [9].

4. Experimental Result and Analysis

Ensemble learning proposed in this paper use three machine learning algorithms: Na ïve Bayes, Support Vector Machine (SVM) and Decision Tree as base learners, and two ensemble methods: Boosting and Bagging are used on each base learner. Firstly the accuracies of these base learners are recorded and then boosting is applied on the base learners, and finally bagging is applied on each of them. At each step, the accuracies are recorded and the comparison of these accuracies is shown in Figure 2.

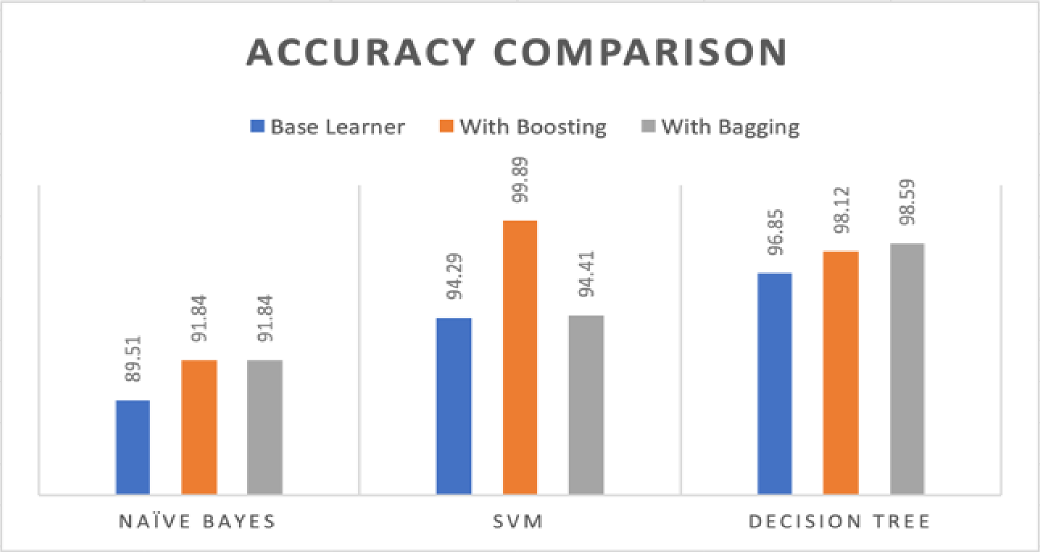

Fig. 2: Accuracy Comparison

This experiments are done with Weka Opening Source Machine Learning Tool running on Java language and the dataset ‘Cervical cancer’ is taken from UCI machine learning repository [10].

This experiment shows that Na ïve Bayes produces least accuracy than other two methods. Although the accuracy of SVM base learner is less than those of Decision Tree, Boosting with SVM produces the most accuracy than other methods. For Bagging method, Decision Tree has the most accuracy than those of other two methods. In all methods, ensemble learning, combining with boosting and bagging, has more accuracy than base learner alone.

5. Conclusion and Further Extension

This research work evaluate the accuracies of various classification models. Cervical cancer dataset obtained from UCI is used as case study. Three base learners or classifiers: Na ïve Bayes, SVM and Decision Tree, and two ensemble methods: Boosting and Bagging are used in this study. The result of this experiment revealed that SVM produces best accuracy for Boosting and Decision Tree produces best accuracy for

Bagging. This experiment also shows that the ensemble methods get better performance than its base learners alone. This finding of experiment are useful for choosing the classification algorithm for future application and ensemble learning can also be used for better accuracy in application. Other classification algorithm, stacking, can be used and it is further aspect of this work.

6. References

[1] World Health Organization - WHO, Cervical Cancer, https://www.who.int/cancer/prevention/diagnosis-screening/cervical-cancer/en

[2] Nongyao Nai-arun and Punnee Sittidech. Ensemble Learning Model for Diabetes Classification, *Trans Tech Publications, Switzerland*,Advanced Materials Research Vols. 931-932 (2014) pp. 1427-1431

[3] Tanmoy Chakraborty. EC3: Combining Clustering and Classification for Ensemble Learning, *IEEE International Conference on Data Mining*, 2374-8486/17, IEEE 2017

[4] Iwan Syarif, Ed Zaluska, Adam Prugel-Bennett and Gary Wills. Application of Bagging, Boosting and Stacking to Intrusion Detection, *School of Electronics and Computer Science, UK and Electronics Engineering Polytechnics Institute of Surabaya, Indonesia.*

[5] Supajittree Boonamnuay, Nittaya Kerdprasop and Kittisak Kerdprasop. Classification and Regression Tree with Resampling for Classifying Imbalanced Data, *International Journal of Machine Learning and Computing,* Vol. 8, No. 4, pp. 336-340, 2018

[6] Wael Ahmad AlZoubi, Mining Medical Databases Using Graph based Association Rules, *International Journal of Machine Learning and Computing,* Vol. 3, No. 3, pp. 294-296, 2013

[7] Rohith Gandhi, Support Vector Machine – Introduction to Machine Learning Algorithms, 2018, http://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

[8] Rutgers, School of Art and Science, Decision Tree Regression, https://www.saedsayad.com/decision_tree_reg.htm

[9] Aporras, What is the Difference Between Bagging and Boosting, 2016, https://quantdare.com/what-is-the-difference-between-bagging-and-boosting/.

[10] UCI Machine Learning Repository, https://archive.ics.uci.edu/ml/index.php