EVALUATING THE EFFECTIVENESS OF ENSEMBLES OF DECISION TREES: BAGGING AND BOOSTING

D. Lavanya[1], Dr. K. Usha Rani[2]

Abstract: The aim of classification learning is to classify the data into different classes by developing a model. For the classification activity the powerful and popular tool is Decision Tree. The machine learning of medical diagnosis would be advantageous. Decision Tree classification has been used for predicting medical diagnoses. Multiple models are used in Ensemble learning methods to achieve better prediction accuracy than single model. The two popular ensemble learning methods Bagging and Boosting are studied on Breast Cancer datasets in this study.

Keywords: Bagging, Boosting, Classification, Decision Trees. Ensemble Decision Trees, Machine learning.

1. INTRODUCTION

Classification is one the most important tasks in data mining which is also an interesting topic to the researchers as it accurately and efficiently classifies the data for knowledge discovery. Classification is an activity that assigns labels or classes to different objects or groups. For the classification activity the powerful and popular tool is Decision Tree [1]. In this knowledge is represented in the form of rules. This makes the user to understand the model in an easy way. In a decision tree each node represents a test on an attribute, the result of the test is a branch and a label or class is represented by a leaf node. To classify an unknown record, its attributes are tested in the tree from the root until a leaf a path is defined. Though each leaf node has an exclusive path from the root, many leaf nodes can make the same classification.

The automation (machine learning) of medical diagnosis would be advantageous. Decision Tree classification has been used for predicting medical diagnoses. Among data mining methods for classification, decision trees have several advantages: Provide human readable rules of classification, Easy to interpret, Construction of decision tree are fast and provide better accuracy. The success of machine learning on medical datasets depends on various factors. The quality of the data is one such factor. Knowledge discovery during training is more difficult if information is irrelevant or redundant or noisy. To overcome the difficulty Data Preprocessing plays an important role. Especially Feature Selection (FS) with decision tree classification greatly enhances the quality of the data in medical diagnosis [2]-[3].

The approach of Ensemble systems is to improve the confidence with which we are making right decision through a process in which various opinions are weighed and combined to reach a final decision. The ensemble based system gain attention and popularity because of enhanced classification accuracy and applicability in wide range or applications. Especially in clinical diagnosis a high rate of reliability

ISBN 978-93-81583-57-0

demands ensemble decision trees more suitable than a base (single) classifier. In this study, a study on two popular ensemble classifiers ie., Bagging and Boosting on Breast Cancer data sets is conducted.

The organization of the paper is as follows: Section 2 deals with brief overview of related work and ensemble systems. In section 3 experiments and evaluation of results are provided. Section 4 presents the conclusion.

2. BACKGROUND

2.1 Ensemble Systems: The misclassification is a common problem that can hinder clinical relevance and practicality. Ensemble techniques have been developed which can decrease classification error rates. By combing individual opinions of several experts the most informed final decision may be reached through automated decision making applications. This procedure also known under various names, such as Committee of Classifiers, Multiple Classifier Systems, Mixture of Experts or Ensemble based Systems produce most favorable results than single-expert systems under a variety of scenarios for a broad range of applications. Some of the ensemble-based algorithms are Bagging, Boosting, AdaBoost, Stacked Generalization and Hierarchical Mixture of Experts.

Two decision committee learning approaches, boosting and bagging, have received extensive attention recently and they have been deeply analyzed. One is Bootstrap Aggregation or Bagging [4] that improves the accuracy of the classification models by averaging the classifiers from a bootstrapped training set. The second is Boosting [5] a method for improving classifiers by assigning weights to the misclassified tuples and then combining classifiers. These two methods with Decision Trees have also been proved to be very successful for many machine learning algorithms. In the analysis of medical data also these methods are proven useful [6]-[7].

***BAGGING*:** Bagging or Bootstrap Aggregating by Breiman is one of the earliest, simplest to implement and most intuitive ensemble based algorithm with good performance [4]. In this raining data sets are randomly drawn and each subset is used to train a different classifier of the same type. Finally, all the classifiers are combined by taking majority vote of their decisions. The class chosen by most classifiers for a particular instance is the ensemble decision. When the data set is of limited size bagging is more suitable. Variations of bagging are Random Forests and Pasting Small Votes.

***BOOSTING*:** Boosting is another ensemble based algorithm which can be used to improve the classification accuracy by resampling the data. Boosting uses a weighted average of results by applying a prediction method to various samples where as bagging uses a simple average of results. In Boosting the samples are drawn differently. The incorrectly predicted cases from a given step (classifier) are given increased weight during the next step. A popular AdaBoost algorithm [8] is a more general version of Boosting algorithm. Variations of AdaBoost are

ISBN - 978-93-81583-57-9

AdaBoost.M1 to handle multiple classes and AdaBoost.R capable of handling regression problems.

Hence these two ensemble techniques are considered in the study to improve CART classifier performance.

2.2. Overview of Related Work: Some of the studies which have been focused on the importance of bagging and boosting ensemble methods in the field of medical diagnosis are reported here. These studies have applied different approaches to classify the data with high classification accuracies.

My Chau Tu et.al. [9] proposed the use of bagging with C4.5 algorithm and Bagging with Naïve Bayes algorithm to diagnose the heart disease of a patient.

My Chau Tu et.al. [10] used bagging algorithm to identify the warning signs of heart disease in patients and compared the results of decision tree induction with and without Bagging.

Tsirogiannis et.al. [11] applied bagging algorithm on medical databases using the classifiers -Neural Networks, SVM'S and Decision Trees. Usage of Bagging proved improved accuracy than without Bagging.

Pan wen [12] conducted experiments on ECG data to identify abnormal high frequency electro cardiograph using decision tree algorithm C4.5 with Bagging.

Kaewchinporn et.al. [13] presented a new classification algorithm TBWC combination of decision tree with bagging and clustering. This algorithm is experimented on two medical datasets: cardiocography1, cardiocography2 and on some other datasets not related to medical domain.

Jinyan LiHuiqing Liu et.al. [14] experimented on ovarian tumor data to diagnose cancer- using C4.5 with and without bagging. Bagging produced better accuracy than without Bagging.

Dong-Sheng Cao et.al. [15] proposed a new Decision Tree based ensemble method combined with Bagging to find the structure activity relationships in the area of Chemometrics related to pharmaceutical industry.

Liu Ya-Qin et.al. [16] experimented on breast cancer data using C5 algorithm with bagging to predict breast cancer survivability.

Tan AC et.al. [17] used C4.5 decision tree and Bagged C4.5 Decision Tree on seven publicly available cancerous micro array data and compared the prediction performance of these methods.

CaiLing Dong et.al. [18] proposed a modified Boosted Decision Tree for breast cancer detection to improve the accuracy of classification.

Jaree Thangkam et.al. [19] performed a work on survivability of patients from breast cancer. The data considered for analysis was Srinagarind hospital databases during the period 1990-2001. In this approach first the data was preprocessed using RELIEFF Attribute (feature) Selection method then AdaBoost algorithm is used with CART as base learner.

Kotsiantis et.al. [20] did a work on Bagging, Boosting and Combination of Bagging and Boosting as a single ensemble using different base learners such as C4.5, Naïve Bayes, OneR and Decision Stump. These were experimented on several benchmark datasets of UCI Machine Learning Repository.

J.R.Quinlan [21] performed experiments with ensemble methods Bagging and Boosting by choosing C4.5 as base learner.

3. EXPERIMENTAL RESULTS

As per the statistics of National Cancer Institute, Breast cancer is a leading cause of death among females in economically developing countries and a second cause in developed countries. Early detection of breast cancer will reduce the death rate. Therefore we considered Breast Cancer data sets for the study of ensembles of decision tree classifiers. The datasets considered in this study are from the publicly available UCI machine learning repository [22]. The description of the data sets is shown in Table 1.

Table 1: Summarization of Breast Cancer Datasets

Dataset	No. of Attributes	No. of Instances	No. of Classes	Missing values
Breast Cancer	10	286	2	Yes
Breast Cancer Wisconsin (Original)	11	699	2	Yes
Breast Cancer Wisconsin (Diagnostic)	32	569	2	No
Breast Cancer Wisconsin (Prognostic)	33	198	2	Yes

For consistency missing values of these data sets are replaced with the mean of the respective attributes. Experiments are conducted using 10-fold cross validation

method. In this study CART decision tree classifier is used as base classifier because it was proved as the best algorithm on medical data in our previous study [23] out of the frequently used decision tree classifiers. To enhance the accuracy of the classifier CART several feature selection methods are used. In our previous study [24] the results exhibited that CART with FS enhances the accuracy than CART alone.

The accuracy of CART with Feature selection and Bagging and CART with Feature Selection and Boosting with 10 models, which were obtained through our previous study [25]-[26] are represented in Table 2. In the previous study we considered first three datasets only. But in the present study we considered one more breast cancer dataset (ie. Breast Cancer Wisconsin(Prognostic)) and the further analysis is carried out.

Table 2: Accuracy(%) of CART with FS and Bagging & CART with FS and Boosting.

Dataset	**CART with FS and Bagging**	**CART with FS and Boosting**
Breast Cancer	74.47	65.03
Breast Cancer Wisconsin (Original)	97.85	95.56
Breast Cancer Wisconsin (Diagnostic)	95.96	95.43
Breast Cancer Wisconsin (Prognostic)	76.76	71.21

It is observed that CART with FS and Bagging has classified the four data sets in higher rates than the CART with FS and Boosting.

In the literature it is stated that "although boosting generally increases accuracy, it leads to deterioration in some data sets" [27]. Hence boosting fails in some cases. There are many reasons for the failure of boosting [27]-[28] such as little data, over training, limited ability to generalize where the data does not include misclassification errors or significant amount of noise and when the classes have no significant overlap.

On increase in committee size (no. of classifiers) boosting usually leads to decrease in prediction error [29]. To verify this fact, in this study the experiment on Boosting is conducted by increasing the committee sizes and the results are tabulated in the Table 3.

Table 3: Accuracy (%) of CART with FS and Boosting with enhanced models

Data Set	Models				
	15	**20**	**25**	**30**	**35**
Breast Cancer	**67.13**	66.78	66.78	66.78	66.78
Breast Cancer Wisconsin (Original)	96.13	96.13	**96.28**	96.13	95.99
Breast Cancer Wisconsin (Diagnostic)	96.66	96.48	**97.01**	96.66	**97.01**
Breast Cancer Wisconsin					
(Prognostic)	74.74	77.27	75.25	77.27	**77.77**

Cart with FS and Bagging with 10-models has better accuracy rates for first two data sets (74.47% and 97.85%) than Cart with FS and Boosting with increased models(67.13% and 96.28%). Cart with FS and Boosting has performed better with 3rd dataset when the number of models increased. And for 4th data set it has better accuracy rates when the number of models increased to 30 and 35 respectively. By observing these Bagging is most preferable to the breast cancer data sets than boosting with enhanced models.

As stated in the study[30], combining greater than 25 classifiers did not show additional significant accuracy gains, here also we observed the same result. To compare the models uniformity should be maintained. Hence, the two hybrid approaches with 10 models (Table 2) are compared and we infer that CART with Bagging is preferable on Breast Cancer datasets than Boosting.

Conclusion: Machine learning is advantageous for medical diagnosis for better classification accuracy. A decision tree classifier CART with feature selection is the best classifier on medical data sets. Ensemble of decision tree classifiers is used to improve classification accuracy of CART classifier. The effectiveness of two ensemble of decision tree classifiers ie., Bagging and Boosting with Boosting with CART the experiments are conducted on four breast cancer data sets as breast cancer is one of major causes of death in women. Through the study it is observed that Bagging is the ensemble algorithm of choice for breast cancer data sets analysis than boosting. By increasing number of models for ensemble also boosting has less prediction rates than bagging with limited number of models.

4. REFERENCES

1. J. Han and M. Kamber, "Data Mining; Concepts and Techniques", Morgan Kaufmann Publishers, 2001.

2. Asha Gowda Karegowda, M.A.Jayaram and A.S. Manjunath, "Feature Subset Selection Problem using Wrapper Approach in Supervised Learning", International Journal of Computer Applications 1(7):13–17, February 2010.
3. Deisy.C, Subbulakshmi.B, Baskar.S and Ramaraj.N, "Efficient Dimen-sionality Reduction Approaches for Feature Selection, Conference on Computational Intelligence and Multimedia Applications", 2007.
4. L. Breiman, "Bagging predictors, Machine Learning", 26, 1996, 123-140
5. J. Friedman, T. Hastie and R. Tibshirani, "Additive Logistic Regression: A Statistical View of Boosting", Stanford University, 1998
6. Alaa M. Elsayad ," Diagnosis of Breast Tumor using Boosted Decision Trees", ICGST-AIML Journal, Volume 10, Issue 1, October 2010
7. My Chau Tu, Dongil Shin and Dongkyoo Shin, "A Comparative Study of Medical Data Classification Methods Based on Decision Tree and Bagging Algorithms" Eighth IEEE International Conference on Dependable, Autonomic and Secure Computing, 2009
8. Y. Freund and R.E. Schapire, "Decision-theoretic Generalization of Online Learning and an Application to Boosting", Journal of Computer and System Sciences, vol. 55, no.1, pp.119-139, 1997
9. My Chau Tu, Dongil Shin and Dongkyoo Shin ,"Effective Diagnosis of Heart Disease through Bagging Approach", 2nd International Conference on Biomedical Engineering and Informatics, 2009.
10. My Chau Tu, Dongil Shin and Dongkyoo Shin, "A Comparative Study of Medical Data Classification Methods Based on Decision Tree and Bagging Algorithms" Eighth IEEE International Conference on Dependable, Autonomic and Secure Computing, 2009.
11. Tsirogiannis G.L, Frossyniotis D, Stoitsis J, Golemati S, Stafylopatis A , Nikita,K.S, "Classification of Medical Data with a Robust Multi-Level Combination scheme", IEEE international joint Conference on Neural Networks, 2004.
12. Pan Wen, "Application of decision tree to identify a abnormal high frequency Electro-cardiograph", China National Knowledge Infrastructure Journal, 2000.
13. Kaewchinporn .C, Vongsuchoto. N and Srisawat. A, " A Combination of Decision Tree Learning and Clustering for Data Classification", 2011 Eighth International Joint Conference on Computer Science and Software Engineering (JCSSE).
14. Jinyan LiHuiqing Liu, See-Kiong Ng and Limsoon Wong," Discovery of Significant Rules for Classifying Cancer Diagnosis Data", Bioinformatics 19(Suppl. 2) Oxford University Press 2003.
15. Dong-Sheng Cao, Qing-Song Xu ,Yi-Zeng Liang and Xian Chen, "Automatic feature subset selection for decision tree-based ensemble methods in the prediction of bioactivity", Chemo metrics and Intelligent Laboratory Systems,2010.
16. Liu Ya-Qin, Wang Cheng and Zhang Lu," Decision Tree Based Predictive Models for Breast Cancer Survivability on Imbalanced Data", 3rd International Conference on Bioinformatics and Biomedical Engineering, 2009.
17. Tan AC, and Gilbert D, "Ensemble Machine Learning on Gene Expression Data for Cancer Classification", Appl Bioinformatics. 2003;2(3 Suppl):S75-83.
18. CaiLing Dong, YiLong Yin and XiuKun Yang, "Detecting Malignant Patients via Modified Boosted tree", Science China Information Sciences, 2010
19. Jaree Thangkam Guandong Xu, Yanchun Zang and Fuchun Huang,"HDKM'08 Proceedings of the second Australian workshop on Health data and Knowledge Management, Vol 80.
20. S.B.Kotsiantis and P.E.Pintelas,"Combining Bagging and Boosting", International Journal of Information and Mathematical Sciences, 1:4 2005.
21. J.R.Quinlan,"Bagging,Boosting and C4.5", In Proceedings Fourteenth National Conference on Artificial ISntelligence",1994.
22. UCIrvine Machine Learning Repository www.ics.uci.edu/~mlearn/ MLRepository.html

23. D.Lavanya, Dr.K.Usha Rani, "Performance Evaluation of Decision Tree Classifiers on Medical Datasets". International Journal of Computer Applications 26(4):1-4, July 2011

24. D.Lavanya, Dr.K.Usha Rani,.," Analysis of feature selection with classification: Breast cancer datasets", Indian Journal of Computer Science and Engineering (IJCSE),October 2011.

25. .Lavanya, Dr.K.Usha Rani," Ensemble decision tree classifier for Breast Cancer data", International journal of Information Technology Convergence and Services (IJITCS) Vol.2, No.1, February 2012.

26. D.Lavanya, Dr.K.Usha Rani," Ensemble Decision Making System for Breast Cancer data" to be published in International Journal of Computer Applications (IJCA) August, 2012 Edition.

27. J.R.Quinlan,"Bagging,Boosting and C4.5", In Proceedings Fourteenth National Conference on Artificial ISntelligence",1994.

28. Clifton D.Sutton, "Classification and Regression Trees, Bagging and Boosting", Handbook of Statistics, vol. 24, 2005

29. Robi Polikar, "Ensemble Based Systems in Decision Making", IEEE Circuits and Systems Magazine, 2006

30. Breiman, Leo, Bagging Predictors, Technical Report, Department of Statistics, University of California at Berkeley, 1994..

****

---

[1]*D. Lavanya, Research Scholar, Dept. of Computer Science,*
Sri Padmavathi Mahila Visvavidyalayam, Tirupati-2, Andhra Pradesh
lav_dlr@yahoo.com
[2]*Dr. K. Usha Rani,, Asso. Professor, Dept.of. Computer Science.*
Sri Padmavathi Mahila Visvavidyalayam, Tirupati-2, Andhra Pradesh
usharanikuruba@yahoo.com