Ensemble Decision Making System for Breast Cancer Data

D. Lavanya
Research Scholar
Sri Padmavathi Mahila Visvavidyalayam
Tirupati-2, Andhra Pradesh

K. Usha Rani
Phd, Dept.of. Computer Science
Sri Padmavathi Mahila Visvavidyalayam
Tirupati-2, Andhra Pradesh

ABSTRACT
Data Mining is a technique to extract the hidden knowledge of information. Among several data mining methods classification is especially useful in the field of medical diagnosis for decision making. In this study, a hybrid approach: CART decision tree classifier with feature selection and boosting ensemble method has been considered to evaluate the performance of classifier. Various Breast cancer data sets are considered for this study as breast cancer is one of the leading causes of death in women.

Keywords
Data Mining, Classification, Decision Trees, Ensemble Systems, Bagging, Boosting, Breast Cancer Datasets.

1. INTRODUCTION
Data Mining [DM] is an interdisciplinary field with a goal of predicting outcomes and to find out relationships in data. Data mining tasks can be descriptive – to discover general interesting patters in the data; and predictive – to predict the behavior of the model on available data. Data mining activities mainly concentrated on the development of models to represent the hidden knowledge contained in the data. These models can be mainly classified as [1]:

- Classification
- Regression
- Clustering
- Rule generation
- Discovering Association rules
- Summarization
- Dependency modeling
- Sequence analysis

Some of the key steps in Data Mining are: Problem definition, Data exploration, Data preparation, Modeling, Evaluation and Deployment [2]. Data mining is interesting due to – falling cost of large storage devices and increasing ease of collecting data over networks; development of robust and efficient machine learning algorithms to process the data; and falling cost of computational power, enabling use of computationally intensive methods for data analysis [3]. The application of DM facilitates systematic analysis of medical data that often contains huge volumes and in unstructured format.

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. Actually, classification process is carried out in two steps. In the first step, through the analysis of the training records of a database a model is constructed. This is known as supervised learning because the class label of each training record is known. In the second step, the constructed model is used for classification.

For the classification activity, the powerful and popular tool is Decision Tree [4]. 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 leaf node represents a label or class. 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.

Medical diagnosis is regarded as an important though complicated task that needs to be executed accurately and efficiently. The automation of this system would be extremely advantageous. Data mining have the potential to generate a knowledge-rich environment, which can help to significantly improve the quality of clinical decisions. 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 organization of paper is as follows. Section 2 deals with brief overview of related work, decision tree construction, feature selection and ensemble systems. In section 3 experiments and evaluation of results are provided. Section 4 presents the conclusion.

2. BACKGROUND

2.1 Overview of Related Work
Several studies have been reported that they have focused on the importance of bagging and boosting ensemble methods in the field of medical diagnosis. These studies have applied different approaches to classify the data with high classification accuracies.

My Chau Tu et.al. [5] 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. [6] 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. [7] 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 [8] conducted experiments on ECG data to identify abnormal high frequency electro cardiograph using decision tree algorithm C4.5 with Bagging.

Kaewchinporn et.al. [9] 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. [10] 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. [11] 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. [12] experimented on breast cancer data using C5 algorithm with bagging to predict breast cancer survivability.

Tan AC et.al. [13] 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. [14] proposed a modified Boosted Decision Tree for breast cancer detection to improve the accuracy of classification.

Jaree Thangkam et.al. [15] 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. [16] 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 [17] performed experiments with ensemble methods Bagging and Boosting by choosing C4.5 as base learner.

2.2 Decision Trees

Decision tree induction is a very popular and practical approach for pattern classification. It is the learning of decision trees from class-labelled training tuples. A decision tree is a flow chart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test and each leaf node holds a class label.

The decision tree classifier has two phases [18]:

i) Growth phase or Build phase.
ii) Pruning phase.

The tree is built in the first phase by recursively splitting the training set based on local optimal criteria until all or most of the records belonging to each of the partitions bearing the same class label. The tree may overfit the data. The pruning phase handles the problem of over fitting the data in the decision tree. The prune phase generalizes the tree by removing the noise and outliers. The accuracy of the classification increases due to the pruning phase.

Pruning phase accesses the fully-grown tree only. The growth phase requires multiple passes over the training data. The time needed for pruning the decision tree is very less compared to build the decision tree. Many decision tree algorithms have been developed.

2.3 Feature Selection

Feature Selection (FS) a preprocessing technique is used to identify the significant attributes, which play a dominant role in the task of classification. This leads to the dimensionality reduction. By applying several search techniques features can be reduced. The reduced feature set improves the accuracy of the classification task in comparison of applying the classification task on the original data set.

2.4 Ensemble Systems in Decision-making

Some important areas such as financial, medical, social, etc., generally people will seek a second (third, sometimes many more) opinion before making a decision because the risk factors are highly influenced. 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.

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.

Some of the reasons for using Ensemble Based Systems [19]:

- Statistical Reasons: Combining the outputs of several classifiers by averaging may reduce the risk of selecting a poorly performing classifier.
- Large Volumes of data: The amount of data is too large to be analyzed effectively by a single classifier.
- Too little data: Resampling techniques can be used to overlap random subsets of inadequate training data and each subset can be used to train a different classifier.
- Divide and Conquer: A particular classifier is unable to solve certain problems. The decision boundary for different classes may be too complex. In such cases, the complex decision boundary can be estimated by combing different classifiers appropriately.
- Data Fusion: A single classifier is not adequate to learn information contained in data sets with heterogeneous features (i.e. data obtained form various sources and the nature of features is different). Applications in which data from different sources are combined to make a more informed decisions are referred as Data Fusion applications and ensemble based approaches are most suitable for such applications.

2.4.1 Bagging

Bagging means Bootstrap aggregation [20] an ensemble method to classify the data with good accuracy. In this method first the decision tree is derived by the base classifiers C_1, C_2, - - - -, C_n on the bootstrap samples D_1, D_2, - - - - D_n, respectively with replacement from the data set D. Later the

final model or decision tree is derived as a combination of all base classifiers C_1, C_2, ---, C_n.

Votes have to be collected from each classifier to a tuple in such a way to classify which class the tuple belongs to. The class label is decided for the tuple based on the maximum number of votes voted for a particular class label. Suppose a class receives the maximum number of votes, equally for more than one class label then any one of the class labels is selected randomly.

Bagging can be applied on any classifier such as Decision trees, Bayesian algorithms, Rule based algorithms, neural networks, Support vector machines, Associative classification, Distance based methods and Genetic Algorithms. Applying Bagging on classifiers especially on Decision Trees and Neural Networks increases accuracy of classification because of their capability to control instability. Bagging plays an important role in the field of medical diagnosis.

2.4.2 Boosting

Boosting also creates an ensemble of classifiers by replacing the data as in Bagging but in this the classifiers are combined by majority voting. Here series of K classifiers is learned iteratively. The k classifiers are represented as C_1, C_2, ---, C_K on datasets D_1, D_2, ---, D_K. For boosting, each training tuple is assigned with a weight. In general, classifier Ci after learning will update the weights so that the subsequent classifier Ci+1 will concentrate on misclassified tuples by Ci. The final classifier C* combines the votes (function of accuracy) of all the classifiers. A popular AdaBoost, a boosting algorithm [21] is considered in this study. .

3. EXPERIMENTAL RESULTS

In this study CART decision tree classifier is considered to classify the breast cancer data sets. As the breast cancer is one of the leading causes of death in women here, we considered three breast cancer data sets. The datasets considered in this study were from UCI machine learning repository [22], which is publicly available. The description of the datasets is shown in Table 1.

Table 1: Summarization of Breast Cancer Datasets

Dataset	No. of Attribu-tes	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

Missing values of these data sets are replaced with the mean of the respective attributes.

Reason to choose the CART classifier is it was proved as best classifier for medical diagnosis in our previous study [23]. To enhance the accuracy of the classifier CART several feature selection methods are used. The results exhibit that CART with FS enhances the accuracy than CART alone. Through our previous study it has been observed that same feature selection may not contribute to the enhancement of detection accuracy for all breast cancer data sets. A particular feature selection method is a best feature selection method for a particular breast cancer data set, which was proved in our previous study [24]. Based on those observations the best feature selection method for each data set is represented in Table 2:

Table 2: Data Sets-Best Feature Selection Methods

Data Set	Feature Selection Method
Breast Cancer	SVMAttributeEval
Breast Cancer Wisconsin (Original)	PrincipalComponentsAttributeEval
Breast Cancer Wisconsin (Diagnostic)	SymmetricUncertAttributesetEval

A hybrid approach that is CART with FS and an ensemble method boosting is experimented on breast cancer datasets in this study. The procedure of hybrid approach is illustrated below: Apply the best feature selection method on breast cancer data sets. Once the reduced data sets are obtained then conduct experiment of boosting with CART algorithm on three Breast Cancer data sets. For boosted ensembling, 10 models are used. Ten fold cross validation is adopted to test the data.

The accuracy of the hybrid approach which is a combination of feature selection and CART with Boosting is tested on the selected breast cancer datasets with best feature selection method. The results are shown in the Table 3.

Table 3: CART with Feature selection and Boosting – Accuracy and Time to build a model

Data Set	Accuracy (%)	Time (sec)
Breast Cancer	65.03	28.36
Breast Cancer Wisconsin (Original)	95.56	1.19
Breast Cancer Wisconsin (Diagnostic)	95.43	1.14

The comparison of the accuracies of three methods- CART algorithm, CART with FS [24] and CART with Feature selection and Boosting is represented in Table 4 and graphical representation in Figure 1. With this comparison, it is clear that by applying hybrid approach the classification accuracy is enhanced for only one (3^{rd}) data set and for the remaining data sets the base classifier i.e., CART with FS has better accuracy rates.

Table 4: Accuracy (%) of CART algorithm, CART with Feature Selection Method and CART with Feature selection and Boosting

Dataset	CART *	CART With FS *	CART With FS and Boosting
Breast Cancer	69.23	**73.03**	65.03
Breast Cancer Wisconsin (Original)	94.84	**96.99**	95.56
Breast Cancer Wisconsin (Diagnostic)	92.97	94.72	**95.43**

*Values are from our previous work [24].

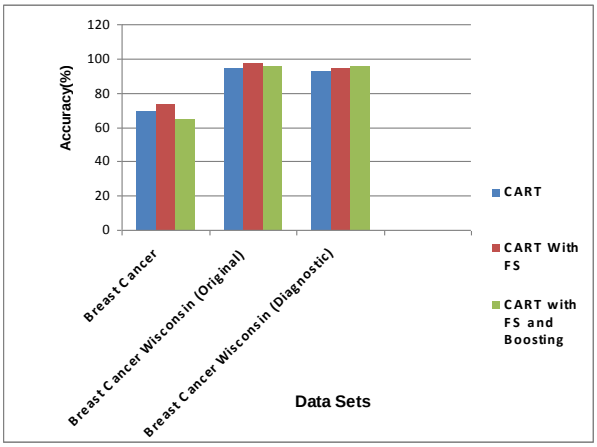

Figure 1. Accuracy (%) of CART algorithm, CART with Feature Selection Method and Hybrid Approach.

In our previous study [25] the three data sets were experimented with an ensemble method Bagging with combination of CART and feature selection as a hybrid approach. The comparison of those accuracies with the present experimental results is tabulated in Table 5.

Table 5: Accuracy (%) of CART with FS and Bagging and 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

* Values are from our previous work [25].

Further the comparison of classification accuracies of both the hybrid approaches are presented as graphical representation in Figure 2.

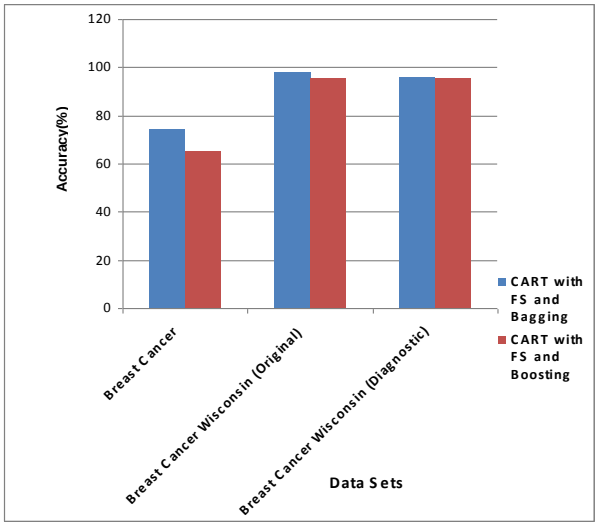

Figure 2. Accuracy (%) of CART with FS and Bagging and CART with FS and Boosting.

It is observed that CART with FS and Bagging has classified the three 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" [17]. Hence boosting fails in some cases. There are many reasons for the failure of boosting [17,26] 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.

By considering the above facts we conclude that bagging is most preferable to the breast cancer data classification along with CART and FS than boosting

4. CONCLUSION

Classification is a popular data mining technique to classify the medical data. Data is preprocessed to remove missing values and feature selection methods are applied to reduce the dataset. The best proved decision tree classifier CART on medical data sets is experimented with feature selection and ensembling techniques. Through this study, it is clear that in the ensemble method Bagging is preferable for diagnosis of breast cancer data than Boosting.

5. REFERENCES

[1] Sushmita Mitra and Pabitra Mitra,"Data Mining in Soft Computing Framework: A Survey", IEEE Transactions on Neural Networks, Vol 13, No 1, January 2002.

[2] The Data Mining Process (online) available: http://publib.boulder.ibm.com/infocenter/db2/uw/v9/index.jsp?topic=/com.ibm.im.easy.doc/c_dm_process.html

[3] T. M. Mitchell, "Machine learning and Data Mining", Commun. ACM, vol. 42, no, 11, 1999.

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

[5] 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.

[6] 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.

[7] 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.

[8] Pan Wen, "Application of decision tree to identify a abnormal high frequency Electro-cardiograph", China National Knowledge Infrastructure Journal, 2000.

[9] 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).

[10] 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.

[11] 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.

[12] 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.

[13] Tan AC, and Gilbert D, "Ensemble Machine Learning on Gene Expression Data for Cancer Classification", Appl Bioinformatics. 2003;2(3 Suppl):S75-83.

[14] CaiLing Dong, YiLong Yin and XiuKun Yang, "Detecting Malignant Patients via Modified Boosted tree", Science China Information Sciences, 2010.

[15] 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.

[16] S.B.Kotsiantis and P.E.Pintelas,"Combining Bagging and Boosting", International Journal of Information and Mathematical Sciences, 1:4 2005.

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

[18] J. Han and M. Kamber, "Data Mining; Concepts and Techniques", Morgan Kaufmann Publishers, 2000.

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

[20] L. Breiman, "Bagging predictors, Machine Learning", 26, 1996, 123-140.

[21] 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.

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