ENSEMBLE DECISION TREE CLASSIFIER FOR BREAST CANCER DATA

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

[1] Research Scholar, Department of Computer Science, Sree Padmavathi Mahila Visvavidyalayam, Tirupati, Andhra Pradesh, India
lav_dlr@yahoo.com

[2] Associate Professor, Department of Computer Science, Sree Padmavathi Mahila Visvavidyalayam, Tirupati, Andhra Pradesh, India
usharanikuruba@yahoo.co.in

ABSTRACT

Data mining is the process of analyzing large quantities of data and summarizing it into useful information. In medical diagnoses the role of data mining approaches increasing rapidly. Particularly Classification algorithms are very helpful in classifying the data, which is important in decision making process for medical practitioners. Further to enhance the classifier accuracy various pre-processing techniques and ensemble techniques were developed. In this study a hybrid approach, CART classifier with feature selection and bagging technique has been considered to evaluate the performance in terms of accuracy and time for classification of various breast cancer datasets.

KEYWORDS

Data mining, Classification, Decision Trees, Ensemble, Bagging, Breast Cancer datasets.

1. INTRODUCTION

In the data mining community, decision tree algorithms are very popular since they are relatively fast to train and In data mining, decision tree algorithms are very popular due to their characteristics such as fast to train and produce transparent models. The machine learning community has produced a large number of programs to create decision trees for classification. Very notable among these for classification of data include ID3, C4.5, C5, CART, CHAID**,** SLIQ, SPRINT, ScalParc and so on. These classifiers provide support for many health care areas in decision making. Out of these CART has been proved to the best classifier for medical data [4]. Classification and regression trees (CART) were initially developed in the 1960's by Morgan and Sonquist92 and have since been refined and popularized by Breiman [3]. Based on recursive partitioning they provide non-parametric flexibility in the prediction of categorical i.e., classification or continuous i.e., regression outcomes.

Accuracy is very important in classification of medical data; a hybrid approach has been presented to enhance the classification accuracy of breast cancer. As the Breast Cancer is one of leading causes of death in women in this study various breast cancer datasets are considered to analyze the performance of CART decision tree algorithm with feature selection and bagging technique for higher classification accuracy and improved diagnosis.

The organization of the paper is paper is: A brief overview of related work , the theory of decision tree, feature selection and bagging algorithm will be given in section 2; The experiments and evaluation are presented in section 3;And Section 4 contains the conclusion of the study.

2. BACKGROUND

2.1 Overview of Related Work

Several studies have been reported that they have focused on the importance of bagging technique in the field of medical diagnosis. These studies have applied different approaches to the given problem and achieved high classification accuracies. Here are some examples:

- My Chau Tu's [14] proposed the use of bagging with C4.5 algorithm, bagging with Naïve bayes algorithm to diagnose the heart disease of a patient.
- My Chau Tu's [15] 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's [19] applied bagging algorithm on medical databases using the classifiers neural networks, SVM'S and decision trees. Results exhibits improved accuracy of bagging than without bagging.
- Pan wen [16] conducted experiments on ECG data to identify abnormal high frequency electrocardiograph using decision tree algorithm C4.5 with bagging.
- Kaewchinporn C's [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 other datasets not related to medical domain.
- Jinyan LiHuiqing Liu's [8] experimented on ovarian tumor data to diagnose cancer using C4.5 with and without bagging.
- Dong-Sheng Cao's [6] proposed a new decision tree based ensemble method combined with feature selection method backward elimination strategy with bagging to find the structure activity relationships in the area of chemometrics related to pharmaceutical industry.
- Liu Ya-Qin's [12] experimented on breast cancer data using C5 algorithm with bagging to predict breast cancer survivability.
- Tan AC's [18] used C4.5 decision tree, bagged decision tree on seven publicly available cancerous micro array data, and compared the prediction performance of these methods.

2.2 Decision Trees

Decision tree [7] is one of the classification methods, which classify the labeled trained data into a tree or rules. Once the tree or rules are derived in learning phase to test the accuracy of a classifier test data is taken randomly from training data. After Verification of accuracy, unlabeled data is classified using the tree or rules obtained in learning phase. The structure of a decision tree is similar to the tree with a root node, a left sub tree and right sub tree. The leaf nodes in a tree represent a class label. The arcs from one node to another node denote the conditions on the attributes. The Tree can be built as:

- The selection of attribute as a root node is done based on attribute splits
- The decisions about the node to represent as terminal node or to continue for splitting the node.
- The assignment of terminal node to a class.

The attribute splits depends on the impurity measures such as Information gain, gain ratio, gini index e.t.c.

Once the tree is built then it is pruned to check for over fitting and noise. Finally the tree is an optimized tree .The advantage of tree structured approach is easy to understand and interpret, handles categorical and numeric attributes, robust to outliers and missing values. Decision tree classifiers are used extensively for diagnosis of diseases such as breast cancer, ovarian cancer and heart sound diagnosis and so on [1], [17], [10],[2].

2.3 Feature Selection

A preprocessing technique feature selection identifies and removes irrelevant attributes that do not play any role in the classification task. Several feature selection methods are available with different search techniques to produce a reduced data set. This reduced data set improves accuracy compared with original dataset. Feature selection does not alter the relevance or meaning the data set. The feature selection methods are categorized as filter, wrapper and hybrid. The result of these methods varies in time and accuracy.

A brief summary of Feature Selection process [13] is as follows: Generate candidate feature subset from the original data set using methods such as complete, heuristic and random and then evaluate the generated candidate feature subset using different evaluation functions such as distance, information, dependency, consistency, classifier error rate. These functions produce relevancy value. This relevancy value used as termination condition to decide the subset as an optimal feature subset or not. If the subset is an optimal feature subset then a validation is performed on that otherwise the above process continues. In the field of medical diagnosis accuracy is very important hence one can incorporate preprocessing techniques in classification tasks.

2.4 Bagging

Bagging means Bootstrap aggregation [11] an ensemble method to classify the data with good accuracy. In this method first the decision trees are derived by building the base classifiers c_1, c_2, ----, c_n on the bootstrap samples D_1, D2, ----, Dn 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 with the majority votes.

Bagging can be applied on any classifier such as neural networks, 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, Neural nets increases accuracy of classification. Bagging plays an important role in the field of medical diagnosis .Many research works in this aspect is depicted in related work.

The process of bagging is shown in the figure 1.

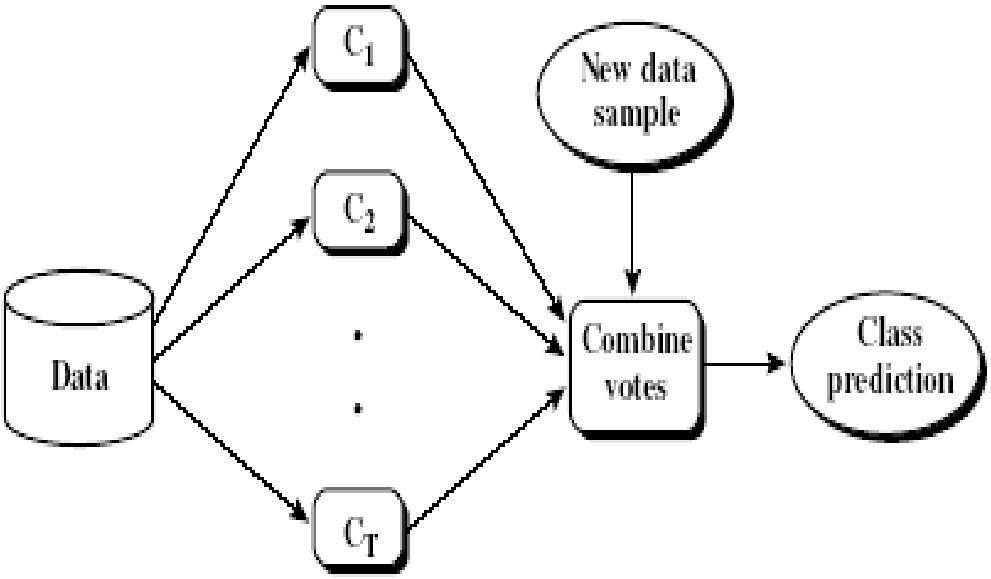

Figure 1. Bagging Process

3. EXPERIMENTS

Experiments are conducted on three breast cancer datasets. In this study CART decision tree algorithm is chosen for analysis of data because it was proved as the best algorithm in our previous study [4] among the frequently used decision tree classifiers ID3 and C4.5 to classify the medical data.

Generally, datasets may contain some attributes that do not have any significance in classification task. Such attributes can be eliminated by feature selection methods. For this purpose, in our previous study, experiments are conducted on breast cancer data sets using several feature selection methods in conjunction with cart algorithm [5]. The results proved that a particular feature selection method is best for a specific breast cancer dataset. Because the data sets chosen differ in the number of instances or records and number of attributes.

The conjunction of CART algorithm with feature selection methods increased the classification accuracy. Further, to enhance accuracy a hybrid approach is considered in this study. The hybrid approach is formed by combining the best feature selection method to a particular breast cancer data set, bagging and cart decision tree algorithm.

The data sets are collected from UCI machine learning repository [www.ics.uci.edu] which is publicly available. The description of the datasets is given in Table 1.

Table 1: Description 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

To obtain consistency the missing values in the datasets are replaced with mean value of the respective attribute. Experiments are conducted using Weka tool and the results are compared with bagging and without bagging using 10-fold cross validation.

The procedure of hybrid approach is illustrated below. Apply the best feature selection method on breast cancer data set. Once the reduced data set is obtained then conduct experiment of bagging with cart algorithm. Thus, the above procedure is performed on three breast cancer data sets iteratively to obtain the classification results. The best feature selection method to each breast cancer data [5] is given in the following Table 2.

Table 2: Best Feature Selection Method for Breast Cancer Datasets

Data Set	**Feature selection Technique**
Breast Cancer	SVMAttributeEval
Breast Cancer Wisconsin (Original)	PrincipalComponentsAttributeEval
Breast Cancer Wisconsin (Diagnostic)	SymmetricUncertAttributesetEval

As accuracy is very important in the field of medical domain, the performance measure accuracy of classification is considered in this study. The results of the hybrid approach are tabulated in Table 3.

Table 3: Hybrid Approach –Accuracy and time to build a model.

Data Set	**Accuracy**	**Time**
Breast Cancer	74.47	28.25
Breast Cancer Wisconsin (Original)	97.85	1.38
Breast Cancer Wisconsin (Diagnostic)	95.96	1.95

The comparison of the accuracies of three methods is represented in Table 4. With this comparison, it is clear that the hybrid approach enhanced the classification accuracy than other approaches for all three Breast cancer data sets.

Table 4: Accuracy of CART algorithm, CART with Feature Selection Method and Hybrid Approach

Dataset	**CART**	**CART With Feature Selection Method**	**Hybrid Approach**
Breast Cancer	69.23	73.03	**74.47**
Breast Cancer Wisconsin (Original)	94.84	96.99	**97.85**
Breast Cancer Wisconsin (Diagnostic)	92.97	94.72	**95.96**

The comparison of the accuracies of the above three methods is presented in the graphical form in figure 2.

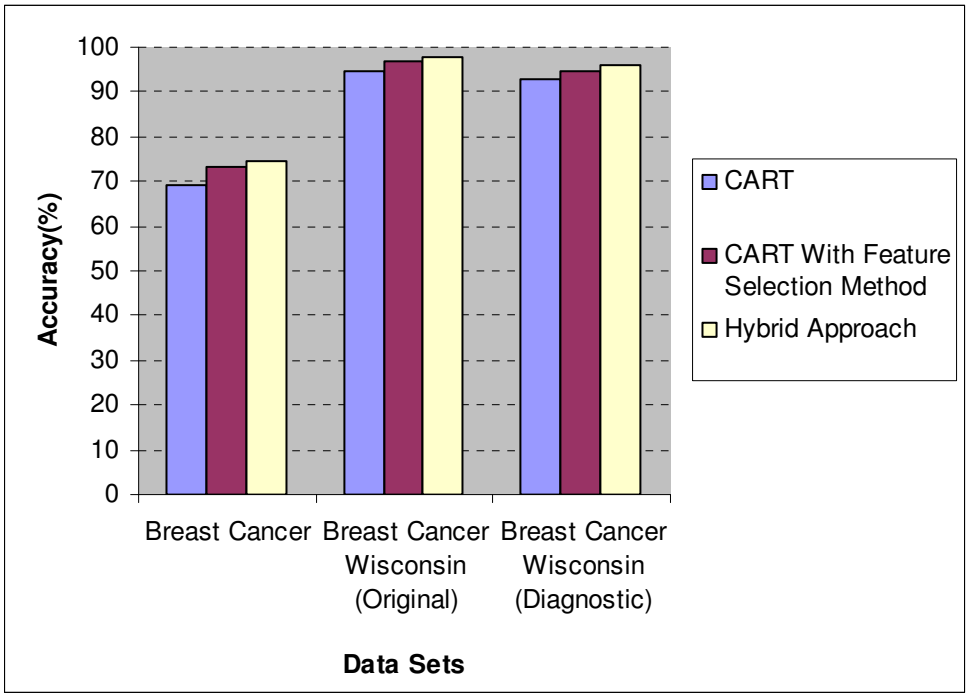

Figure 2. Accuracy of CART algorithm, CART with Feature Selection Method and Hybrid Approach

4. CONCLUSION

For medical diagnosis various data mining techniques are available. In this study, for classification of medical data we employed decision tree algorithm because it produce human readable classification rules which are easy to interpret. A hybrid method is proposed to enhance the classification accuracy of Breast Cancer data sets. The training data is tested with 10-fold cross validation. The data sets are preprocessed to remove missing values. The feature selection methods used to eliminate those attributes that have no significance in the classification process. Bagging the training dataset is one of the most common methods of improving decision tree.

The experimental results of a hybrid approach with the combination of preprocessing, bagging with cart demonstrated the enhanced classification accuracy of the selected data sets.

REFERENCES

1. Antonia Vlahou, John O. Schorge, Betsy W.Gregory and Robert L. Coleman, "*Diagnosis of Ovarian Cancer Using Decision Tree Classification of Mass Spectral Data",* Journal of Biomedicine and Biotechnology • 2003:5 (2003) 308–314.
2. Aruna, Dr S.P. Rajagopalan and L.V.Nandakishore," A*n Empirical Comparison of Supervisedlearning algorithms in Disease Detection*". International Journal of Information Technology Convergence and Services (IJITCS) Vol.1, No.4, August 2011.
3. Breiman, Friedman, Olshen, and Stone. "*Classification and Regression Trees*", Wadsworth, 1984, Mezzovico, Switzerland.
4. 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.
5. 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.
6. Dong-Sheng Cao, Qing-Song Xu ,Yi-Zeng Liang, Xian Chen, *"Automatic feature subset selection for decision tree-based ensemble methods in the prediction of bioactivity*", Chemometrics and Intelligent Laboratory Systems.
7. J. Han and M. Kamber, "*Data Mining; Concepts and Techniques"*, Morgan Kaufmann Publishers, 2000.
8. 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.
9. Kaewchinporn .C, Vongsuchoto. N, 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. Kuowj, Chang RF,Chen DR and Lee CC," *Data Mining with decision trees for diagnosis of breast tumor in medical ultrasonic images*" ,March 2001.
11. L. Breiman, *"Bagging predictors"*, Machine Learning, 26, 1996, 123-140.
12. Liu Ya-Qin, Wang Cheng, Zhang Lu*," Decision Tree Based Predictive Models for Breast Cancer Survivability on Imbalanced Data*" , 3rd International Conference on Bioinformatics and Biomedical Engineering , 2009.
13. Mark A. Hall, Lloyd A. Smith," *Feature Subset Selection: A Correlation Based Filter Approach",* In 1997 International Conference on Neural Information Processing and Intelligent Information Systems (1997), pp. 855-858.
14. My Chau Tu, Dongil Shin, Dongkyoo Shin *,"Effective Diagnosis of Heart Disease through Bagging Approach",* 2nd International Conference on Biomedical Engineering and Informatics,2009.
15. My Chau Tu, Dongil Shin, 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.
16. Pan Wen, *"Application of decision tree to identify a abnormal high frequency electro-cardiograph"*, China National Knowledge Infrastructure Journal, 2000.

17. Stasis, A.C. Loukis, E.N. Pavlopoulos, S.A. Koutsouris, D. "*Using decision tree algorithms as a basis for a heart sound diagnosis decision support system",* Information Technology Applications in Biomedicine, 2003. 4th International IEEE EMBS Special Topic Conference, April 2003.

18. Tan AC, Gilbert D. *"Ensemble machine learning on gene expression data for cancer classification",* Appl Bioinformatics. 2003;2(3 Suppl):S75-83.

19. 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.