Evaluation of Decision Tree Classifiers and Boosting Algorithm for Classifying High Dimensional Cancer Datasets

Abid hasan

Abstract—**Cancer datasets contains large number of gene expression values with a limited number of samples. Classifying these datasets using different classification algorithm is one of the most challenging tasks for the researchers because of their high dimensionality and enormous size. Extracting predictive features for accurately classifying these datasets requires choosing appropriate classification algorithm. Along with the feature selection capability embedded in the classifiers, some additional feature selection method can be useful for better classification accuracy for cancer datasets. Decision tree classifiers are good candidates for this purpose while in this paper we have used the boosting algorithms AdaBoost as a boosting algorithm for classifying along with the decision tree classifiers for evaluating their performance for different cancer datasets with different size and number of features (genes). In this paper, one of the previously proposed methods of feature selection has been used along with some conventional feature selection methods to obtain predictive features for classification and the performances on the accuracy of classifying. The time to build the model for decision three induction classifiers and for the Boosting algorithm is also analyzed.**

Index Terms—**Boosting algorithm, cancer datasets, classification, data mining, decision tree induction.**

I. Introduction

Classification is playing an important role in the field of data mining as well as in the studies of machine learning, neural network, statistics and many expert systems over many years[1], [2]. Different classification algorithm has been successfully implemented in various applications. Among them some of the popular implications of classification algorithms are scientific experiments, credit approval, weather prediction, fraud detection, medical diagnosis, image processing, target marketing and lots more [3], [4]. In the recent years medical data classification especially cancer data classification caught a huge interest amidst the researchers. The significance of this work can be verified by the recent researches on cancer classification by many researchers. [5] - [7].

Medical dataset such as cancer dataset contains large number of gene expression values. These can be obtained from different cancer patients whose gene expression level are measured by several modern technology such microarray technology [5]. Using microarray a large number of gene expression value can be obtained in a short time with great accuracy. However this large amount of data creates significant challenges for researchers with regard to computational complexity and data manipulation. Microarray data also faces challenges of feature selection, noise, background and special effects. [5] All the datasets that has been used here are obtained from different microarray chips. Selecting a subset of discriminatory genes as feature for classification is critical for ensuring both accuracy and speed of classification. Of the tens of thousands of genes involved in an experiment, only a small number of them show predictive quality for classifying the datasets. The high dimensionality (known as *curse of dimensionality*) of these datasets poses a great challenge for the classifiers to find suitable features for accurate classification of the datasets.

To perform any classification, the input is a dataset that taken as an input which contains several training records where each records has several attributes [8]. These attributes can be categorized into two domains: Categorical and numerical based on the discreetness and continuous property while a class level distinguishes the attributes. Different classification methods used to model the classifier that predicts the class levels of unknown objects. In this paper five cancer datasets has been chosen: Acute Lymphoblastic Leukemia (ALL) dataset [14], Breast cancer dataset [15], High Grade Glioma dataset [16], Prostate cancer dataset [17] and Lung cancer dataset [18]. Research for performance analysis of decision tree classifiers on cancer datasets has been done [8], while in this paper the boosting algorithms has been compared with the decision tree classifiers as classification methods in the process of classifying cancer datasets. Among the decision tree classifiers C4.5 [10], CART [11] and Random Tree [13] has been chosen while for the Boosting algorithm, AdaBoost [13] has been used.

Second and third section of this paper presents the review of the decision tree induction classifiers and Boosting algorithm (AdaBoost). Fourth section presents the short description of the cancer datasets used for the experiment in this paper. Feature Selection and Method of experiment has been described in the next two sections while the performance comparison on cancer datasets for chosen classifiers and time to build the model for these classifiers is presented in section seven and concludes in the last section.

Manuscript received March 5, 2012; revised April 5, 2012.

Abid Hasan is with the Department of Computer Science and Information Technology, Islamic University of Technology, Gazipur 1704, Bangladesh (e-mail: aabid@iut-dhaka.edu).

II. Decision Tree Induction

Applying decision tree for classification has been one of the most popular choices by the researchers in recent days

where the decision is obtained after learned from a class-labeled training tuples. In decision trees, test on an attribute is denoted by each node, outcome of a test is denoted by each branch while the leaf node represents the class label. There are two phases of a decision tree, [1] based on local optimal criteria the Growth phase or Build phase built the tree by recursively splitting the data set until all or most of the records belonging to each of the partitions bears the same class label. The second phase of decision tree classifier is known as pruning phase in case the accuracy of the classification by removing noise and outliers. After observing several popular and most frequently used decision tree algorithms C4.5, CART and Random Tree are chosen for performance analysis [9].

A. C4.5

C4.5 algorithm was developed by Quinlan Ross [10]. It is an extension to Iterative Dichotomiser 3 (ID3) developed by Quinlan which was introduced in 1986. Required attributes for building decision tree this algorithm uses Gain Ratio as an attribute selection method. The attribute with highest normalized information gain is selected to make decision. It is based on Hunt's algorithm which can handle missing values. Handling categorical and continuous attributes in the building process of decision tree, C4.5 splits the attributes into two parts based on the threshold as two parts of data falls at the two sides of threshold values. It uses pessimistic pruning. One of the base cases of this algorithm is the samples belong to the same class. If it happens, it creates a leaf node for the decision tree saying to choose that class. On the other case if none of the features provides any information gain then the algorithm creates a decision node higher up the tree using the expected value of the class.

B. CART

Classification and Regression Tree (CART) was introduced by Breiman [11] and has become a common basic method for building statistical models from simple feature data. It is also based on Hunt's algorithm. CART is powerful as it can deal with incomplete data, multiple types of features both in input features and predicted features, and the tree it produces often contain rules which are humanly readable. Decision three contains a binary question (yes/no answer) about some features at each node in the tree. The leaves of the tree contain the best prediction based on the training data. Decision list are a reduced form of this where one answer to each question leads directly to a leaf node. A tree's leaf node may be a single member of some class, a probability function, a predicted mean value for a continuous feature or a gaussian (mean and standard deviation for a continuous value). The basic algorithm is given a set of samples; finding the question about some features which splits the data minimizing the mean impurity (designed to capture how similar the samples are to each other) of the two partition. Applying this splitting on each partition until some stop criteria is reached.

C. Random Tree

Random tree was introduced by Leo Breiman and Adele Cutler [12]. It is one of the most accurate learning algorithms. This algorithm runs efficiently on large databases and can handle thousands of inputs variables without variable deletion. An effective method is included which can estimate the missing data and maintains accuracy when a large portion of data are missing as well as a method for estimating what variables are important for classification. It computes proximities between pairs of cases that can be used in clustering and locating outliers. Both classification and regression problems can be handles by Random tree. It is a collection of tree predictors. The classifier takes feature vector as input and classifies the vector with each tree in the forest. The class label that received majority of 'votes' is classified as the designated class. All the trees trained with same parameters on different training datasets generated from original training set using bootstrap procedure. Error estimation procedure is not necessary in random tree as the errors are estimated during the training internally.

III. Boosting Algorithm

Boosting is a machine learning meta-algorithm used for supervised learning. Boosting algorithm is not algorithmically restricted. This algorithm works by iterative learning weak algorithm with respect to a distribution and adding them to a final strong algorithm. In this process, the weights are added to the weak learners' accuracy and reweighted after a weak learner is added. If misclassified, weight is increased and loses weight if classified correctly. Thus weak learners focus more on the examples that previous weak learners misclassified. Among many of the boosting algorithms we have chosen two of them: AdaBoost and Multi Boosting Method.

A. AdaBoost

AdaBoost (Adaptive Boosting) is similar to SVM which works by combining several "votes". Instead of using support vector (i.e. important examples), AdaBoost uses weak learners. It is a meta-algorithm that can be used in conjunction with other learning algorithms to improve their performance. However, occasionally this algorithm can be less capable to the overfitting problem than most learning algorithms. The algorithm uses weak classifiers (more than 50% correct result, better than random). Classifiers with higher error rate (more than 50%) from a random classifier will be useful as well since they will have negative coefficient in final linear combination thus behave like inverse. AdaBoost was formulated by Yoav Freund and Robert Schapire in 1995 [13]. It is adaptive in the sense that it calls a weak classifier repeatedly in a series of rounds from a total number of classifiers and subsequent classifiers built are fine tuned for those instances that were misclassified by previous classifiers. AdaBoost is sensitive to noisy data and outliers.

IV. Datasets

The rapid advance of microarray technology provides scientists the opportunity of observing various gene in a genome simultaneously measuring the expression levels of the tens of thousands of genes in massive experiments. Analysis of large-scale genomics data in order to extract biologically meaningful information presents unprecedented

opportunities and challenges for data mining in area such as classification. The expression level of the same sets of genes under study are normally measured from different samples or under different conditions and eventually recorded in a data matrix. A typical microarray has the following characteristics: (1) high dimensionality due to tens of thousands of genes; (2) severely limited amount of samples – usually in tens or at most couple of hundreds due to the expense of obtaining microarray samples; and (3) abundance of redundancy among genes. This paper presents a performance evaluation of different classification algorithm on high dimensional medical datasets (cancer datasets). Here for our experiment we have chosen some of the popular cancer datasets.

A. Breast Cancer Dataset

One of the dataset was the lymph-node-negative primary breast cancer data, which was used in [14]. This dataset contains 286 lymph-node-negative patient tumor samples. Among them 180 were lymph-node-negative relapse free patients and 106 were lymph-node-negative patients that have developed a distant metastasis. The experiment was done using a Affymetrix Human U133a GeneChip which recorded the expression of around 22,000 transcripts.

B. ALL (Acute Lymphoblastic Leukemia)

ALL dataset was used in [15]. This datasets was prepared from 128 different individuals who had Acute Lymphoblastic Leukemia (ALL). Among these patients 95 of them had B-cell ALL while 33 of them had T-cell ALL. This experiment was performed on a HGU95AV2 Affymetrix single channel microarray chip which had 12,625 genes. Some additional attributes were also available like patients id, date of diagnosis, sex of patients, age of the patients, type and stage of the disease. B indicates B-cell ALL while T indicates T-cell ALL respectively.

C. High Grade Glioma Dataset

The third dataset was the High Grade Glioma (Brain Tumor) dataset used in [16]. Microarray analysis was used to determine the expression of approximately 12,000 genes from 50 glioma patients. Among these patients 28 of them had glioblastomas and 22 are anaplastic oligodendrogliomas. The experiment was performed on Affymetrix Human Genome U133B Array.

D. Prostate Cancer

Prostate cancer dataset was used by [17]. Among 102 patients, 52 of them had prostate tumor samples and 50 non-tumor prostate samples with around 12,600 genes. Affymetrix Human Genome U133 Plus 2.0 Array had been used for experimenting this dataset.

E. Lung Cancer Dataset

This dataset was used [18]. The dataset contains two subtypes of lung cancer: malignant pleural mesothelioma (MPM) and adenocarcinoma (ADCA) of lung. There were 181 tissue samples where 31 of them had MPM and 150 of them had ADCA. Each sample was described by 12,533 genes.

V. Feature Selection

Due to the enormous number of genes (features) in the medical datasets, feature selection before classification algorithm can be used on the dataset for classifying the datasets is an essential part of the process. Along with the feature selection properties embedded in the classification algorithm, it is required to apply some additional feature selection methodology for better classification accuracy.

As mentioned earlier from the huge feature set only limited number of features has that predictive ability to help classify the cancer datasets with greater accuracy. The feature selection method that has been applied in this paper is originally used in [19]. The genes are ranked based on their *trustworthiness value* calculated using the covariance of two subclass of the dataset. According to [19] the trust value will be high if the average of the covariance between the two subtypes of the dataset is low and difference between the two subtypes are high. After this method has been applied, all the genes are associated with a trust value against them. The list of genes is then sorted based on their trust value in the order of highest value at the top and lower values accordingly.

In the next step of feature selection, genes (features) above some certain threshold is being selected as top ranked feature and the features below the threshold are considered to be too noisy to trust. The threshold has been set in an order that no potential features will be discarded and unnecessary genes will not be include in the list of top ranked features. It is to be considered that the genes above the threshold possess sufficient trust to be input to a conventional feature selection algorithm. A number of feature selection methods can be applied on this list of genes. The resultant list of genes can be considered as both discriminative and trustworthy.

VI. Method Of Experiment

Data from each of these datasets were preprocessed and normalized using Robust Multi-array Average (RMA) expression measure. It contains three preprocessing steps: convolution background correction, quantile normalization and summarization based on multi-array model fit robustly by median polish algorithm [20]. In order to remove the Non informative (i.e. either redundant or irrelevant) genes, first house-keeping genes (i.e. probe names that start with AFFX) were removed. The feature selection method discussed above is then applied on the list of genes. The feature selection method ranks the genes of all the cancer datasets. After the genes are arranged in the order of their highest rank the threshold has been determined by repeated experiments and set to a value empirically. The number of genes initially selected form the ranked gene list was 10000 for Breast cancer dataset. For ALL dataset this number was 8000 and for High Grade Glioma Dataset the number of set to 8000 genes as well. For Prostate cancer dataset and Lung Cancer dataset this number was 7000 and 6500 respectively. Afterwards some selected non specific filtering approaches were imposed. The requirements were, each gene must have an expression level greater than log(200) in at least 25% of the samples. The median expression level must be greater than log(300) and have an IQR (Inter Quartile Range) more than 0.5. A supervised attribute filter *cfsSubsetEval* in weka

[21] has been used on the gene lists obtained after the non-specific filters have been used. The supervised attribute filter has been used to evaluate the discriminative quality of a subset of attribute by considering each gene's predictive ability along with the degree of redundancy. All these filtering methods are being selected considering the characteristics of the gene expression data (microarray data) of each cancer datasets. After performing all these filtering approach the number of genes has been reduced considerably. The dataset section shows the total number of genes in each dataset. For Breast cancer dataset the numbers of features are reduced to 45 from the list of 22,000 genes while for ALL dataset the number is reduced to 16 from 12,625 genes. There were 6 genes as top ranked features had been selected for High Grade Glioma dataset and for Prostate cancer dataset and Lung Cancer dataset the number of features selected for classifying these datasets were 14 and 19 respectively. The reduced number of genes then works as features for the classifiers chosen for classification of the cancer datasets. Dataset classification has been also performed on weka tool.

VII. Experimental Results

The result was analyzed using Weka tool. 10-fold corss validation has been used on the dataset to test the accuracy of classification and the time taken for building model for the chosen classifiers. Table I show the number of feature (genes) selected for classifying the dataset while Table II shows the classification accuracy for decision tree classifiers and boosting algorithms as classifiers. The other objective of this paper is to compare the time to build the model for each of the classifiers for classifying using the selected number of features. Table 3 shows the model building time (in seconds) for each of the classifiers.

TABLE I: Number of Features

Dataset	Total number of Gene	Number of genes selected as features for classification algorithms
Breast Cancer Dataset	~22,000	45
ALL Dataset	12,625	16
High Grade Glioma Dataset	~12,000	6
Prostate Cancer Dataset	~12,600	14
Lung Cancer Dataset	12,533	19

TABLE II: Classification Accuracy

Dataset	Accuracy (%)			
	C4.5	CART	Random Tree	Adaboost
Breast Cancer	91.275	90.604	76.5101	96.6443
ALL	94.531	91.406	89.0625	94.5313
High Grade Glioma	84.00	82.00	82.00	85.00
Prostate Cancer	77.451	76.470	67.6401	86.2745
Lung Cancer	93.288	96.644	92.6174	97.6325

After applying the proposed method of feature selection in [19] and efficient use of some of the conventional filtering methods, the reduced number of features has been recorded in Table I. It is important that few good features that can classify the datasets with greater accuracy are a vantage for the classification algorithms.

From Table II it is evident that the classification accuracy of the Boosting algorithm is significantly higher than all the proposed decision tree classifiers (C4.5, CART and Random Tree). Fig. 1 shows the comparison for different classifiers' accuracy of classification on all the chosen cancer datasets where AdaBoost classifier shows the highest accuracy rate for classifying all the cancer datasets.

TABLE III: Time to Build Model

Dataset	Time (in seconds)			
	C4.5	CART	Random Tree	Adaboost
Breast Cancer	1.77	9.91	0.05	5.7
ALL	0.71	7.39	0.04	0.51
High Grade Glioma	0.24	3.85	0.04	0.25
Prostate Cancer	0.91	7.15	0.05	3.3
Lung Cancer	1.36	8.46	0.03	6.38

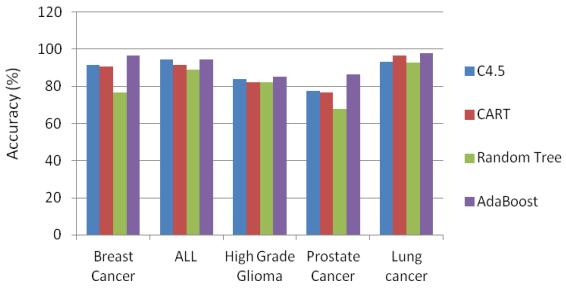

Fig. 1. Comparison of classification accuracy.

Next we take into consideration the time to build the model for each of the classifiers. The times taken to build the models have been obtained from the weka tool with the built-in classifiers provided by weka tool. The model building time for each classifier has been recorded into Table 3. The time was taken in seconds. Fig 2. Shows the comparison of time required for model building.

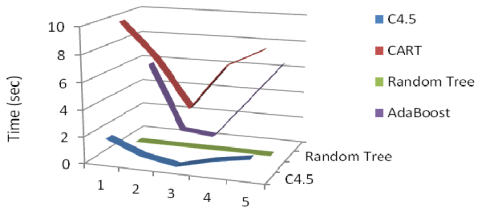

Fig. 2. Time comparisons for the classifiers to build the model.

Among all the classifiers Random tree classifier took the least amount of time to build the classification model. For all these datasets Random tree classifier shows the same result in the case of model building. After Random tree classifier the

least amount of time taken to build model is by another decision tree classifiers; C4.5. The boosting algorithm (AdaBoost) took considerably higher time to build the classification model in compare to the decision tree classifiers. Although Random tree classifier took the least amount of time to build the model, however from Table 2 we see that Random Tree classifier provides least classification accuracy. Thus we need to reconsider our objective to choose the boosting algorithm as a better one than the decision tree classifiers in classifying the cancers datasets.

What is more important in the classification of huge datasets like cancer datasets is the classification accuracy. How efficiently and how accurately classification has been done is all that counts. And in that respect AdaBoost classifier outperformed the decision tree classifiers. In modern age, where the memory and processing power of the computers (even personal computers) is so great that the slight different in the computational complexity is easily ignorable. As long as the classification accuracy is higher by the boosting algorithm we can consider this approach as a better classification approach than classifying high dimensional cancer datasets using decision tree classifiers.

VIII. Conclusion

In this paper we tried to evaluate the classification capability of boosting algorithm and the decision tree classifiers for various cancer datasets and show that the accuracy rate of classifying by AdaBoost is better than some of the popular decision tree classifiers. Microarray data such as cancer datasets are contained with huge amount of data with a very little amount of information that can be directly extracted from the datasets. Use of different efficient data mining techniques can provide us that necessary information for classification of these high dimensional data. For that purpose determining efficient classification algorithm is a great challenge.

In this paper we have used one of the proposed feature selection methods for initially rank the genes based on their predictive capability and afterwards using some conventional filtering methods the number of features for these datasets has been reduced considerably. This reduced number of features then act as the classifying features for the classification algorithms. Our main goal was to find a better classification algorithm providing a feature set, and in that respect boosting algorithm outperformed some of the frequently used decision tree classifiers. Although the amount of time to build the classification model is higher for AdaBoost than some of the decision tree classifiers, however as long as the classification accuracy is grater the boosting algorithm is a better choice as classifiers for classifying high dimensional cancer datasets.

References

[1] J. Han and M. Kamber, Data Mining, Concepts and Techniques, *Morgan Kaufmann Publishers*, 2000.

[2] T. Mitchell, Machine Learning, *MacGraw Hill*, 1997.

[3] R. Brachman, T. Khabaza, W. Kloesgan, G. Piatetsky-Shapiro, and E. Simoudis, "Mining Buisness Databases," *Comm. ACM*, vol. 39, no. 11, pp. 42-48, 1996.

[4] U. M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From Data Mining to Knowledge Discovery in Databases," *AI Magazine*, vol. 17, pp. 37-54, 1996.

[5] F. A. Ubaudi, P. J. Kennedy, D. R. Catchpoole, D. Guo, and S. J. Simoff, Miroarray *Data Mining*: Selecting Trustworthy Genes with Gene Feature Ranking.

[6] J. J. Liu, G. Culter, W. Li, Z. Pan, S. Peng, T. Hoey, L. Chen, and X. B. Ling, "Multiclass Cancer Classification and biomarker discovery using GA-based algorithms," *Bioinformatics of Oxford Journals*, vol. 21, no. 11, pp. 2691-2697, 2005.

[7] C. H. Ooi and P. Tan, "Genetic Algorithms Applied to Multi-classPrediction for the Analysis of Gene Expression Data," *Bioinformatics of Oxford Journals* vol. 10, no. 1, 2003.

[8] D. Lavanya and K. Usha Rani, "Performance Evaluation of Decision Tree Classifiers on Medical Datasets," *International Journal of computer Applications* vol. 26, no. 4, 2011.

[9] G. Stasis, A. C. Loukis, E. N. Pavlopoulos, and S. A. Koutsouris, D. "Using decision tree algorithms as a basis for a heart sound diagnosis decision support system," Presented at Information Technology Application in Biomedicine, *4th International IEEE EMBS Special Topic Conference*, 2003.

[10] J. R. Quinlan, "Induction of decision trees," *Journal of Machine Learning* vol. 1, pp. 81-106, 1986.

[11] Breiman, Friedman, Olshen and Stone, Classification and Regression Trees, Wadsworth, 1984, Mezzovico, Switzerland.

[12] *Wald I*, Machine Learning, July 2002.

[13] Y. Freund, R. E. Schapire, *a Decision-Theoretic Generalization of on-Line Learning and an Application to Boosting*, 1995.

[14] C. Sabina, X. Li, R. Gentleman, A. Vitale, K. S. Wang, M. F Foá R,and J. Ritz, Gene Expression Profiles of B-lineage Adult Acute Lymphocytic Leukemia Reveal Genetic Patterns that Identify Lineage Derivation and Distinct Mechanism of Transformation.

[15] Y. Wang, G. M. J. Klijin, and Y. Zhang, *et al.* "Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer," Lancet 2005; 365: 671-79.

[16] C. L. Nutt, D. R. Mani, and R. A. Betensky, *et al*, Gene Expression-based classification of malignant gliomas correlated better with survival than hitological classification.

[17] T. A. Stamey, J. N. Kabalin, J. E. McNeal, I. M. Johnstone, F. Freiha, E. A. Redwine, and N. Yang, "Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate: II. Radical prostatectomy treated patients," *Journal of Urology*, vol. 141, no. 5, pp. 1076–1083, 1989.

[18] G. J. Gordon *et al*, "Translation of Microarray Data into Clinically Relevant Cancer Diagnostic Tests Using Gege Expression Ratios in Lung Cancer And Mesothelioma," *Cancer Research*, vol. 62, pp. 4963-4967, 2002.

[19] A. H. Golam, M. M, M. D. Shareef, A. A. M. Hawlader, and K. Paul, "Cancer Classification from Microarray Data Using Gene Feature Ranking," *International Journal of Data Mining and Emerging Technologies* vol. 1. no. 2. November 2011

[20] J. D. Emerson and D. C. Hoaglin, "Analysis of two-way tables by medians. In: Hoaglin D. C, Mosteller F. and Turkey J.W. (ed.), Understanding Robust and Exploratory Data Analysis," New York: John Wiley & Sons. ISBN 0471384917, 1983, pp. 165-210.

[21] I. H. Witten and E. Frank, Data Mining: Practical machine learning tools and techniques, San Francisco: *Morgan Kaufmann Publishers*, 2005.

Abid Hasan was born in Dhaka, Bangladesh in the year of 1989. He completed his Bachelor of Science and Computer Science and Information Technology (CIT) from the department of Computer Science and Information Technology, Islamic University of Technology, Gazipur, Dhaka, Bangladesh in the year of 2010.

After completing his B.Sc degree, he joined as a Lecturer in Islamic University of Technology from the same year of graduating from the same university. He has some previous publications in International Conference on Computer and Computational Intelligence (ICCCI 2011) and International Journal of Data Mining and Emerging Technologies (IJDET). His area of interest is Bioinformatics, Pattern recognition and machine learning. Currently he is working on feature selection methodology for better classifying high dimensional Microarray data.