Volume 5, No. 3, March-April 2014

ISSN No. 0976-5697

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

A Comparative Analysis of Decision Tree Methods to Predict Kidney Transplant Survival

Yamuna N R
Department of Mathematics,
SRM University,
Chennai, India

Venkatesan P
Department of Statistics,
Tuberculosis Research Centre (ICMR),
Chennai-600 031, India

Abstract: The decision tree is one of the recent developments of sophisticated techniques for exploring high dimensional databases. In data mining, a decision tree is a predictive model which can be used to represent both classification and regression. The aim of this study is to classify kidney transplant patient's response based on the set of predictor variables using ensemble methods. This paper also compares the performance of decision tree algorithms (ID3, C4.5 and CART), and ensemble methods such as Random forest, Boosting and Bagging with C4.5 and CART as a base classifier. The result shows that CART with Boosting shows the better result than other methods.

Keywords: CART; C4.5; ID3; Boosting; Bagging; Random forest.

I. INTRODUCTION

Data mining is an integral part of knowledge discovery in databases, which is used to find interesting patterns, trends, statistical models, relationships in databases [1]. Decision tree is the most popular classification algorithm in data mining, which can be used for both classification as well as regression. The outcome of decision tree is a flow chart like structure model where each internal node denotes a test on an attribute, each branch represents the outcome of the test, and the bottom nodes of the decision tree is called leaf or terminal node which denotes a class prediction. At each node, the decision tree algorithm always selects the best attribute to split the data into individual classes. Once the decision rules have been determined, it is possible to use the rules to predict new node values based on unseen data. Decision tree consists of variety of algorithms such as Iterative Dichotomizer 3 (ID3) [2, 3], C4.5 [4], Chi-square automatic interaction detection (CHAID) [5], Classification and regression trees (CART) [6], C5.0 [7], etc. The most commonly used decision tree algorithm is C4.5 and CART which recently had been ranked best algorithm in the "Top 10 algorithms in data mining" [8]. Many researchers have investigated the technique of combining the predictions of several classifiers to generate a single classifier [9]. Since decision tree is an unstable method, ensemble method is used to improve the performance of the base learning algorithms. Boosting [10], Bagging [11] and Random Forest [12] are most popular ensemble methods. In boosting, AdaBoost [13] is a powerful algorithm to improve weak classifier. In 1999, Opitz and Maclin compared the ensemble methods such as Bagging, AdaBoost and Arcing.

Empirical study on these ensemble methods for decision tree has shown that Boosting and Random forests are the best ensemble methods for decision tree in situation without noise [14, 15]. The major difference between Boosting and Bagging are: Boosting uses a function of the performance of a classifier as a weight for voting, while Bagging uses equal weight voting. Boosting algorithms are stronger than Bagging on noise-free data [16].

Kidney transplantation is the organ transplant of a kidney into a patient with end-stage renal disease. The prevalence rate for end-stage renal disease is increasing day-by-day. In comparison to dialysis, kidney transplant is a better treatment method due to a healthier survival rate of the patient. However, the success rate of kidney transplant depends on several factors. Hemodialysis treatment is an effective treatment means for renal failure patients. Predicting the outcome of kidney transplant is not an easy problem in Medical research. The purpose of this paper is to compare performances of classification techniques to classify the kidney transplant patient's response based on the set of predictor variables.

The rest of this paper is organized as follows. Section 2, describes the well-known decision tree algorithms (ID3, CART and C4.5) and ensemble methods (Boosting, Bagging and Random forest). Section 3 presents the statistical measure used to evaluate the classification performance for all the methods and the experimental results. Finally, conclusions and discussions are in section 4.

II. DECISION TREES AND ENSEMBLE METHODS

A. Iterative dichotomizer 3 (ID3) Algorithm:

ID3 decision tree algorithm is based on the concept learning system. The construction of decision tree algorithm starts with the whole data set and it picks the best attribute as the root node. After the best attribute selection, it splits that node into two subgroups with the same feature value. If all objects in a subgroup have the same classification, then the process stops for that branch, and the algorithm returns a terminal node with that classification. If the subgroup contains multiple classifications, and there are no more features to test, the algorithm returns a leaf node with the most frequent classification. This algorithm use information gain to find the best feature to split the dataset. Information gain measures the reduction in impurity for a specific attribute. The partition with the maximum information gain is chosen as the decision for this node. The information gain measure is based on the entropy function from information theory.

If the response variable takes on *n* different values, then the entropy of *S* is defined as,

$$Entropy(S) = \sum_{j=1}^{n} - p_j \log_2 p_j \quad (1)$$

Where p_j is the frequency of the value *j* in *S*. The information gain, *Gain (S, A)* of an attribute *A*, relative to a collection of examples *S* is defined as,

$$Gain\ (S, A) = Entropy\ (S) - \sum_{v=1}^{n} \frac{|S_v|}{|S|} Entropy(S_v) \quad (2)$$

where S_v is the subset of *S* for which attribute *A* has value *v*.

B. Classification and Regression Tree (CART):

CART algorithm based on statistical methodology developed for classification with categorical outcomes and regression with continuous outcomes [6]. It is a data-mining tool based on binary recursive partitioning. The construction of CART algorithm is similar to that with ID3, with the exception of the information gain measure. In classification tree, the impurity measure *i(t)* is computed using Gini criterion which is used to find the best split. The goodness of a split can be defined as the reduction in impurity

$$\Delta i\ (t) = i(t) - p(t_l)\ i(t_l) - p(t_r) i(t_r) \quad (3)$$

$$i(t) = 1 - \sum_j p_j^{\,2} \quad (4)$$

Where *i(t)* denote the impurity of the node *t* and $p(t_l)$ and $p(t_r)$ are the probability that the object falls into the left and right daughter node of node *t*. p_j is the proportion of cases in category *j*. $i(t_L)$ and $i(t_R)$ are the impurities of the left and right nodes respectively. Select the predictor variable and split point with the highest reduction in impurity and perform the split of the parent node into two nodes based on the selected split point. Repeat the process using each node as a new parent node until the tree has the maximum size. After generating the maximal tree CART uses the pruning technique to select the optimal tree.

The pruning procedure develops a sequence of smaller trees and computes cost-complexity for each tree. Based on the cost-complexity parameter, the pruning procedure determines the optimal tree with high accuracy. Complexity is given by the following equation:

$$R_\alpha = R(T) + \alpha |\tilde{T}| \quad (5)$$

Where *R (T)* is the resubstitution estimated error, $|\tilde{T}|$ is the number of terminal nodes of the tree, which determines the complexity of the tree, and α is the cost-complexity associated with the tree. *R (T)* is given by the misclassification error is computed by the following equation:

$$R(T) = \frac{1}{N} \sum_{i=1}^{N} X\left(d(x_n) \neq j_n\right) \quad (6)$$

Where X is the indicator function, which is equal to 1 if the statement $X(d(x_n) \neq j_n$ is true and 0 if it is false and *d(x)* is the classifier. The value of the complexity parameter in the pruning usually lies between 0 and 1. The pruning procedure develops a group of trees using different values of complexity parameter, giving different sizes of tree. According to Breiman et al. (1984) among a group of trees of different sizes, for a value of α, only one tree of smaller size has high accuracy.

The optimal tree is one that has the smallest prediction error for new samples. Prediction error is measured using either independent test set or cross validation (CV). When the data set is not large enough to split the data into training and testing data, V-fold cross validation is used. Cross validation is repeated V times, considering each time different sub sets of training and test data, and thus developing V number of varied trees. Among the V different trees, the simplest tree that has the lowest cross validation error rate (CV error) is selected as the optimal tree.

C. C4.5:

The construction of this algorithm is similar to ID3 algorithm. Over –fitting problem is the main issue in ID3 decision tree algorithm. The C4.5 decision tree algorithm addresses this using tree pruning techniques to prune the tree generated by ID3. At each point of the decision tree, the attribute showing the largest gain ratio is selected to divide the decision tree. Gain ratio for attribute *A* is defined as

$$Gain\ Ratio\ (S, A) = Gain\ (S, A) / Split\ Info\ (S, A) \quad (7)$$

$$Split\ Info(S, A) = - \sum_{v=1}^{n} \frac{|S_v|}{|S|} Entropy(S_v) \quad (8)$$

C4.5 algorithm removes the biasness of information gain when there are many outcome values of an attribute. Moreover it uses pessimistic pruning to remove unnecessary branches in the decision tree to improve the accuracy of classification.

D. Boosting:

Boosting method has proved to be an effective method to improve the performance of base classifiers, both theoretically and empirically. It is used to adaptively change the distribution of training examples. Boosting assigns a weight to each training example and may adaptively change the weight at the end of each boosting round. A sample is drawn according to the sampling distribution of the training examples to obtain a new training set. Next a classifier is induced from the training set and used to classify all the examples in the original data. The weights of the training examples are updated at the end of each boosting round, examples that are classified incorrectly will have their weights increased, while those that are classified correctly will have their weights decreased. This forces the classifier to focus on examples that are not easy to classify in subsequent iterations. The final ensemble is obtained by aggregating the base classifiers obtained from each boosting round.

E. Bootstrap Aggregation (Bagging):

Boostrap aggregation technique repeatedly selects the samples from a dataset according to a uniform probability distribution. Each bootstrap sample has the same size as the original data. Because the sampling is done with replacement, some instances may appear several times in the same training set, while others may be omitted from the

training set. The basic procedure for bagging is summarized as follows:

Algorithm: Bagging

a. Let *v* be the number of bootstrap samples
b. For *i=1 to v* do
c. Create a bootstrap sample of size N, D_i
d. Train a base classifier C_i on the bootstrap sample D_i.
e. End for
f. $C^*(x) = \arg\max \sum_i \delta(C_i(x) = y)$

{δ(.) =1 if its argument is true and otherwise}

After training the *v* classifiers, a test instance is assigned to the class that receives the highest number of votes. Bagging improves generalization error by reducing the variance of the base classifiers. The performance of bagging depends on the stability of the base classifier.

F. Random Forest:

A random forest is a collection of unpruned decision trees [12]. It combines many tree predictors, where each tree depends on the values of a random vector sampled independently. Moreover, all trees in the forest have the same distribution. In order to construct a tree, assume that *m* is the number of training observations and "*a*" is the number of attributes in a training set. In order to determine the decision node at a tree, choose *m*<<*a* as the number of variables to be selected. Select a bootstrap sample from the *m* observations in the training set and use the rest of the observations to estimate the error of the tree in the testing phase. Randomly choose *m* variables as a decision at a certain node in the tree and calculate the best split based on the *m* variables in the training set. Trees are always grown and never pruned compared to other tree algorithms.

III. RESULTS AND DISCUSSION

The dataset used in this paper was obtained from a kidney transplant database [17]. Data set consists of 469 cases and ten attributes including the response variable such as age, sex, duration of hemodialysis prior to transplant (Dialy), diabetes (DBT), number of prior transplants (PTX), amount of blood transfusion (blood), mismatch score (MIS), use of ALG-an immune suppression drug (ALG), duration time starting from transplant (MONTH) and status of the new kidney (FAIL). Status of the new kidney was used as the response variable for fitting CART, C4.5 and ID3 classification to multiple explanatory variables. The response variable was classified into two categories – new kidney failed (40.9%) and new kidney functioning (59%). The top six ranked attributes are age, dialy, blood, MIS, ALG and MONTH are considered for building the classification model.

In this study we used gini impurity measure for categorical target attributes. 10-fold cross validation was carried out for each algorithm. Table 1 shows the accuracy comparison of different data mining algorithms. When all the nine factors were considered to find accuracy of data mining algorithm, it was found that CART model showed the highest specificity of 77.3%.

Table 1: Accuracy comparison of different data mining algorithms

Algorithms		All variable (%)	Selected variable (%)	Sensitivity (%)	Specificity (%)
ID3		62.8	63.1	55.9	69.8
C4.5	Pruned	72.4	73.5	71.7	74.4
	Unpruned	69.5	72.9	67.7	76.2
CART	Pruned	71.22	72.2	65.5	77.3

The selected variables alone were used to find sensitivity and specificity of the data mining algorithms. When nine factors are used, classification accuracy turns to be 62.8%, 72.4% and 71.2% for ID3, C4.5 and CART respectively. In C4.5 pruned decision tree, the accuracy rate is higher than the unpruned decision tree.

When six factors are used, classification accuracy turns to be 63.1%, 73.5% and 72.2% for ID3, C4.5 and CART respectively. Since classification accuracy for all variables is lower than that of six variables, we did not carry on further analysis. C4.5 had the highest sensitivity and CART had the highest specificity. ID3 had the worst accuracy, sensitivity and specificity compared to other methods. C4.5's correct classification rate is 73.5%.

Table 2 shows the accuracy, sensitivity and specificity comparison of decision tree and ensemble methods. For the dialysis dataset, we found that the decision tree CART with boosting achieved a classification accuracy of 76.74% with a sensitivity of 82.6% and a specificity of 70.0%. Next to CART with boosting, C4.5 with bagging model achieved a classification accuracy of 75.2% with a sensitivity of 75.6% and a specificity of 75.0%. The CART with bagging achieved a classification accuracy of 74.42% with a sensitivity of 82.6% and a specificity of 65.0%.

Table 2: Accuracy comparison of different ensemble methods

Ensemble methods	Percentage success		Sensitivity (%)	Specificity (%)
	All variable (%)	Selected variable (%)		
Random Forest	73.01	74.04	72.2	75.5
CART with Boosting	74.42	76.74	82.6	70.0
CART with Bagging	81.3	74.42	82.6	65.0
C4.5 with Boosting	71.04	72.0	68.0	74.3
C4.5 with Bagging	74.02	75.2	75.6	75.0

The Random forest achieved a classification accuracy of 74% with a sensitivity of 72.2% and a specificity of 75.5%. C4.5 with boosting performance was worst compared to the other technique in selected variables. Bagging using CART as a base learner may decrease the misclassification rate in prediction with respect to using a single CART.

The produced decision tree by C4.5 algorithm is given in Figure 1. Prepruning involves deciding when to stop developing subtrees during the tree building process. The minimum number of observations in a leaf can determine the size of the tree. After a tree is constructed, the C4.5 rule induction program can be used to produce a set of equivalent rules. Pruning produces fewer, more easily interpreted results. More importantly, pruning can be used as a tool to correct for potential overfitting. In C4.5 decision tree, the number of leaves is 6 and size of tree is 11.

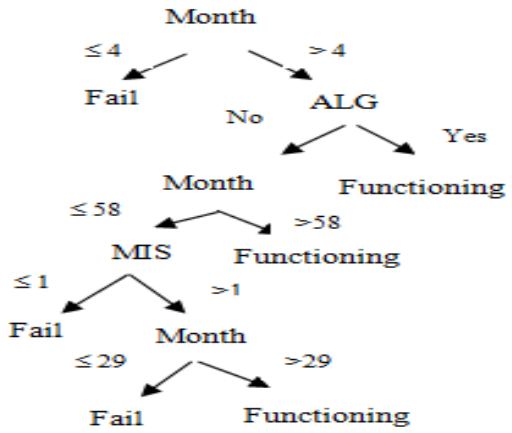

Figure 1: C4.5 decision tree

For evaluation of the fitted classification model, classification accuracy and ROC chart are used. ROC chart displays the sensitivity against 1-specificity of a classifier for a range of cutoffs. The cutoff choice represents a trade-off between sensitivity and specificity. Ideally one would like to have high values for both sensitivity and specificity, so that the model can accurately classify an outcome of events.

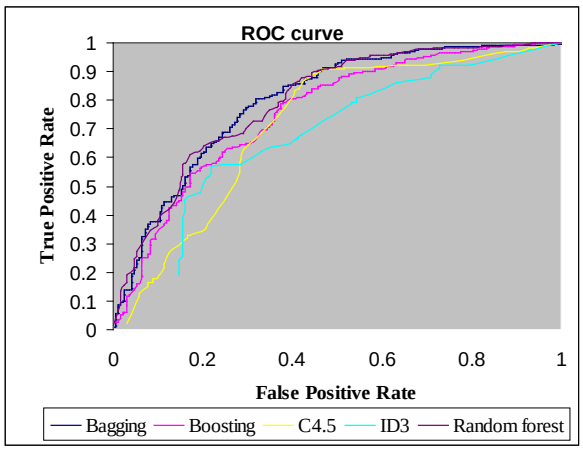

Figure 2: ROC curve for the Boosting, Bagging, Random forest, C4.5 and ID3

IV. CONCLUSION

Data Mining is gaining its popularity in almost all applications of real world. One of the data mining techniques i.e., classification is an interesting topic to the researchers as it is accurately and efficiently classifies the data for knowledge discovery. This technique enable knowledge to be extracted from data in the form of statistical models in order to see how variables relate to each other and to better understand the underlying phenomena among them. Many predictive data mining techniques generate one model that can be used to make predictions for new examples. Ensembles are combinations of several models whose individual predictions are combined in some manner. Many researchers have shown that ensembles often outperform their base models if the base models perform reasonably well on novel examples and tend to make errors on different examples. Numerous techniques have been proposed over the years for constructing ensembles which result in an increased predictive performance, and hence, they have become very popular.

Decision trees tend to perform better when dealing with categorical features. Many researchers have found that decision tree learning such as ID3, C4.5 and CART perform well in data classification. Quinlan (1996) conducted experiments to compare the performance of bagging, boosting and C4.5 and concluded that both bagging and boosting can substantially improve the performance of C4.5 whereas boosting shows greater benefits [18]. Many authors have reported that the method with the best classification performance may differ from one data to another. Endo et al. (2008) compared the seven important algorithms to predict breast cancer survival [19]. In that study, logistic regression model showed the highest accuracy compared to other algorithm. But decision tree model showed high sensitivity. Finally he concluded that the optimal algorithm might be different by the predicted objects and dataset.

In this paper we have studied the data mining algorithms to classify kidney transplant dataset using ID3, CART, C4.5, Boosting, Bagging and Random forest. Compared to ID3 and CART, C4.5 classifier methods obtain a good result. CART with boosting obtains higher results than C4.5 with bagging, CART with bagging and also Random forest shows good results but C4.5 with boosting did not perform well to classify the experimental dataset compare to other methods. Adaboost can perform poorly when the training data is noisy. CART with ensemble methods shows the high sensitivity and random forest shows the high specificity. The experimental results show that boosting with CART algorithm as base classifiers and also C4.5 with bagging is the best algorithm for classification of this medical data. The result suggested that decision tree CART with ensemble method could derive a better prognosis model in practice. Further studies are needed to confirm the findings.

V. ACKNOWLEDGEMENT

This research has been supported by a grant from the University Grants Commission. We wish to thank Director, Tuberculosis Research Centre (ICMR) Chennai for permitting us to use randomized clinical trial data.

VI. REFERENCES

[1]. Fayyad, U. M., Piatetsky-Sapiro, G and Smyth, P. “From data mining to knowledge discovery”, Advances in knowledge discovery and data mining, 1996, pp. 2-34, AAAI press.

[2]. Quinlan, J. R. “Discovering rules by induction from large collections of examples”, Expert Systems in the Micro Electronic Age, Edinburgh University Press, 1979, pp.168–201.

[3]. Quinlan, J. R. “Induction of Decision Trees”, Machine Learning, 1986, Vol 1, pp. 81-106.

[4]. Quinlan, J. R. “C4.5: Programs for Machine Learning”, Morgan Kaufmann Publishers, San Mateo, CA, 1993.

[5]. Kass, G. V. “An Exploratory Technique for Investigating Large Quantities of Categorical Data”. Applied Statistics, Vol. 29, No. 2, 1980, pp.119–127.

[6]. L. Breiman, J. Friedman, R. Olshen and C. Stone. "Classification and Regression Trees", Wadsworth International Group, Belmont, CA, 1984.

[7]. Quinlan, J.R. (2003). "C5.0 Online Tutorial", http://www.rulequest.com.

[8]. Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Yu, P. S., Zhou, Z., Steinbach, M., Hand, D. J and Steinberg, D (2008). "Top 10 Algorithms in Data Mining", Knowledge and Information Systems, 14 **(**1): 1-37.

[9]. Wolpert, D. (1992). "Stacked generalization", Neural Networks, 5: 241-259.

[10]. Schapire, R. (1990). "The strength of weak learnability", Machine Learning, 5(2): 197-227.

[11]. Breiman, L. (1996a). "Bagging Predictors", *Machine Learning*, 24(2): 123-140.

[12]. Breiman, L (2001). "Random Forests". *Machine Learning* 45 (1): 5–32.

[13]. Freund, Y. Schapire, R. (1996). "Experiments with a new boosting algorithm", In Proceedings of the Thirteenth International Conference on Machine Learning, 148-156 Bari, Italy.

[14]. Dietterich, T. G. (2000). "An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting and randomization". Machine learning, 40: 139-157.

[15]. Kotsiantis, S and Pintelas, P. (2004). "Local Boosting of Weak Classifiers", Proceedings of Intelligent Systems Design and Applications (ISDA 2004), August 26-28, Budapest, Hungary.

[16]. Opitz, D and Maclin, R (1999) "Popular Ensemble Methods: An Empirical Study", 11: 169-198.

[17]. Chap T. Le (1997). "Applied survival analysis", Wiley, New York.

[18]. Quinlan, J. R. (1996) "Bagging, Boosting and C4.5", AAAI/IAAI, 1: 725-730.

[19]. Endo, A, Shibata, T and Tanaka, H (2008) "Comparison of Seven Algorithms to Predict Breast Cancer Survival", Biomedical Soft Computing and Human Sciences, 13(2), pp.11-16.

[20]. Banfield, R.E, Hall, L.O, Bowyer, K.W and Kegeimeyer, W. P. "A comparison of decision tree ensemble creation techniques" (2007), IEEE Transactions on pattern analysis and machine intelligence, 29: 173-180.

Short Bio Data for the Authors

Corresponding author: Dr. P.Venkatesan
E.mail: venkaticmr@gmail.com
Tel. No: 9444057487

Author: N.R.Yamuna
E.mail: nryamuna@yahoo.co.in
Tel.No: 9789848284