An Efficient Classification Tree Technique for Heart Disease Prediction

S.Vijiyarani
Assistant Professor
Department of Computer Science
School of Computer Science and Engineering
Bharathiar University
Tamil Nadu, India

S.Sudha
M.Phil Research Scholar
Department of Computer Science
School of Computer Science and Engineering
Bharathiar University
Tamil Nadu, India

ABSTRACT
The data mining can be defined as discovery of relationships in large databases automatically and in some cases it is used for predicting relationships based on the results discovered. Data mining plays a vital role in various applications such as business organizations, e-commerce, health care industry, scientific and engineering. In the health care industry, the data mining is mainly used for predicting the diseases from the datasets. Various data mining techniques are available for predicting diseases namely Classification, Clustering, Association rules and Regressions. This paper analyzes the classification tree techniques in data mining. The aim of this paper is to investigate the experimental results of the performance of different classification techniques for a heart disease dataset. The classification tree algorithms used and tested in this work are Decision Stump, Random Forest, and LMT Tree algorithm. Comparative analysis is done by using Waikato Environment for Knowledge Analysis or in short, WEKA. It is open source software which consists of a collection of machine learning algorithms for data mining tasks.

Keywords
Classification, Decision Stump, LMT, Random forest, Heart disease and Weka.

1. INTRODUCTION
Data mining is an extraction of useful knowledge from large data repositories. Compared with other data mining application areas, medical data mining plays an important role and it has some unique characteristics. In medical domain, the medical data mining has the high potential for extracting the hidden patterns in the datasets [1]. These patterns are used for clinical diagnosis. The medical data are widely distributed, voluminous and heterogeneous in nature. The data is collected and then integrated to provide a user oriented approach to novel and hidden patterns of the data. A major problem in bioinformatics analysis or medical science is in attaining the correct diagnosis of certain important information. For the ultimate diagnosis, normally, many tests generally involve the clustering or classification of large scale data.

The test procedures are said to be necessary in order to reach the ultimate diagnosis. However, on the other hand, too many tests could complicate the main diagnosis process and lead to the difficulty in obtaining the end results, particularly in the case where many tests are performed. This kind of difficulty could be resolved with the aid of machine learning which could be used directly to obtain the end result with the aid of several artificial intelligent algorithms which perform the role as classifiers. Classification is one of the most important techniques in data mining. If a categorization process is to be done, the data is to be classified, and/or codified, and then it can be placed into chunks that are manageable by a human [2].

Consider an example, rather than dealing with 3.5 million merchants at a credit card company, if we could classify them into 100 or 150 different classifications that were virtually dead on for each merchant, a few employees could manage the relationships rather than needing a sales and service force to deal with each customer individually. Likewise, at a university, if an alumni group treats its donors according to the classifications, part-time students might be the representatives with minor donors and full-time professionals might receive incoming calls from the donors in which the names appear on buildings on campus.

This paper describes classification tree algorithms and it also analyzes the performance of these algorithms. The performance factors used for analysis are accuracy and error measures. The accuracy measures are TP rate, F Measure, ROC area and Kappa Statistics. The error measures are Mean Absolute Error, Root Mean Squared Error, Relative Absolute Error and Relative Root Squared Error.

The rest of this paper is organized as follows. Section 2 describes the review of literature. Section 3 discusses the classification tree algorithms used for predicting the heart disease. Experimental results are analyzed in Section 4 and Conclusions and References are given in Section 5 and 6.

2. LITERATURE REVIEW
In [11], the heart disease dataset is analyzed by neural network approach which includes variable length rate with back propagation algorithm and the momentum. The author concluded that the efficiency is increased in classification process by applying parallel approach which is included in the training phase.

In [8] the heart attack is predicted by applying association rule mining technique. The proposed algorithm CBARBSN is based on clustering and sequence numbers of the transactional database. As a result CBARBSN performed well than the existing ARNBSN algorithm. Based on the execution time the performance is compared. In [7], the cardiovascular heart disease is predicted by the classification techniques namely RIPPER, decision tree, artificial neural networks and support vector machine. As a result the author concluded that the support vector machine performs better when compared to other algorithm because it attains highest accuracy and least error rate.

In [9], the decision making process of coronary artery disease is effectively diagnosed by rotation forest algorithm. It uses the Levenberg-Marquardt back propagation algorithm of ANN to ensemble the base classifiers. As a result, without random forest algorithm the highest accuracy is obtained in Levenberg-Marquardt.

In [6] the classification data mining techniques is used for analyzing the performance. The algorithms used are naïve bayes, WAC and Apriori. As a result the performance is evaluated by using classification matrix.

In [10] based on the probability of decision support the heart disease is predicted. As a result the author concluded that decision tree performs well and sometimes the accuracy is similar in Bayesian classification.

3. HEART DISEASE PREDICTION

Heart disease plays an important role in data mining because in worldwide most of the death occur in heart diseases. Medical diagnosis plays vital role and yet complicated task that needs to be executed efficiently and accurately. To reduce cost for achieving clinical tests an appropriate computer based information and decision support should be aided. Comparative studies of various techniques are available for a good efficient and accurate implementation of automated systems**.** [4]

The World Health Organization (WHO) analyzed that twelve million deaths occurs worldwide due to Heart diseases. Many of the deaths occurs in United States and other developed countries based on cardio vascular diseases. Heart disease was the major causes of different countries include India. In every 34 seconds the heart disease kills one person. There are different categories in Heart disease but it mainly focuses on three types namely Cardiovascular Disease, Cardiomyopathy and Coronary heart disease.

3.1Data Source

To compare these data mining classification techniques Cleveland cardiovascular disease dataset from UCI repository was used [6]. The dataset has 13 attributes and 303 records. The attributes are Age - Age in years, Sex Male=1,Female=0, cp - Chest pain type, Blood pressure - Resting Blood pressure upon hospital admission Cholesterol - Serum Cholesterol in mg Fasting blood sugar - Fasting blood sugar>120 mg/dl true=1 and false=0 Resting ECG - Resting electrocardiographic results Thalach - Maximum Heart Rate, Induced Angina - Does the patient experiment angina as a result of exercise Old peak ST depression induced by exercise relative to rest Slope - Slope of the peak exercise ST segment CA - Number of major vessels colored by fluoroscopy Thal - Normal ,fixed defect, reversible defect[6]

3.2 Classification Tree Algorithms

3.2.1Decision Stump

The decision stump was coined by Wayne Iba and Pat Langley in 1992. A decision stump is a supervised learning model and it consists of decision tree in one level. That is the decision tree with root node (internal) and it is connected to its leaf node (external), based on the value of single input feature the decision stump makes a prediction. Decision stump is also referred as 1-ruler [12]. Several variations are possible based on the type of the input feature. The decision stump may be build with possible feature value as leaf node for nominal features, or two leaves with a stump, one for chosen corresponding category and other category for all other leaf. These two schemes are identical for binary features. Some threshold feature value is selected for continuous features, and the decision stump contain two leaves, one for values below the threshold and other for value above the threshold. Three or more leaves of the stump may be chosen by multiple thresholds. The bagging and boosting techniques used as components of the decision stumps.

3.2.2 LMT

LMT or Logistic Model Trees consist of a decision tree structure with logistic regression function at the leaves [13]. As in decision tree, the tested attributes is associated with every inner node. The attributes with k values, the node has k child nodes for nominal attributes and depending on the value of the attribute the instances are sorted down. For the attributes of numeric, the node has two child nodes and comparing the attributes of tested value to a threshold (the instances are sorted down based on threshold [14]. LMT uses pruning of cost complexity. Compared to other algorithm, it is slower to compute.

3.2.3 Random Forest

Random forest is an ensemble classifier that consists of many decision trees. The output of the classes is represented by individual trees. The random forest inducing algorithm is developed by Leo Breiman and Adele Cutler and it is their trademark. It is derived from random decision a forest that was proposed by Tin Kam Ho of Bell Labs in 1995 [15]. This method combines with random selection of features to construct a decision trees with controlled variations. The tree is constructed using algorithm discussed below.

i) Let N be the number of training classes and M be the number of variables in classifier.
ii) The input variable m is used to determine the node of the tree. Note that m<M
iii) Choosing n times of training sets with the replacement of all available training cases N. by predicting the classes, estimate the error of the tree.
iv) Choose m variable randomly for each node of the tree and calculate the best split.
v) At last the tree is fully grown and it is not pruned.

The tree is pushed down for predicting a new sample. When the terminal node is ends up the label is assigned the training sample [16]. This procedure is iterated over all trees ad it is reported as random forest prediction.

4. EXPERIMENTAL RESULTS

4.1 Accuracy Measure

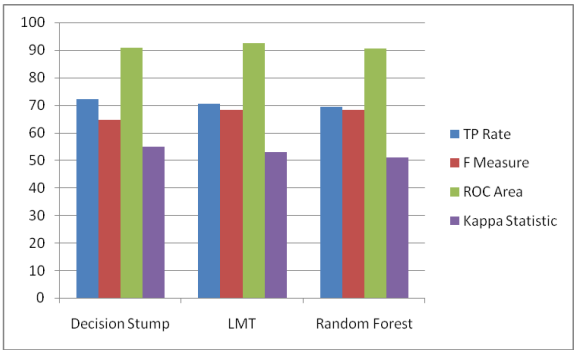
The following table shows the accuracy measure of classification techniques. They are the True Positive rate, F-measure, Receiver Operating Characteristics (ROC) Area and Kappa Statistics. The TP Rate is the ratio of play cases predicted correctly cases to the total of positive cases. . It is a probability corrected measure of agreement between the classifications and the true classes. It is calculated by taking the agreement expected by chance away from the observed agreement and dividing by the maximum possible agreement. F Measure is a way of combining recall and precision scoresbinto a single measure of performance. Recall is the ratio of relevant documents found in the search result to the total of all relevant documents [18]. Precision is the

proportion of relevant documents in the results returned. ROC Area is a traditional to plot this same information in a normalized form with 1-false negative rate plotted against the false positive rate.

Table 1

Algorithm	TP Rate	F Measure	ROC Area	Kappa Statistic
Decision Stump	72.27	64.5	90.8	55.45
LMT	70.6	68.3	92.4	53.63
Random Forest	69.3	68.3	90.6	51.56

Table 1.1



From the graph, we analyzed that, TP rate accuracy of decision stump performs better when compared to other algorithms. When compared to F Measure accuracy both LMT and Random forest have produced better results than decision stump. The ROC Area of the point lies on Random Forest, but it attains the highest accuracy in LMT algorithm. At last the accuracy measure of Kappa statistics performs better in decision stump algorithm compared to LMT and Random Forest. In order to find the accuracy of classification tree techniques decision stump outperforms well when compared to LMT and random forest algorithm.

4.2Error Rate

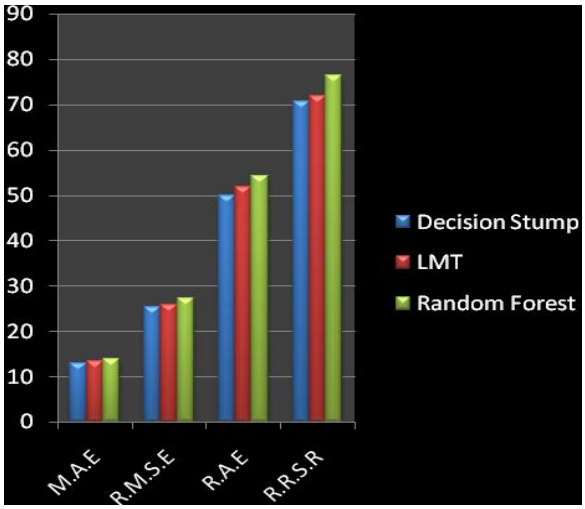
The table 2 shows the Error rate of classification techniques. They are the Mean Absolute Error (M.A.E), Root Mean Square Error (R.M.S.E), Relative Absolute Error (R.A.E) and Root Relative Squared Error (R.R.S.R) [19]. The mean absolute error (MAE) is defined as the quantity used to measure how close predictions or forecasts are to the eventual outcomes. The root mean square error (RMSE) is defined as frequently used measure of the differences between values predicted by a model or an estimator and the values actually observed. It is a good measure of accuracy, to compare the forecasting errors within a dataset as it is scale-dependent. Relative error is a measure of the uncertainty of measurement compared to the size of the measurement.

The root relative squared error is defined as a relative to what it would have been if a simple predictor had been used. More specifically, this predictor is just the average of the actual values. Thus, the relative squared error manipulates by taking the total squared error and normalizes it by dividing by the total squared error of the simple predictor. One reduces the error to the same dimensions as the quantity by taking the square root of the relative squared error is being predicted. From the analysis of all error rates we concluded that Decision Stump algorithm requires least error rate when compared to other algorithm.

Table: 2

Algorithm	M.A.E	R.M.S.E	R.A.E	R.R.S.R
Decision Stump	12.86	25.37	49.83	70.76
LMT	13.36	25.86	51.74	71.95
Random Forest	13.99	27.39	54.17	76.39

Table: 2.1



From the graph, in all error measure we analyze that decision stump performs well because it contains least error rate when compared to LMT and Random Forest techniques.

5. CONCLUSION

There are different data mining techniques that can be used for the identification and prevention of heart disease among patients. In this paper, three classification tree techniques in data mining are compared for predicting heart disease. They are tree based Decision Stump techniques, LMT and Random forest Techniques. By analyzing the experimental results, it is observed that the decision stump classification tree technique turned out to be best classifier for heart disease prediction because it contains more accuracy and least error rate. In future we tend to improve performance efficiency by applying other data mining techniques.

6. REFERENCES

[1] Ian H. Witten and Eibe Frank. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2nd edition, 2005.

[2] "Data mining: Introductory and Advanced Topics" Margaret H. Dunham

[3] K.P Soman, Shyam Diwakar, V.Vijay "Insight into Data mining theory and practice"

[4] Ruben D. Canlas Jr.,"DATA MINING IN HEALTHCARE: CURRENT APPLICATIONS AND ISSUES", August 2009

[5]."Cleveland heart disease dataset" sci2s.ugr.es/keel/dataset.php?cod=57

[6] N. Aditya Sundar, P. Pushpa Latha, M. Rama Chandra, "Performance Analysis of Classification Data Mining Techniques over Heart Disease Data base" [IJESAT] international journal of engineering science & advanced technology ISSN: 2250–3676, Volume-2, Issue-3, 470 – 478

[7] Esra Mahsereci Karabulut & Turgay İbrikçi "Effective Diagnosis of Coronary Artery Disease Using The Rotation Forest Ensemble Method" June 2011 / Accepted: 30 August 2011 / Published online: 13 September 2011 # Springer Science+Business Media, LLC 2011

[8] MA.JABBAR, Dr. PRITI CHANDRA, B.L.DEEKSHATULU "Cluster Based Association Rule Mining For Heart Attack Prediction" JTAIT Vol. 32 No.2 October 2011.

[9] Milan Kumari, Sunila Godara "Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction", IJCST Vol. 2, Issue 2, June 2011

[10] Dr. D. Raghu. T. Srikanth, Ch. Raja Jacub, "Probability based Heart Disease Prediction using Data Mining Techniques" IJCST Vol. 2, Issue 4, Oct - Dec. 2011, ISSN: 0976-8491 (Online) | ISSN: 2229-4333(Print)

[11] Dr. K. Usha Rani "Analysis of Heart Diseases Dataset Using Neural Network Approach" (IJDKP) Vol.1, No.5, September 2011

[12] "Local Additive Regression of Decision Stumps" www.math.upatras.gr/~esdlab/en/members/.../B-13%20setn06-2.pdf

[13] N. Landwehr, M. Hall, and E. Frank. "Logistic model trees", 2003.

[14] Niels Landwehr, Mark Hall and Eibe Frank, "Logistic Model Trees" Institute for Computer Science, University of Freiburg, Freiburg, Germany. Department of Computer Science, University of Waikato, Hamilton, New Zealand.

[15] L. Breimam and A. Cutler, "Random Forest (2001)", http://stat-www.berkeley.edu/users/breiman/RandomForests/cc_home.htm

[16] L.Breiman, Random Forests, Machine Learning (2001), 45, 5-32

[17] Canran Liu , Paul Frazier, Lalit umar, "Comparative assessment of the measures of thematic classification accuracy" Remote Sensing of Environment 107 (2007) 606–616 Received 28 March 2006; received in revised form 20 October 2006; accepted 21 October 2006

[18]"Binary classification performances measure cheat sheet" www.damienfrancois.be/blog/files/modelperfcheatsheet.pdf

[19] B.Nithyasri, K.Nandhini, Dr. E.Chandra "CLASSIFICATION TECHNIQES IN EDUCATION DOMAIN" (IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 05, 2010, 1679-1684

AUTHOR'S PROFILE Mrs. S.Vijayarani has completed MCA and M.Phil in Computer Science. She is working as Assistant Professor in the School of Computer Science and Engineering, Bharathiar University, Coimbatore. Her fields of research interest are data mining, privacy, security issues and data streams. She has published papers in the International journals and presented research papers in international and national conferences research papers in international and national conferences.

Ms. S.Sudha has completed M.Sc in Software Systems. She is currently pursuing her M.Phil in Computer Science in the School of Computer Science and Engineering Bharathiar University, Coimbatore. Her fields of interest are privacy in data mining.