Predicting Presence of Heart Disease in Patients using Decision Tree Classification Techniques

Rajiv Gupta
Master of Science in IT
Hood College
Frederick, MD, USA

INTRODUCTION

Heart disease is the leading cause of death in the world over the past 10 years (World Health Organization 2007). The European Public Health Alliance reported that heart attacks, strokes and other circulatory diseases account for 41% of all deaths (European Public Health Alliance 2010).
Several different symptoms are associated with heart disease, which makes it difficult to diagnose it quicker and better. Working on heart disease patients databases can be compared to real-life application. Doctor's knowledge to assign the weight to each attribute. More weight is assigned to the attribute having high impact on disease prediction. Therefore it appears reasonable to try utilizing the knowledge and experience of several specialists collected in databases towards assisting the diagnosis process. It also provides healthcare professionals an extra source of knowledge for making decisions.

The healthcare industry collects large amounts of healthcare data and that need to be mined to discover hidden information for effective decision making. Motivated by the world-wide increasing mortality of heart disease patients each year and the availability of huge amount of patients' data from which to extract useful knowledge, researchers have been using data mining techniques to help health care professionals in the diagnosis of heart disease (Helma, Gottmann et al. 2000).
Data mining is the exploration of large datasets to extract hidden and previously unknown patterns, relationships and knowledge that are difficult to detect with traditional statistical methods (Lee, Liao et al. 2000). Thus data mining refers to mining or extracting knowledge from large amounts of data. Data mining applications will be used for better health policy-making and prevention of hospital errors, early detection, prevention of diseases and preventable hospital deaths (Ruben 2009).
Heart disease prediction system can assist medical professionals in predicting heart disease based on the clinical data of patients [3]. Hence by implementing a heart disease prediction system using Data Mining techniques and doing some sort of data mining on various heart disease attributes, it can able to predict more probabilistically that the patients will be diagnosed with heart disease.
This paper presents a new model that enhances the Decision Tree accuracy in identifying heart disease patients. It uses the different algorithm of Decision Trees.

APPROACH AND METHODOLOGY

The following objectives are set for this heart prediction system.
The prediction system should not assume any prior knowledge about the patient records it is comparing.
The chosen system must be scalable to run against large database with thousands of data.
This chosen approach is implemented using WEKA tool. WEKA is an open source software tool which consists of a collection of machine learning algorithms for Data Mining tasks. It contains tools for data preprocessing, classification, regression, clustering, association rules, and visualization [9]. For testing, the classification tools and explorer mode of WEKA are used. Decision Tree classifiers with Cross Validation 10-fold in Test mode is considered for this study.
The following steps are performed in WEKA.
Start the WEKA Explorer.
Open .CSV dataset file & save in .ARFF format.
Click on Classify tab & select J48 etc (from Trees) from choose button.
Select appropriate Test mode option.
Click on Start button & result will be displayed.

DATA

For comparing various Decision Tree classification techniques, Cleveland dataset from UCI repository is used, which is available at http://archive.ics.uci.edu/ml/datasets/
Heart+Disease. The dataset has 76 attributes and 303 records. However, only 13 attributes are used for this study & testing
In this paper, three decision tree algorithms namely J48 algorithm, logistic model tree algorithm and Random Forest decision tree algorithm are used for comparison. The proposed methodology involves reduced error pruning, confident factor and seed parameters to be considered in the diagnosis of heart disease patients. Reduced error pruning has shown to drastically improve decision tree performance. These three decision tree algorithms are then tested to identify which combination will provide the best performance in diagnosing heart disease patients.

Classification Tree Algorithms Used
J48 algorithm:
J48 is an open source Java implementation of the C4.5 algorithm in the WEKA tool. This algorithm uses a greedy technique to create decision trees for classification and uses reduced-error pruning [8]. Decision tree is built by analyzing data nodes, which are used to evaluate significance of existing features. J48 algorithm is an extension of ID3 algorithm and possibly creates a small tree. It uses divide and conquers approach to growing decision trees. [5]. At each node of the tree, the algorithm chooses an attribute that can further split the samples into subsets. Each leaf node represents a class or decision.

J48 with Reduced error Pruning:
Pruning is very important technique to be used in tree creation because of outliers. It also addresses overfitting. Datasets may contain little subsets of instances that are not well defined. To classify them correctly, pruning can be used. Separate and Conquer rule learning algorithm is basis to prune any tree. This rule learning scheme starts with an empty set of rules and the full set of training instances. Reduced-error pruning is one of such separate and conquer rule learning scheme.
There are two types of pruning i.e.
Post pruning (performed after creation of tree)
Online pruning (performed during creation of tree).

Logistic Model Tree Algorithm:
Logistic Model Tree is the classifier for building 'logistic model trees', which consist of a decision tree structure with logistic regression function at the leaves. The algorithm can deal with binary and multi-class target variables, numeric and nominal attributes and missing values [11]. A combination of learners that rely on simple regression models if only little and/or noisy data is available and add a more complex tree structure if there is enough data to warrant such a structure. LMT uses cost-complexity pruning. This algorithm is significantly slower than the other algorithms.
Random Forest Algorithm:
Random forest is an ensemble classifier that consists of many decision trees. The output of the classes is represented by individual trees. It is derived from random decision a forest that was proposed by Tin Kam Ho of Bell Labs in 1995 [12]. This method combines with random selection of features to construct a decision trees with controlled variations.
It is easy to use, simple and easily parallelized.
It does not require models or parameters to select except for the number of predictors to choose at random at each node.
It runs efficiently on large databases; it is relatively robust to outliers and noise

EVALUATION OF CLASSIFICATION ALGORITHMS
The performance of classification algorithms is usually examined by evaluating the sensitivity, specificity, and accuracy of the classification. The sensitivity is proportion of positive instances that are correctly classified as positive (i.e. the proportion of patients known to have the disease, who test positive for it). The specificity is the proportion of negative instances that are correctly classified as negative (i.e. the proportion of patients known not to have the disease, who test negative for it). The accuracy is the proportion of instances that are correctly classified.

BACKGROUND

Millions of people are getting some sort of heart disease every year and heart disease is the biggest killer of both men and women in the United States and around the world. The World Health Organization (WHO) analyzed that twelve million deaths occurs worldwide due to Heart diseases. In almost every 34 seconds the heart disease kills one person in world.
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. Data mining is the use of software techniques for finding patterns and consistency in sets of data. Also, with the advent of data mining in the last two decades, there is a big opportunity to allow computers to directly construct and classify the different attributes or classes.
Knowledge of the risk factors associated with heart disease helps health care professionals to identify patients at high risk of having heart disease. Statistical analysis has identified risk factors associated with heart disease to be age, blood pressure, total cholesterol, diabetes, hyper tension, family history of heart disease, obesity and lack of physical exercise, fasting blood sugar etc [7].
Researchers have been applying different data mining
Techniques to help health care professionals with improved accuracy in the diagnosis of heart disease. Neural network, Naive Bayes, Decision Tree etc. are some techniques used in the diagnosis of heart disease.
Applying Decision Tree techniques has shown useful accuracy in the diagnosis of heart disease. But assisting health care professionals in the diagnosis of the world's biggest killer demands higher accuracy. Our research seeks to improve diagnosis accuracy to improve health outcomes.
Decision Tree is one of the data mining techniques that cannot handle continuous variables directly so the continuous attributes must be converted to discrete attributes. Couple of Decision Tree use binary discretization for continuous-valued features. Other important accuracy improving is applying reduced error pruning to Decision Tree in the diagnosis of heart disease patients. Intuitively, more complex models might be expected to produce more accurate results, but which techniques is best? Seeking to thoroughly investigate options for accuracy improvements in heart disease diagnosis this paper systematically investigates comparing multiple classifiers decision tree technique.
This research uses Waikato Environment for Knowledge Analysis (WEKA). The data of UCI repository often presented in a database or spreadsheet. In order to use this data for WEKA tool, the data sets need to be in the ARFF format (attribute-relation file format).
WEKA tool is used for to preprocess the dataset. After reviewing all these 76 different attributes, the unimportant attributes is dropped and only the important attributes (i.e. 14 attributes in this case) is considered for analysis to yield more accurate and better results. The 14th one is basically a predicted attribute, which is referred as Class. With thorough comparison between different decision tree algorithms within WEKA tool and deriving the decisions out of it, would help the system to predict the likely presence of heart disease in the patient and will definitely help to diagnose heart disease well in advance and able to cure it in right time.

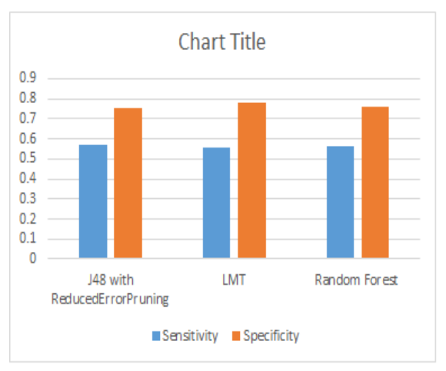
RESULTS

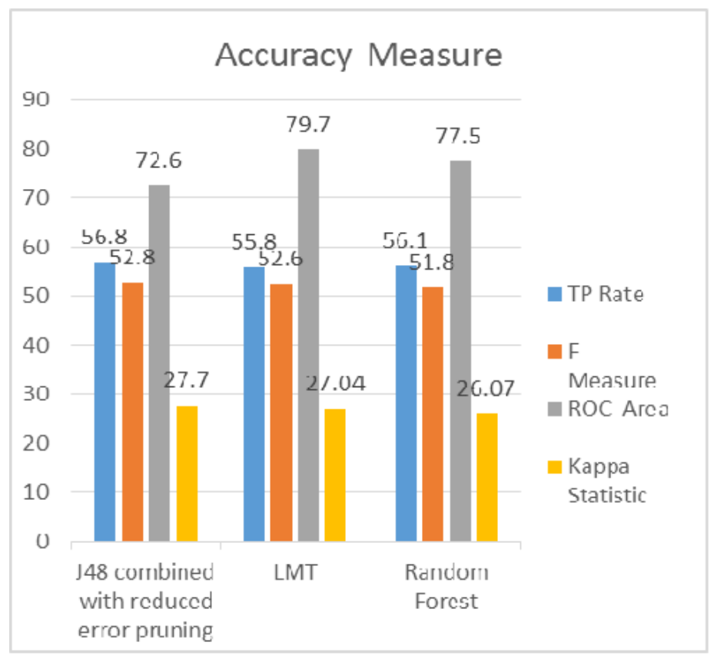
When comparing the results with LMT and Random Forest algorithm, J48 algorithm achieved higher sensitivity and accuracy while LMT achieved higher specificity than J48 and Random Forest algorithm. So overall from Table 10 and Table 11, it is concluded that J48 (with ReducedErrorPruning) has got the best overall performance.
Also, J48 algorithms use reduced-error pruning build less number of trees. The LMT algorithm builds the smallest trees. This could indicate that cost-complexity pruning prunes down to smaller trees than reduced-error pruning, but it also indicate that the LMT algorithm does not need to build large trees to classify the data. The LMT algorithm seems to perform better on data sets with many numerical attributes, whereas for good performance for 3 algorithms, the data sets with few numerical attributes gave a better performance. We can see from the results that J48 is the best classification tree algorithm among the three with pruning method.

Accuracy Measure
The following table shows the accuracy measure of classification techniques. They are the TP rate, F-measure, ROC Area, Kappa Statistics and FP rate.

From the graph, we analyzed that, TP rate accuracy of J48 performs better when compared to two algorithms. When compared to F-Measure accuracy both LMT and J48 have produced better results than Random Forest. The ROC Area of Random Forest is good, but it attains the highest accuracy in LMT algorithm. At last the accuracy measure of Kappa statistics performs better in J48 algorithm compared to LMT and Random Forest. So in terms of final accuracy among classification tree techniques, J48 outperforms well when compared to LMT and Random forest algorithm.
Error Rate
The table 6 shows the Error rate of classification techniques. 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 differences between values predicted by a model and the values actually observed. It is a good measure of accuracy. The relative error (RAE) is a measure of the uncertainty of measurement compared to the size of the measurement. The relative squared error (RRSR) manipulates by taking the total squared error and normalizes it by dividing by the total squared error of the simple predictor. From all error rate analysis, it is concluded that LMT algorithm performs well because it contains least error rate when compared to other two algorithms.
When comparing the results with LMT and Random Forest algorithm, J48 algorithm achieved higher sensitivity and accuracy while LMT achieved higher specificity than J48 and Random Forest algorithm. So overall from Table 10 and Table 11, it is concluded that J48 (with ReducedErrorPruning) has got the best overall performance.

Also, J48 algorithms use reduced-error pruning build less number of trees. The LMT algorithm builds the smallest trees. This could indicate that cost-complexity pruning prunes down to smaller trees than reduced-error pruning, but it also indicate that the LMT algorithm does not need to build large trees to classify the data. The LMT algorithm seems to perform better on data sets with many numerical attributes, whereas for good performance for 3 algorithms, the data sets with few numerical attributes gave a better performance. We can see from the results that J48 is the best classification tree algorithm among the three with pruning method.





Algorithm	M.A.E %	R.M.S.E %	R.A.E	R.R.S.R
J48 with reduced error pruning	20.71	34.93	79.94	97.22
LMT	19.75	32.78	76.20	91.24
Random Forest	20.12	33.52	77.66	93.31

Algorithm	TP Rate %	F- Measure %	ROC Area %	Kappa Statistic %
J48 with Reduced error pruning	56.8	52.8	72.6	27.7
LMT	55.8	52.6	79.7	27.04
Random Forest	56.1	51.8	77.5	26.07

Algorithm	Sensitivity %	Specificity %	Accuracy %
J48 with ReducedErrorPruning	0.568	0.752	56.76
LMT	0.558	0.781	55.77
Random Forest	0.561	0.760	56.1

CONCLUSION

By analyzing the experimental results, it is concluded that J48 tree technique turned out to be best classifier for heart disease prediction because it contains more accuracy and least total time to build. We can clearly see that highest accuracy belongs to J48 algorithm with reduced error pruning followed by LMT and Random Forest algorithm respectively. Also observed that applying reduced error pruning to J48 results in higher performance while without pruning, it results in lower performance. The best algorithm J48 based on UCI data has the highest accuracy i.e. 56.76% and the total time to build model is 0.04 seconds while LMT algorithm has the lowest accuracy i.e 55.77% and the total time to build model is 0.39 seconds
In conclusion, as identified through the literature review, we believe only a marginal success is achieved in the creation of predictive model for heart disease patients and hence there is a need for combinational and more complex models to increase the accuracy of predicting the early onset of heart disease.

FUTURE WORK

There are many possible improvements that could be explored to improve the scalability and accuracy of this prediction system. Due to time limitation, the following research/work needs to be performed in the future.
Like to make use of testing different discretization techniques, multiple classifiers Voting technique and different Decision tree types like information gain, gain ratio and Gini index. Eg. Experiment need to perform on use of Equal Frequency Discretization Gain Ratio Decision Trees by applying nine Voting scheme in order to enhance the accuracy and performance of diagnosis of heart disease.
This paper proposes a framework using combinations of support vector machines, logistic regression and decision trees to arrive at an accurate prediction of heart disease. Further work involves development of system using the mentioned methodology to be use for checking the imbalance with other data mining models.
Like to explore different rules such as Association, Clustering, K-means etc for better efficiency and ease of simplicity.
To make use of Multivariate Decision Tree approach on smaller and larger amount of data.

REFERENCES

Pang-Ning Tan, Michael Steinbach, Vipin Kumar, "Introduction To Data Mining" Addison Wesley 2006
[2] Ian Witten, Eibe Frank and Mark Hall, "Data Mining ….Practical Machine Learning Tools and Techniques", Morgan Kaufmann, 2011
[3] Arthur Asuncion and David Newman, "UCI Machine Learning Repository", 2007 http://archive.ics.uci.edu/ml/datasets/Heart+Disease
[4] Taylor Kulp "book-parser-recommender paper", 2013 https://code.google.com/p/book-parser-recommender/ source/browse/#svn%2Ftrunk%2FPaper
[5] Ian Witten "Data Mining with Weka",Weka MOOC 2013 https://www.youtube.com/playlist?list=PLm4W7_iX_v4NqP_UjceOGd-OKNVO4c_cPD
[6]. INTRODUCTION OF DATA MINING, "Data Mining: What is Data Mining" http://www.anderson.ucla.edu/faculty/ jason.frand/teacher/technologies/palace/datamining.htm.
[7] ESCAP (2010) Retrieved 7-February-2011 http://www.unescap.org/stat/data/syb2009/9.Healthrisks-causes-of-death.asp.
[8] Esposito, F., D. Malerba, et al. (1997). "A Comparative Analysis of Methods for Pruning Decision Trees." IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE VOL. 19
[9] Utgoff, Paul E. "Linear Machine Decision Tree." (1991): 6 Feb. 2013.
10] C.L. Blake and C.J. Merz. UCI repository of machine learningdatabases,1998.www.ics.uci.edu/~mlearn/MLRepository.html
[11]WEKA available at http://www.cs.waikato.ac.nz/ml/weka/ 2006
[12] J.Quinlan C4.5: Programs for Machine Learning. Morgan Kaufmann, 1992.
[13]Ruben D Canlas Jr,"DATA MINING IN HEALTHCARE CURRENT APPLICATIONS AND ISSUES" August 2009