International Journal of Engineering Research and General Science Volume 2, Issue 6, October-November, 2014
ISSN 2091-2730

Heart Disease Prediction Using Classification with Different Decision Tree Techniques

K. Thenmozhi [1], P.Deepika[2]

[1]Asst. Professor, Department of Computer Science, Dr. N.G.P. Arts and Science College, CBE

[2]Asst. Professor, Department of Computer Science and Applications, Sasurie College of Arts and Science, CBE

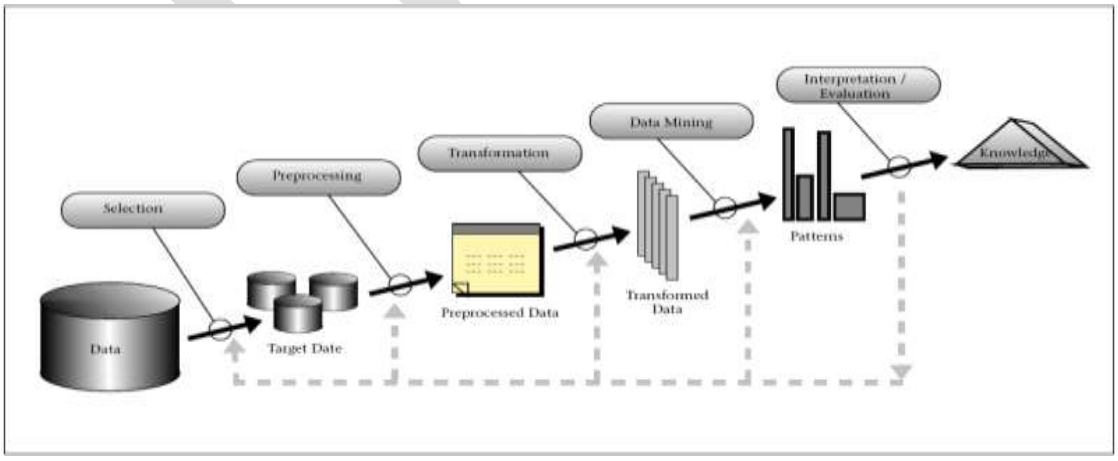
ABSTRACT: Data mining is one of the essential areas of research that is more popular in health organization. Data mining plays an effective role for uncovering new trends in healthcare organization which is helpful for all the parties associated with this field. Heart disease is the leading cause of death in the world over the past 10 years. Heart disease is a term that assigns to a large number of medical conditions related to heart. These medical conditions describe the irregular health condition that directly affects the heart and all its parts. The healthcare industry gathers enormous amount of heart disease data which are not "mined" to discover hidden information for effective decision making. Data mining techniques are useful for analyzing the data from many different dimensions and for identifying relationships. This paper explores the utility of various decision tree algorithms in classify and predict the disease.

KEYWORDS : Data mining, KDD, Classification, decision tree, ID3, C4.5,C5.0, J48

INTRODUCTION

Data mining is one of the most vital and motivating area of research with the objective of finding meaningful information from huge data sets. Now a day, Data mining is becoming popular in healthcare field because there is a need of efficient analytical methodology for detecting unknown and valuable information in health data. Data mining tools performs data analysis and may also uncover important data patterns contributing greatly to Knowledge bases, Business strategies, Scientific and Medical Research. Data mining is a more convenient tool to assists physicians in detecting the diseases by obtaining knowledge and information regarding the disease from patient's data.

Data mining and KDD (Knowledge Discovery in Databases) are related terms and are used interchangeably. According to Fayyad et al., the knowledge discovery process are structured in various stages whereas the first stage is data selection where data is collected from various sources, the second stage is preprocessing of the selected data, the third stage is transformation of the data into appropriate format for further processing, the fourth stage is Data mining where suitable Data mining technique is applied on the data for extracting valuable information and evaluation is the last stage



CLASSIFICATION

Classification is a process that is used to find a model that describes and differentiate data classes or concepts, for the purpose of using the model to predict the class of objects whose class label is unknown.

TOOLS FOR CLASSIFICATION

Some of the major tools used for constructing a classification model include Decision tree, Artificial Neural Network and Bayesian Classifier.

DECISION TREE

Berry and Linoff defined decision tree as "a structure that can be used to divide up a large collection of records into successive smaller sets of records by applying a sequence of simple decision rules. With each successive division, the members of the resulting sets become more and more similar to one another."

Decision tree is similar to the flowchart in which every non-leaf nodes denotes a test on a particular attribute and every branch denotes an outcome of that test and every leaf node have a class label. The node at the top most labels in the tree is called root node. Using Decision Tree, decision makers can choose best alternative and traversal from root to leaf indicates unique class separation based on maximum information gain[4].

Decision trees are produced by algorithms that are used to identify various ways of splitting a data set into segments. These segments form an inverted decision tree. That decision tree originates with a root node at the top of the tree

ID3

ID3 the word stands for Iterative Dichotomiser 3. ID3 is one of the decision tree model that builds a decision tree from a fixed set of training instances. The resulting tree is used to classify the future samples.

C4.5

C4.5 is the latest version of ID3 induction algorithm. It is an extension of ID3 algotithm. This builds a decision tree like the ID3. It builds a decision tree from training dataset using Information Entropy concept. So that C4.5 is often called as Statistical Classifier. This C4.5 is a widely used free data mining tool.

C5.0

This model is an extension of C4.5 decision tree algorithm. Both C4.5 and C5.0 can produce classifiers expressed as either decision tree or rulesets. In many applications, ruleset are preferred because they are simpler and easier to understand. The major differences are tree sizes and computation time. C5.0 is used to produce smaller trees and very fast than C4.5.

J48

J48 decision tree is the implementation of ID3 algorithm developed by WEKA project team. J48 is a simple C4.5 decision tree for classification. With this technique, a tree is constructed to model the classification process. Once the tree is build, it is applied to each tuple in the database and the result in the classification for that tuple.

DECISION TREE TYPES

There are many types of Decision trees. The Difference between them is mathematical model that is used to select the splitting attribute in extracting the Decision tree rules. Three most commonly used research tests types: 1) Information Gain, 2) Gini index and 3) Gain ratio Decision Trees.

INFORMATION GAIN

The entropy word stands for the meaning of information gain. This approach selects the splitting attribute that minimizes the value of entropy, thus maximizing the information gain. To identify the splitting attribute of decision tree, one must calculate the information gain for each and every attribute. Then they select the attribute that maximizes the Information Gain. It is the difference between the original information and the amount of information needed.

GINI INDEX

The Gini Index is used to measure the impurity of data. The Gini index is calculated for every attribute that is available in the dataset

GAIN RATIO

To reduce the effect of the bias resulting from the use of Information Gain, a variant known as Gain Ratio. The information Gain measure is biased toward test with many outcomes. That means, it prefers to select the attributes having a large number of values. Gain Ratio adjusts the Information Gain for each attribute to allow for the breadth and uniformity of the attribute values.

Gain Ratio = Information Gain / Split Information

Where the split information is a value based on the column sums of the frequency table.

PRUNING

After extracting the decision tree rules, reduced error pruning is pruning the extracted decision rules. Reduced error pruning is one of the efficient and fastest pruning methods and it is used to produce both accurate and small decision rules. Applying reduced error pruning provides more compact decision rules and reduces the number of extracted rules.

PERFORMANCE EVALUATION

To evaluate the performance of each combination the sensitivity, specificity and accuracy were calculated. To measure the stability of performance the data is divided into training and testing data with 10-fold cross validation.

Sensitivity = True Positive/ Positive

Specificity = True Negative/ Negative

Accuracy = (True Positive + True Negative) / (Positive + Negative)

FOCUS ON THE SURVEY:

Atul Kumar Pandey et al. proposed a Novel frequent feature selection method for heart disease prediction[7]. The Novel feature selection method algorithm which is the Attribute Selected Classifier method including CFS subset evaluator and Best First

search method followed by J48 Decision Tree then integrating the Repetitive Maximal Frequent Pattern Technique for giving better accuracy.

Atul Kumar Pandey et al. proposed a prediction model with 14 attributes[8]. They developed that model using j48 Decision Tree for classifying Heart Disease based on the Clinical features against unpruned, pruned and pruned with reduced error pruning method. They shown the result that the accuracy of Pruned J48 pruned Decision Tree with Reduced Error Pruning Approach is more better than the simple Pruned and Unpruned Approach. They proposed the prediction model to the clinical data of heart disease where training instances 200 and testing instances 103 using split test mode.

Nidhi Bhatla et al. proposed that the observations reveal that the Neural Networks with 15 attributes has outperformed over all other data mining techniques[2]. Another conclusion from the analysis is that Decision Tree has shown good accuracy with the help of genetic algorithm and feature subset selection. This Research has developed a prototype Intelligent Heart Disease Prediction system using data mining techniques namely Decision Tree, Naïve Bayes and Neural Network. A total of 909 records were obtained from the Cleveland Heart Disease database. These records were equally divided into two datasets. That are Training dataset with 455 records and Testing dataset with 454 records. Various techniques and data mining classifiers are defined in this work which has emerged in recent years for efficient and effective heart disease diagnosis. In this, Decision tree has performed well with 99.62% accuracy by using 15 attributes. Moreover, in combination with genetic Algorithm and 6 attributes, Decision tree has shown 99.2% efficiency.

Classification Techniques	**Accuracy with**	
	13 attributes	**15 attributes**
Naive Bayes	94.44	90.74
Decision Tree	96.66	99.62
Neural Network	99.25	100

Chaitrali S. Dangare et al. analyzed prediction system for Heart disease using more number of attributes[3]. This paper added two more attribute obesity and smoking. They expressed a number of factors that increase the risk of Heart disease. That are , High Blood Cholesterol, Smoking, Family History, Poor Diet, Hyper Tension ,High Blood Pressure, Obesity and Physical inactivity. The data mining classification techniques called Decision Tree, Naïve Bayes and Neural Network are analyzed on Heart Disease database. The performance of these techniques are compared based on their accuracy. They used J48 algorithm for this system. J48 algorithm uses pruning method to built a tree. This technique gives maximum accuracy on training data. And also they used Naïve Bayes classifier and Neural Network for predicting the Heart Disease. They compared the accuracy for both 13 input attribute and 15 input attribute values.

V.Manikandan et al. proposed that association rule mining are used to extract the item set relations[6]. The data classification is based on MAFIA algorithms which result in accuracy, the data is evaluated using entropy based cross validation and partition techniques and the results are compared. MAFIA stands for Maximal Frequent Itemset Algorithm. They used C4.5 algorithm to show the rank of heart attack with Decision Tree. Finally, the Heart Disease database is clustered using K-means clustering algorithm, which will remove the data applicable to heart attack from the database. They used a dataset with 19 attributes. And the goal was to have high accuracy, igh precision and recall metrics

Techniques	Precision	Recall	Accuracy(%)
K-Mean based on MAFIA	0.78	0.67	74%
K-Mean based on MAFIA with ID3	0.80	0.85	85%
K-Mean based on MAFIA with ID3 and C4.5	0.82	0.92	92%

CONCLUSION

Heart Disease is a fatal disease by its nature. This disease makes a life threatening complexities such as heart attack and death. The importance of Data Mining in the Medical Domain is realized and steps are taken to apply relevant techniques in the Disease Prediction. The various research works with some effective techniques done by different people were studied. The observations from the previous work have led to the deployment of the proposed system architecture for this work. Though, various classification techniques are widely used for Disease Prediction, Decision Tree classifier is selected for its simplicity and accuracy. Different attribute selection measures like Information Gain, Gain Ratio, Gini Index and Distance measure can be used.

REFERENCES:

[1] Bramer, M., Principles of data mining. 2007: Springer.

[2] Nidhi Bhatla, Kiran Jyoti, " An Analysis of Heart Disease Prediction using Different Data Mining Techniques" International Journal of Engineering and Technology Vol.1 issue 8 2012.

[3] Chaitrali S. Dangare and Sulabha S. Apte, " Improved Study Of Heart Disease Prediction Using Data Mining Classification Techniques", International Journal Of Computer Applications, Vol. 47, No. 10, pp. 0975-888, 2012.

[4] Apte & S.M. Weiss, Data Mining with Decision Tree and Decision Rules, T.J. Watson Research Center, http://www.research.ibm.com/dar/papers/pdf/fgcsap tewe issue_with_cover.pdf,(1997).

[5] Divya Tomar and Sonali Agarwal, "A survey on Data Mining Approaches for Healthcare", International Journal of Bio-Science and Bio-Technology, Vol. 5, No. 5(2013), pp. 241-266.

[6] V.Manikandan and S.Latha, " Predicting the Analysis of Heart Disease Symptoms Using Medical Data Mining Methods" ,International Journal of Advanced Computer Theory and Engineering, Vol. 2, Issue. 2, 2013.

[7] Atul Kumar Pandey, Prabhat Pandey, K.L. Jaiswal and Ashok Kumar Sen, " A Novel Frequent Feature Prediction Model For Heart Disease Diagnosis", International Journal of Software & Hardware Research in Engineering, Vol. 1, Issue. 1, September 2013.

[8] Atul Kumar Pandey, Prabhat Pandey, K.L. Jaiswal and Ashok Kumar Sen, " A Heart Disease Prediction Model using Decision Tree", IOSR Journal of Computer Engineering, Vol. 12, Issue.6, (Jul. – Aug. 2013), pp. 83-86.

[9] Dr. Neeraj Bhargava, Dr. Ritu Bhargava, Manish Mathuria, "Decision tree analysis on j48 algorithm for data mining", International journal of Advanced research in Computer Science and Software Engineering, Vol. 3, Issue. 6, June 2013.

[10] Tina R. Patil, Mrs. S.S. Sherekar, " Performance Analysis of Naïve Bayes and J48 Classification algorithm for Data Classification" , International Journal Of Computer Science and Applications, Vol. 6, No.2, Apr 2013.