IJCSN International Journal of Computer Science and Network, Volume 5, Issue 6, December 2016
ISSN (Online) : 2277-5420 www.IJCSN.org
Impact Factor: 1.02

Prediction of Heart Disease using Decision Tree a Data Mining Technique

[1] **Mudasir Manzoor Kirmani**, [2] **Syed Immamul Ansarullah**

[1] SKUAST-K, J&K, India

[2] MANUU, Hyderabad, India

Abstract - Data mining is the process of discovering interesting patterns and knowledge from mammoth size of data. Heart disease or cardiovascular disease is the class of diseases that involve the heart or blood vessels (arteries and veins). Today most countries face high and increasing rates of heart disease and it has become a leading cause of debilitation and death worldwide. In many countries heart disease is viewed as a "second epidemic" replacing infectious diseases leading to the main cause of death. Making a diagnosis of heart disease includes taking a complete medical evaluation, history, physical examination and early diagnosis of heart disease can help in reducing the rate of mortality (Thaksin University, 2006). One of the best ways to diagnose a heart disease is by using decision tree algorithm. Most researchers have applied J48 Decision Tree based on Gain Ratio and binary discretization. Gini Index and Information Gain are two successful types of Decision Trees that are less used in the diagnosis of heart disease. Some of the discretization techniques like voting method and reduced error pruning are known to produce more accurate Decision Trees. This research work investigates the results after applying a range of techniques to different types of Decision Trees in order to get better performance in heart disease diagnosis. To evaluate the performance of the alternative Decision Trees the sensitivity, specificity, and accuracy are calculated. This research work proposes a model that performs better than J48 Decision Tree and Bagging algorithm in the diagnosis of heart disease.

Keywords - *Data Mining, Decision Tree, Discretization, Heart Disease.*

1. Introduction

Heart disease has become a leading cause of debilitation and deaths worldwide. Today in many countries heart disease is viewed as a "second epidemic," replacing infectious diseases as the leading cause of death (Gale Nutrition Encyclopedia, 2011) [1]. 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) [2]. One fifth of Asian countries as per the Economical and Social Commission of Asia and the Pacific lives are lost to non-communicable disease such as cardiovascular diseases, (ESCAP 2010) [3].

The availability of huge amount of clinical data where vital information is hidden is seldom visited and remains untapped, researchers use data mining techniques to help health care professionals in the diagnosis of heart disease. 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. The huge amounts of data generated by healthcare transactions are too complex and voluminous to be processed and analyzed by traditional methods. Data mining provides the methodology and technology to transform these mounds of data into useful information for decision making with the aim to identify valid, novel, potentially useful correlations and patterns in data by combing through copious data sets to sniff out patterns that are too subtle or complex for humans to detect [4].

This research work presents a model that improves the accuracy of Decision Tree in identifying heart disease in patients.

2. Background

The detection of heart disease from various factors or symptoms is a multi-layered issue which is not free from false presumptions and is often accompanied by unpredictable effects. Thus the effort to utilize knowledge and experience of numerous specialists and clinical screening data of patients collected in databases to facilitate the diagnosis process is considered a valuable option. Poor clinical decisions may lead to disasters and

hence are seldom entertained. Based on the Statistical analysis of risk factors associated with heart disease like age, blood pressure, smoking habit, total cholesterol, diabetes, hypertension, family history of heart disease, obesity, and lack of physical activity.

Researchers have been applying different data mining techniques to help health care professionals with improved accuracy in the diagnosis of heart disease using neural network, Naive Bayes, Genetic algorithm and Decision Tree classifications. The researchers used pattern recognition and data mining methods in predicting models in the domain of cardiovascular diagnoses. The experiments were carried out using classification algorithms Naïve Bayes, Decision Tree, K-NN, Neural Networks and the results justifies that Naive Bayes technique outperformed other used techniques [5].

The researchers used the data mining algorithms decision trees, Naive bayes, neural networks, association classification and genetic algorithm for predicting and analyzing heart disease from the dataset [6]. Three popular data mining algorithms (support vector machine, artificial neural network and decision tree) were employed by the researchers [7] to develop a prediction model using data from 502 different cases. SVM became the best prediction model followed by artificial neural networks [7]. The researchers [8] uses decision trees, naïve bayes, and neural network to predict heart disease with 15 popular attributes as risk factors listed in the medical literature [8]. The researcher has analyzed prediction systems for Heart disease using more number of input attributes. The system uses 13 medical attributes to predict the likelihood of patient getting a Heart disease. This research paper added two more attributes i.e. obesity and smoking. The data mining classification techniques, namely Decision Trees, Naive Bayes, and Neural Networks are analyzed on Heart disease database. The performance of these techniques is compared, based on accuracy. As per our results accuracy of Neural Networks, Decision Trees, and Naive Bayes are 100%, 99.62%, and 90.74% respectively. Our analysis shows that out of these three classification models Neural Networks predicts Heart disease with highest accuracy [9]. Sitair-Taut et al. [10] used the WEKA tool to investigate applying Naïve Bayes and J4.8 Decision Trees for the detection of coronary heart disease. The results showed that there is no significant difference between Naïve Bayes and Decision Trees in the ability to realize a correct prediction of coronary heart disease (Sitar-Taut, Zdrenghea et al. 2009). Tu. et. al. [10] used the bagging algorithm in the WEKA tool and compared it with J4.8 Decision Tree in the diagnosis of heart disease. The bagging algorithm showed the better accuracy of 81.41% while the Decision Tree showed an accuracy of 78.91% (Tu, Shin et al. 2009) [10]. Hlaudi Daniel Masethe and Mosima Anna Masethe [11] applied J48, Naïve Bayes, REPTREE, CART, and Bayes Net in this research for predicting heart attacks. The research result shows prediction accuracy of 99% [11].

This research work is aimed to improve diagnosis accuracy to improve health outcomes. Some of the Decision Tree technique types used like J4.8 and C4.5 Decision Trees are based on Gain Ratio in the extraction of Decision Tree rules. However there are other Decision Tree types such as Information Gain and Gini Index that have been less used in the diagnosis of heart disease. This paper systematically investigates applying multiple classifiers voting technique with different multi-interval discretization methods such as equal width, equal frequency, chi merge and entropy with different types of Decision Tree such as Information Gain, Gini Index, and Gain Ratio.

3. Methodology

The two main issues that affect the performance of decision tree are data discritization method and type of decision tree used. Reduced error pruning is used to further improve decision tree performance. The proposed methodology involves systematically testing different discretization techniques, multiple classifiers, voting technique and different Decision Trees type in the diagnosis of heart disease patients. Different combinations of discretization methods, decision tree types and voting are tested to identify which combination will provide the best performance in diagnosing heart disease patients.

3.1 Data Discretization

Discretization is the process of converting continuous valued variables to discrete values where limited numbers of labels are used to represent the original variables. The discrete values can have a limited number of intervals in a continuous spectrum, whereas continuous values can be infinitely many (Olson and Delen, 2008) [12]. Olson and Delen (2008) [12] stated a number of reasons why researchers prefer using discrete values as opposed to continuous ones in developing prediction models for both users and experts discrete features are easier to understand, use and explain. Some of the machine learning algorithms such as Rough sets [4] can only work with discrete valued variables. The Discretization methods can be categorized in two types supervised data discretization method and unsupervised data discretization methods.

3.1.1 Supervised Data Discretization

The supervised data discretization methods use the class labels for carrying out discretization process such as chi-

square based methods and entropy based methods. The chi merge discretization method uses x2 statistic to determine the independence of the class from the two adjacent intervals, combining them if they are dependent and allowing them to be separate (Kerber 1992) [16]. This algorithm merges the pair of intervals with the lowest value of x2 as long as the number of intervals is more than predefined maximum number of intervals.

The Entropy is an information-theoretic measure of the 'uncertainty' contained in a training set (Han and Kamber 2006) [4]. It evaluates candidate cut points through an entropy-based method to select boundaries for discretization. Instances are sorted into ascending numerical order and then the entropy for each candidate cut point is calculated. Cut points are recursively selected to minimize entropy until a stopping criterion is achieved. In this model the stop criterion is achieved using five intervals of the attributes.

3.1.2 Unsupervised Discretization

The unsupervised discretization methods do not make use of class membership information during the discretization process. In unsupervised discretization, equal-width interval and equal-frequency methods are used. The equal-width discretization algorithm determines the minimum

The equal-frequency algorithm determines the minimum and maximum values of the discretized attribute, sorts all values in ascending order, and divides the range into a user-defined number of intervals so that every interval contains the same number of sorted values (Dougherty, Kohavi et al. 1995) [13].

3.2 Voting

Voting involves dividing the training data into smaller equal subsets of data and building a Decision Tree classifier for each subset of data. Voting is based on plurality or majority voting; each individual classifier contributes a single vote (Hall, Bowyer et. al. 2000) [17]. Applying voting to classification algorithms has showed successful improvement in the accuracy of these Classifiers (Paris, Affendey et al. 2010) [18].

3.3 Decision Tree

Han and Kamber (2006) [4] defined decision tree as a flowchart like tree structure, where each internal node (non-leaf node) denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (or terminal node) holds a class label. The topmost node in a tree as given in figure 1 is the root node.

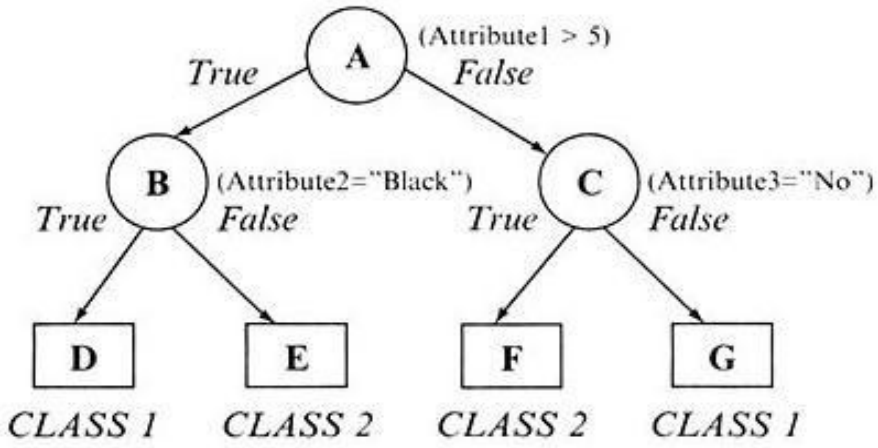

Figure 1: A simple Decision Tree

3.3.1 Decision Tree Type

There are many types of Decision Trees. The difference between them is the mathematical model that is used in selecting the splitting attribute in extracting the Decision Tree rules. The research tests were conducted on the three most commonly used types like Information Gain, Gini Index and Gain Ratio Decision Trees.

3.3.2 Information Gain

The entropy (Information Gain) approach selects the splitting attribute that minimizes the value of entropy, thus maximizing the Information Gain. To identify the splitting attribute of the Decision Tree, one must calculate the Information Gain for each attribute and then select the Attribute that maximizes the Information Gain. The Information Gain for each attribute is calculated using the following formula (Han and Kamber 2006; Bramer 2007) [4].

$$E = \sum_{i=1}^{k} \pi \log 2 \pi \quad \ldots\ldots\ldots\ldots\ldots (1)$$

Where k is the number of classes of the target attributes, π is the number of occurrences of class 'i' divided by the total number of instances (i.e. the probability of 'i' occurring).

3.3.3 Gini Index

The Gini Index measures the impurity of data. The Gini Index is calculated for each attribute in the data set. If there are k classes of the target attribute with the probability of the ith class being π, the Gini Index is defined as (Bramer 2007) [19]

$$\text{Gini Index} = 1 - \sum_{i=1}^{k} \pi 2 \quad \ldots\ldots\ldots\ldots (2)$$

3.3.4 Gain Ratio

The Information Gain measure is biased toward tests with many outcomes. Gain Ration technique prefers to select attributes having a large number of values (Han and Kamber 2006) [4]. Gain Ratio adjusts the Information Gain for each attribute to allow for the breadth and uniformity of the attribute values.

$$\text{Gain Ratio} = \frac{\text{Information Gain}}{\text{Split Information}} \quad \ldots\ldots\ldots (3)$$

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

3.4 Pruning

After extracting the decision tree rules, reduced error pruning was used to prune the extracted decision rules. Reduced error pruning is one of the fastest pruning methods and known to produce both accurate and small decision rules (Esposito, Malerba et al. 1997) [20]. Applying reduced error pruning provides more compact decision rules and reduces the number of extracted rules.

3.5 10-Fold Cross Validation

In 10-fold cross validation, the complete dataset is randomly split into 10 mutually exclusive subsets of approximately equal size dataset. The classification model is trained and tested 10 times. Each time it is trained on nine folds and tested on the remaining single fold.

3.6 Confusion Matrix

In classification problems, the primary source of performance measurements is a confusion matrix (Coincidence matrix, classification matrix or a contingency table). Given 'm' classes a confusion matrix is a table of at least size 'm' by 'm' (Olson and Delen, 2008) [12].

- If the instance is positive and it is classified as positive, it is counted as a true positive (TP);
- If it is classified as negative, it is counted as a false negative (FN);
- If the instance is negative and it is classified as negative, it is counted as a true negative (TN); and
- If it is classified as positive, it is counted as a false positive (FP).

These terms are useful when analyzing a classifier's ability and same is given below.

Table 1: Confusion Matrix

		Predicted Class	
Predicted Class		**C1**	**C2**
	C1	True Positives	False Negatives
	C2	False Positives	True Negatives

To evaluate the performance of each combination the sensitivity, specificity, and accuracy were calculated. The sensitivity is proportion of positive instances that are correctly classified as positive. The specificity is the proportion of negative instances that are correctly classified as negative. The accuracy is the proportion of

instances that are correctly classified. To measure the stability of the performance of the proposed model the data is divided into training and testing data with 10-fold cross validation.

$$Sensitivity = \frac{True\ Positive}{Positive} \quad \text{..... (4)}$$

$$Specificity = \frac{True\ Negative}{Negative} \quad \text{..... (5)}$$

$$\text{Accuracy} = \frac{(\text{True Positive} + \text{True Negative})}{(\text{Positive} + \text{Negative})} \quad \text{..... (6)}$$

4. Data

The data used in this study is taken from Cleveland Clinic Foundation Heart disease data set available at [14]. The data set has 76 raw attributes. However, all of the published experiments only refer to 13 attributes which are given in table 2:

Table 2: Selected Cleveland Heart Disease Data Set

Name	**Type**	**Description**
Age	Continuous	Age in years
Sex	Discrete	1 = male 0 = female
Cp	Discrete	Chest pain type: 1 = typical angina 2 = atypical angina 3 = non-anginal pain 4 =asymptomatic
Trestbps	Continuous	Resting blood pressure (in mm Hg)
Chol	Continuous	Serum cholesterol in mg/dl
Fbs	Discrete	Fasting blood sugar > 120 mg/dl: 1 = true 0 = false
Restecg	Discrete	Resting electrocardiographic results: 0 = normal 1 = having ST-T wave abnormality 2 =showing probable or define left Ventricular hypertrophy by Estes' criteria
Thalach	Continuous	Maximum heart rate achieved
Name	**Type**	**Description**
Exang	Discrete	Exercise induced angina: 1 = yes 0 = no
Old peak ST	Continuous	Depression induced by exercise relative to rest
Slope	Discrete	The slope of the peak exercise segment : 1 = up sloping 2 = flat 3= down sloping
Ca	Discrete	Number of major vessels colored by fluoroscopy that ranged between 0 and 3.
Thal	Discrete	3 = normal 6= fixed defect 7= reversible defect
Diagnosis	Discrete	Diagnosis classes: 0 = healthy 1= patient who is subject to possible heart disease

5. Results

After applying Info Gain, Gini Index, Gain Ratio decision trees by setting discritization parameters to Equal Width, Equal Frequency and ChiMerge the results of sensitivity, specificity, and accuracy for the diagnosis of heart disease is given in table 3 and table 4 respectively.

Table 3: Without Voting Decision Tree Results

Algorithm	Discretization	Sensitivity	Specificity	Accuracy
Info Gain	Equal width	**78.1%**	**79.4%**	**79.1%**
Gini Index		76.4%	83.4%	78.8%
Gain Ratio		66.1%	80.5%	75.5%
Info Gain	Equal frequency	75.5%	80.5%	78.3%
Gini Index		75.5%	75.7%	78.1%
Gain Ratio		76.6%	77.7%	77.6%
Info Gain	ChiMerge	71.7%	77.5%	77.5%
Gini Index		73.7%	78.4%	76.3%
Gain Ratio		65.3%	81.5%	78.6%
Info Gain	Entropy	78.1%	79.5%	77.1%
Gini Index		77.1%	81.1%	77.8%
Gain Ratio		68%	82.5%	76.7%

Below given Table 4 shows the results of Sensitivity, Specificity and Accuracy without applying voting to the Decision Tree. The highest accuracy achieved is 79.1% by the equal width discretization Information Gain Decision Tree.

Table 4: After voting decision tree results

Algorithm	Condition	Sensitivity	Specificity	Accuracy
Info Gain	Equal width	78.1%	79.4%	82.6%
Gini Index		76.4%	83.4%	82.9%
Gain Ratio		66.1%	80.5%	78.5%
Info Gain	**Equal frequency**	75.5%	75.7%	82.6%
Gini Index		**78.6%**	**80.5%**	**85.3%**
Gain Ratio		73.2%	77.7%	83.6%
Info Gain	ChiMerge	71.7%	77.5%	80.5%
Gini Index		73%	78.4%	78.3%
Gain Ratio		63.3%	81.5%	80.8%
Info Gain	Entropy	78.1%	79.5%	79.1%
Gini Index		77.1%	81.1%	80.8%
Gain Ratio		68%	82.5%	77.7%

After applying different partitions of voting to the data. The highest accuracy achieved is 85.3% by the Equal Frequency discretization Gini Index Decision Tree. The table 5 shows the difference in accuracy when applying the nine subsets voting scheme. The highest increase in the accuracy is achieved by the Equal Frequency discretization Gini Index Decision Tree that is 7.2%.

Table 5:Difference in accuracy after applying 9 subsets of voting scheme

Algorithm	Discretization	Increase in Accuracy after applying Nine subsets voting Scheme
Info Gain	Equal width	3.5%
Gini Index		4.1%
Gain Ratio		3.0%
Info Gain	Equal frequency	5.8%
Gini Index		**7.2%**
Gain Ratio		6.0%
Info Gain	ChiMerge	3.0%
Gini Index		2.0%
Gain Ratio		2.2%
Info Gain	Entropy	2.0%
Gini Index		3.0%
Gain Ratio		1.0%

Table 6: Comparing Proposed Gini Index Decision Tree Model with Previous Models and their Results

Algorithm	Sensitivity	Specificity	Accuracy
J4.8 Decision Tree	72.01%	84.48%	78.9%
Bagging Algorithm	74.93%	86.64%	81.41%
Equal Frequency Discretization Gain Ratio Decision Tree	77.9%	85.2%	84.1%
Proposed Model (using Nine voting Equal Frequency Discritization with Gini Index Decision Tree)	**78.6%**	**80.5%**	**85.3%**

Most of the researchers have used binary discretization with Gain Ratio Decision Tree in the diagnosis of heart disease; applying multi-interval equal frequency discretization with nine voting Gini Index Decision Tree provides better results in the diagnosis of heart disease patients.

6. Conclusion

Decision Tree is one of the best data mining techniques used in the diagnosis of heart disease; but compared to other data mining algorithms its accuracy is not perfect. This research work systematically tested decision tree type and voting to identify a more robust, more accurate method. Applying voting shows increase in the accuracy of different types of Decision Tree. Gini Index Decision Tree can enhance the accuracy of the diagnosis of heart disease.

7. Future Scope

The application of Decision tree in predicting heart disease will help in managing health of individuals. However, the research work needs to be tested for larger volumes of data from different medical databases. The research work has given an insight into possibility of using decision tree

algorithms in predicting heart disease and the same can be extended in different horizontal and vertical domains of medical science to serve the society in general and individuals in particular.

References

[1] Gale Nutrition Encyclopedia (2011). Heart Disease. Available at http://www.answers.com/topic/ischaemic-heart-disease (Accessed 25 February 2011)

[2] European Public Health Alliance. (July 2010-Febuary 2011). [Online]. Available: http://www.epha.org/a/2352.

[3] ESCAP Available: http://www.unescap.org/stat/data/syb2009/9.Health-risks-causes-of-d eath.asp

[4] Han, J. and Kamber, M. (2006). Data Mining: Concepts and Techniques. Second Edition, Morgan Kaufmann Publishers, San Francisco

[5] G.Subbalakshmi et al."Decision Support in Heart Disease Prediction System using Naive Bayes";Indian Journal of Computer Science and Engineering (IJCSE) Vol. 2 No. 2 Apr-May 2011

[6] Aditya Methaila et.al. EARLY HEART DISEASE PREDICTION USING DATA MINING TECHNIQUES, CCSEIT, DMDB, ICBB, MoWiN, AIAP – 2014.

[7] Mythili T et. al. "A Heart Disease Prediction Model using SVM-Decision Trees-Logistic Regression (SDL)"; International Journal of Computer Applications (0975 – 8887) Volume 68– No.16, April 2013

[8] Nidhi Bhatla and Kiran Jyoti ,"An Analysis of Heart Disease Prediction using Different Data Mining Techniques"; International Journal of Engineering Research & Technology (IJERT) Vol. 1 Issue 8, October – 2012

[9] Chaitrali S. Dangare Sulabha S. Apte, "Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques"; International Journal of Computer Applications Volume 47– No.10, June 2012.

[10] Sitar-Taut et.al." Using Machine Learning Algorithms in Cardiovascular Disease Risk Evaluation".

[11] Hlaudi Daniel Masethe, Mosima Anna Masethe, "Prediction of Heart Disease using Classification Algorithms"; Proceedings of the World Congress on Engineering and Computer Science 2014 Vol. II WCECS 2014, 22-24 October, 2014, San Francisco, USA.

[12] ` David L. Olson and Dursun Delen, "Advanced Data Mining Techniques" springer.com 2008

[13] James Dougherty, Ron Kohavi and Mehran Sahami, "Supervised and Unsupervised Discretization of Continuous Features".

[14] Cleveland Clinic Foundation Heart disease data set available at http://archive.ics.uci.edu/ml/datasets/Heart+Disease

[15] Mai Shouman, Tim Turner and Rob Stocker," Using Decision Tree for Diagnosing Heart Disease Patients"; Proceedings of the 9-th Australasian Data Mining Conference (AusDM'11), Ballarat, Australia, CRPIT Volume 121 - Data Mining and Analytics 2011

[16] Kerber, R. (1992). "ChiMerge: Discretization of Numeric Attributes." In Proceedings of the Tenth National Conference on Arterial Intelligence.

[17] Hall, L. O., K. W. Bowyer, et al. (2000). "Distributed Learning on Very Large Data Sets." In Workshop on Distributed and Parallel Knowledge Discover.

[18] Paris, I. H. M., L. S. Affendey, et al. (2010). "Improving Academic performance Prediction using Voting Technique in Data Mining." World Academy of Science, Engineering and Technology 62.

[19] Bramer, M. (2007). Principles of data mining, Springer.

[20] 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, NO. 5.