Heart Disease Detection using Datamining

I. Jeevitha[1] R. Darshna[2] R. Sruthi[3]
[1]Assistant Professor [2,3]Student
[1,2,3]Department of BCA and MSc SS
[1,2,3]Sri Krishna College of Arts and Science, Coimbatore, India-641008

Abstract— Data Mining refers to using a variety of techniques to classify the propose of information or decision making knowledge in the database .The healthcare industry collects huge amounts of healthcare records which, unfortunately, are not "mined" to discover hidden information for effective decision making .The data from medical history has been found as diverse data and it seems that the various forms of data should be interpret to envisage the heart disease of a patient. Various techniques in Data Mining have been applied to predict the patients of heart disease .Using data mining classification techniques such as Decision Tree Algorithm, Naïve Bayes, and Neural Network. etc., the patient risk level is classified .Accuracy of the risk level is high when using more number of attributes. Results show that each technique has its unique strength in realizing the objectives of the defined mining goals .It can predict the likelihood of patients getting a heart disease using medical profiles such as age, sex, blood pressure and blood sugar.

Key words: Decision Tree Algorithm, Naïve Bayes, Neural Network, Data Mining, Heart Disease

I. INTRODUCTION

The identification of the heart disease from diverse features or signs is a profound problem that is not free from false assumptions and is frequently accompanied by spontaneous effects. The health care industry collects huge amount of health care data which unfortunately are "not mined" to discover hidden info for effective decision making-NN classifier results show promising in nature for removing the redundancy of data and to improve the accuracy of classifier as compared with other classifiers of supervised and un-supervised learning methods in data mining.. The World Health Organization (WHO) has estimated that 12 million of deaths occurred worldwide every year due to the cardiovascular diseases. The factors that have been shown that increases the risk of Heart disease include Family history, Smoking, Poor diet, High blood pressure.

II. RESEARCH OBJECTIVES

The KNN classifier, decision tree algorithm and neural networks can discover and extract hidden knowledge (patterns and relationships) associated with heart disease from a historical heart disease database. It can answer complex queries for diagnosing heart disease and thus assist healthcare practitioners to make intelligent clinical decisions which traditional decision support systems cannot. By providing the effective treatments, it also helps to reduce treatment costs. To enhance visualization and ease of interpretation, it displays the results both in tabular and graphical forms.

III. DECISION TREE TECHNIQUE

Decision tree technique is one of the data mining techniques that cannot handle continuous variables directly so the continuous attributes must be converted to isolated attributes, a process called discretization. Decision Trees use binary discretization for continuous-valued features. However, multi-interval discretization methods are known to produce more accurate Decision Trees than binary discretization. There are many types of Decision Trees. The difference between those types is the mathematical model that is used in selecting the splitting attribute in extracting the Decision Tree rules. The research tests the three most commonly used types: Information Gain, Gini Index, and Gain Ratio Decision Trees. There are many popular decision tree algorithms CART, ID3, C4.5, CHAID and J48. The overall concept is to build a tree that provides balance of flexibility and accuracy.

A. Data Discretization

Discretization process is recognized to be one of the generally significant data pre-processing tasks in data mining. This method is categorised as supervised or unsupervised. The unsupervised discretization methods do not make use of class membership information during this process. Supervised binning methods transform numerical variables into categorical counterparts and refer to the target (class) information when selecting discretization cut points. They use the group labels for functioning discretization process such as chi-square based methods and entropy based methods. All the discretization methods are used as a pre-processing step to convert the constant attributes in the data set toisolated attributes. The amount of intervals used by the discretization technique is five. Each method was used to pre-process the standard data set for trials of each decision tree type.

B. Information Gain

The Information Gain approach selects the split attribute that minimize the significance of entropy, thus maximising the Information Gain. To identify the splitting attribute of the Decision Tree, 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:

$E= \sum^{k}_{i=1} Pi \log_2 Pi$, Where k is the number of classes of the target attribute.

Pi is the number of occurrences of class i divided by the total number of instances (i.e. the probability of I occurring).

C. Gini Index

The Gini Index is used to measure 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 Pi, the Gini Index is defined as: Gini Index = $1-\sum^{k}_{i=1} Pi \log_2 Pi^2$.The splitting attribute is the

attribute with the largest reduction in the value of the Gini Index.

D. Gain Ratio

A variant known as Gain Ratio was introduced to reduce the effect of the bias resulting from the use of Information Gain. The Information Gain measure is biased toward tests with many outcome. That is, it prefers to select attributes having a large number of values. Gain Ratio adjusts the Information Gain for all attribute to permit 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.

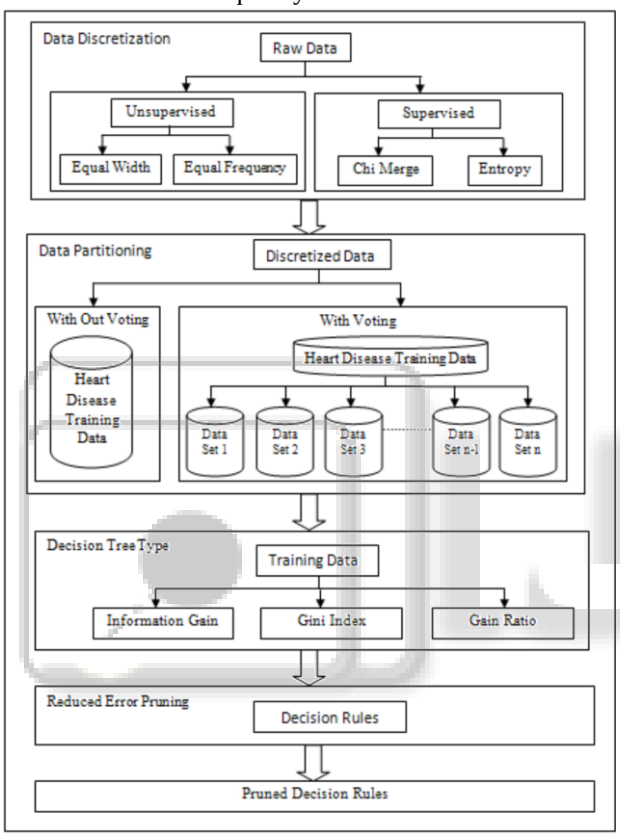

Fig. 1: Research Process used to Assess Alternative Decision Tree Techniques

IV. NAÏVE BAYES TECHNIQUE

Naïve Bayes classifier is based on Bayes theorem. The Bayes theorem is as follows: let X={x1,x2,..,xn} be a set of n attributes. In Bayesian, X is considered as evidence and H be some hypothesis means, the data of X belongs to specific class C. We have to determine P (H|X), the probability that the hypothesis H given evidence i.e. data sample X. According to Bayes theorem the P(H|X) is expressed as P (H|X) P (X|H) P (H) / P (X).

The naïve Bayes classifier, or simple Bayes classifier, consists of two main components, namely, a training set of tuples and their associated class label. Blood and urine test results from the clinical laboratory database were used as a training dataset, while class labels were defined based on the results of the interview sessions.

Naïve Bayes model for cardiovascular disease risk's level detection. Evaluation sessions were conducted in the same private hospital with cardiologists, an internist, and the head nurse of the catheterization laboratory. Four categories of questions were designed to be scaled by five-level opinions (Liker scale), namely, strongly agree, agree, neither, disagree, and strongly disagree, based on the hospital's medical procedures. The resulting application has benefits for doctors and other medical personnel to support medical analysis related to cardiovascular disease with the same level of accuracy or accuracy very similar to that achieved when manually conducted by a cardiologist or internist, especially for adults.

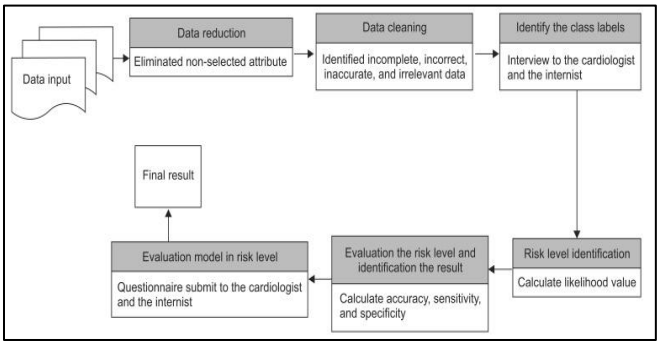

Fig. 2: Evaluation of Data Model

Identify characteristics of patients with heart disease. Only Naïve Bayes model identifies the characteristics of patients with heart disease. It shows the probability of each input attribute for the predictable state. Figure 3 shows that 80% of the heart disease patients are males (Sex = 1) of which 43% are between ages 56 and 63.

Other significant characteristics are: high probability in fasting blood sugar with less than 120 mg/dl reading, chest pain type is asymptomatic, slope of peak exercise is flat, etc.

Figure 4 shows the characteristics of patients with no heart disease with high probability in fasting blood sugar with less than 120 mg/dl reading, no exercise induced, number of major vessels is zero, etc. These results can be further analyzed.

Attributes	Values	Probability %
FastingBloodSugar	FastingBloodSugar = 0	86.179
Exang	Exang = 0	83.74
CA	ca = 0	80.488
Thal	thal = 3	79.268
Oldpeak	Oldpeak < 0.63	67.073
Slope	slope = 1	65.854
Restecg	Restecg = 0	57.724
Sex	Sex = 1	56.911
Sex	Sex = 0	43.089
Restecg	Restecg = 2	41.463
Chest	ChestPainType = 3	41.057
ThalachMaxHeartRate	ThalachMaxHeartRate >= 167.58	38.211
1 2 3 4		

Fig. 3: Naïve Bayes Attribute Characteristics Viewerin Descending Order for Patients with Heart Disease.

Attributes	Values	Probability %
FastingBloodSugar	FastingBloodSugar = 0	86.179
Exang	Exang = 0	83.74
CA	ca = 0	80.488
Thal	thal = 3	79.268
Oldpeak	Oldpeak < 0.63	67.073
Slope	slope = 1	65.854
Restecg	Restecg = 0	57.724
Sex	Sex = 1	56.911
Sex	Sex = 0	43.089
Restecg	Restecg = 2	41.463
Chest	ChestPainType = 3	41.057
ThalachMaxHeartRate	ThalachMaxHeartRate >= 167.58	38.211
1 2 3 4		

Fig. 4: Naïve Bayes Attribute Characteristic Viewer in Descending Order for Patients with No Heart Disease

V. NEURAL NETWORK TECHNIQUE

Neural Network model (Figure 5) shows that the most significant attribute value that nepotism patients with heart disease is " Old peak = 3.05 – 3.81" (98%). Other attributes that favour heart disease include " Old peak >= 3.81" , " CA=2" , " CA=3" , etc. Attributes like " Serum Cholesterol >= 382.37" , " Chest Pain Type = 2" , " CA =0" , etc. also support the predictable state for patients with refusal heart disease.

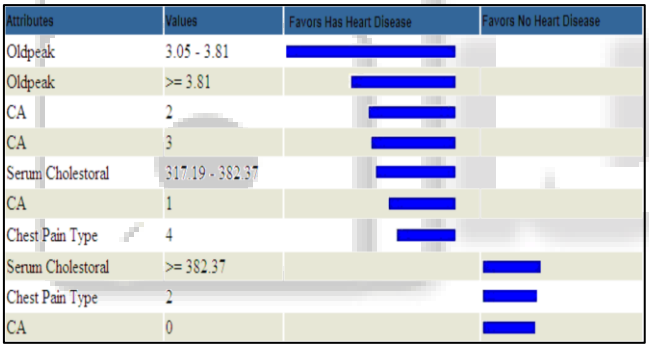

Fig. 5: Attribute Discrimination Viewer in Descending Order for Neural Network

VI. VALIDATING THE DATA MINING MODELS

Using the Lift Chart and classification matrix methods the validation of the data mining methods can be identified.

A. Lift Chart

A lift chart plots the outcome of prediction queries from a testing dataset beside known principles for the predictable column that exist in the dataset. It displays the results of the mining model, together with a representation of the results that an ideal model would produce, and a representation of the results of random guessing. Any advancement over the random line is called lift. The more lift that the model demonstrates, the more effective the model is. Only mining models that contain discrete predictable attributes can be compared in a lift chart.

We can create a lift chart by using the Input Selection tab to configure the target model and choose a test data set. Then, click the Lift Chart tab to view the completed chart.

B. Classification Matrix

A classification matrix is an additional way of proving how precisely the mining models in a structure creates predictions. To build a classification matrix, Analysis Services counts the number of good and bad predictions, using the actual values that exist in the testing dataset. The matrix is an exclusive tool since it not only shows how commonly the model correctly predicted a value, but also shows which values the model predicted incorrectly. A classification matrix shows the actual count of true positives, false positives, true negatives, and false negatives for each predictable attribute.

You can generate a classification matrix in the Mining Accuracy Chart tab of Data Mining Designer. First, use the Input Selection tab to configure the target model and choose a test data set. Then, click the Classification Matrix tab. The chart is automatically displayed, with no further configuration required.

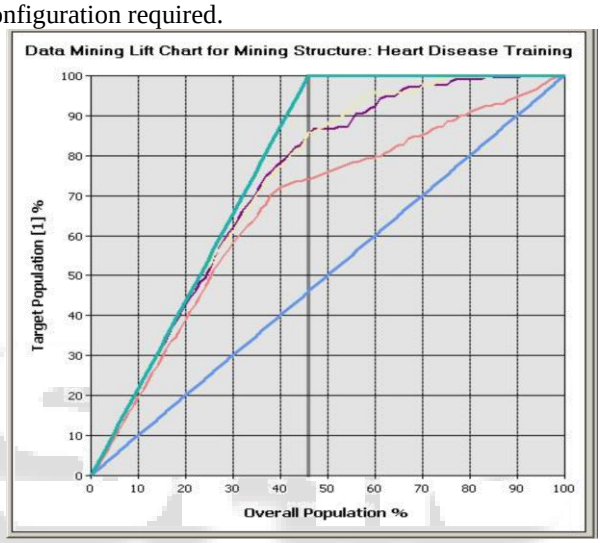

Fig. 6: Data Mining Lift Chart for Mining Structure

VII. CONCLUSION

The main goal of our effort is to provide a study of diverse data mining technique that can be employed in automated heart disease prediction systems. Various techniques and data mining classifiers are defined in this work which has emerged in recent years for efficient heart disease diagnosis. The future scope of this prediction using data mining technique is very much useful. The analysis shows that different technologies are used in all the papers with taking different number of attributes.

REFERENCES

[1] V. Krishnaiah, G. Narsimha, and N. Subhash Chandra: " Heart Disease Prediction System Using Data Mining Technique by Fuzzy K-NN Approach" , © Springer International Publishing Switzerland 2015 S.C. Satapathy et al. (eds.), Emerging ICT for Bridging the Future – Volume 1

[2] Sellappan Palaniappan, Rafiah Awang: " Intelligent Heart Disease Prediction System Using Data Mining Techniques" ,IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.8, August 2008.

[3] Mai Shouman, Tim Turner, Rob Stocker: ” Using Decision Tree for Diagnosing Heart Disease Patients” ,CRPIT Volume 121 - Data Mining and Analytics 2011.
[4] Anbarasi, M., E. Anupriya, et al. (2010). "Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm." International Journal of Engineering Science and Technology Vol. 2(10).
[5] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4981580/
[6] https://msdn.microsoft.com/enus/library/ms174493(v=sql.105).asp
[7] http://www.saedsayad.com/supervised_binning.htm
[8] https://docs.microsoft.com/en-us/sql/analysis-services/data-mining/discretization-methods-data-mining
[9] Perner, P. and S. Trautzsch (1998). "Multi-Interval Discretization Methods for Decision Tree Learning."Advances in Pattern Recognition, A. Amin, D. Dori, P. Pudil, and H. Freeman (Eds.),LNCS 1451, Springer Verlag S. 475-482.Podgorelec, V., P. Kok
[10] Kaur, H., Wasan, S. K.: “Empirical Study on Applications of Data Mining Techniques in Healthcare”, Journal of Computer Science 2(2), 194-200, 2006.

