International Journal of Data Engineering (IJDE)
Singaporean Journal of Scientific Research(SJSR)
Vol.7.No.1 2015 Pp.371-379
available at:www.iaaet.org/sjsr
Paper Received :28-12-2014
Paper Accepted:23-01-2015
Paper Reviewed by: 1Prof. M.Binod Kumar 2. Chai Cheng Yue
Editor : Dr. Chu Lio

A STUDY ON EFFECTIVE DATA MINING ASSOCIATION RULES FOR HEART DISEASE PREDICTION SYSTEM

R.Thanigaivel
Research Scholar
Department of Computer Science and Engineering
Hindustan University
Chennai, Tamilnadu

Dr. K.Ramesh Kumar
Associate Professor
Department of Information Technology
Hindustan University
Chennai, Tamilnadu.

Abstract

Mining of medical diagnoses of data is very difficult task in current data mining approach. The heart disease data is collective information of blood pressure, Cholesterol problem, diabetes with Lipid Profile and another complex diseases. The relational of one disease to another is rare so classification task is very difficult and identify. So prediction of heart disease is very critical in the process of data mining rule based classification technique used for prediction. The rule based classification technique based on association rule mining. The better rule mining technique the better classification and predication of heart disease. In this survey explained the data mining association rule for detection of heart diseases, but all the techniques has some drawbacks, so i am going to a new direction to invent new techniques.

Keywords: Data Mining, Heart Disease, Association Rule, Classification Techniques, Simulation.

1. INTRODUCTION

Data Mining is a dynamic area and one of the most popular approaches to do data mining is discovering Medical Decision Support(MDS) systems are designed to support clinicians in their diagnosis. The prediction of heart disease pattern with classification algorithms is explained in the Literature review. It is essential to find the best fit classification algorithm that has greater accuracy on classification in the case of heart disease classification. A dimensionality of the data is reduced by attribute selection methods. This cleaned data is classified by different classification algorithms such as MAFIA, K-Means, C4.5 Algorithm. This kind of classification be likely to optimize the use of data storage for numerous purposes - practical, directorial, legal, data can be classified according to any criteria, not only based on the relative position or regularity of use.

A. *Causes of casualties*

Heart disease was the major cause of casualties in the many countries including India. Heart diseases kill one person every 20 seconds in India. The term heart disease applies to a number of illnesses that affect the circulatory system, which consists of heart and blood vessels. It is intended to deal only with the condition commonly called "Heart Attack" and the factors, which lead to such condition.

The healthcare industry collects huge amounts of healthcare data and that need to be mined to discover hidden information for effective decision making. Discover of hidden patterns and relationships often go unexploited[6]. Clinicians and patients need reliable information about an individual's risk of developing Heart Disease. Ideally, they would have entirely accurate data and would be able to use a perfect model to estimate risk. Such a model would be able to categorize people with heart disease and others. Indeed, the perfect model would even be able to predict the timing of the disease's onset. The risk factors for heart disease can be divided into modifiable and non-modifiable. Modifiable risk factors include obesity, smoking, lack of physical activity and so on. The non-modifiable risk factors for heart disease are like age, gender, and family history. Many people have at least one heart disease risk factor.

Some kinds of heart disease are cardiovascular diseases, heart attack, coronary heart disease and Stroke. Stroke is a type of heart disease it is caused by narrowing, blocking, or hardening of the blood vessels that go to the brain or by high blood pressure [12][13].

B. Symptoms of heart attack can include:

- Discomfort, pressure, heaviness, or pain in the chest, arm, or below the breastbone.
- Discomfort radiating to the back, jaw, throat, or arm.
- Fullness, indigestion, or choking feeling (may feel like heartburn).
- Sweating, nausea, vomiting, or dizziness.
- Extreme weakness, anxiety, or shortness of breath.
- Rapid or irregular heartbeats

C. Association Rules

Let D = {T1, T2, . . . ,Tn} be a set of n transactions and let I be a set of items, I = {i1, i2 . . . im}. Each transaction is a set of items, i.e. Ti . An association rule is an implication of the form X ε Y, where X, Y ,and X ∩ Y = ε; X is called the antecedent and Y is called the consequent of the rule. In general, a set of items, such as X or Y , is called an itemset. In this work, a transaction is a patient record transformed into a binary format where only positive binary values are included as items. This is done for efficiency purposes because transactions represent sparse binary vectors. Let P(X) be the probability of appearance of itemset X in D and let P(Y |X) be the conditional probability of appearance of itemset Y given itemset X appears. For an itemset X ε I, support(X) is defined as the fraction of transactions Ti ε D such that X ε Ti. That is, P(X) = support(X). The support of a rule Xε Y is defined as support (X * Y) = P(X * Y). An association rule X ε Y has a measure of reliability called confidence(X * Y) defined as P(Y |X) = P(X*Y)/P (X) = support(X*Y)/support(X). The standard problem of mining association rules [1] is to find all rules whose metrics are equal to or greater than some specified minimum support and minimum confidence thresholds. A k-itemset with support above the minimum threshold is called frequent. We use a third significance metric for association rules called lift [25]: lift (X * Y) = P(Y |X)/P (Y) = confidence(X * Y)/support (Y).Lift quantifies the predictive power of Xε Y ; we are interested in rules such that lift (X * Y) > 1.

D. Decision Trees

In decision trees [6] the input data set has one attribute called class C that takes a value from K discrete values 1, . . . , K, and a set of numeric and categorical attributes A1, . . . , Ap. The goal is to predict C given A1, . . . , Ap. Decision tree algorithms automatically split numeric attributes Ai into two ranges and they split categorical attributes Aj into two subsets at each node. The basic goal is to maximize class prediction accuracy P(C = c) at a terminal node (also called node purity) where most points are in class c and c ∈ {1, . . . , K}. Splitting is generally based on the information gain ratio (an entropy based measure. The splitting process is recursively repeated until no improvement in prediction accuracy is achieved with a new split. The final step involves pruning nodes to make the tree smaller and to avoid model overfit. The output is a set of rules that go from the root to each terminal node consisting of a conjunction of inequalities for numeric variables (Ai <= x,Ai > x) and set containment for categorical variables (Aj ε {x, y, z}) and a predicted value c for class C. In general decision trees have reasonable accuracy and are easy to interpret if the tree has a few nodes. Detailed discussion on decision trees can be found in [18,19,31]. to one item. Our first constraint is the negation of an attribute, which makes search more exhaustive. If an attribute has negation then additional items are created, corresponding to each negated categorical value or each negated interval. Missing values are assigned to additional items, but they are not used. In short, each transaction is a set of items and each item corresponds to the presence or absence of one categorical value or one numeric interval.

E. The Risk Factor for Heart Disease

Family history of heart disease: - most people identify that the heat disease can run in families. That if anybody has a family history of heart disease, he/she may be at greater risk for heart attack, stroke and other heart diseases.

Smoking: - smoking is major cause of heart attack, stroke and other peripheral arterial disease. Nearly 45% of all people who die from smoking tobacco and Chewing Guthka do so due of heart and blood vessel diseases. A smoker's risk of heart attack reduces rapidly after only one year of not smoking.

Cholesterol: - abnormal levels of lipids (fats) in the blood are risk factor of heart diseases. Cholesterol is a soft, waxy substance found among the lipids in the bloodstream and in all the body's cells. High level of triglyceride (most common type of fat in body) combined with high levels of LDL (low density lipoprotein) cholesterol speed up atherosclerosis increasing the risk of heart diseases.

High blood pressure: - High blood pressure also known as HBP or hypertension is a widely misunderstood medical condition. High blood pressure increase the risk of the walls of our blood vessels walls becoming overstretched and injured. Also increase the risk of having heart attack or stroke and of developing heart failure, kidney failure and peripheral vascular disease and heart related disease.

Obesity:-the term obesity is used to describe the health condition of anyone significantly above his or her ideal healthy weight. Being obese puts anybody at a higher risk for health problem such as heart disease, stroke, high blood pressure, diabetes and more.

Lack of physical exercise: -lack of exercise is a risk factor for developing coronary artery disease (CAD). Lack of physical exercise increases the risk of CAD, because it also increases the risk for diabetes and high blood pressure.

Diabetes: Diabetes if not controlled can lead to significant heart damage including heart attack and death.

Eating Habits: Healthy diet, intake of low salt in diet, saturated fat in body, Tran's fat, cholesterol and refined sugars will lower our chances of getting heart disease.

Stress: Poorly controlled stress and danger can lead to heart attacks and strokes.

2. LITERATURE REVIEW

In 2013, **Senthil Kumar et al.,** [14], proposed a method that uses components of fuzzy logic like Fuzzification, Advanced Fuzzy Resolution Mechanism and defuzzification. The fuzzification is a process to transfer crisp values into fuzzy values. In the analysis of heart disease a fuzzy resolution mechanism uses predicted value with five layers, each layer has its own nodes. The results are tested with Cleveland heart disease dataset. Fuzzy Resolution Mechanism was developed using MATLAB. Defuzzification process converts the fuzzy set into crisp values. Then the error signal is back propagated in the network. The actual response of the network moves nearer to the desired response by adjusting the synaptic weights in a statistical sense in the network. The generalized delta rule which minimizes the error is used for the weight adjustment in the network. Thus a medical decision support system can be developed particularly in the diagnosing of heart disease.

In 2013, **Syed Umar Amin et al.,** [43], developed genetic neural network hybrid system. This system uses the global optimization advantage of genetic algorithm for initialization of neural network weights. A back propagation algorithm is used to train the networks with optimize initialization of synaptic weights by Genetic Algorithm.

Jesmin Nahar, Tasadduq Imam, Kevin S Tickle and Yi-Ping Phoebe Chen [3] presented a rule extraction model on heart disease using Apriori and Predictive Apriori techniques. They had considered some attributes like age, sex, chest pain, old peak. Based on these attributes they had developed rules for male and female which tells the healthy patterns observed in them. They had demonstrated rule mining to determine interesting knowledge which would analyze the factors causing heart disease.

In 2013, **Nabeel Al-Milli.,** [10], developed heart disease prediction system that uses the back propagation algorithm technique to develop multilayer neural networks in a supervised manner. The error-correction learning rule is the basis for the back propagation algorithm. The algorithm uses a forward pass and a backward pass through the different layers of the network. The forward pass use to fix the synaptic weights of the networks. In the backward pass, the synaptic weights are all adjusted in accordance with an error-correction rule. Error signal is calculated as the difference between the desired output and the actual response of the network.

I.S.Jenzi, P. Priyanka, Dr. P. Alli [11] proposed a new system based on Data Mining for predicting Heart Disease. They had collected patterns from medical data for finding heart disease. They did user friendly application for predicting the disease. They found that decision tree was easy to interpret and had a good accuracy. They found two difficulties like large dataset and user interface should be efficient enough to support data in different format instead of ARTF which can be taken care of in next research.

Abdullah A. Aljumah, Mohammed Gulam Ahamad, Mo-hammad Khubeb Siddique[12] focuses on predictive analysis of diabetic treatment using regression based data mining technique. Authors had collected data from Saudi Arabia consisting of variables like drug, weight, smoke intake and age. They designed a GUI based application which was easy to operate and helpful for providing training to doctors and medical assistants.

Dr. K. Rameshkumar [13] developed a model using ARM (Association Rule Mining) to extract valuable information from database. Author has proposed a new algorithm which would take care of missing values for detecting HIV AIDS. With the help of this proposed algorithm author could able to extract information about CD4 cell counts, RNA levels and treatment given for various patient. The model is lacking of handling data with a very good accuracy.

N. Deepika et al. proposed Association Rule for classification of Heart-attack patients [33]. The extraction of significant patterns from the heart disease data warehouse was presented. The heart disease data warehouse contains the screening clinical data of heart patients. Initially, the data warehouse preprocessed to make the mining process more efficient. The first stage of Association Rule used preprocessing in order to handle missing values. Later applied equal interval binning with approximate values based on medical expert advice on Pima Indian heart attack data. The significant items were calculated for all frequent patterns with the aid of the proposed approach. The frequent patterns with confidence greater than a predefined threshold were chosen and it was used in the design and development of the heart attack prediction system. The, Pima Indian Heart attack dataset used was obtained from the UCI machine learning repository. Characteristics of the patients like number of times of chest pain and age in years were recorded. The actions comprised in the preprocessing of a data set are the removal of duplicate records, normalizing the values used to represent information in the database, accounting for missing data points and removing unneeded data fields. Moreover it might be essential to combine the data so as to reduce the number of data sets besides minimizing the memory and processing resources required by the data mining algorithm [33]. In the real world, data is not always complete and in the case of the medical data, it is always true. To remove the number of inconsistencies which are associated with data we use Data preprocessing.

K. Srinivas et al. presented Application of Data Mining Technique in Healthcare and Prediction of Heart Attacks [34]. The potential use of classification based data mining techniques such as Rule based, Decision tree, Naïve Bayes and Artificial Neural Network to the massive Volume of healthcare data. Tanagra data mining tool was used for exploratory data analysis, machine learning and statistical learning algorithms. The training data set consists of 3000 instances with 14 different attributes. The instances in the dataset are representing the results of different types of testing to predict the accuracy of heart disease. The performance of the classifiers is evaluated and their results are analyzed. The results of comparison are based on 10 tenfold cross-validations. According to the attributes the dataset is divided into two parts that is 70% of the data are used for training and 30% are used for testing. The comparison made among these classification algorithms out of which the naive Bayes algorithm considered as the best performance algorithm. The performance of various algorithms is listed below [1].

Table.1. Performance Study of Data mining Algorithms

The Algorithms Used	Accuracy	Time Taken
Naïve Bayes	54.33%	613ms
Decision Tree	53.10	816ms
KNN	46.78%	1005ms

Diagnosis of heart disease was used Naïve Bayes, K-NN, Decision List in this Naïve Bayes has taken a time to run the data for accurate result when compared to other algorithms.

Sudha et al. [21] to propose the classification algorithm like Naïve Bayes, Decision tree and Neural Network for predicting the stroke diseases. The classification algorithm like decision trees, Bayesian classifier and back propagation neural network were adopted in this study. The records with irrelevant data were removed from data warehouse before mining process occurs. Data mining classification technology consists of classification model and evaluation model. The classification model makes use of training data set in order to build classification predictive model. The testing data set was used for testing the classification efficiency. Then the classification algorithm like decision tree, naive Bayes and neural network was used for stroke disease prediction. The performance evaluation was carried out based on three algorithms and compared with various models used and accuracy was measured. While comparing these classification algorithms, the observation shows the neural network performance was more than the other two algorithms.

Latha Parthiban and R. Subramanian presented Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm [4]. Adaptable based fuzzy inputs are adapted with a modular neural network to rapidly and accurately approximate complex functions. The CANFIS model combined the neural network adaptive capabilities and the fuzzy logic quantitative approach then integrated with genetic algorithm to diagnosis the presence of the disease. Coactive neuro-fuzzy inference system model has good training performance and classification accuracies. Dataset of heart disease was obtained from UCI Machine Learning Repository .Coactive Neuro-fuzzy modeling was proposed as a dependable and robust method developed to identify a nonlinear relationship and mapping between the different attributes.

Dangare et al. proposed Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques [6]. Prediction System for heart disease used system contains huge amount of data, used to extract hidden information for making intelligent medical diagnosis. The main objective of this research was to build Intelligent Heart Disease Prediction System that gives diagnosis of heart disease using historical heart database. To develop the system, medical terms such as sex, blood pressure, and cholesterol like 13 input attributes are used. To get more appropriate results, two more attributes i.e. obesity and smoking, as attributes were considered as important attributes for heart disease. A Multi-layer Perceptron Neural Networks (MLPNN) that maps a set of input data onto a set of appropriate. It consists of 3 layers input layer, hidden layer & output layer. There is connection between each layer & weights are assigned to each connection. The primary function of neurons of input layer is to divide input into neurons in hidden layer. The dataset consists of total 573 records in heart disease database. The total records are divided into two data sets one is used for training consists of 303 records & another for testing consists of 270 records. Initially dataset contained some fields, in which some value in the records was missing. These were identified and replaced with most appropriate values using Replace Missing Values filter. The Replace Missing Values filter scans all records & replaces missing values with mean mode method known as Data Preprocessing. After pre-processing the data, data mining classification techniques such as Neural Networks, is used for classification. Many problems in business, science, industry, and medicine can be treated as classification problems. Owing to the wide range of applicability of ANN and their ability to learn complex and nonlinear relationships including noisy or less precise information, neural networks are well suited to solve problems in biomedical engineering. So here use for the neural network technique is classification of medical dataset 14 attributes by considering the single and multilayer neural network models [19].

Olatubosun Olabode et al. [32] to classify the Cerebrovascular disease by using artificial neural network with back propagation error method. The Multi-layer perceptions artificial neural networks with back-propagation error method were feed-forward nets with one or more layers of nodes between the input and output nodes. These additional layers contain hidden units or nodes that were not directly connected to both the input and output nodes. The neural network was trained using back propagation algorithm with sigmoid function on one hidden layer with the 16 input attributes. Predictive models were used in variety of domains for the diagnosis. Dataset for this work were collected 100 records (60 males and 40 females) from federal medical fields. The input values obtained from the records of the forms the input variables in the input layer with 16 nodes. The neural network weights were initialized randomly. This work range of the weights was between [-0.5 and 0.5] and the learning rate was set between 0.1 and 0.9. The training, validation, generalization accuracy was measured.

Table 2: Table shows different data mining techniques used in the diagnosis of Heart Disease over different Heart Diseases datasets.

Author	Year	Techniques Used	Attribute
Carlos et al.,	2001	Association Rules	25
Dr. K. Usha Rani	2011	Classification	13
		Neurel Network	
Jesmin Nahar et al.,	2013	Apriori	14
		Predictive Apriori	
		Tertius	
Latha et al.,	2008	Genetic Algorithm	14
		CANFIS	
Majabbar et al.,	2011	Clustering	14
		Association Rule Mining	
		Sequence Mining	
Ms. Ishtaket et al.,	2013	Decision Tree	15
		Naïve Bayes	
		Neural Network	
Nan-Chen et al.,	2012	(EVAR)	
		Machine Learning	
		Markov Blanket	
Oleg et al.,	2012	Artificial Neural Network	
		Genetic Polymorphism	
Shadab et al.,	2012	Naïve Bayes	15
Shantakumar et al.,	2009	MAFIA	13
		Clustering	
		K-Means	

3. DATA MINING MODEL

In the describe survey CART, ID3 and decision table have been used to predict attributes such as age, sex, blood pressure and blood sugar for chances of a patient getting heart disease. The data is analyzed and implemented in WEKA ("Waikato Environment for Knowledge Analysis") tool. It is open source software which consists of a collection of machine learning algorithms for data mining tasks. Data mining finds out the valuable information hidden in huge volumes of data. Weka tool is a collection of machine learning algorithms for data mining techniques, written in Java. We have used 10 folds cross validation to minimize any bias in the process and improve the efficiency of the process. The three classifiers like CART (Classification and Regression Tree), ID3 (Iterative Dichotomized 3) and decision table

(DT) were implemented in WEKA. The results show clearly that the proposed method performs well compared to other similar methods in the literature, taking into the fact that the attributes taken for analysis are not direct indicators of heart disease.

4. COMPARATIVE STUDY

Here, we analyze heart data set visually using different attributes and figure out the distribution of values. Figure 1 shows the distribution of values of Heart disease patients. Table 3 shows the experimental result. Here carried out some experiments in order to evaluate the performance and usefulness of different classification algorithms for predicting heart patients.

Table 3: Performance of the Classifiers

Evaluation Criteria	Classifiers		
	CART	ID3	Decision Tree(DT)
Timing to Build Model(in sec)	0.28	0.03	0.05
Correctly Classified Instances	257	225	259
Incorrectly Classified Instances	51	77	55
Accuracy %			

Here show that CART classifier has more accuracy than other classifiers. The percentage of correctly classified instances is often called accuracy or sample accuracy of a model. Kappa statistic, mean absolute error and root mean squared error will be in numeric value only. It is also show the relative absolute error and root relative squared error in percentage for references and evaluation. The results of the simulation are shown in Table 4.

Table 4: Training and Simulation Error

Evaluation Criteria	Classifiers		
	CART	ID3	Decision Table(DT)
Kappa Statistics	0.0148	0.0266	-0.0045
Mean Absolute Error	0.2345	0.2661	0.2664
Root Mean Squared Error(RMSE)	0.3348	0.4992	0.3687
Relative Absolute Error(RAE)	88.09	103.52	99.12
Root Relative Squared Error(RRSE)	94.45	141.02	99.12

Comparison of detailed accuracy by class is shown in table 5.

Table 5: Comparison Of Accuracy Measures

Classifier	TP	FP	Precision	Recall	Class
CART	0.988	0.978	0.842	0.989	Healthy
	0.021	0.013	0.265	0.024	Possible Heart Disease
ID3	0.849	0.823	0.864	0.845	Healthy
	1.179	0.152	0.176	0.180	Possible Heart Disease
Decision Table(DT)	0.975	0.974	0.845	0.974	Healthy
	0.025	0.143	0.13	0.025	Possible Heart Disease

The performance of the learning techniques is highly dependent on the nature of the training data. Confusion matrices are very useful for evaluating classifiers. The columns represent the predictions, and the rows represent the actual class. To evaluate the robustness of classifier, the usual methodology is to perform cross validation on the classifier.

Table 6: Confusion Matrix

Classifier	Healthy	Possible Heart Disease	Class
CART	252	3	Healthy
	47	1	Possible Heart Disease
ID3	213	38	Healthy
	37	8	Possible Heart Disease
Decision Table	249	6	Healthy
	47	1	Possible Heart Disease

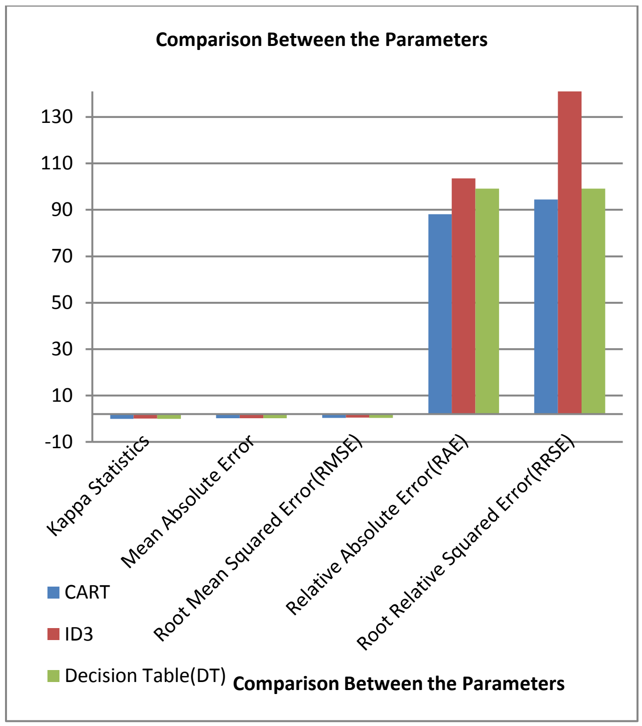

Figure 1. Comparison in between the parameters

Based on the above Figures 3, 4 and Table 6, we can clearly see that the highest accuracy is 83.49% and the lowest is 72.93%. The other algorithm yields an accuracy of 82.50%. In fact, the highest accuracy belongs to the CART Classifier. An average of 241 instances out of total 303 instances is found to be correctly classified with highest score of 253 instances compared to 221 instances, which is the lowest score. The total time required to build the model is also a crucial parameter in comparing the classification algorithm.

In this simple experiment, from Table II, we can say that a ID3 and DT requires the shortest time which is around 0.02 and 0.03 seconds consecutive with compared to CART which requires the longest model building time which is around 0.23 seconds. Kappa statistic is used to assess the accuracy of any particular measuring cases, it is usual to distinguish between the reliability of the data collected and their validity. The average Kappa score from the selected algorithm is around -0.0041 - 0.0261. From figure 4, we can observe the differences of errors resultant from the training of the three selected algorithms. This experiment implies a very commonly used indicator which is mean of absolute errors and root mean squared errors. Alternatively, the relative errors are also used. Since, we have two readings on the errors, taking the average value will be wise.

To better understand the importance of the input variables, it is customary to analyze the impact of input variables during heart disease prediction, in which the impact of certain input variable of the model on the output variable has been analyzed. Tests were conducted using three tests for the assessment of input variables: Chi-square test, Info Gain test and Gain Ratio test. Different algorithms provide very different results, i.e. each of them accounts the relevance of variables in a different way. The average value of all the algorithms is taken as the final result of variables ranking, instead of selecting one algorithm and trusting it. The results obtained with these values are shown in Table VI.

Table 7: Results of Test and Average Rank

Variable	Chi-Squared	Info Gain	Gain Ratio	Average Rank
Age	3.7548	0.015896	0.005689	1.287526
Sex	4.6542	0.059864	0.018545	1.542351
chest pain	29.4875	0.087554	0.043184	10.21548
Trestbps(resting blood pressure)	7.3687	0.018956	0.007853	2.458776
Cholesterol	0.2845	0.000887	0.000552	0.089084
fasting blood sugar	0.6798	0.005481	0.002821	0.228917
Restecg(resting electrocardiographic results)	13.2598	0.025848	0.026584	4.054512
Thalach(maximum heart rate achieved)	2.9654	0.005498	0.035448	0.985452
Exang(exercise induced angina	20.3485	0.048758	0.047134	6.624872
Slope	25.8854	0.068955	0.042215	9.021456

The aim of this analysis is to determine the importance of each variable individually. Table 7 shows that attribute *cp (Chestpain)* impacts output the most, and that it showed the best performances in all of the three tests. Then these attributes follow: *slope (The slope of the peak exercise segment), Exang (Exercise induced angina),* and *Restecg (Resting electrocardiographic)*. Figure 2 shows the importance of each attributes.

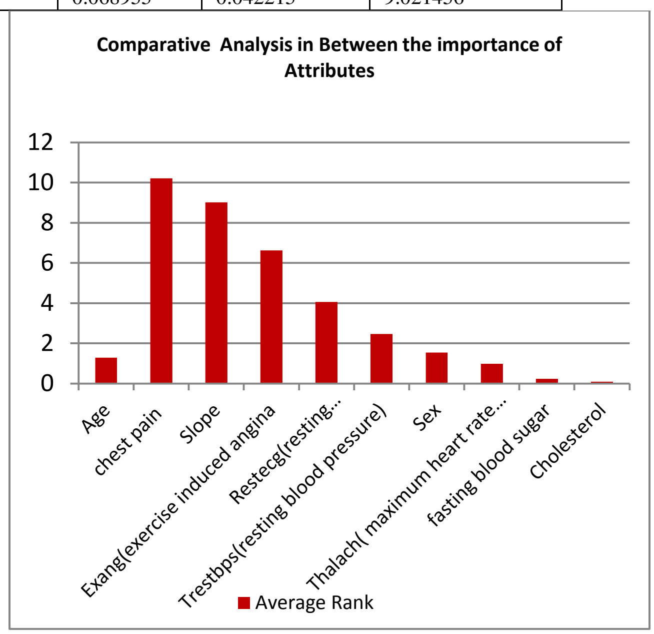

Figure 2:Comparative Analysis in Between the importance of Attributes

5. CONCLUSION

Heart disease is one of the leading causes of death worldwide and the early prediction of heart disease is important. The computer aided heart disease prediction system helps the physician as a tool for heart disease diagnosis. Some Heart Disease classification system is reviewed in this paper. From the analysis it is concluded that, data mining plays a major role in heart disease classification. Neural Network with offline training is a good for disease prediction in early stage and the good performance of the system can be obtained by preprocessed and normalized dataset. The classification accuracy can be improved by reduction in features.

Reference

[1] Matjaz' Kukar, Analysing and improving the diagnosis of ischaemic heart disease with machine learning, Elsevier, 1999.
[2] Humar Kahramanli, Novruz Allahverdi, Design of a hybrid system for the diabetes and heart diseases, Elsevier, 2008.
[3] Jesmin Nahar, Tasadduq Imam, Computational intelligence for heart disease diagnosis: A medical knowledge driven approach, Elsevier, 2013.
[4] Nan-Chen Hsieh & Lun-Ping Hung & Chun-Che Shih, Intelligent Postoperative Morbidity Prediction of Heart Disease Using Artificial Intelligence Techniques, J Med Syst, 2012.
[5] Adebayo Peter Idowu, Data Mining Techniques for Predicting Immunize-able Diseases: Nigeria as a Case Study, International Journal of Applied Information Systems, 2013.
[6] Monali Dey, Siddharth Swarup Rautaray, Study and Analysis of Data mining Algorithms for Healthcare Decision Support System, International Journal of Computer Science and Information Technologies(2014).
[7] R. Bhuvaneswari and K. Kalaiselvi, "Naïve Bayesian Classification Approach in Healthcare Applications", International Journal of computer Science and Telecommunication", vol. 3, no. 1, pp. 106-112, Jan 2012.
[8] M. Shouman, T. Turner, and R. Stocker, "Applying K-Nearest Neighbour in Diagnosing Heart Disease Patients", International Journal of Information and Education Technology, vol. 2, no. 3, pp. 220-224, Jun 2012.
[9] A. S. Kumar, "Diagnosis of Heart Disease Using Advanced Fuzzy Resolution Mechanism, International Journal of Science and Applied Information Technology, vol. 2, no. 2, pp. 22-30, Mar 2013.
[10] N. Al-Milli, "Backpropagation Neural Network for Prediction of Heart Disease", Journal of Theoretical and Applied Information Technology, vol. 56, no. 1, Oct 2013.
[11] I.S.Jenzi , P.Priyanka, Dr. P. Alli "A reliable Classifier Model Using Data Mining Approach for Heart Disease Prediction" International Journal of Advanced Research in Computer Science and Software Engineering. Volume 3, Issue 3, pp. 20-24, March 2013.
[12] Abdullah A. Aljumah, Mohammed Gulam Ahamad, Mohammad Khubeb Siddique "Application of data mining: Diabetes helth care in young and old patients" Journal of King Saud University- Computer and Information Sciences,25 pp. 127-136, 2013.
[13] Dr. K Rameshkumar "Association Rules Mining from HIV/AIDS patient's case history database with missing values" International Journal on Data Mining and Intelligent Information Technology Applications" Vol.2 , No. 1,pp. 18-24, March 2012.
[14] Sean N. Ghazavi, Thunshun W. Liao. "Medical data mining by fuzzy modeling with selected features Artificial Intelligence in Medicine 43,pp. 195-206 , 2008.
[15] M. Anbarasi, E. Anupriya and N.CH.S.N. Iyenga, "Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm", *International Journal of Engineering Science and Technology*, Vol. 2, No. 10, pp. 5370 - 5376, 2010.
[16] Chaitrali S. Dangare and Sulabha S. Apte, "Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques", *International Journal of Computer Applications*, Vol. 47, No. 10, pp. 0975 – 888, 2012.
[17] D. Shanthi, G. Sahoo and Dr. N. Saravanan, "Designing an Artificial Neural Network Model for the Prediction of Thrombo-embolic Stroke", *International Journal of Biometric and Bioinformatics*, Vol. 3, No. 1, pp. 250 - 255, 2008.
[18] P. K. Anooj, "Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules", *Journal of King Saud University Computer and Information Sciences*, Vol. 11, pp. 309 - 314, 2011.
[19] J. C. Obi and A. A. Imainvan, "Decision Support System for the Intelligent Identification of Alzheimer using Neuro Fuzzy logic", *International Journal on Soft Computing* , Vol. 2, No. 2, pp. 25 - 38, 2011.
[20] D. Shanthi, G. Sahoo and Dr. N. Saravanan, "Evolving Connection Weights of Artificial Neural Network Using Genetic Algorithm With Application to the Prediction Stroke Diseases", *International Journal of Soft Computing*, Vol. 2, pp. 95 - 101, 2009.
[21] A. Sudha, P. Gayathiri and N. Jaisankar, "Effective Analysis and Predictive Model of Stroke Disease using Classification Methods", *International Journal of Computer Applications*, Vol. 43, No. 14, pp. 0975 – 8887, 2012.
[22] Tom Dent, "*Predicting the risk of coronary heart disease*", PHG foundation publisher, 2010.
[23] World Health Organization, "*Global status report on no communicable diseases*", 2010.
[24] World Health Organization, "*Pocket Guidelines for Assessment and Management of Cardiovascular Risk*", 2007.
[25] D. Shanthi, G. Sahoo and N. Saravanan "Input Feature Selection using Hybrid Neuro-Genetic Approach in the Diagnosis of Stroke Disease", *International Journal of Computer Science and Network Security*, Vol. 8, No.12, pp. 99 - 106, 2008.

[26] M A. Jabbar, Priti Chandra and B. L. Deekshatulu, “Cluster based association rule mining for heart attack prediction”, *Journal of Theoretical and Applied Information Technology*, Vol. 32, No.2, pp. 197 - 201, 2011.
[27] Jyoti Soni, Ujma Ansari and Dipesh Sharma, “Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction”, *International Journal of Computer Applications*, Vol. 17, No. 8, pp. 43 – 48, 2011.
[28] K. Srinivas, G. Raghavendra Rao and A. Govardhan, “Survey on prediction of heart morbidity using data mining techniques”, *International Journal of Data Mining & Knowledge Management Process*, Vol. 1, No. 3, pp. 14 - 34, 2011.
[29] K. Usha Rani, “Analysis of heart diseases dataset using neural network approach”, *International Journal of Data Mining and Knowledge Management Process* ,Vol. 1, No. 5, pp. 1 - 8, 2011.
[30] E. P. Ephzibah and V. Sundarapandian, “A neuro fuzzy expert system for heart disease diagnosis”, *An International Journal Computer Science & Engineering*, Vol. 2, No. 1, pp. 17 - 23, 2012.
[31] Mai Shouman, Tim Turner and Rob Stocker, “Integrating Decision Tree and K-Means Clustering with Different Initial Centroid Selection Methods in the Diagnosis of Heart Disease Patients”, *Proceedings of the International Conference on Data Mining*, 2012.
[32] Olatubosun Olabode and Bola Titilayo Olabode, “Cerebrovascular Accident Attack Classification Using Multilayer Feed Forward Artificial Neural Network with Back Propagation Error”, *Journal of Computer Science*, Vol. 8, No. 1, pp.18 - 25, 2012.
[33]N. Deepika and K. Chandra shekar, “Association rule for classification of Heart Attack Patients”, *International Journal of Advanced Engineering Science and Technologies*, Vol. 11, No. 2, pp. 253 – 257, 2011.
[34]K. Srinivas, B. Kavitha Rani and Dr. A. Govrdhan, “Application of Data Mining Techniques in Healthcare and Prediction of Heart Attacks”, *International Journal on Computer Science and Engineering*, Vol. 02, No. 02, pp. 250 - 255, 2011.
[35]Asha Rajkumar and B. Sophia Reena, “Diagnosis Of Heart Disease Using Data mining Algorithm” , *Global Journal of Computer Science and Technology*, Vol. 10, No. 10, pp. 38 - 43, 2010.
[36]Latha Parthiban and R. Subramanian, “Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm", International Journal of Biological, Biomedical and Medical Sciences 3; 3, 2008.
[37]T. John Peter, K. Somasundaram, study and development of novel feature selection framwork for Heart disease prediction, International Journal of Scientific and Research Publications, 2012.
[38] Indira S. Fal Dessai, Intelligent Heart Disease Prediction System Using Probabilistic Neural Network, International Journal on Advanced Computer Theory and Engineering, 2013.
[39] M.Akhil jabbar, Dr.Priti Chandra, Dr.B.L Deekshatulu, Heart Disease Prediction System using Associative Classification and Genetic Algorithm, International Conference on Emerging Trends in Electrical, Electronics and Communication Technologies, 2012.
[40] Chaitrali S. Dangare,Sulabha S. Apte, Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques, International Journal of Computer Applications, 2012.
[41] Rashedur M. Rahman, Farhana Afroz, Comparison of Various Classification Techniques Using Different Data Mining Tools for Diabetes Diagnosis, Journal of Software Engineering and Applications, 2013.
[42] Venkatadri.M, Dr. Lokanatha C. Reddy a review on data mining from past to the future. International Journal of Computer Applications, 2011.
[43]Syed Umar Amin, Kavita Agarwal, Dr. Rizwan Beg, “Genetic Neural Network Based Data Mining in Prediction of Heart Disease Using Risk Factor”, in Proc. IEEE Conference on Information and Communication Technologies(ICT), 2013.