A Comparative Study on Prediction of Heart Disease using Cluster and Rank based Approach

K. Aravinthan[1], Dr. M. Vanitha[2]

Research Scholar, Department of Computer Science, JJ College of Arts and Science (Autonomous), Pudukkottai[1]

Assistant Professor, Department of Computer Science & Engineering, Alagappa University, Karaikudi[2]

Abstract: Data mining have great potential for healthcare industry to enable health systems to systematically use data and identify the efficiency and improve care with reduce cost. Due to complexity of healthcare and slower improvement of technology adoption needs to implement effective data mining. In data mining there are variety of techniques and methods to suggest or decision making knowledge in database and extracting or support prediction of data. With growing research on heart disease prediction system, this research work has developed a Heart disease prediction with comparative study of several data mining algorithms. Many data mining prediction tools are there in worldwide to make prediction for medical data. In this study it is observed that compare several algorithms to analysis the performance, accuracy of the heart disease prediction system. The commonly used techniques of heart disease prediction and their analytics, summaries are summarized in this research work.

Keywords: Data mining, Heart disease, J48, Naïve Bayes, Neural Network, Prediction.

I. INTRODUCTION

Heart disease remains the death of the primary reasons for the past two decades. In recent years, computer technology and machine learning techniques to develop software to help physicians make decisions in the first stage of heart disease. Diagnosis of heart disease mainly relies on clinical and pathological data. System can help predict heart disease medical professionals predict data based on the patient's clinical status of heart disease. In the field of biomedical data mining played an important role in disease diagnosis in predictive medicine, the information provided by the patient signs and symptoms may include redundancy and interrelated, especially when patients from more than one type of disease suffering from the same category. The doctor could not be able to diagnose correctly.

Intelligent data mining algorithms can be used to solve the problem tackle group has many entries health information to predict. Now, artificial neural networks have been used during the day to work hard and difficult. Neural networks are Training and hope to find the missing historical data dependencies, and can be used as Prediction. Feed forward neural network BP training to become the standard classification and forecasting activities.

DATA MINING

Although data mining has been more than two decades, its potential is only realized now. Combined with statistical analysis of data mining, machine learning and database technology to extract hidden patterns and relationships from large databases. Fayyad Data mining is defined as "hidden from the data stored in the database, the previous non-trivial extraction unknown and potentially useful information about the process." Zhu Brindisi defines it as "the process of selecting, exploring and modeling large amount of data found the product to obtain the relationship between the law of the owner of the database clear and

useful results of the first purpose unknown." Data mining uses two strategy: monitoring and unsupervised learning. In supervised learning, the training set is used to study the model parameters, whereas in unsupervised learning not to use the training set (for example, K- means clustering is unsupervised).

Each technology providers mining depends on the purpose of modeling different purposes. The two most common Classification and prediction models destination. Classification model to predict the classification and labeling (discrete, Unordered), but the forecast models predict a continuous value function. Application of decision tree and neural network, the regression association rules classification and clustering algorithm using prediction algorithms.

Naive Bayes or Bayes rule is a method of machine learning and data mining much foundation rules (algorithm) Used to create a predictive power of the model. It provides a new way to explore and understand data. It learn by calculating the relationship between the target (if relevant) from between the "test" to learn another (i.e. Argument) [1].

HEART DISEASE

Coronary heart disease is narrowing of small blood vessels that supply blood and oxygen to the heart. This is also known as coronary artery disease. Coronary heart disease is caused by atherosclerosis called state, when the fatty materials and substances called plaque in the artery walls caused by the accumulation occurs. This makes them shrink. Due to coronary artery stenosis, blood flow to the heart can slow down or stop, causing chest pain, breathing, heart attack symptoms such as shortness. Men under 40 years of coronary heart disease than women of higher risk, but as a woman ages, her risk increases, so that it is almost equal to a person's risk. The main risk factors for coronary

heart disease are: 1) Diabetes 2) hypertension 3) low-density lipoprotein (bad cholesterol) 4) reduce low-density lipoprotein (good cholesterol) 5) cholesterol do not get enough physical activity 6) Obesity 7), smoking.

India is undergoing rapid epidemiological transition, the result of economic and social change; cardiovascular disease has become an increasingly important cause of death. The pattern of the disease in India is in the past decade significant changes. According to the World Health Organization report, 10 people were killed in India at present, eight by non-communicable diseases, such as cardiovascular disease, diabetes, and rural Indian cities india. In caused by non-infectious causes six deaths is how 10 disease. Registrar General of India data showed that a heart attack death in 30% of the rural population mainly in Andhra Pradesh india. in is due to heart disease and death [2].

II. LITERATURE SURVEY

As data mining techniques such as decision trees to help build predictive models of intelligent systems for heart disease, naive Bayes and neural networks proposed Palaniappan Nga wang and method used CRISP-DM Set up a data mining model in a group of heart disease database obtained Cleveland 3. The results showed that the external force of each method to achieve the goal of targeting mining data. Decided to predict heart disease intelligence system able to answer the question, the traditional support systems failed. Created important knowledge, for example, provides a pattern, Between medical factors related to the relationship between heart disease [3].

Carlos implementation of efficient diagnosis and heart disease compared to the search tree, association rules [4]. A new technology to develop linear and nonlinear characteristics VFC function of multiple parameters Kui Hean presented by Lee et al. [5] heart disease. Help build an intelligent system model prediction data under Mining techniques such as decision trees, naive Bayes and neural networks proposed Palaniappan et sellappan [6]. Constrained association rules to determine the prediction of heart disease problems have been studied Carlos Ordonez [7] .MA.jabbar, Priti Chandra, B.L.Deekshatulu proposed evolutionary algorithm heart disease Prediction. Enhanced human heart disease It was suggested that the use of prediction by M.Ambarasi et al [8] and. Intelligent Forecasting System of heart attack effective use of data mining and AINN proposed by the genetic algorithm child heart disease [9] of the set of options. They are perceived as having a back-propagation algorithm for training multilayer neural networks. The method according to the chart to predict heart disease has been suggested MA.jabbar, B.L.Deekshatulu and Priti Chandra [10]. The method is based on the largest group and Mining of weighted association rules. Predict heart disease associated classification proposed MA.jabbar, B.L.Deekshatulu and Priti Chandra [11]. They use classification based on the Gini coefficient prediction of heart disease. Mining proposed cluster-based prediction of heart attack by the

association rules MA.jabbar, B.L.Deekshatulu and Priti Chandra [12]. The method is a sequence of numbers and clustering. The entire database is divided into equal-sized partitions and association rules drawn from each partition.

III. METHODOLOGY USED

Naïve Bayes Algorithm [13]

Bayesian classifier is based on Bayes' theorem. The classification algorithm conditional independence, this means that, assuming that a given attribute value

Bayesian classifier is based on Bayes' theorem. The classification algorithm conditional independence, this means that, assuming that a given attribute value class is independent of the value of other properties in. The Bayes Theorem is follows:

Let $X=\{x1,x2,\ldots.,xn\}$ 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.

To determine P(H|X), the probability that the hypothesis H holds given evidence

According to ayes theorem the P (H|X) is expressed as

$$P (H|X) = P (X|H) P (H) / P(X)$$

J48 Algorithm [14]

J48 is a data mining tool Weka open source Java implementation C4.5 algorithm. C4.5 decision tree is created based on a set of data items marked program. The algorithm developed by Ross Quinlan. Generated by the C4.5 decision tree can be used for classification, and for this reason, C4.5 is often called a statistical classifier ("C4.5 (J48)", Wikipedia).

Algorithm Steps:

I. For the case, they belong to the same class Tree said sheet so that the sheet to be returned Labeled with the same class.

II. The information is calculated for each potential Property, by giving the test properties. Then, Information gain calculation will lead from the test property.

III. Next, the best properties are based on this selection criteria and selection of properties Branch.

NN (Neural Network) Algorithm [15]

Artificial Neural Network (ANN), often referred to as "neural network" (ANN) is a mathematical model or computational model based on biological neural networks, in other words, it is a simulation of biological neural systems. It is calculated by the artificial neural information interconnection and treated using a joint way. In most cases, the neural network is based on an adaptive systems in the learning stage external or internal information flowing through the network, changing its structure.

IV. EXPERIMENTAL RESULTS

In this section to analysis the result to predict heart disease dataset. The actual dataset will be gathered from UCI Machine Repository. In Heart Disease dataset contains 14 attributes and 305 instances. According to these values to

eliminate missing values and prepare the training data set for evaluating the results.

Parameters: P1 - Age P2 - Gender P3 – CP (chest pain) P4- trestbps: resting blood pressure P5 – cholesterol P6 – fbs: fasting blood sugar>120 ? yes=1,no=0 P7 – restecg: resting electrocardiographic results 0,1,2 P8 – thalach : maximum heart rate achieved P9 – exang : exercise induced angina (1= yes ; 0= no) P10 – oldpeak = ST depression induced by exercise relative to rest P11 – slope : the slope of the peak exercise ST segment P12 – ca: no. of major vessels (0 to 3) colored flurosopy P13 – thal :3 =normal ,6=fixed defect ,7= reversable defect P14 – diagnosis of heart disease.
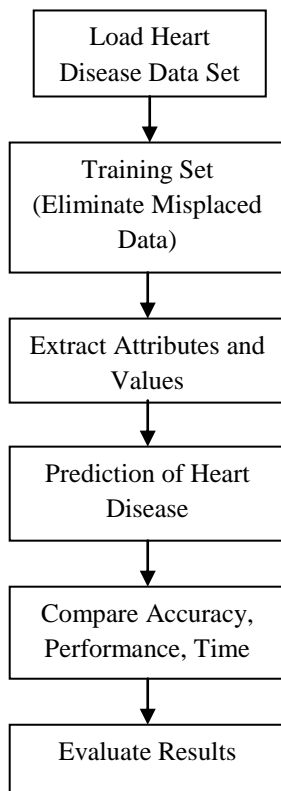
Load Heart Disease Data Set

↓

Training Set (Eliminate Misplaced Data)

↓

Extract Attributes and Values

↓

Prediction of Heart Disease

↓

Compare Accuracy, Performance, Time

↓

Evaluate Results

Figure1. Working Architecture for proposed work

PERFORMANCE EVALUATION
In this section has shown the performance analysis of the different data mining algorithms. The accuracy measure using formula is:

$$TA = \frac{(TP + TN)}{TP + TN + FP + FN}$$

$$RA = \frac{(TP + FP) * (TN + FN) + (FN + TP) * (FP + TP)}{(Total * Total)}$$

TA → Total Accuracy   RA →Random Accuracy TP → True Positive TN → True Negative FP → False Positive FN → False Negative
Table 1 shows the results of the algorithm quite remarkable compared to other algorithms in Data Mining. The comparing with other two algorithms NN will performs better of 82% of accuracy.

Table1. Performance comparison for different algorithms

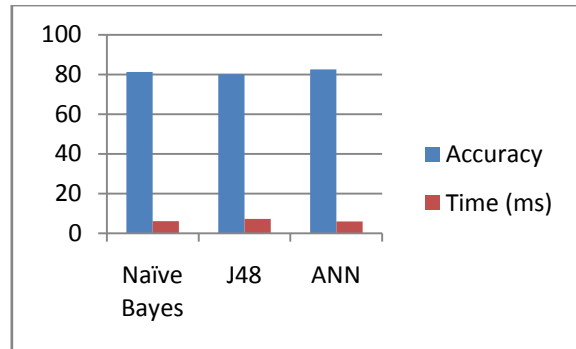|	Accuracy	Time (ms)
Naïve Bayes	81.3021	609
J48	80.099	719
ANN	82.5677	589

Figure 2 Performance Diagram of different algorithms

Table 2 shows the error rate of the algorithm compared with the all other algorithms in data mining. It clearly shown the NN will be better than the Naïve Bayes and J48 algorithm clearly.

Table2. Error Rate Analysis for different algorithms

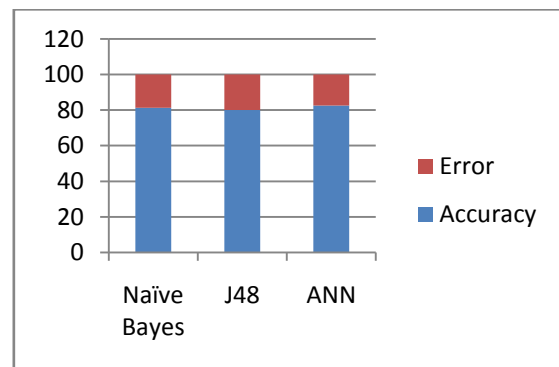|	Accuracy	Error
Naïve Bayes	81.3021	18.6979
J48	80.099	19.901
ANN	82.5677	17.4323

Figure3. Diagram for Error Rate Analysis for different algorithms

V. CONCLUSION

This study has presents a new approach from medical predict heart disease Patients. UCI Heart disease data set has been used purpose of the experiment. It has come to the information of patients and non-heart disease patients. Neural Network algorithm was used to improve accuracy data mining process. The Weka data mining tools, It has been used as a modification to produce API in MATLAB. The results indicate a significant in the proposed algorithm improvements. The results show that the algorithm can achieve accuracy up to 82.57%.

REFERNCES

[1] 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

[2] The Times of India.14th august 2011

[3] Palaniappan, S., Awang, R., Intelligent Heart Disease Prediction System Using Data Mining Techniques, IJCSNS International Journal of Computer Science and Network Security. 8(8): 343-350 (2008).

[4] Carlos Ordonez.: Comparing association rules and decision trees for heart disease prediction, ACM, HICOM (2006)

[5] Hean Gyu Lee et al.:Mining bio signal data :CAD Diagnosis using linear and non linear features of ARV,LNA I481 9pp 56-66(2007)

[6] Sellappan Palaniappan et al.: Intelligent heart disease prediction on system using data mining techniques.IJCSNS Vol 8 no 8 (Aug 2008)

[7] Carlos Ordonez.: Improving Heart Disease Prediction using constrained association Rule, seminar presentation at TOKYO (2004).

[8] M.Ambarasi etc al.:, Enhanced Prediction of Heart Disease with Feature subset selection using Genetic Algorithm, IJESI, Vol 2(10) (2010)

[9] Shantakumar B patil etc all.: Intelligent and effective heart attack prediction system using data mining and artificialneura lnetwork, Europian journal of scientific research vol 31,No 4 (2009)

[10] MA.Jabbar, B.L.Deekshatulu and Priti Chandra.: Graph based approach for heart disease prediction. In: proceedings of ITC 2012, Bangalore, Springer-Verlag (2012)

[11] MA.Jabbar, B.L.Deekshatulu and Priti Chandra.: Knowledge Discovery using Associative Classification for Heart Disease Prediction. In: International symposium on Intelligent Informatics (ISI 2012) (Springer)

[12] MA.Jabbar,PritiChandra, .L.Deekshatulu.:Cluster based association rule mining for heart attack prediction,JATIT,vol 32,no 2,(Oct 2011)

[13] K.Sudhakar, Dr. M. Manimekalai ,"Study of Heart Disease Prediction using Data Mining" , International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 1, January 2014

[14] Korting, Thales Sehn. "C4. 5 algorithm and Multivariate Decision Trees." Image Processing Division, National Institute for Space Research—INPE, 11 Nov 2011.

[15] Overview of Bayesian network approaches to model gene environment interactions and cancer susceptibility Chengwei Su, Angeline Andrew, Margaret Karagas, Mark E. Borsuk, mar 21, 2013.