ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 20, April 2016

A Novel Method For Prediction Of Heart Disease Using Naïve Bayes

K. Aravinthan[#1] and Dr. M.Vanitha[*2]
# Research Scholar, PG and Research Department of Computer Science,
J.J.College of Arts and Science (Autonomous), Pudukkottai, Tamilnadu, India.
[1]aravinthk83@gmail.com
*Assistant professor, Department of Computer Science and Engineering,
Alagappa University, Karaikudi, India.
[2]mvanitharavi@gmail.com

***Abstract*— Data mining is by far not yet knows nontrivial extraction of implicit data; the information is assumed useful data. Data mining is a necessary process in which intelligent methods to extract the data pattern applied. Using data mining can be assessed, we can make informed decisions about future use and can render a better way to extract knowledge model. Data mining refers to various techniques used for identifying information or knowledge in a database and extracts them so that they can put such decisions, forecasting, and the predicted values calculated by the regional decisions. Health agency medical data collected huge amounts of money, unfortunately, no "pull" to discover hidden information, effective decision; it found that the connection between the parameters in the database is the subject of data mining. This research work has developed a decision support system using data mining that is naive Bayes prediction of heart disease modeling techniques. In the medical use of profiles, such as age, sex, blood pressure, blood glucose, ECG chest pain and other graphics can predict the likelihood of heart disease patients. It is implemented as needed in MATLAB medical testing parameters as input application. It can be used as a training tool to train nurses and doctors diagnose patients with heart disease**.

***Keywords*— Data mining, Decision support, heart disease, naïve bayes, prediction.**

I. INTRODUCTION

Data mining has been established as a new field of hidden patterns in large data sets to explore. Medicine is the use of different clinical reports, and generate another area of the patient symptoms such large amounts of data. Data mining can also be largely used in the same purpose sets of medical data as well. Explore these hidden patterns in medical data sets can be used for clinical diagnosis. However, medical data collection dispersed, heterogeneous, and nature is huge. These data sets should be organized with the hospital management and system integration.

Health agency medical data collected huge amounts of money, unfortunately, no "pull" to discover hidden and effective decision-making information. Discover hidden patterns and relationships often go mining. Advanced data mining technology can help improve this situation. This study developed a system using data mining techniques, namely, decision tree, naive Bayes and neural network forecasting prototype of heart disease. The results showed that each technology in mining goals established goals for its unique capabilities. Heart Disease Prediction System (HDPS) can answer complex queries, if the system cannot support traditional decision-making. In the medical use of profiles, such as age, sex, blood pressure, blood sugar, you can predict the likelihood of heart disease patients. Significant knowledge is possible, for example, relationship patterns, medical factors associated with heart disease between to build. HDPS is web-based, easy to use, scalable, reliable and scalable.

Clinical decision is based on data in the database of knowledge hidden wealth intuition and experience, doctors do not often performed. Quality of service over this practice can lead to undesirable deviations, errors and influence to provide patients with health care costs. Based computers can reduce medical errors, improve patient safety and reduce unnecessary changes in practice and improve patient outcomes patient record system proposed integrated clinical decision support. This proposal is promising as data modeling and analysis tools, such as data mining, it is possible to produce a wealth of knowledge that can significantly contribute to the quality of clinical decision-making environment in the possibility of improvement.

II. RELATED WORKS

Deepika N. et. Association rules proposed for patients with heart disease [1] attack classification. Extract showed significant patterns from the data warehouse heart disease. The clinical data repository contains data for heart disease screening a patient's heart. Initially, the data storage pretreatment efficiency of the extraction process. The first stage of association rules to deal with missing values using pretreatment. Later, based on the views of data Pima Indians of heart attack medical specialists approximation also applies to group interval. The project is significant computing with the help of this method are under frequent pattern. With greater than a predetermined threshold of confidence frequent mode is selected and heart attack prediction system design and development. Pima whole, from the use of UCI machine learning library for a heart attack India data. Patients with chest pain and age and number of years of characteristics are recorded. Included in the pre-processing operation data set to eliminate duplicate records, indicating in the database, this data point represents the missing information, and delete the data field values normalized necessary. It may also be necessary to reduce the number of data sets, in addition, to minimize the memory resources, and data mining algorithms needed to process the data combined. In the real world, the data is incomplete, and in the case of medical data is always true. Eliminating the pre-treatment times inconsistent with our data associated with the data.

K. Srinivas et al. Application of data mining technology in the health sector and proposed to predict heart attack [2]. Possible use of data mining technology on the basis of classification, as a rule bulky health data base, decision trees, naive Bayes and neural networks. For exploratory data analysis, machine learning and statistical learning algorithm predicts using data mining tools. Training data set includes the case of 3000, with 14 different attributes. In the instance of the data set shows the results of different types of tests to predict the accuracy of heart disease. The classifier performance was evaluated, and the results were analyzed. Result of the comparison is based on the October 10-fold cross-validation. According to the property of the data that is divided into two portions of data are used for training are 70% and 30% were used for testing. The comparison is carried out between the classification algorithm, which, naive Bayes algorithm as the best performance of the algorithm [2].

Sudha et al. [3] proposed classification algorithm, Naive Bayes, decision trees and neural networks predict disease and stroke. Decision tree classification algorithm, Bayesian classifier and BP neural network was adopted in this study. And non-relevant data record is deleted before it happens mining process from data storage. Classification is the classification and evaluation of model data mining models. Training data classification model used to build predictive classification model settings. Test data sets are used to test the efficiency of classification. Then the ranking algorithm as a decision tree, naive Bayesian neural network is used to predict motion sickness. Performance evaluation conducted on the basis of three algorithms and a variety of models and used, and measurement accuracy compared. The classification algorithm to compare observed performance of the neural network is shown in more than the other two algorithms.

Mai Shouman et al. k-means and decision tree method [4] proposed to predict heart disease cluster. In his works they recommend choosing centroid KMEANS group, in order to improve efficiency in several ways. Enter the attribute data obtained from the establishment of 13 Cleveland Clinic Heart Foundation. The sensitivity, specificity and accuracy of different initial centroid calculation and selection methods and different number of groups. And stochastic properties of random line method, ten races run, and calculate the average value and the best of each method. The Group and the integration of comparative KMEANS tree previously, the integration of k-means clustering and decision tree by tradition the same set of data applications can improve the accuracy of decision tree in the diagnosis of heart disease patients. In addition, the integration of the k-means clustering and decision tree can be more accurate than the search algorithm of patients diagnosed with heart disease. Achieve an accuracy of 83.9%, and two clusters of activation process.

Olatubosun Olabode et al [5] back cerebrovascular disease classification using artificial neural networks and error propagation method. With back-propagation method of artificial neural network multilayer perceptron is a feedforward network, between the input and output nodes with one or more layers. These additional layers comprises not directly connected to the input node and the node or implied outlet cells. The BP neural network algorithm uses a hidden input with S-shaped function trained attribute 16. Using prediction models in a variety of diagnostic-level domain. Data collection for this work 100 records (60 males and 40 females) were collected federal medical and other fields. Enter the value obtained from the recording layer is formed in the input node 16 input variables. Right neural network weights are randomly initialized. The weight of this work range between [-0.5 0.5] and the learning rate is set at between 0.1 and 0.9. Training, validation, summarized accurately measured.

III. RESEARCH OBJECTIVES

Diagnosis of the disease is a significant medical and tedious task. Detecting cardiac disease or symptoms of a variety of factors that are a plurality of layers, which is not free from the false assumption that is often accompanied by unpredictable impacts. Therefore, to collect clinical screening of patients with knowledge and experience of many experts and data from the database, in favor of the efforts of the diagnostic process is considered a valuable option. At low prices provide a valuable service is a major constraint medical institutions (hospitals, medical centers) encountered. Quality service valuable epresented accurately diagnose patients and provide effective treatment. Poor clinical decisions may lead to disaster, so little entertainment. It is also important to reduce hospital costs of clinical trials. Computer-based information and / or to determine the appropriate support system can achieve lower cost of clinical trials to help. Naive Bayes rule or Bayesian machine learning and data mining methods much foundation. Rules (algorithm) is used to create models and predictive ability. It provides a new way to explore and understand data. You can learn by calculating the target (if relevant) and other variables (ie, independent) from the correlation between the "test."

IV. SCOPE OF THE RESEARCH

Here, the scope of the project is a computer-based clinical decision support and can reduce medical errors, improve patient safety and reduce unnecessary changes in practice, and improve the prognosis of the patient's medical history to integrate patients [7]. This proposal is promising as data modeling and analysis tools, such as data mining, you must generate a rich environment or knowledge, can help to significantly improve the quality of clinical decision-making potential .The main objective of this study was to develop a prototype of heart disease forecasting system using three data mining, namely, decision tree, naive Bayes and neural networks HDFS modeling technique [6]. By providing effective treatment, but also help to reduce the cost of treatment and improve the visualization and ease of explanation, a huge knowledge and accurate data in the field. Big companies invest heavily in this type of activity in order to focus on the risks involved and possible events. This work brought on as a reasonable expectation for the future development of the basis for all previous and current data together.

V. DATASOURCE

The accuracy and precision of the test data with actual classification comparison made by the percentage of correct prediction model.

Total accuracy = Correct prediction recording / Total no. training record

In this experiment, it is considered a medical data related to heart disease. The data set obtained from Cleveland database. This is a set of data publicly available on the Internet. Cleveland data set is normal and abnormal human person with respect to heart disease classification. Data Representation: Case: 414 properties: 13 and class attributes.

A. Class

Class0: Normal Person.   Class1: first stroke
Class2: second stroke   Class3: end of life Attribute

TABLE 1

ATTRIBUTES DESCRIPTION FOR CLEVELAND HEART DISEASE DATASET

Attribute Name	Type	Description
Age	Continuous	Age in Years
Gender	Discrete	1= male 0 = female
Cp	Discrete	Chest Pain Type: 1= typical angina 2 = atypical angina 3 = non-angina 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
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

B. Predictable attribute

To diagnosis heart disease if value 0: < 50% diameter narrowing (no heart disease); value 1: >50% diameter narrowing (Heart disease predict).

VI . IMPLEMENTATION OF NAÏVE BAYES

Dimension naive Bayes classifier technology is mainly used when the input is high. Despite its simplicity, Naive Bayes can often outperform more sophisticated classification methods. Naive Bayes model feature recognition cardiac disease. The probability of each input attribute in the foreseeable state of the display.

A. Naïve Bayes Algorithm

One of the Bayesian approach is naive Bayes classifier, which uses probability formula:

$$P(A/B) = \frac{P(B/A) \times P(A)}{P(B)}$$

Where in, A and B are two events (for example, the train arrival time, because the probability of rainy weather). Naive Bayes classifier using this to find an invisible instance of the most likely classification probability theory (unrated). The algorithm and classification of active data, but worse of digital data in the tra ining set.

VII. EXPERIMENTAL RESULTS

This dataset contains information on the diagnosis of heart disease. These data are collected from the Cleveland Clinic Foundation, and in UCI library. Including six cases of missing values have been deleted from the original database. This dataset have 13 attributes of data and 303 of instances for observations.

A. Classification Measures

Negatives are correctly classified by a marked negative tuple. False positives (F_Pos) is incorrect classification label, false negative, they are negative tuple classification marked by false positive tuples. Recognition rate sensitivity means that the positive rate or meet the sensitivity and specificity of the approach can be used to mark the percentage of "Yes" in the compute performance and accuracy of the sample is used. The measures are defined as,

$$\text{Sensitivity} = \frac{\text{True\_Pos}}{\text{Pos}}$$

Specificity refers to the true negative rate. True_Pos is (that is, "the present" samples were correctly classified) by the number of true positives and POS is positive samples.

$$\text{Specificity} = \frac{\text{True\_Neg}}{\text{Neg}}$$

True negative is the number of true negatives (sample or "absent" that were correctly classified) and Neg is the number of negative samples and F_Pos is false (sample "lost" is incorrectly marked as "number is").

$$\text{Precision} = \frac{\text{True\_Pos}}{\text{True\_Pos} + \text{F\_Pos}}$$

$$\text{Accuracy} = \text{Sensitivity}\frac{\text{Pos}}{\text{Pos} + \text{Neg}} + \text{Specificity}\frac{\text{Neg}}{\text{Pos} + \text{Neg}}$$

True positive, true negative, false positive and false negative also assess the costs and benefits (or risks and benefits) classification model is useful.

TABLE 2
PERFORMANCE MEASURES FOR HEART DISEASE PREDICTION

Data Set	Accuracy	Error Rate
n=1	86.6667	13.3333
n=2	85.1852	14.8148
n=3	77.7778	22.2222
n=4	81.4815	18.5185
n=5	85.1852	14.8148
n=6	85.1852	14.8148
n=7	84.0741	15.9259
n=8	77.7778	22.2222
n=9	88.8889	11.1111
n=10	84.0741	15.9259
Accuracy Rate	**83.6296**	**16.3703**

Figure 1 specifies the accuracy measure for heart disease prediction using naïve bayes classifier. This performance will calculate using T_Pos, F_Pos, T_Neg, F_Neg of the system.

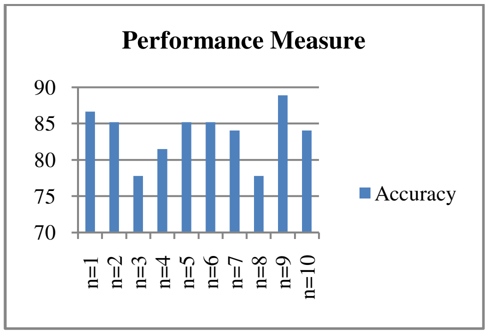

Fig 1. Performance measure for heart disease prediction

Figure 2 specifies the error rate of the proposed work. This error rate will measure based on accuracy of the proposed work

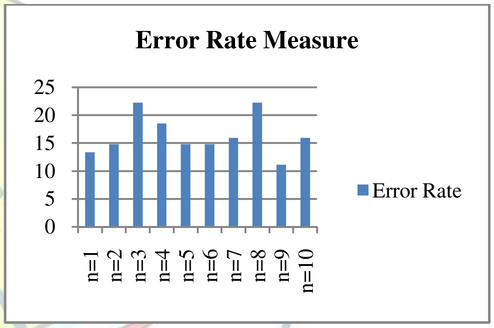

Fig. 2 Error rate measure for heart disease prediction

VIII. CONCLUSION & FUTURE WORK

The proposed works helps to predict heart disease on the system decided to use these developed soften Naive Bayes classifier. The system extracts the implicit knowledge of heart disease history database. Bayesian forecasting softening technology than patients with heart disease more effectively. The model is able to respond to complex queries, each with its own strengths relaxed interpretation model, and easy access to detailed information and accuracy. The system is a greater number of records or property in this sense can be incorporated and can generate significant new use rule based scalable data mining technology. The current health care system has been used in the diagnosis of 13 properties. You can also combine other data mining techniques to predict and additional attributes. The results showed that the Bayesian achieve the highest accuracy, sensitivity, specificity and precision performance measures. Naive Bayesian classification algorithm prediction model obtained in this study the best use of heart disease. Therefore, it is the development and implementation of testing in the clinical decision support system to help doctors and health professionals the right person in the diagnosis of heart disease.

In the future, it is being investigated in combination with each algorithm to another, and get accurate algorithm for combining the accuracy of the measurement data set of heart disease.

REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] Intelligent Heart Disease Prediction System Using Data Mining Techniques-Sellappan Palaniappan, Rafiah Awang 978-1-4244-1968-5/08/ ©2008 IEEE.

[7] Obenshain, M.K: “Application of Data Mining Techniques to Healthcare Data”, Infection Control and Hospital Epidemiology, 25(8), 690–695, 2004.

[8] Shantakumar B.Patil, Y.S.Kumaraswamy, Intelligent and Effective Heart Attack Prediction System Using Data Mining and Artificial Neural Network, European Journal of Scientific Research ISSN 1450-216X Vol.31 No.4 (2009), pp.642-656 © EuroJournals Publishing, Inc. 2009

[9] J. Sandhya et al., "Classification of Neurodegenerative Disorders Based on Major Risk Factors Employing Machine Learning Techniques," International Journal of Engineering and Technology, vol.2, no. 4, 2010.

[10] Sellappan Palaniappan, Rafiah Awang, Intelligent Heart Disease Prediction System Using Data Mining Techniques, 978-1-4244-1968-5/08/$25.00 ©2008 IEEE.

[11] Han, J., Kamber, M.: “Data Mining Concepts and Techniques”, Morgan Kaufmann Publishers, 2006.

[12] Tang, Z. H., MacLennan, J.: “Data Mining with SQL Server 2005”, Indianapolis: Wiley, 2005.