ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 11, November 2014

Predicting the Risk of Heart Attacks using Neural Network and Decision Tree

S.Florence[1], N.G.Bhuvaneswari Amma[2], G.Annapoorani[3], K.Malathi[4]

PG Scholar, Indian Institute of Information Technology, Srirangam, Tiruchirappalli, India[1]

Faculty, Indian Institute of Information Technology, Srirangam, Tiruchirappalli, India[2,4]

Assistant Professor, University College of Engineering, BIT Campus, Tiruchirappalli, India[3]

ABSTRACT: The healthcare environment is more and more data enriched, but the amount of knowledge getting from those data is very less, because lack of data analysis tools. We need to get the hidden relationships from the data. In the healthcare system to predict the heart attack perfectly, there are some techniques which are already in use. There is some lack of accuracy in the available techniques like Naïve Bayes. Here, this paper proposes the system which uses neural network and Decision tree (ID3) to predict the heart attacks. Here the dataset with 6 attributes is used to diagnose the heart attacks. The dataset used is acath heart attack dataset provided by UCI machine learning repository. The results of the prediction give more accurate output than the other techniques.

KEYWORDS: Naïve Bayes, Neural Network, Decision tree, ID3, Heart attack

I. INTRODUCTION

The heart attack is a common problem in all human beings with the age above 30. The cholesterol level is another one major problem which leads to heart attack. The knowledge discovery in databases is a well-defined process containing several distinct steps to get the perfect accuracy. Data mining is the core step, which results in the discovery of hidden information with useful knowledge. The discovered knowledge will be used by the healthcare administrators to predict some of the diseases and problems like heart attacks. Predicting patient's behaviour in the future is the main application of data mining techniques. A formal definition of knowledge discovery in databases is given as follows: "Data mining is the non-trivial extraction of implicit previously unknown and potentially useful information about data" [1].

Medical diagnosis is an important yet complicated task that needs to be done accurately and efficiently. The automation of this system is very much needed to help the physicians to do better diagnosis and treatment [2]. Poor clinical decisions can lead to disastrous consequences which are therefore unacceptable. The major challenge of the healthcare system nowadays is to predict the diseases in quality manner.

The clinical decisions made by doctors may be error prone and leads to some problems for patients. This system should be automated in the manner that to predict the diseases in the accurate one. One of the main problems in healthcare system is to predict the heart attacks of the patients previously. There some techniques available to predict these things in an accurate manner. The available data mining techniques are not used properly to predict the diseases in the healthcare systems [10][11].

II. RELATED WORK

There are many approaches and algorithms have been used to predict the heart attacks. Hai et al. in their work proposed neural based learning classifier system for classifying data mining tasks. They conducted experiments on 13 different datasets from the University of California, Irvine repository and one artificial dataset. They showed that neural based learning classifier system performs equivalently to supervised learning classifier system on five datasets, significantly good performance on six datasets and significantly poor performance on three datasets [17].

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 11, November 2014

Shantakumar and Kumaraswamy, in their work proposed an intelligent and effective heart attack prediction system using data mining and artificial neural network. They also proposed extracting significant patterns for heart disease prediction. They used K-means clustering to extract the data appropriate to heart attack from the warehouse.They used MAFIA algorithm to mine the frequent patterns[3][4]. Niti et al. in their work proposed a decision support system for heart disease diagnosis using neural network. They trained their system with 78 patient records and the errors made by humans are avoided in this system [18].

Anbarasi et al. in their work proposed an enhanced prediction of heart disease with feature subset selection using genetic algorithm. They predicted more accurately the presence of heart disease with reduced number of attributes. They used Naïve Bayes, Clustering, and Decision Tree methods to predict the diagnosis of patients with the same accuracy as obtained before the reduction of attributes. They concluded that the decision tree method outperforms the other two methods [15].

Comparing to the works discussed above, this work is different with the use of neural network and decision tree algorithm in an integrative manner to predict heart attacks with high amount of accuracy [9].

III. PROPOSED ALGORITHM

There are various data mining techniques available with their suitability dependent on the domain application. Statistics provide a strong fundamental background for quantification and evaluation of results. However, algorithms based on statistics need to be modified and scaled before they are applied to data mining. We now describe a few Classification data mining techniques with illustrations of their applications to healthcare.

Decision tree include CART (Classification and Regression Tree), ID3 (Iterative Dichotomized 3) and C4.5. These algorithms differ in selection of splits, when to stop a node from splitting, and assignment of class to a non-split node [5]. CART uses Gini index to measure the impurity of a partition or set of training tuples [6]. It can handle high dimensional categorical data. Decision Trees can also handle continuous data (as in regression) but they must be converted to categorical data [16].Neural network will be trained using the training datasets and then the results are tested using the test data. It will show the high level of accuracy because of the number of neurons implemented [15].
The heart attack dataset provided by the university of California machine learning repository [13]. The dataset contains 6 attributes like age, sex, cardiac duration, signal, possibility of attack. The final one is the class label. Depends upon the attribute values present in the dataset the corresponding class label that is the prediction is happening at the final stage. The heart attack dataset which is available in UCI repository used to for the experimental purpose [12]. Preprocessing is an important step in the knowledge discovery process, as real world data tend to be incomplete, noisy, and inconsistent. Data preprocessing includes data cleaning, data integration, data transformation, and data reduction. Data cleaning routines attempt to fill in missing values, smooth out noise while identifying outliers and correct inconsistencies in the data. In this proposed work, the most probable value is used to fill in the missing values. Data transformation routines convert the data into appropriate forms for mining.

Normalization is useful for classification purpose. By normalizing the input values for each attribute measured in the training tuples will speed up the learning process. In this work, the normalization technique used is min-max normalization. The min-max normalization given in equation (1) discussed in [17] is defined as follows:

$$v^1 = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) \quad (1)$$

Replace Missing Values filter will scan all (or selected) nominal and numerical attributes and replace missing values with the modes and mean. Discretization filter is designed to convert numerical attributes into nominal ones; however the unsupervised version does not take class information into account when grouping instances together. There is always a risk that distinctions between the different instances in relation to the class can be wiped out when using such a filter.

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 11, November 2014

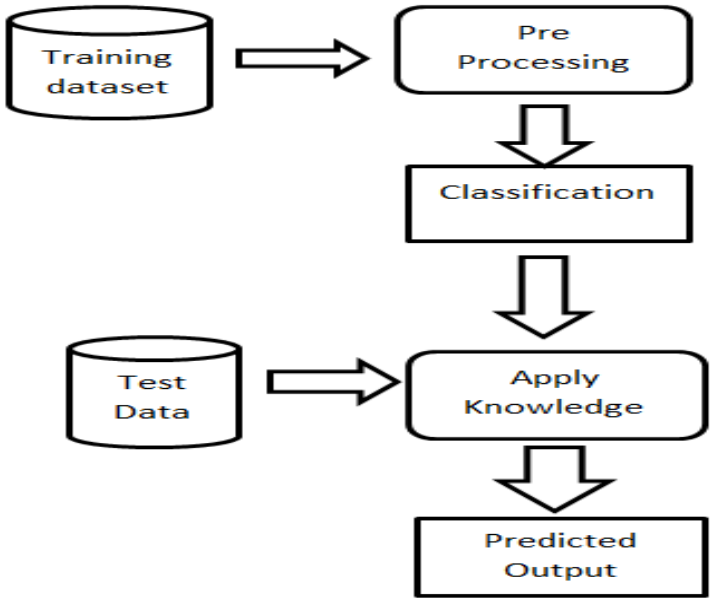

Fig 1. Block diagram of the proposed system

The block diagram of the proposed work is shown in Fig 1. The dataset used for training and testing is taken from UCI Machine Learning Repository. The data is preprocessed and given for classification. The dataset is divided into two parts. 75% of the data is used for training and 25% is used for testing the system. The knowledge obtained from the classification is used to test the system. For classification purpose, we used neural networks and decision tree. The proposed neural network architecture is shown in Fig 2.

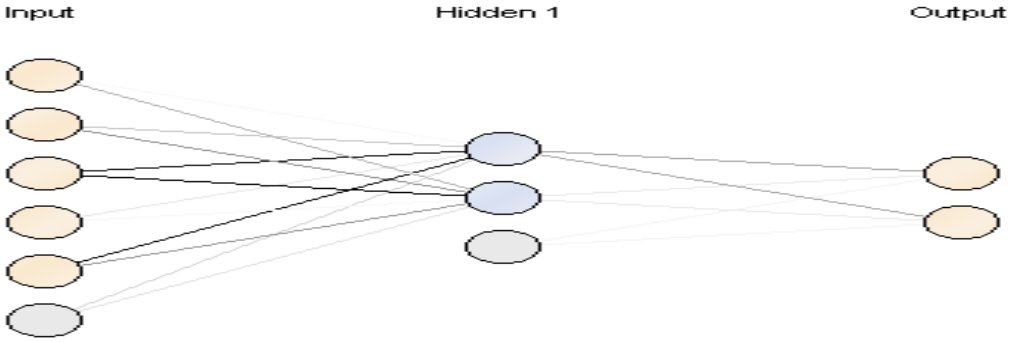

Fig 2. Proposed Neural Network Architecture

The input layer consists of 6 nodes, the hidden layer consists of 3 nodes and the output layer consists of 2 nodes. There is only one hidden layer which is processed based on the input layers. Finally it shows 2 outputs, that is the possibility of heart attacks. The prediction is done using the tool called Rapid Miner Studio. The output is received using the neurons available in the neural network. Table 1 shows the summary of heart attack dataset. The dataset consists of 6 attributes including the class label.

Table 1. Summary of heart attack dataset

Attributes	Description	Domain Values
Sex	Sex	Male(1) Female(0)
Age	Age in years	30 to 80
Car.dur	Cardiac duration	50 to 115

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 11, November 2014

Cholesterol	Level of cholesterol in blood	126 to 564 mg/dl
Sigdz	Signal level	Normal(0) Abnormal(1)
Tvdlm	Possibility of attack	Yes No

IV. SIMULATION RESULTS

The heart attack dataset taken from the UCI machine learning is used for training and testing the medical diagnosis system. The prediction is done using the tool called Rapid Miner Studio.

Neural network description

Hidden 1
========

Node 1 (Sigmoid)
----
sex: 0.325
age: -3.160
cad.dur: -13.333
choleste: -1.314
sigdz: -14.236
Bias: -2.526

Node 2 (Sigmoid)
----
sex: 4.452
age: -6.265
cad.dur: -14.734
choleste: -0.421
sigdz: -5.972
Bias: -1.693

Output
======

Class 'yes' (Sigmoid)

Node 1: -5.036
Node 2: -1.179
Threshold: 0.575

Class 'no' (Sigmoid)

Node 1: 5.036
Node 2: 1.179
Threshold: -0.575

The activation function used for training is the sigmoid function. The threshold and bias are set based on trial and error. The sample rules generated from the decision tree are as follows:

Rule 1: If age is old and sex is female and cholesterol is high then tvdlm is yes
Rule 2: If age is young and sex is female and Sigdz is normal then tvdlm is no
Rule 3: If age is old and sex is male and cholesterol high then tvdlm is yes

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 11, November 2014

The statistical analysis of heart attack dataset is shown in Table 2. This table gives the minimum value, maximum value, mean and standard deviation of all the attributes. The statistical diagram for the output generated is shown in Fig 3.

Table 2. Statistical analysis of heart attack dataset

Name	Min	Max	Mean	Deviation
Age	18	81	53.652	9.935
Sex	0	1	0.317	0.465
Car.dur	50	120	38.093	52.035
Cholesterol	126	564	117.8	82.872
Sigdz	0	1	0.683	0.465
Confidence(yes)	0.001	0.827	0.303	0.236
Confidence(no)	0.173	0.999	0.697	0.236

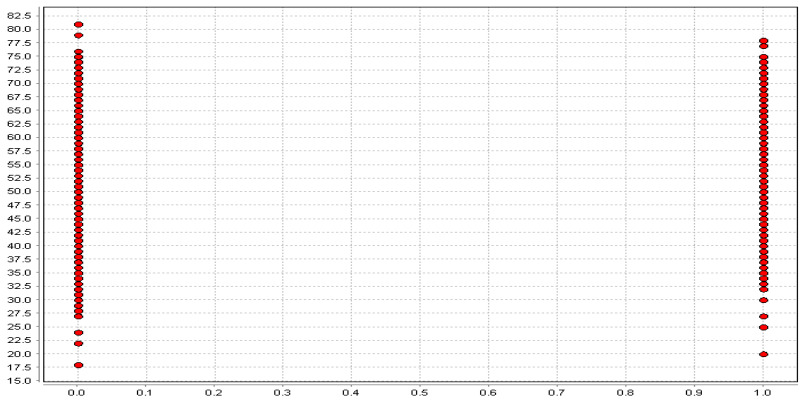

Fig 3. Statistical diagram

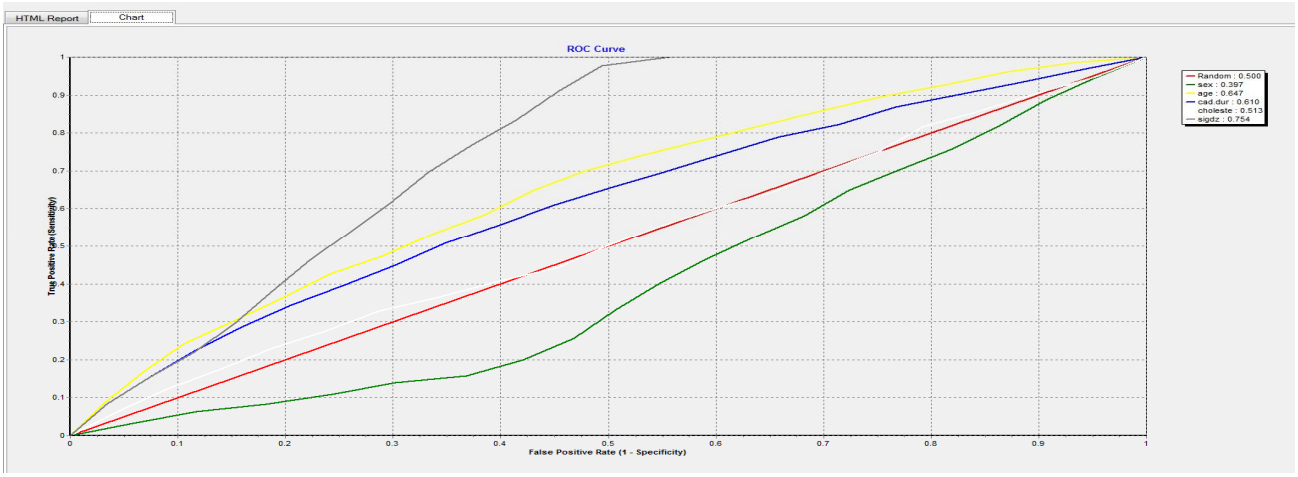

Fig 4 ROC Curve

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 11, November 2014

Fig 4. shows the receiver operating characteristic (ROC) curve of the proposed work. It is a graphical plot that illustrates the performance of a binary classifier system as its discrimination threshold is varied. The curve is created by plotting the true positive rate against the false positive rate at various threshold settings.

V. CONCLUSION

We studied the problem of constraining and summarizing different algorithms of data mining. We focused on using different algorithms for predicting combinations of several target attributes. In this paper, we have presented an intelligent and effective heart attack prediction methods using data mining. In our future work, this can further enhanced and expanded. For predicting heart attack significantly 15 attributes are listed. Besides the 15 listed in medical literature we can also incorporate other data mining techniques, e.g., Time Series, Clustering and Association Rules.

REFERENCES

[1] Frawley and Piatetsky-Shapiro, "Knowledge Discovery in Databases: An Overview". The AAAI/MIT Press, Menlo Park, C.A 1996.
[2] K.Srinivas, "Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks". (IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 02, 2010, 250-255, 2010.
[3] Shantakumar B.Patil and Y.S.Kumaraswamy, "Intelligent and Effective Heart Attack Prediction System Using Data Mining and Artificial Neural Network", European Journal of Scientific Research, Vol.31, No.4, pp.642-656, 2009.
[4] Shantakumar B.Patil and Y.S.Kumaraswamy, "Extraction of Significant Patterns from Heart Disease Warehouses for Heart Attack Prediction", International Journal of Computer Science and Network Security ,Vol.9, No.2, pp.228-235, 2009.
[5] "Hospitalization for Heart Attack, Stroke, or Congestive Heart Failure among Persons with Diabetes", New Mexico Special report: 2001 – 2003,
[6] "Heart disease" from http://wikipedia.org
[7] Rumelhart, D.E., McClelland, J.L., and the PDF Research Group , Parallel Distributed Processing, MA: MIT Press, Cambridge 1996.
[8] Heckerman, D., "A Tutorial on Learning With Bayesian Networks", Microsoft Research, 1995.
[9] Neapolitan, R., "Learning Bayesian Networks". , London: Pearson Printice Hall, 2006.
[10] Krishnapuram, B., et al., "A Bayesian approach to joint feature selection and classifier design", IEEE transactions on Pattern Analysis and Machine Intelligence, Vol.**6**, No.9, pp. 1105-1111, 2004.
[11] Sellappan Palaniappan, Rafiah Awang, "Intelligent Heart Disease Prediction System Using Data Mining Techniques", International Jornal of Computer Science and Network Security, Vol.8, No.2, pp.235-241, 2009.
[12] Pedro Domingos , Michael Pazzani , "On the Optimality of the Simple Bayesian Classifier under Zero-One Loss, Machine Learning", Kluwer Academic Publishers. Manufactured in The Netherlands, pp.103–130, 1997.
[13] "Heart Disease dataset", www.ucirepository.com
[14] Constantinos Koutsojannis et. al., "Using a Neurofuzzy Approach in Medical Application", Springer-Verlag Berlin Heidelberg, 2007.
[15] M.Anbarasi, E.Anupriya, N.CH.S.N.Iyengar, ―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] Jiawei Han and Micheline Kamber, ―Data Mining Concepts and Techniques‖, Morgan Kaufman Publishers, 2009.
[17] Hai H.Dam, Hussain A.Abbass and Xin Yao, "Neural – Based Learning Classifier Systems", IEEE Transactions on Knowledge and Data Engineering, Vol.20, No.1, pp.26-39, 2008.
[18] Niti Guru, Anil Dahiya and Navin Rajpal, " Decision Support System for Heart Disease Diagnosis using Neural Network", Delhi Business Review,Vol.8, No.1, pp.99-101,2007.