ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Special Issue 9, May 2016

Different Data Mining Approaches for Predicting Heart Disease

Ashwini Shetty A[1], Chandra Naik [2]

PG Student, Department of Computer Science, NMAMIT College, Nitte, Udupi District, Karnataka, India[1]

Assistant Professor, Department of Computer Science, NMAMIT College, Nitte, Udupi District, Karnataka, India [2]

ABSTRACT: Data mining techniques are used in most of the health care organization, for predicting different diseases. Data mining techniques helps to collect hidden information and make effective decision related to data. In this paper, we have analyzed algorithms like Neural Network and Genetic Algorithm, which plays the major role in predicting heart disease. The hybrid system uses global optimization which is advantage of Genetic Algorithm for the initializing the Neural Network weight.

KEYWORDS: Data mining, Heart disease, Neural Network, Genetic Algorithm.

I. INTRODUCTION

"Data mining is process of extracting the useful information from huge pile of data repositories". Data mining includes exploring, analysing and extracting data of different aspects and discovers hidden information from it i.e. raw data is converted into useful information. This information can be used in different fields like healthcare industry, financial estimation, weather forecasting and study of organic composite. It provides different technique to find pattern and discovers hidden information from the given data. Medical data mining requires lot of accuracy, attention and uncertainty. Healthcare administrators will make use of the values discovered by the data. Providing quality service at an affordable charge is major challenges. Quality must be maintained while providing the service i.e. analysing disease and provides effective medical care to patients. Decisions based on doctor's experience may fail in some cases. Poor interpretation can lead to some fatal effects which are unacceptable. Data mining in healthcare is an intelligent diagnostic tool.

Mr. KK Aggarwal, President of HealthCare Foundation of India said that "2.4 million Indians are dying due to the heart disease every year. The number will continue to increase due to things like stress, unhealthy eating habits, lack of physical exercise, lack of sleep and dependence on alcohol and cigarettes".

The heart is considered to be very important organ in human body; it pumps the blood through the numerous arteries and veins called the cardiovascular system. If heart doesn't function properly, it may cause an impact on other body parts like kidney, brain etc and if heart stops functioning, it will lead to death within few minutes. There are various factors which might increase the possibility of heart diseases are:

- Family history
- High blood pressure
- Lack of physical exercise
- Lack of sleep
- Hypertension
- Cholesterol

Cardiovascular disease, Arrhythmia, Congestive heart failure, Cardiomyopathy and Coronary artery disease are some form of heart diseases. "Cardio Vascular Disease" is a broad term referred to a wide variety of conditions that have an effect on the heart. Cardiovascular disease occurs mainly due to excessive work load through which stress and other problems may lead to several diseases, incompetence and death. Due to change in life styles Cardio Vascular Disease has become a leading cause of deaths globally, heart disease caused more deaths than from any other disease.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Special Issue 9, May 2016

Data mining plays key role in healthcare industry by predicting various disease and better understanding of medical related data. This study intended on developing intelligent system using different technique of data mining based on Neural Network (NN) and Genetic Algorithms (GA) for prediction (risk indicator) of heart disease.

II. RELATED WORKS

Several methods of data mining techniques have been studied and applied for predicting heart disease and have obtained different results by applying different methods.

Niti Guru et al [1] developed a system for Diagnosis of Heart Disease by using Neural Network as aid of prediction. 13 input attributes are included in each patient record of the dataset records. The records are tested and validated using back propagation algorithm. A technique is proposed by Heon gyu Lee et al.[2] to build multiparametric feature and to achieve used classifiers like support vector machine,C4.5, Naïve Bayes and Classifier based on multiple association rule.

Sellappan Palaniappan et al. [3] proposed system using some of the data mining techniques. Decision trees, Naive Bayes and Neural Networks are the techniques they used for their system. Each procedure has its own ability in achieving the suitable result. Complicated what if queries are answered by IHDPS which decision support systems cannot. Risk factors like blood pressure, age, gender, blood cholesterol level and blood sugar level etc are used to predict the possibility of patients getting heart disease. .NET is the platform used for implementation of IHDPS.

K.Srinivas et al. [4] proposed system where techniques of data mining i.e. Artificial Neural Network, Naive Bayes, and Decision tree are used. Tanagra data mining tool proposes Statistical learning, analysis of data and machine learning algorithms.Apte et al. [5] compared the performance of various classification models and showed that neural network technique is best suitable for the prediction. They used more than 13 attributes for their work. They analyzed the prediction system for cardiac diseases using more number of input attributes and to attain this system, they compared the efficiency of the different techniques.

Liao et al. [6] author report about data mining techniques and application, development through a survey of literature, form 2000 to 2011. Paper surveys three areas of data mining research: knowledge types, analysis types, and architecture types. A discussion deals with future progress in social science and Engineering methodologies implement data mining techniques and the development of applications in problem- oriented

Shouman et al. [7] proposed method to predict heart disease using k-means clustering with the decision tree and used centroid selection method to obtain better result. Integrated k-means clustering and decision tree provide better accuracy compare to paging the algorithm. Polatet et al. [8] proposed system using k-nearest neighbor approach and hybrid fuzzy for prediction of heart disease.

III.METHODOLOGY

Data Analysis

Healthcare information system is being used in almost all of the hospitals so as to manage health care data; as the system consists of enormous amount of data used to extract concealed data to building intelligent medical diagnosis. The pronominal aim is to develop prediction system which will diagnose heart disease by making use of patient's medical dataset (publicly available) [9]. To build this system 13 risk factors (attributes) are used as input i.e. age of the person, gender of the person, blood sugar level, blood cholesterol level, current blood pressure etc. shown in TABLE 1.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Special Issue 9, May 2016

TABLE 1: Description of Risk Factors

Sl.No	Attributes	Values
1	Gender	Male=0 Female=1
2	Age	>79=-2 61-78=1 51-60=0 35-50=-1 20-34=-2
3	Resting Heart Rate	In mm-Hg [low=-1,normal=0,high=1]
4	Blood Pressure	In mm-Hg [low=-1,normal=0,high=1]
5	Resting Electrocardiographic Result	Normal=0 ST-T wave abnormality=1 left ventricle hypertrophy=2
6	Exercise Induced Angina	Yes=0 No=1
7	Blood Sugar	Yes=0 No=1
8	Cholesterol	In mm-Hg [low=-1,normal=0,high=1]
9	Chest Pain Type	Typical =4 Atypical =3 Non angina=2 Asymptomatic=1
10	Number of Color Vessels By fluoroscopy	Values 0-3
11	Solpe	Upsloping =1 Flat=2 Downsloping=3
12	ST depression	Yes=0 No=1
13	Thalassemia	Normal=3 Fixed defect=6 Reversible defect=7

Analysis of data, since unnecessary noise will be present in the database, this step is considered to be critical. Data cleaning and data integration need to be performed after analysing the data. This is done to find the missing values and remove excessive data. These missing fields will lead to incorrect outputs.

Neural Network Architecture

Neural Network uses 13 input layers, 10 hidden layers and 2 output layers. Inputs depend on the values which are mentioned above in TABLE 1. Number of hidden nodes must be specified such that it will decide the best output and training. 'Configure' function initialize weights of neural network. Levenberg-Marquardt back propagation algorithm used for training and testing.trainlm is fastest back propagation algorithm.trainlm is a network training function. This function will be used to update weight and bias value. Default value must be set to 100 to maximize the number of epochs. When the errors are modified in the network weights the learning will stop and adjusted to the optimal quality at which classification is accurate. For optimizing according to the fitness function the configured weights are given to GA. This system determines whether Heart Disease is present or not for a patient.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Special Issue 9, May 2016

Parameter Setting

Optimization Toolbox used to implement this system. 'configure' function is used to NN with each weight between -2 to 2 .Fitness function that is being used in the GA is the Mean Square Error(MSE). The three parameter that are pass to GA are weights obtained, the interconnected weight and thresholds of the trained NN. Multilayered feed forward network include input layer, hidden layer and output layer takes 13, 10 and 2 neurons. The training of the weights in NN is such that each weight has values between -2 to 2.Each link are assigned the weights. GA is used adjustment of weight with population size=20. Weight and the bias values of the network in this application are being represented by the chromosomes. Based on MSE fitness function will be calculated for each chromosome. Once selection is done crossover & mutation in GA, replaces the chromosome with lower adaption with the better values , and fitter strings obtained by optimize solution which corresponds to interconnecting weights & threshold of NN are produced. The resulting lower values close to zero, represents the generalized format of the network which is ready for classification problem. Amongst several set of weight vectors simultaneous GA search takes place. Initial population is randomly generated to achieve weight efficiency and performance, suitable parameter including selection probability of crossover and mutation, initial population etc.

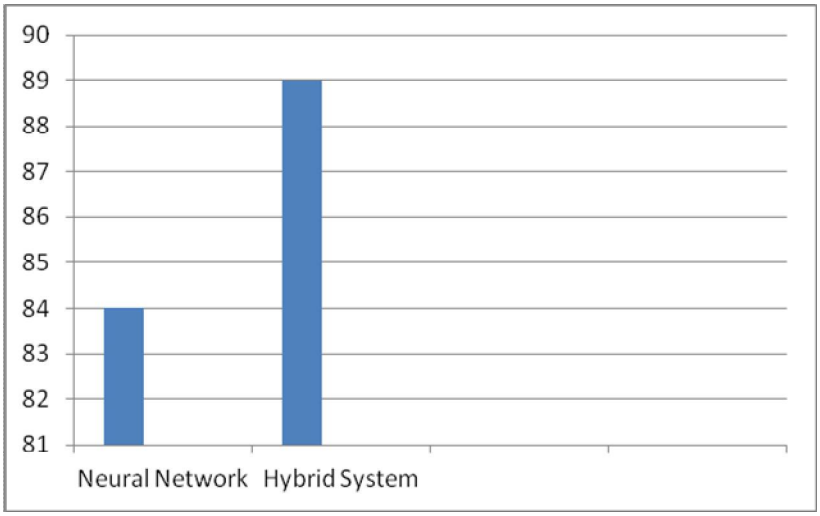
IV. RESULTS AND DISCUSSION

Weka tool is used to find the missing value in the records; it identifies and replaces by suitable value using ReplaceMissingValues filter. The system calculates the accuracy by producing the confusion matrix using MATLAB. Confusion matrix can be calculated using the formula as shown in equation (1)

$$\text{Accuracy} = (TP + TN)/(TP + FP + TN + FN) \quad (1)$$

Here TP is true positive, TN is true negative, FN is false negative, FP is false positive. The accuracy of Neural Network and Hybrid System determined is 84% and 89% .Figure 1 shows the accuracy for each method with 13 inputs is plotted on graph.

Figure 1: Accuracy for each method.



V. CONCLUSIONS AND FUTURE WORK

Algorithms are studied and analysed in order to find the most suitable algorithm to predict diseases related to heart. To enhance the performance of NN, a new hybrid model of NN and Genetic Algorithm to optimize the weight of neural network. As the part of future work, many more attributes can be added to the already existing set of attributes.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Special Issue 9, May 2016

Methods like clustering and association can also be used for the prediction. Unstructured data from the medical records can be derived by text mining method.

REFERENCES

[1] Niti Guru, Anil Dahiya, Navin Rajpal, "Decision Support System for Heart Disease Diagnosis Using Neural Network", Delhi Business Review, Volume 8, No. 1, January - June 2007.
[2] Heon Gyu Lee, Ki Yong Noh, Keun Ho Ryu, "Mining Biosignal Data: Coronary Artery Disease Diagnosis using Linear and Nonlinear Features of HRV," LNAI 4819: Emerging Technologies in Knowledge Discovery and Data Mining, pp. 56-66, May 2007.
[3] Sellappan Palaniappan and Rafiah Awang, "Intelligent heart disease prediction system using data mining techniques", in IEEE/ACS International Conference, Doha, pp. 108–115, 2008.
[4] K. Srinivas, B. K. Rani, and D. A. Govrdhan, "Application of data mining techniques in healthcare and prediction of heart attacks", International Journal on Computer Science and Engineering, vol. 2, no. 2, pp. 250–255, 2011.
[5] Shulabha S.Apte, Chaitrali S.Dangare, "Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques", International Journal of Computer Applications (0975 – 888) ,Volume 47,No.10,PP 44-48, June 2012.
[6]S. H. Liao, P. H. Chu, and P. Y. Hsiao, "Data mining techniques and applications - a decade review from 2000 to 2011," Elsevier Expert Systems with Applications, vol. 39, no. 1, pp. 11 303–311,2012.
[7]M. Shouman, T. Turner, and R. Stocker, "Integrating decision tree and k-means clustering with different initial centroid selection methods in the diagnosis of heart disease patients," in Proceedings of the International Conference on Data Mining, 2012.
[8]Polat K,S Sahan, and S Gunes, "Automatic Detection of Heart Disease Using an Artificial Immune Recognition System(AIRS) with Fuzzy Resource Allocation Mechanism and k-nn(nearest neighbor) Based Weighting Preprocessing," Expert System With Application, pp 625-631, 2007.
[9] Statlog- http://archive.ics.uci.edu/ml/machine-learning- databases/statlog/heart/