Prediction of Heart Disease Using Data Mining Technique

Sanavar Bangi[1], Pooja Gadakh[2], Pradnya Gaikwad[3], Pratiksha Rajpure[4]
[1, 2, 3, 4] Department of Information Technology

Abstract- *Prediction of heart disease is most complicated and challenging task in the field of medical science. Heart disease is the most threatening one among various diseases as it can not be detected easily. Bad clinical decisions would cause death of a patient. In our project the heart disease can be detected by using classifier. K-means clustering algorithm is used for the normalizing the data and trained label data set is detected by classifier. Prediction of heart diseases would be carried out by using multilayer perceptron neural network. The neural network in the system accepts clinical features as input and it is trained using clustering algorithm to predict that there is presence or absence of heart disease in the patient with highest accuracy of 98% comparaitive to other system. By applying data mining techniques, valuable knowledge can be extracted from the health care system. After prediction of heart disease the result of patient can be check by using Android Application.We propose efficient ANN algorithm hybrid with the back propagation technique approach for heart disease prediction. So there is necessity of creating an excellent project which will help practitioners predict the heart disease before it occurs.*

Keywords:- Heart diseases prediction, ANN Classifier, K-Means Clustering, Back Propogation, Multilayer Perceptron.

I. INTRODUCTION

Prediction of heart diseases is most complicated and challenging task in the field of the medical science. Heart is one of the most common reason of death in India or other Asian countries. In 2003 approx 17.3 million people died all over globe and out of this,10 million were only due to coronary heart diseases. Along without changing lifestyle there are many such factors such as smoking, alcohol, obesity. High blood pressure, diabetes etc. which are responsible for the risk of having heart problem.

However, with the resent studies, with the introduction of artificial intelligence and medical sciences, we can actually help in preventing any such kind of disease. For making a good decision, machine learning helps in extracting relevant data from huge database which are available in hospital.Various kind of techniques which have been applied in the prediction of a heart diseases or classification has been discussed and a proposed methodology of hybrid technique has been given which can be implemented in future to have an accuracy of almost 100% or least error.

In today's world, large number of population is suffering from different types of heart diseases and the count of patients suffering and dying from these disease is increasing day by day. So there is a need of accurate and early detection of heart disease with proper and adequate treatment which can save the life of many patients. But unfortunately, due to the complicated processes and different symptoms and pathological tests the correct diagnosis of heart diseases is a difficult task and causes delay in the proper treatment. Hence, there is a need to develop the prediction systems for heart disease which can help the medical experts in the early and accurate diagnosis of heart disease.

II. LITERATURE SURVEY

In this paper data mining technique are helpful in extracting and analyzing the complicated medical data using various kind techniques. Researchers have been applying various techniques of machine learning such as artificial neural network, BP[Back Propagation Algorithm] genetic Algorithm optimization purpose.[7]

k-means clustering algorithm used for the prediction of the heart diseases. It is a method of cluster analysis which aims to partition 'n' observations to 'k' clusters.[6]

In the system of heart diseases prediction,multilayer perceptron architecture of neural network is used.Prediction system gives the improved result with highest accuracy of 98.58% for 20 neurons in hidden layer with same Cleveland heart diseases database. Decission supports system with improved multilayer perceptron.[3]

This algorithm is generally used to train multilayer perceptron and many other neural networks. In back-propagation algorithm, the output obtained is compared with the target or expected output and the error is computed. This computed error is then again given to the neural network (fed back or back propagated) and weights are adjusted using this error so that the resulting output will get closer to the target or

expected output. This process is repeated for number of times such that at each iteration the error value gets reduced and the output gets more and more closer to the target or expected output. This process is known as ”training”.[5]

The Back propagation algorithm was used to train the ANN architecture and the same had been tested for the various categories of stroke disease. [8]

III. METHODOLOGY

In the system for heart disease prediction, multilayer perceptron architecture of neural network is used. The system consists of two steps, in the first step 13 clinical attributes are accepted as input and then the training of the network is done with training data by back propagation learning algorithm[11].

A. Multilayer Perceptron Neural Network

The multilayer perceptron neural network, as its name indicates that it is made up of multiple layers. The single layer perceptron solve only linearly separable problems but many of the complex problems are not linearly separable so to solve such problems one or more layers are added in single layer perceptron hence it is known as multilayer perceptron. Multilayer perceptron network is known as feed forward neural network having one or more hidden layers as shown in Fig.1.They are generally used for pattern recognition, classification of input patterns, prediction based on the input information and approximation.

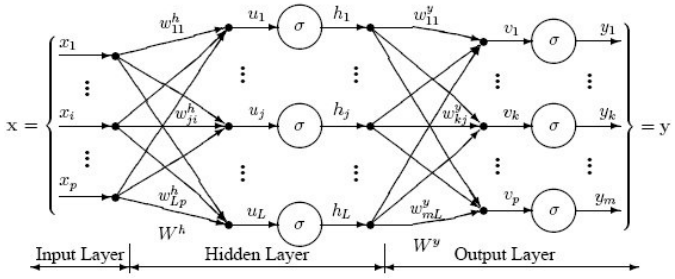


Fig. 1. Multi-layer Perceptron Neural Network Architecture

It is independent of the orientation and size of the image.Color histogram is the most popular method for extracting the color information from the image. It gives the information about distribution of different colors that are present in an image. In case of digital images, histogram is the number of pixels that have colors values in each of a fixed list of color ranges that cover the image’s color space. The color histogram can be built for any type of color space, but it is often used for three-dimensional spaces like RGB or HSV. In case of monochromatic images, the term intensity histogram may be used instead [1].

Above network have an input layer with three neurons, one hidden layer(in the middle) with three neurons and an output layer with three neurons [12].

- **Input Layer -** The input layer accepts the input vector (x1...xp) and standardizes the values of each variable in the range of -1 to 1. Then the distribution of these standardized values along with constant input called bias of value 1 is given to each hidden layer neurons by input layer this ICICES2014 - S.A.Engineering College, Chennai, Tamil Nadu, India ISBN No.978-1-4799-3834-6/14/$31.00©2014 IEEE bias value is then multiplied by a weight and added to the sum that is going into the neuron.
- **Hidden Layer -** At each neuron in the hidden layer, a weight (wji) is multiplied to the value from each input neuron. Then a combined value uj is produced by adding the resulting weighted values from each hidden layer neuron. This weighted sum(uj) is then given to the a transfer function ,producing the outputs of value hj . The combined outputs obtained from the hidden layer neurons are then given to the neurons in output layer.
- **Output Layer -** At each output layer neuron weight (wkj) is multiplied to the value that is obtained from each hidden layer neuron, and then a combined value vj is produced by adding the resulting weighted values. The weighted sum (vj) is then given to the transfer function, which outputs a value yk. The y values are the outputs of the network.

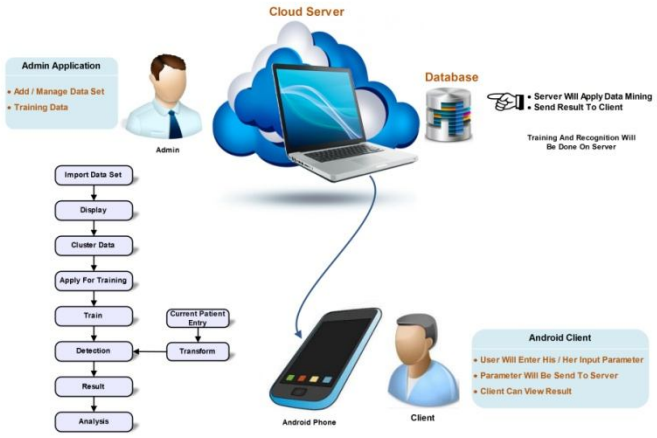
B. Back-Propagation Network

The back-propagation algorithm is the popular algorithm for the training of the neural network. This algorithm is generally used to train multilayer perceptron and many other neural networks. In back-propagation algorithm, the output obtained is compared with the target or expected output and the error is computed. This computed error is then again given to the neural network (fed back or back propagated) and weights are adjusted using this error so that the resulting output will get closer to the target or expected output. This process is repeated for number of times such that at each iteration the error value gets reduced and the output gets more and more closer to the target or expected output. This process is known as ”training” of neural network.

In this prediction system the Cleveland heart disease database is used to fed the input to neural network. The network is having three layers and feed forward neural network model. The back propagation learning algorithm with learning rate of momentum and adaptive learning is used to train the neural network. In the input layer of the network

there are 13 neurons that accept the 13 values of clinical information from the heart disease database. The hidden layer neurons can be varied in order to reduce error and increase accuracy and the output layer consists of single neuron that indicates whether the heart disease is present or absent.

IV. PROPOSED SYSTEM



In heart disease we consider input dataset as a patient detail like Age, Gender, Blood Pressure, Sugar level. The input can be normalized by using K-means clustering algorithm. The input data are cluster which analysis to partition. The clustering trained label data are classified by ANN(Artificial Neural Network) classifier. After classification the heart disease are predicted then data are stored on cloud server and report of patient can be check by using Andriod Application.

1. Android client

User will give the input parameter like(age, gender, BP, cholesterol etc) to the central server. after that user will get result.

2. Central server

Predict the disease and training and recognition can be done on server then result will be send to the user.

3. Admin

Admin will add and manage the data set and train the given data.

4. Android Application

The result can be check on android phone.

V. EXPERIMENTAL RESULTS

a. Data Set

The Data set is taken from Data mining repository of University of California, Irvine (UCI). Data set from Cleveland data set, Hungary Data set, Switzerland Data set, long beach and Statlog data set are collected. Cleveland, Hungary, Switzerland and va long beach data set contains 76 attributes in all. But only 14 attributes are used. Among all those Cleveland data set and Statlog data set are the most commonly used data set. Because all the other has missing values. sample of data set collected from the UCI repository.

Age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	num
63	1	1	145	233	1	2	150	0	2.3	3	0	6	0
67	1	4	160	286	0	2	108	1	1.5	2	3	3	2
67	1	4	120	229	0	2	129	1	2.6	2	2	7	1
37	1	3	130	250	0	0	187	0	3.5	3	0	3	0
41	0	2	130	204	0	2	172	0	1.4	1	0	3	0
56	1	2	120	236	0	0	178	0	0.8	1	0	3	0
62	0	4	140	268	0	2	160	0	3.6	3	2	3	3
57	0	4	120	354	0	0	163	1	0.6	1	0	3	0
63	1	4	130	254	0	2	147	0	1.4	2	1	7	2
53	1	4	140	203	1	2	155	1	3.1	3	0	7	1
57	1	4	140	192	0	0	148	0	0.4	2	0	6	0
56	0	2	140	294	0	2	153	0	1.3	2	0	3	0
56	1	3	130	256	1	2	142	1	0.6	2	1	6	2
44	1	2	120	263	0	0	173	0	0	1	0	7	0

b. Key Attribute

KEY ID	KEY ATTRIBUTE
1	PatientId – Patient's identification number
2	Age in Year
3	Sex (value 1: Male; value 0: Female)
4	Chest Pain Type (value 1: typical type 1 angina, value
5	typical type angina, value
6	non-angina pain, value 4: Asymptomatic)
7	Fasting Blood Sugar (value 1: >120 mg/dl; value 0:
8	Serum Cholesterol (mg/dl)
9	Restecg – resting electrographic results (value 0: normal; value 1: having ST-T wave Abnormality; value 2: showing probable or definite left ventricular hypertrophy)
10	Maximum Heart Rate Archieved; value (0.0) :> 0.0 and <=80, value (1.0) : >81 and <119, value (2.0):=120;
11	Fasting Blood Sugar; 120
12	Exang - exercise induced angina (value 1: yes; value 0: no)
13	Old peak – ST depression induced by exercise
14	Slope – the slope of the peak exercise ST segment (value 1: unsloping; value 2: flat; value 3: down sloping)
15	CA – number of major vessels colored by floursopy (value 0-3)
16	Thal (value 3: normal; value 6: fixed defect; value 7: reversible defect)

VI. CONCLUSION

Heart disease is one of the leading causes of death worldwide and the early prediction of heart disease is important. Instead of going for a number of tests, predicting heart disease with less number of attributes is a challenging task in Data Mining. Neural Network with training data 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 neural network in system accepts clinical features as input and it is trained using clustering algorithm to predict that there is presence or absence of heart disease in the patient with highest accuracy of 98 % comparative to other system. The classification accuracy can be improved by reduction in features. From the analysis it is concluded that, data mining plays a major role in heart disease classification. The overall objective is to study the various data mining techniques available to predict the heart disease and to compare them to find the best method of prediction. After prediction of heart disease the result of patient can be check by using Android Application.

REFERENCES

[1] "Medline plus: Heart diseases," http://www.nlm.nih.gov/medlineplus/heartdiseases.html.

[2] A. Rajkumar and G. S. Reena, "Diagnosis of heart disease using datamining algorithm", Global Journal of computer Science and Technology, vol. 10, pp. 38– 43, December 2010.

[3] M. Gudadhe, K. Wankhade, and S. Dongre, "Decision Support system for heart disease based on support vector machine and artificial neural network", In proceedings of IEEE International Conference on Computer and Communication Technology (ICCCT), pp. 741–745, November 2010.

[4] "Dtreg", http://www.dtreg.com/mlfn.htm.

[5] S.N.Sivanandam and S.N.Deepa, "Principles of soft computing, Wiley India Edition, 2007.

[6] Bala Sundar V, "Development of data Clustering Algorithm for predicting heart", IJCA vol 48(7), June 2012, pp 23-26.

[7] Anita Dewan , Meghana Sharma "Prediction of heart Disease using a hybrid technique in data mining Classification"IEEE International Conference on advanced engineering march 2015.

[8] 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.