A SYSTEM FOR HEART DISEASE PREDICTION USING DATA MINING TECHNIQUES

Minal Zope
Department of Computer Engineering, Savitiribai Phule Pune University.

Sagar Birje
Department of Computer Engineering, Savitiribai Phule Pune University.

Lijo Johns
Department of Computer Engineering, Savitiribai Phule Pune University.

Amit Vasudevan
Department of Computer Engineering, Savitiribai Phule Pune University.

Nishant Salunkhe
Department of Computer Engineering, Savitiribai Phule Pune University.

ABSTRACT

The diagnosis of heart disease is most complicated and tedious task in the field of medical science. Thus there is a need for development, a support system that will help medical practitioners to detect heart disease of a patient. Heart disease is something that cannot be detected by physical observation, but by analyzing different constraints that is associated with this disease. The diagnosis depends on the careful analysis of different clinical and pathological data of the patient by medical experts, which is a complicated process. We propose efficient algorithm hybrid with ANN (Artificial Neural Network) and K-mean technique approach for heart disease prediction. The main objective of our model is to develop a prototype which can determine and extract known knowledge related with heart disease from the past heart disease database record. After implementing it and comparing with other techniques which have been used previously, our prediction came out to be around 87.35% which is way better compared with other techniques.

INTRODUCTION

Data mining is a technique used for the extraction of hidden predictive information from sets of databases and is a powerful technology with great potential and useful to both IT companies and medical fields to emphasize on the most valuable information in their data warehouses. Data mining tools are designed to deal with behaviours and future movements, allowing businesses to make zealous knowledge-driven decisions. The main functionality of data mining involves classification, association and clustering. Due to its increasing demand, various data mining techniques are applied for better decision making in the field of medicine. Many medical organizations are facing a big challenge that is the providing the quality services like diagnosing patients correctly and providing treatment where common man can afford its costs. Data mining techniques simplifies several important and critical questions related to health. In India or many other Asian countries Heart disease is the most common reasons of death. In 2003 approximately 17.3 million people died around the globe and out of this, 9 million were only due to the coronary heart disease. 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 a heart problem. However nowadays, we can avoid such kind of diseases by advance techniques and better decision making at early stage. In this paper, we have implemented heart disease prediction using two techniques K means and Neural Network. The use of this hybrid technique propels our accuracy to 87-90%. The proposed system is implemented using Net-beans 8.1 and Java Serialization.

RELATED WORK

There are number of systems for prediction of different diseases which are proposed and implemented by using different techniques and methods, Our research in this area shows that, there are not many systems developed more than one techniques.

Tsai and Watanabe have classified myocardial heart disease from ultrasonic images by optimizing the fuzzy membership functions by using genetic algorithm m based method.

Usha Rani [1] has implemented ANN in heart disease database using feed forward method and back propagation algorithm.

Anbarasi.M and et Al [2] has used Genetic Algorithm (GA) to determine the attributes for the diagnosis of heart disease.

In [3], Subbulakshmi and et al., used Naïve Bayes algorithm for prediction of Decision Support in heart disease prediction System.

Chen A.H [4] used Artificial Neural Network (ANN) algorithm for classifying the heart disease based on input. Learning Vector Quantization (LVQ) is a prototype model based on supervised Classification Algorithm. In [5], Milan Kumari compares Ripper, Support Vector Machine, Decision Tree and Artificial Neural Network based on Sensitivity, Specificity, Accuracy, Error Rate, True Positive Rate and False Positive Rate to predict whether the person is infected or not infected. In [6], Feed forward back propagation neural network is used as a classifying algorithm to distinguish between infected or not in both the cases.

In [7], Shouman and et. Al., used K-Nearest Neighbor (K-NN) in classification problem.

In [8], Bala Sundar and et. al., used K-mean Clustering Algorithm for the heart disease prediction. [9] K.Srinivas in 2010 made a comparative analysis of popular data mining technologies namely decision tress, Naive Bayes & Neural Network for classifying heart disease dataset. [10] M.Akhil Jabbar, B.L Deekshatulua , Priti Chandra made classification of heart disease using k nearest neighbour and genetic algorithm. [11] In Asha Rajkumar, the data mining Classification is based on supervised machine(SVM) learning Algorithm.

METHODOLOGY

In the system for heart disease prediction, we have implemented both the techniques Clustering and Classification. The system consists of three steps, in the first step 13 clinical attributes are accepted as input which are loaded through UCI repository, in the second step clustering is done by which we get clustered dataset. In the third step neural network technique is used with the help of back propagation algorithm to train the system.

PARSING OF DATA SET

In this part of parsing a dataset, the dataset we will be implementing is loaded through UCI repository. UCI is a website available on internet through which we can acquire standard datasets of various diseases such as cancer, heart, brain tumour etc. The dataset available initially is stored in the CSV format. Parsing of this CSV file is done by using Java Serialization. Labelled dataset is generated by implementing serialization.

K MEAN CLUSTERING

K means is used to solve clustering problem and is known as unsupervised learning algorithm. The process is very efficient way to classify the data set into number of clusters (assume k clusters). The k-means algorithm takes the input for the number of clusters (say k) and divides a set of n objects into *k* clusters so that the emerging inter-cluster similarity is low but the intra-cluster similarity is high. According to the mean value of the object present in a cluster, cluster similarity is measured.

WORKING

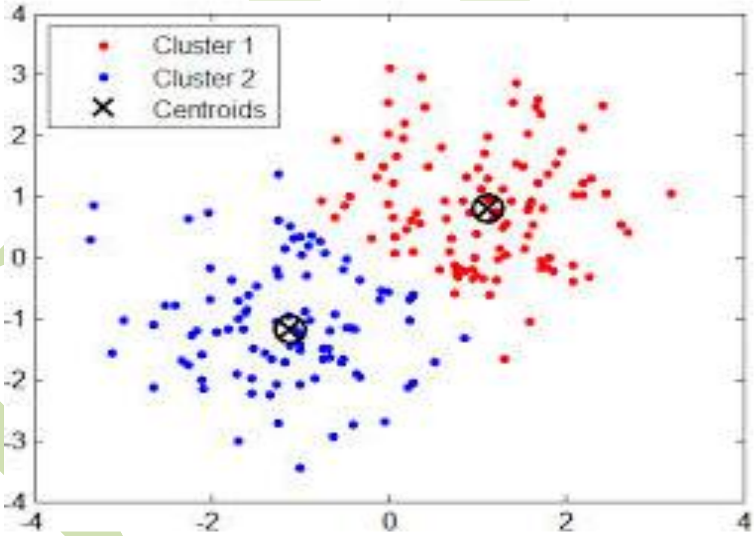

Fig.1 K means Clustering

ALGORITHM: Each cluster's centre is defined by the midpoint value of the objects in the cluster.

INPUT:

- k: Number of clusters,
- D: Dataset containing n objects

OUTPUT: Set of k clusters

METHOD:

- Choose k objects from D as the initial clusters
- Repeat
- (Re)assign each object to the cluster to which the object is mostly alike

- restore the cluster means, based on the midpoint/mean value of the objects in the cluster (i.e. calculate the mean value of the objects for each cluster)
- Until no change.

ARTIFICIAL NEURAL NETWORK

Artificial Neural Network is widely used machine learning algorithm, which is equivalent to the Neurons in Humans. In ANN, artificial neurons and process information are interconnected using suitable connections for computation. It is a learning system that changes its structure based on information (i.e. external or internal) that progresses through the network in the phase of learning.
The Working of ANN is processed in two parts:

- Feed Forward Network
- Back propagation Algorithm

BACK PROPAGATION ALGORITHM

Back propagation, or propagation of error both , is a known method of instructing artificial neural networks how to execute a given job.
The back propagation algorithm is commonly used in layered feed forward ANNs method.
This proposes that the artificial neurons are organized in layers, and send their signals “forward”, and then the errors are reciprocated backwards. The main concept of the back propagation algorithm is to minimise this error, until the ANN is drilled completely in training data set. It means back propagation algorithm is used after feed forward method is applied.

STEPS:

```
▪ 1. Initialize the weights (Randomly)
▪ 2. Repeat
* for each example ‘n’ in the training set do
 i. O = neural-net-output(network, n) ; forward pass
ii. T = teacher output for n
iii. estimate error (T - O) at the output units
iv. Compute delta W for every weights from hidden layers to output layer ; backward pass
v. Compute delta W for every weights from input layer to hidden layers ; backward pass continued
vi. restore the weights in the network
 * end
▪ 3. Until all prototypes are classified correctly
▪ 4. Return (network).
```

IMPLEMENTED WORK

After training our dataset with the help of back propagation algorithm, we ran some tests by adding multiple entries and comparing them with other techniques our accuracy was much better than the other techniques like decision tree, SVM etc.

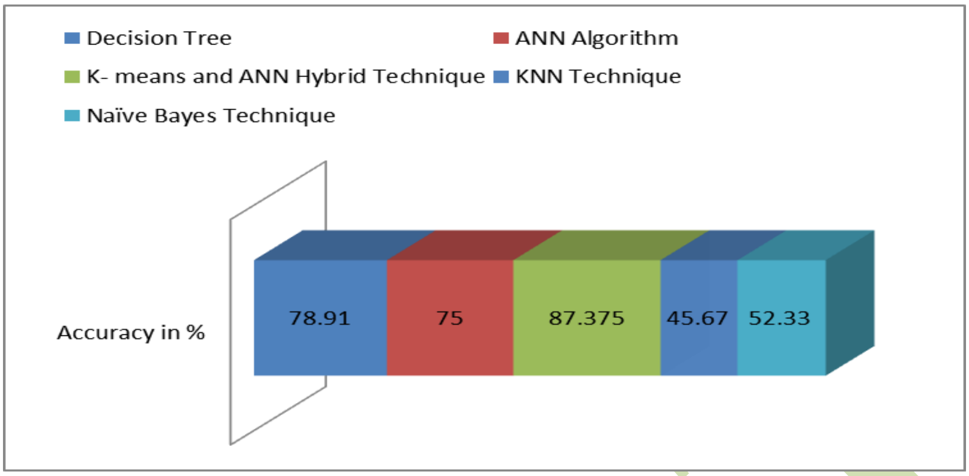

Fig.2 Comparison with our approach.

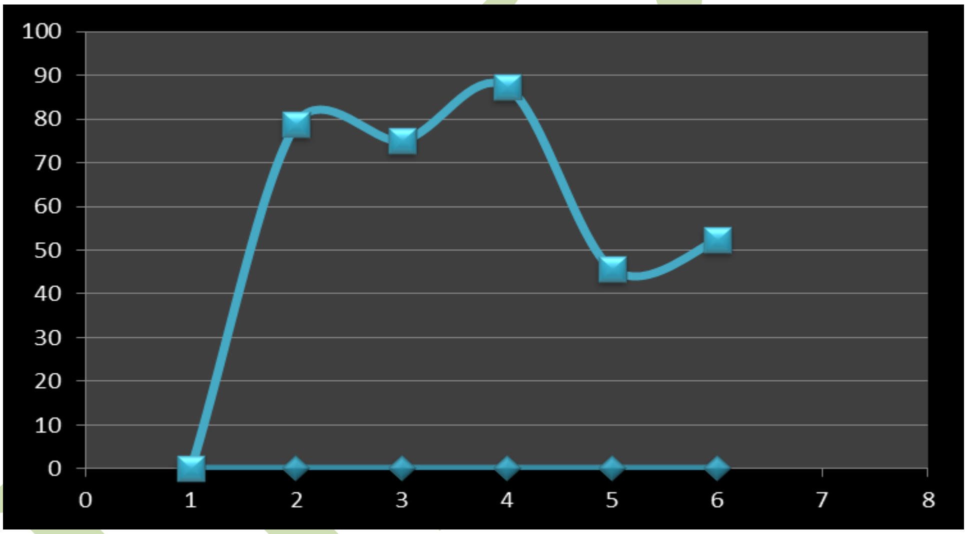

Fig.3 Graph of Tested Results

CONCLUSION

Our paper proposes a useful and much improved technique that can be used for early detection and prediction of heart diseases, previous study showed us that the prediction was made by using only one of the techniques which was either clustering or classification. The results of our model are much better than the previous proposed models. The proposed model is trained under k means and ANN algorithms which provides fast and much improved output and would be helpful for doctors to sedate, mentor and examine their patients. Only drawback currently of this system is it is a standalone system. In further work we would like to develop a utility which is client based system.

REFERENCES

[1] Usha. K Dr, "*Analysis of Heart Disease Dataset using Neural network approach", IJDKP, Vol 1(5), Sep 2011.*

[2] Anbarasi.M, Anupriya and Iyengar "*Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm", International Journal of Engineering and Technology, Vol 2(10), 2010,pp 5370-5376.*

[3] Subbulakshmi, Ramesh and Chinna Rao "*Decision Support in Heart Disease Prediction System using Naïve Bayes", IJCSE, ISSN 0976- 5166, Vol 2(2), May 2011.*

[4] Chen A.H., "*HDPS: Heart Disease Prediction System", Computing in Cardiology, ISSN 0276-6574, pp 557-560, 2011.*

[5] Milan Kumari and Sunila Godara, "*Comparative Study of Data Mining Classification Methods in Cardio-Vascular Diseasen Prediction", IJCST, Vol 2(2), June 2011.*

[6] Qeethara Kadhim Al. Shayea, "*Artificial neural network in Medical Diagnosis", IJCSI, Vol 3(2), March 2011.*

[7] Shouman.M, Turner.T and Stocker.R, "*Applying K-Nearest Neighbour in diagnosing Heart Disease Patients", International Journal of Information and Education Technology, Vol 2(3), June 2012.*

[8] Bala Sundar V, "*Development of Data Clustering Algorithm for predicting Heart", IJCA, Vol 48(7), June 2012, pp 8-13.*

[9] K.Srinivas, Dr.G.Raghavendra Rao, Dr. A.Govardhan, "*Analysis of Coronary Heart Disease and Prediction of Heart Attack in Coal Mining Regions Using Data Mining Techniques", The 5th International Conference on Computer Science & Education Hefei, China. August 24–27, 2010.*

[10] Jabbar M.A., "*Knowledge discovery from mining association rules for Heart disease Prediction", JATIT, Vol 41(2), pp 166-174, 2012.*

[11] Asha Rajkumar and Mrs. Sophia Reena, " *Diagnosis of Heart Disease using Data Mining Algorithms, Global Journal of Computer Science and Technology, vol. 10(10), 2010, pp 38-43.*