Scientific Journal of Impact Factor(SJIF): 3.134

e-ISSN(O): 2348-4470
p-ISSN(P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 2,Issue 10, October -2015

# Heart Disease Prediction using Data Mining Techniques

Minal Zope[1], Amit Vasudevan[2], Sagar Birje[3], Lijo Johns[4], Nishant Salunkhe[5]

[1,2,3,4,5]*Department of Computer Engineering, SavitribaiPhule Pune University, Ganeshkhind, Pune, India*

***Abstract:*** *Prediction of Heart Diseases is the most complicated task in Medical Science. Thus there is a need for development, a support system that will help medical practitioners to detect heart disease ofa patient. Heart disease is something that cannot be detectedby physical observation, but by analyzing different constraints that is associated with this disease. It has been seen that, this disease comes instantly when its limitations are reached. To avoid such incidents, a well planned diagnosis is required. 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. It can be used to solve queries for detecting heart diseases and help doctors to make smart clinical decisions.*

***Keyword:****Heart Disease, Machine Learning, Data Mining, Artificial Neural Network, K-mean.*

## I. INTRODUCTION

Data mining is used for the extraction of hidden predictive information from sets of databases and is a powerful technology with great potential to bothIT companies and medical fields to concentrate on the most valuable information in their data warehouses. Data mining tools are designed to envisionbehaviors 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 bigchallenge 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.

The term heart disease means different problems that affect the normal functioning of circulatory system, which consists of heart and blood vessels. There are different categories of heart diseases like cardiovascular disease in which the heart and blood vessels are affected and as a result of which the blood is not pumped and circulated properly through all the body parts. In myocardium (heart muscle) disease the heartdoes not get sufficient blood that it requires because ofcholesterol and fat that is deposited inside the wall of thecoronary arteries that supply the blood to heart.The disease diagnosis process in the medical field can be considered as a decision making process in which the diagnosis of a new and unknown cases is made by medical practitioner from the information that is available from medical data and from experience in medical field. In order to make this decision making process less costly, easy, faster, more accurate and efficient, the process can be automated.

In today's world, large number of people both young and old are suffering from different types of heart diseases and the count of patients suffering and dying from these diseases are increasing.So there is a need of accurate and early detection of heart disease with proper medical treatment.

## II. RELATED WORK

The number of systems for prediction of different diseases are proposed and implemented by using different techniques and methods, Our review in this area shows that, there are not many systems developed using hybrid approach i.e. two algorithms/techniques clustering and classification.

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.

The Significance of our methodology is that we are applying hybrid approach by using two techniques of data mining which are clustering and classification.

## III. PROPOSED SYSTEM

In previous section, westudied that no other system had implemented a hybrid approach for heart disease prediction. We propose a model which will use both clustering and classification techniques.It provides efficient and accurate prediction of heart disease and will help medical practitioners to make accurate decisions. Our Proposed System isdivided into 3 main modulesas follows:

Module 1: Parsing the dataset from the UCI machine learning repository,

Module 2: Clustering the data which is parsed,

Module 3: Classify the clustered data setsand predict whether patient is infected or not infected.

### 3.1 PARSING OF DATASET

In this module, the dataset which we will be using is taken from UCI machine learning repository. The dataset is available on the internet through which we can load the dataset related to various diseases such as heart disease, cancer, dermatology etc. The loaded dataset will be in CSV (Comma Separated Values) format. Since we will be using java as our base language there are multiple ways of parsing and reading csv files in java. On parsing the csv file dataset, a labelled dataset is generated.

### 3.2 CLUSTERING USING K-MEAN

K-meansis an unsupervised learning algorithmswhich is known to solve clustering problem. The process is a easy and efficient way to classify the data set through a some number of clusters (assume k clusters) used as a fixed apriori. 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. Cluster similarity is measured in terms to the mean value of the objects present in a cluster, which can be percieved as the cluster's centroid or center.

K-mean algorithm is given as follows :

First, Random objects are selected (say k), each of which at the start represents a cluster mean/midpoint. For each of the left objects, an object is assented to the cluster to which it is the most correlative, according to the distance in the object and the cluster mean/midpoint. It then enumerates the new mean or midpoint for each cluster. This process repeats until the criterion function converges. The square-error criterion is used, which is defined as

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} |p - m_i|^2$$

where,
E is the addition of the square error for every objects used in the dataset; p is the point in space characterized a given object; and $m_i$ is the mean of cluster $c_i$ (both p and $m_i$ are multidimensional). For every object in each cluster, the

 

Euclidean distance or the interval from the object to its cluster center is squared, and those Euclidean intervals are added. This paradigm tries to make the k clusters as impermeable and as serrate as possible. The k-means procedure is summed up in following figure.

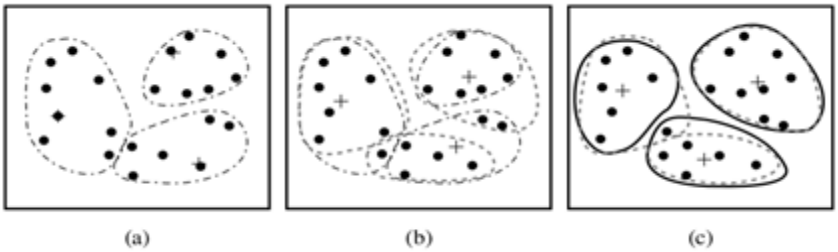


***Fig.1Clustering by k-mean. (the mean is marked by a "+")***

**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 ailke
- 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.

### 3.3 ARTIFICIAL NEURAL NETWORK

Artificial Neural Network is well known machine learning algorithm, which is similar 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.

This algorithm has 3 layers:

a) Input Layer
b) Hidden Layer
c) Output Layer

 

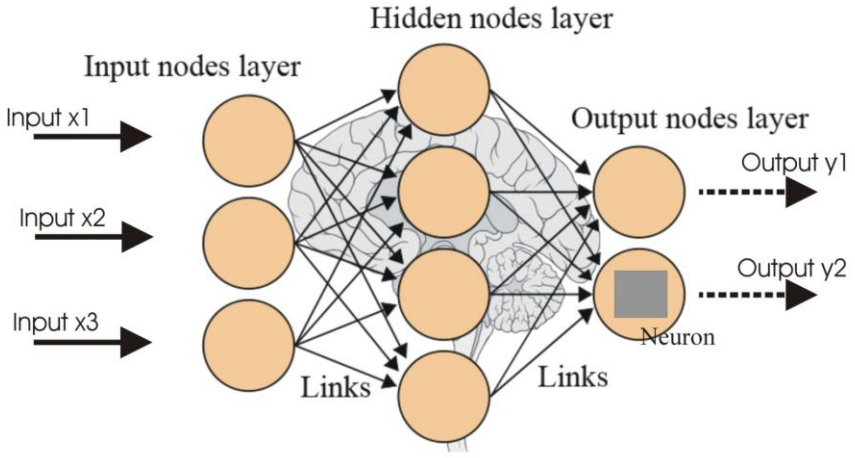


***Fig.2Structure of ANN***

**Input Layer :**The input layer will accept the output values from clustered dataset and will 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 thisbias 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 weightis multiplied to the value from each input neuron. Then a combined value isproduced by adding the resulting weighted values from each hidden layer neuron. This weighted sum is then given to the a transfer function producing the an outputvalue. The combined outputs obtained from the hidden layer neurons are then given to the neurons in output layer.
Hidden Layers = 'number of input layers + number of output layers + 1'.

**Output Layer :**The output from each hidden layer is multiplied by weight and the resulting values are summed to give a final value. This value is the output of the Network.

**A.** The working of ANN algorithm is processed in two parts :
1. Feed Forward network
2. Back propagation algorithm

**B. Back propagation Algorithm :**

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

**C. 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. retsore the weights in the network
* end

- 3. Until all prototypes are classified correctly
- 4. Return(network)

## IV. CONCLUSION

Our paper proposes a useful and accurate technique that can be used for early detection and prediction of heart diseases, previous study showed us that the prediction was made only on the basis of one algorithm i.e. either clustering or classification. The precision of our model is much better than the previous models. The proposed model will be a trained by k mean and ANN algorithms which will provide fast and accurate output and would be helpful for doctors to sedate and mentor their patients. We also reviewed that the earlier proposed models used SVM , Naïve Bayes but found out that neural network is best among classification technique used in our domain. 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.

## V. 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 ChinnaRao "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 SunilaGodara, "Comparative Study of Data Mining Classification Methods in Cardio-Vascular DiseasenPrediction", IJCST, Vol 2(2), June 2011.
[6] QeetharaKadhim 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] BalaSundar V, "Development of Data Clustering Algorithm for predicting Heart", IJCA, Vol 48(7), June 2012, pp 8-13.
[9] K.Srinivas, Dr.G.RaghavendraRao, 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.

 