

Heart Disease Prediction Using ANN and Improved K-Means

Ankita R. Mokashi¹, Madhuri N. Tambe², Pooja T. Walke³

Computer science and Engineering Department, ICOER, Pune University, Pune, India^{1,2,3}

Abstract: Data mining is the computer based process to analyze large sets of data and then extract the meaningful data. Data mining tools predict future trends, allow for business to make knowledge-driven decisions. Heart disease is most challenging disease for reducing patient number. There are many data mining techniques like decision tree, Naive Bayes and neural network. In this paper, we use the improved k-means and ANN techniques for improving accuracy. We have 13 parameters like age, sex, chest pain, blood pressure, cholesterol, fasting blood sugar, slope, ca etc as input to the system and using this attributes and algorithms we can predict the heart disease will occur or not. We suggest the medicines in future work.

Keywords: Data mining, Heart disease, Artificial intelligence, Clustering, Artificial Neural Network.

I. INTRODUCTION

There are number of factors which increase the risk of Heart disease. Now-a-days, in the world Heart disease is the major cause of death in less span of time. The World Health Organization (WHO) has calculated that 12 million deaths occur in world, every year due to the Heart diseases. Prediction by using data mining techniques gives us accurate result of Heart Diseases. IHDP (intelligent heart disease prediction system) can discover, extract the hidden knowledge associated with heart disease from historical dataset.

Data Mining: Is the extracting useful information from data. There are variety of tools & technology used in data mining. Number of terms like data classification, clustering, data integration, data regression for normalizing the data clustering or classification is used [2]. Prediction of heart disease based on the patient's database. In Biomedical diagnosis the data mining plays an important role for predicting heart disease by using information regarding symptoms. Sometime physicians may not be able to diagnosis correctly at correct time hence because of this it is difficult to predict disease.

Artificial Intelligence (AI) Artificial intelligence is the study of how to make computers do things that people are better at or would be better at if:

- They could extend what they do to a World Wide Web-sized amount of data and
- Not make mistakes.

It is the science and engineering of making intelligent machines, especially intelligent computer programs.

In other words, it is related to the similar task of using computers to understand human intelligence, but AI does not have to confine itself to methods that are biologically observable.

II. HEART DISEASE

Heart is the vital organ of our body. Life is depend o heart. Now-a-day's a major task of healthcare association like hospitals, medical focuses are diagnosing the patients accurately. Heart disease is the most challenging task for reducing patient's number.

III. LITERATURE SURVEY

In this section, Data mining techniques used for decision making in heart disease are analyzed.

Few research works has been carried out, results for diagnosis of various diseases using data mining are.

- Sellappan et al. Propose Intelligent Heart Disease Prediction System Using Data Mining Techniques[3]

This research uses three data mining techniques they are Decision Trees, Naïve Bayes and Neural Network. Using these techniques IHDP can discover & extract hidden knowledge associated with heart disease from historical heart disease database. All three models could answer complex queries, each with its own strength with respect to ease of model interpretation, access to detailed information and accuracy.

- K. Srinivas et al. Propose Analysis of Coronary Heart Disease and Prediction of Heart Attack in Coal Mining Regions Using Data Mining Techniques [4].They have presented the effective heart disease prediction method using data mining techniques. They have presented automatic and effective heart attack prediction methods.

For predicting heart attack significantly 15 attributes are listed in medical literature. Besides this list, they have incorporated other attributes which will effect on results such as financial status like stress, pollution and previous medical history.

- A. Khemphila et al. propose Heart disease Classification using Neural Network and Feature Selection [5], introduced the multi-layered perception and Back Propagation algorithm. ANN is used to classify the dataset. It improves the classification accuracy. This research shows that feature selection helps to increase computational efficiency while improving classification accuracy. The output of this has attained an accuracy of training data set as 89.56% and validation data set as 80.99%.

- M A. Jabbar et al. presented Classification of Heart Disease using Artificial Neural Network and Feature Subset Selection [6], introduced a classification approach which uses ANN and feature subset selection for the classification of heart disease. This experiment results

indicate that on an average with ANN and feature subset selection provides the average better classification accuracy and dimensionality reduction.

- Dangare et al. Propose Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques [7]. It used the multilayer perceptron neural network (MLPNN). This paper uses Neural Networks, Decision Trees, Naive Bayes. The system uses medical terms such as sex, blood pressure, cholesterol like 13 attributes to predict the likelihood of patient getting a Heart disease. But this paper added 2 more attributes they are obesity and smoking. From results conclusion is Neural network is more efficient.

- Anupama Chadha and Suresh Kumar proposed An Improved K-Mean Clustering Algorithm: A Step Forward for Removal of Dependency on K [10]. K-means algorithm depends on k clusters, so to remove k dependency, modified or improved k-means was proposed. Also the results show the quality of clusters is not compromised compared to k-means algorithm.

IV. PROPOSED SYSTEM

Health care systems manage huge amount of data to extract knowledge for making medical diagnosis. The main objective of this research is to build a system that predicts the heart disease that is in which level the disease is and suggest some medicines according to the level of the disease. To develop this system, medical terms such as sex, blood pressure, and cholesterol like 14 input attributes are used.

System Architecture:

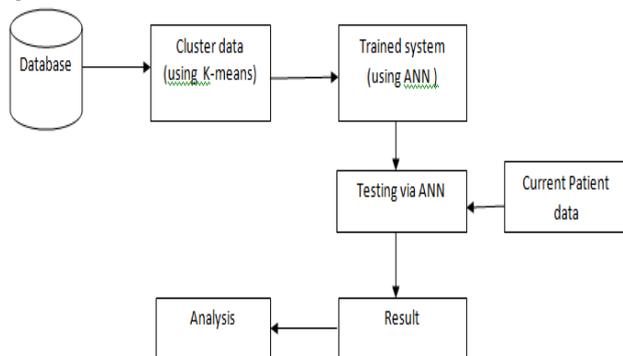


Fig. Proposed System Architecture

In our system there are six modules, they are explained as follows:

Module 1: Database

The publicly available database on internet that is mostly preferred by many researchers is used for prediction. The input database contains CSV (Comma Separated Vector). The database contains 303 records.

Input attributes

Sr.No	Attribute	Description	Condition
1	Age	Age in years	Continuous
2	Sex	Male or female	0 = male 1 = female
3	CP	Chest pain type	0 = typical type 1 1 = typical type angina

			2 = non-angina pain 3 = asymptomatic
4	Thestbps	Resting blood pressure	Continuous value in mm hg
5	Chol	Serum cholesterol	Continuous value in mm/dl
6	Fbs	Fasting blood sugar	0 ≥ 120 mg/dl 1 ≤ 120 mg/dl
7	Restecg	Resting electrographic results	0 = normal 1 = having ST_T wave abnormal 2 = left ventricular hypertrophy
8	Thalach	Maximum heart rate achieved	Continuous value
9	Exang	Exercise induced angina	0 = no 1 = yes
10	Oldpeak	ST depression induced by exercise relative to rest	Continuous value
11	Slope	Slope of the peak exercise ST segment	0 = up sloping 1 = flat 2 = down sloping
12	Ca	Number of major vessels colored by floursopy	0-3 value
13	Thal	Defect type	0 = normal 1 = fixed 2 = reversible defect
14	Num	Diagnosis of heart disease	0 =<50% 1 =>50%

Table 1. Description of 13 input attributes

Module 2: Cluster Data

In this module input attributes are normalized. For this normalization process we use the Improved K-means.

Improved K-Means

Improved K-Means algorithm is extension over K-Means algorithm. K-Means algorithm highly depends on the selection of initial cluster centers. K-Means algorithm does not guarantee to provide same result for different runs on same data set. Therefore improved k-means algorithm is selected for clustering the dataset that does not require selecting initial clusters as input.

Algorithm 1:

Input: Data set x contains n data points; the number of cluster is k.

Output: k clusters of meet the criterion function convergence.

Program process:

Step1: Initialize the cluster center.

Step 1.1: Select a data point xi from data set X, set the identified as statistics and compute the distance between xi and other data point in the data set X. If it meet the

distance threshold, then identify the data points as statistics, the density value of the data point x_i add 1.

Step 1.2: Select the data point which is not identified as statistics, set the identified as statistics and compute its density value. Repeat Step 1.2 until all the data points in the data set X have been identified as statistics.

Step 1.3: Select data point from data set which the density value is greater than the threshold and add it to the corresponding high-density area set D .

Step 1.4: Filter the data point from the corresponding high-density area set D that the density of data points relatively high, added it to the initial cluster center set. Followed to find the $k-1$ data points, making the distance among k initial cluster centers are the largest.

Step 2: Assigned the n data points from data set X to the closet cluster.

Step 3: Adjust each cluster center K by the formula (3).

Step 4: Calculate the distance of various data objects from each cluster center by formula (4), and redistribute the n data points to corresponding cluster.

Step 5: Adjust each cluster center K by the formula (3).

Step 6: Calculate the criterion function E using formula (1), to determine whether the convergence, if convergence, then continue; otherwise, jump to Step 4.

The output of this module goes to next i.e. Trained dataset module.

Module 3: Trained Dataset Using ANN

There are three input layers in ANN, they are: input layer, intermediate (called the hidden layer) and output. Several hidden layers can be placed between the input and output layers.

- **Input Layer** – The activity of the input units represents the raw information that is fed into the network.
- **Hidden Layer** - The activity of each hidden unit is determined by the activities of the input units and the weights on the connections between the input and the hidden units.
- **Output Layer** - The behavior of the output units depends on the activity of the hidden units and the weights between the hidden and output units.

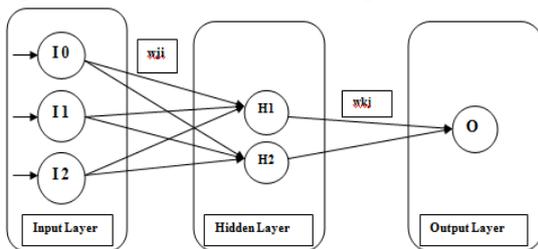


Fig.4: Artificial Neural Network

In above figure, hidden layer accepts data from the input layer. It uses input values and modifies them using some weight value, this new value is then send to the output layer but it will also be modified by some weight from connection between hidden and output layer. Output layer process information received from the hidden layer and produces an output. This output is then processed by activation function.

A) Back-Propagation Neural Network:-

Back-propagation, an abbreviation for "backward propagation of errors", is a common method of training artificial neural networks used in conjunction with an optimization method such as gradient descent. The method calculates the gradient of a loss function with respects to all the weights in the network.

B) Feed Forward Neural Network:-

In a feed forward network, information flows in one direction along connecting pathways, from the input layer via the hidden layers to the final output layer. A feed forward neural network is an artificial neural network where connections between the units do not form a cycle. The feed forward neural network was the first and simplest type of artificial neural network devised.

ANN application areas:

- Tax form processing to identify tax fraud
- Enhancing auditing by finding irregularities
- Bankruptcy prediction
- Customer credit scoring
- Loan approvals
- Credit card approval and fraud detection
- Financial prediction
- Energy forecasting
- Computer access security (intrusion detection and classification of attacks)
- Fraud detection in mobile telecommunication networks

Module 4: Testing Dataset

In this module we are taking the entries from client as attributes specified. Comparing the attributes with trained dataset the system predicts output as per attributes. We can test multiple as well as single entry. Multiple entries used by Admin while single entry used by admin as well as client.

Module 5: Result and Analysis

In this module the output from testing is analyzed. This analysis is divided in five levels as follow:

- 1)Level 1: Nitroglycerine
- 2)Level 2: Angioplasty
- 3)Level 3: Medicine 2
- 4)Level 3: Medicine 3

V. RESULTS

The database contains 303 records taken from UCI standard repository. Initially the dataset is loaded and managed by the authorized user. Then at the backend the dataset is normalized according to the attributes and their corresponding values with the help of improved k-means clustering method. This normalized database can be seen by the user and then it is trained by using ANN freed forward and back propagation algorithms. After training of the database the admin can analyze multiple entries or analyze a single entry.

A confusion matrix is obtained to calculate the accuracy of classification. A confusion matrix shows how many instances have been assigned to each class.

TP (True Positive): It denotes the number of records classified as true while they were actually true.

FN (False Negative): It denotes the number of records classified as false while they were actually true.

FP (False Positive): It denotes the number of records classified as true while they were actually false.

TN (True Negative): It denotes the number of records classified as false while they were actually false.

Database containing 303 records are used and following result is obtained with percentage accuracy.

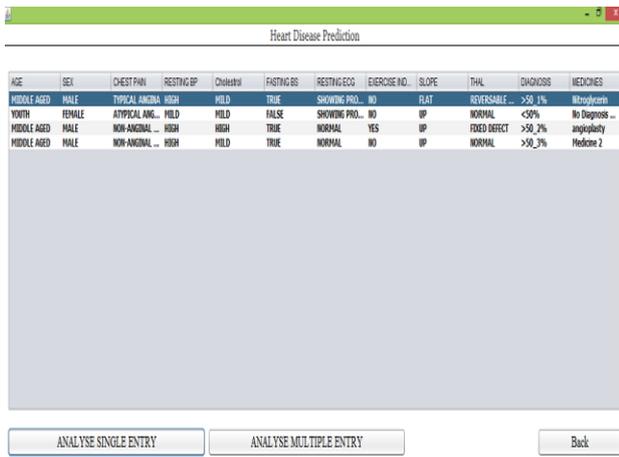
True Positive Count: 62

True Negative Count: 200

False Positive Count: 39

False Negative Count: 0

- [7] Chaitrali S. Dangare and Sulabha S. Apte, "Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques", International Journal of Computer Applications, Vol. 47, No. 10, 2012, pp. 0975 – 888.
- [8] MA.Jabbar, B.L Deekshatulu, Priti Chandra, "Heart disease prediction system using associative classification and genetic algorithm" pp. 183-192 Elsevier- 2012.
- [9] Sonawane, "Prediction of Heart Disease Using Multilayer Perceptron Neural Network" ICICES2014 – S. A. Engineering College, Chennai, Tamil Nadu, India
- [10] Anupama Chadha and Suresh Kumar, "An Improved K-Means Clustering Algorithm: A Step Forward for Removal of Dependency on K", 2014 International Conference on Reliability, Optimization and Information Technology – ICROIT.



AGE	SEX	CHEST PAIN	RESTING BP	Cholesterol	FATIGUE	RESTING ECG	EXERCISE NO.	SLOPE	THAL	DIAGNOSIS	MEDICINES
MIDDLE AGED	MALE	TYPICAL ANGINA	HIGH	MILD	TRUE	SHOWING PRO.	NO	FLAT	REVERSIBLE	>50 %	Nitroglycerin
YOUTH	FEMALE	ATYPICAL ANGINA	MILD	MILD	FALSE	SHOWING PRO.	NO	IP	NORMAL	<50%	No Diagnosis
MIDDLE AGED	MALE	NON-ANGINAL	HIGH	HIGH	TRUE	NORMAL	YES	IP	FIXED DEFECT	>50 %	angioplasty
MIDDLE AGED	MALE	NON-ANGINAL	HIGH	MILD	TRUE	NORMAL	NO	IP	NORMAL	>50 %	Medicine 2

Fig:- Predicted output with given attributes

VI. CONCLUSION

Some Heart disease classification techniques are reviewed in this paper. From the analysis it is concluded that artificial neural network algorithm is best for classification of knowledge data from large amount of medical data. Good performance with increase in efficiency is obtained from neural network when provided with normalized data. The data is normalized using a clustering. It is supportive system to the doctor's decision. The Artificial neural network is one of the best for heart disease prediction. Improved K-means is also best rather than original K-means.

REFERENCES

- [1] Stuart Russell and Peter Norvig (1995), "Artificial Intelligence: A Modern Approach," Third (3rd) edition, Pearson, 2003.
- [2] Jiawei Han, Micheline Kamber, "Data mining: concepts and techniques", Morgan Kaufmann Publisher, second edition.
- [3] Sellappan Palaniappan, Rafiah Awang, "Intelligent Heart Disease Prediction System Using a Data Mining Techniques", International Journal of Computer Science and Network Security (IJCSNS), VOL.8 No.8, August 2008.
- [4] K. Srinivas, B. Kavitha Rani and Dr. A. Govrdhan, "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 IEEE, China. August 24–27, 2010.
- [5] Anchana Khemphila, Veera Boonjing, "Heart disease Classification using Neural Network and Feature Selection", 2011 21st International Conference on Systems Engineering.
- [6] M A. Jabbar, Priti Chandra and B. L. Deekshatulu, "Classification of Heart Disease using Artificial Neural Network and Feature Subset Selection", Journal of Theoretical and Applied Information Technology, Vol. 32, No.2, 2011, pp. 197 - 201.