Hybrid Approach for Heart Disease Detection Using Clustering and ANN

Neha Chikshe[#1], Tejasweeta Dixit[#2], Rashmi Gore[#3], Prerana Akade[#4]

[1#] Dept. of Information Technology, JSCOE, Hadapsar, Pune, India
nehachikshe1792@gmail.com

[2#] Dept. of Information Technology, JSCOE, Hadapsar, Pune, India
tejasweeta01@gmail.com

[3#] Dept. of Information Technology, JSCOE, Hadapsar, Pune, India
rashmigore786@gmail.com

[4#] Dept. of Information Technology, JSCOE, Hadapsar, Pune, India
preranaakade@gmail.com

Abstract- Data mining is a process of extracting data from data set and transforming it into understandable structure for further use. Data mining techniques have been applied magnificently in many fields including business, science and bio informatics, and on different types of data like textual, visual, spatial, and real-time and sensor data. Heart disease prediction is treated as most difficult task in the field of medical sciences. Heart disease detection using data mining can answer complicated queries for diagnosing heart disease and thus assist healthcare practitioners to make intelligent clinical decisions which traditional decision support systems cannot. By providing effective treatments, it also helps to reduce treatment costs. The aim of this study is to develop an artificial neural networks-based diagnostic model for heart disease using a complex of traditional and genetic factors of this disease.

Keywords — *Data mining, Heart disease, K-means algorithm, Artificial Neural Network.*

_____*****_____

I. INTRODUCTION

Heart disease is the dominant cause of death in developed countries and is one of the main contributors to disease strain in developing countries like INDIA. Several studies reveal that heart disease was the leading cause of mortality accounting 32 percent deaths, as high as Canada and USA. People go to healthcare practitioners but the prediction made by them is not 100 percent accurate. Quality service implies diagnosing patients correctly and administering treatments that are impressive. Poor clinical decisions can lead to disastrous consequences which are therefore unacceptable.[2] Hence there is a need to design and establish a clinical decision support for classification of heart disease. Hospitals must also reduce the cost of clinical tests. They can achieve these results by employing appropriate computer based knowledge and/or decision support systems. The healthcare industry collects huge amounts of healthcare data which are not "mined" to determine hidden information for impressive decision making. Discovery of hidden patterns and relationships often goes unexploited. Advanced data mining techniques can help remedy this situation.[5]

Data mining the extraction of hidden predictive data from large databases is a powerful new technology with great probability to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and performance allowing businesses to make proactive, knowledge-driven decisions. The automated, proposed analyses offered by data mining move after the analyses of past events provided by retrospective tools typical of decision support systems.[2] Data mining tools can answer business questions that traditionally were too time consuming to resolve. They cleanse databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.

II. DATA MINING TECHNIQUES

a. ASSOCIATION

Association is one of the finest known data mining technique. In association, a pattern is discovered based on a relationship between items in the same transaction. That is the reason why association technique is also known as relation technique. The association method is used in market basket analysis to identify a set of products that customers frequently purchase together.

b. CLASSIFICATION

Classification is a standard data mining technique based on machine learning. Basically classification is used to classify each item in a input of data into one of predefined set of classes or groups.[7] Classification method makes use of mathematical techniques such as decision trees, continues programming, neural network and statistics. In classification, we establish the software that can learn how to classify the data items into groups. For example, we can apply classification in application that "given all records of employees who left the company; predict who will probably leave the company in a future period." In this case, we divide the records of employees into two groups that named "leave" and "stay". And then we can

ask our data mining software to classify the employees into separate groups.

c. CLUSTURING

Clustering is a data mining technique that makes relevant or useful cluster of objects which have similar characteristics using automatic technique. Clustering is a process of partitioning a set of data (or objects) into a set of meaningful sub-classes, called clusters.[8] The clustering technique describe the classes and puts objects in each class, while in the classification techniques, objects are allow into predefined classes. To make the concept clearer, we can take book management in library as an example.

d. PRIDICTION

The prediction, as it name implicit, is one of a data mining techniques that discovers relationship communication dependent and independent variables. For instance, the prediction analysis technique can be used in sale to think profit for the future if we consider sale is an independent variable, profit could be a dependent variable. Then established on the historical sale and profit data, we can draw a fitted regression curve that is used for profit prediction.

III. DATA MINING TECHNIQUES USED FOR HEART DISEASE DETECTION

a. NAVIE BAYES

A Naive Bayes classifier predicts that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. This classifier is very simple, efficient and is having a good performance. Sometimes it often outperforms more refined classifiers even when the assumption of independent predictors is far. [1] This advantage is especially produced when the number of predictors is very large. One of the most important disadvantages of Naive Bayes is that it has strong feature independence assumptions. Naive Bayes classifier is based on Bayes theorem. This classifier algorithm uses conditional independence, means it assumes that an attribute value on a given class is independent of the values of other attributes. The algorithm performs positively with categorical data but poorly if we have numerical data in the training set. [8] One of the most important disadvantages of Naive Bayes is that it has strong feature independence assumptions.

b. DECISION TREE

Decision Trees (DTs) are a non-parametric supervised learning method used for classification.
The decision tree approach is more vigorous for classification problems.[8] There are two steps in this techniques building a tree & applying the tree to the dataset. Decision tree is one of the most used data mining techniques because its model is easy to understand for users. In decision tree technique, the root of the decision tree is a simple question or condition that has multiple answers. Each answer then leads to a set of questions or conditions that help us determine the data so that we can make the final decision based on it.[1] There are many decision tree algorithms and among them the most popular is J48 which uses a pruning technique to build a good decision tree. Pruning is a method which tries to eliminate the over fitting data which is not so relevant in making a decision and leads to poor prediction. At last, a tree is build to provide flexibility and accuracy. Some of the advantages are they can be easily know and interpret, robust, perform well with huge datasets, able to handle both numerical and categorical data.[1] Decision-tree learners can create over complicated trees that do not generalize well from the training data is one the limitation.

c. NEURAL NETWORKS

An artificial neural network (ANN), often just labelled a "neural network" (NN), is a mathematical model or computational model established on biological neural network. In other words, it is an emulation of biological neural system. A Multi-layer perceptron Neural Networks (MLPNN) is used. It consists of 3 layers– input layer, output layer and a hidden layer with some interconnecting weights. The output is calculated as a function. [1]

$$O_j = f(\sum W_{ji} X_i)$$

Where O_j is output neuron
X_i is input neuron
W_{ji} is the weight connecting X_i and O_j
f is sigmoidal function

It maps a set of input data onto a set of convenient output data. It consists of 3 layers input layer, hidden layer & output layer. There is connection between each layer & weights are assigned to each connection. The primary function of neurons of input layer is to divide input xi into neurons in hidden layer.[5] Neuron of hidden layer adds input signal xi with weights Wji of respective connections from input layer.

d. K- NEAREST NEIGHBOR

In k-NN classification, the output is a class participation. An object is classified by a majority vote of its neighbours, with the object being allow to the class most common among its k nearest neighbours (k is a positive integer, typically small). If k = 1, then the object is simply assign to the class of that single nearest neighbour.[4] K-NN is a type of instance-based learning, or lazy learning, where the function is only almost locally and all computation is deferred until classification. The k-NN algorithm is among the simplest of all machine learning algorithms. The main disadvantage of the KNN algorithm is that it is a lazy learner

IV. LITERATURE SURVEY

System Under Study	Method Used	Result Obtained
Prediction of Heart Disease Using a Hybrid Technique in Data Mining Classification	Data mining and Artificial neural network	neural network is best among all the classification techniques when we talk about prediction or classification of a non-linear data.
Prediction system for heart disease using naive bayes	Naive bayes	Heart Disease Prediction System is established using Naive Bayesian Classification technique. The system extracts hidden knowledge from a actual heart disease Database. This is the most effective model to predict patients with heart disease
Classification of Heart Disease Using K-Nearest Neighbor and Genetic Algorithm	K-NN and Generic algorithm	Nearest neighbor (KNN) is very simple, most suitable, highly efficient and effective algorithm for pattern recognition.KNN is a sincere classifier, where samples are classified based on the class of their nearest neighbor. Medical data bases are high volume in nature
Heart Disease Diagnosis Using Predictive Data mining	Decision tree and Naïve bayes	In Future Genetic algorithm will be used in order to reduce the actual data size to get the optimal subset of attribute sufficient for heart disease prediction. Prediction of the heart disease will be evaluated give to the result produced from it.

V. CONCLUSION

The overall objective of this paper is to study different data mining techniques available to predict the heart disease and to compare them to find the best method of prediction. We can use Naive Bayes, decision tree, classification algorithms to detect heart disease, but as these are having some of the limitations we prefer to use ANN and k-means algorithms. We have done study on k-mean and ANN algorithm. ANN and k-mean algorithm when combined takes less time for execution. Both these algorithms support each other and overcome the limitations of other algorithms as well. K-means clustering is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining. After analyzing both the algorithms to create Heart disease diagnostic models with appropriate analytical characteristics using ANNs and k-means.

REFRENCES

[1] Ankita Dewan, Meghna Sharma, "Prediction of Heart Disease Using a Hybrid Technique in Data Mining Classification", ITM University Gurgaon, INDIA,2015,PP 704-706.

[2] Sivagowry .S1, Dr. Durairaj. M2 and Persia.A3 1 and 3 Research Scholar, 2 Assistant Professor,"An Empirical Study on applying Data Mining Techniques for the Analysis and Prediction of Heart Disease ", Journal of School of Computer Science and Engg. Bharathidasan University, Trichy, Tamilnadu, India.

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

[4] M.Akhil jabbar* B.L Deekshatulua Priti Chandra b,"Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm", International Conference on Computational Intelligence: Modeling Techniques and Applications (CIMTA) 2013, pp 85 94.

[5] Chaitrali S. Dangare , Sulabha S. Apte, PhD. ,"Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques" ,International Journal of Computer Applications (0975 888) Volume 47 No.10, June 2012

[6] Vishwanath Bijalwan1, Vinay Kumar2, Pinki Kumari3 and Jordan Pascual4 ,"KNN based Machine Learning Approach for Text and Document Mining ",3Bansathali University, Rajasthan, India 4 Department of Computer Science, University of Oviedo, Spain ,International Journal of Database Theory and Application Vol.7, No.1 (2014), pp.61-70 .

_____

[7]   Shadab   Adam   Pattekari   and   Asma
       Parveen ,"PREDICTION SYSTEM FOR HEART
       DISEASE USING NAIVE BAYES ", International
       Journal of Advanced Computer and Mathematical
       Sciences ISSN 2230-9624. Vol 3, Issue 3, 2012, pp
       290-294 .

[8]   B.Venkatalakshmi, M.V Shivsankar" Heart Disease
       Diagnosis Using Predictive

[9]   Data mining" 2014 International Conference on
       Innovations   in   Engineering   and   Technology
       (ICIET'14) On 21st&22ndMarch, Organized by

[10]   K.L.N. College of Engineering, Madurai, Tamil Nadu, India

_____