International Journal of Intelligent Computing and Technology
ISSN: 2457-0249 (Online)
IJICT

Website: https://www.ijict.com **June (2021) - Vol. 5, Iss.1 Pages:27-34**

PERFORMANCE ANALYSIS ON HYBRID DATA MINING TECHNIQUES FOR HEART DISEASE PREDICTION

[1]G Priyanka, [2] Dr P Revathi

[1]M.C.A. Student, [2]Assistant Professor,

Holy Cross College, Trichy, Tamilnadu

Article History- Received: April 2021; Published: June 2021

Abstract

In Machine Learning Environment the data stores for health care services are generally perceived as being 'Data Mines' and 'Information Rich' but yet 'Knowledge Poor' for diagnosing and treating purposes. There is a huge quantitative content of data available within the healthcare systems. However, there is a lack of analytical tools to discover hidden relationships and trends in data related to Health issues. Knowledge discovery and Data mining have found numerous applications in medical field with scientific domain. Valuable knowledge can be discovered from application of data mining techniques in healthcare system. Heart disease is one of the primary causes of mortality in the human illness and cannot be envisaged easily. Prediction of cardio vascular disease is a critical challenge in the area of clinical data analysis. This proposed analysis briefly examine the potential use of classification-based data mining techniques such as Naïve Bayesian Classification and Cluster analysis based Partitioning method like k-Means Partitioning approach for heart disease prediction. This proposal also submits the clinical data relevance to heart disease consumptions, through the complexity measures the performance of the techniques are analyzed.

Keywords: Data mining, Machine Learning, Predictive model, Classifications, Cluster Analysis

INTRODUCTION

In the current days of society, data mining is used in a massive area and many off- the-shelf data mining tools, techniques and procedures are available and sphere of influence data mining application software's are reachable, but data mining in healthcare dataset is a comparatively an infantile research field. Now a day's data mining concept and techniques used to resolve the healthcare problems. In this paper it has been discussed about how data mining techniques are applied in healthcare field. Heart attack is leading causes of death globally. The World Health Organization (WHO) estimates that 7.3 million deaths globally were due to coronary heart disease in 2008 [1]. In recent period of survey, day- to-day numbers of patient consult doctors are increasing. Hence, the doctors, patients, Government and researchers are tiresome to put extra attempt and use numerous techniques in healthcare prediction. As an effect, the data generated in the field of healthcare and the data enhanced day by day. Smart heart disease prediction system can discover and extract hidden knowledge (pattern and relationship) associated with heart disease from an enhanced database generated by the field of healthcare. It can answer complex queries for diagnosing heart disease and thus assist healthcare predictioners to make smart clinical decisions which traditional decision support system cannot. The healthcare industry collects large amounts of data which unfortunately are not "mined" to discover hidden patterns and a relationship often goes unexploited. Advanced Data Mining techniques and Machine Learning algorithms can help remedy this situation.

Medical diagnosis is regarded as in important yet complicated task that needs to be executed accurately and efficiently. The automation of this system would be extremely advantageous. Regrettably all doctors do not possess expertise in every sub specialty and moreover there is a shortage of resource person at certain places. Therefore, an automatic medical diagnosis system would probably be exceedingly beneficial by bringing all of them together [2]. This prototype Heart Disease Prediction using data mining techniques namely, k-Means and Naïve Bayes are reviewed. The system can answer complex "what if" or "why though" queries which traditional decision support system cannot. Using medical profiles such as age, sex, chest pain, blood sugar, blood pressure, cholesterol, ECG measurement, maximum heart rate, thalassemia and number of major vessels it can predict the likelihood of patients getting a heart disease.

It enables significant knowledge e.g. Patterns, relationships between medical factors related to heart disease to be established.

There are number of industries that are using it on a regular basis. Some of these organizations include retail stores, banks, insurance companies, engineering, medicine, crime analysis, expert prediction and web mining [3]. Providing precious services at affordable costs is a major constraint encountered by the healthcare organizations (hospital medical centers). Valuable quality service denotes the accurate diagnosis of patients and providing efficient treatment. Poor clinical decisions may lead to disasters and hence are seldom entertained. Besides, it is essential that the hospitals decrease the cost of clinical test. Appropriate computer – based information and/or decision support system can aid in achieving clinical tests at a reduced cost [4]. Therefore, Data Mining has developed into a vital domain in healthcare. It is possible to predict the efficiency of medical treatment by building the data mining applications. The detection of a disease from several factors or symptoms is a multi-layered problem and might lead to false assumptions frequently associated with erratic effects. Therefore, it appears reasonable to try utilizing the knowledge and experience of several specialists collected in databases towards assisting the diagnosis process [5, 6].

DATA MINING

Data mining refers to extracting or "mining" knowledge from large amounts of data. Many other terms carry a similar or slightly different meaning to data mining, such as knowledge mining from data, knowledge extraction, data/pattern analysis, data archaeology, and data dredging [7]. It uses two strategies: supervised and unsupervised learning.

In supervised learning, a model is able to predict with the help of labelled dataset whereas in unsupervised learning, the algorithm is trained using data that is unlabeled. Each data mining technique serves a different purpose depending on the modeling objective. The two most common modeling objectives are classification and prediction. Classification models predict categorical labels (discrete, unordered) while prediction models predict continuous-valued functions [8]. Decision Tree use classification algorithm such as Naive Bayes while clustering use partitioning method such as k-Means [7].

K-MEANS ALGORITHM

The k-Means Method used for clustering. “Clustering is the process of dividing the dataset into groups, consisting of similar data-points”. An algorithm takes the input parameter, k, and partitions a set of n objects into k clusters so that the resulting intracluster similarity is high but the intercluster similarity is low. Cluster similarity is measured in regard to the mean value of the objects in a cluster, which can be viewed as the cluster’s centroid or center of gravity [7].

Procedure for the Method
Step 1: Select the number of clusters to be identified. i.e., select a value for k=3 in this case.
Step 2: Randomly select 3 distinct data point. Step 3: Measure the distance between the first point and selected three clusters
Step 4: Assign the first point to nearest cluster.
Step 5: Calculate the mean value including the new point for the nearest cluster.

According to the k-Mean algorithm it iterates over again and again unless and until the data points within each cluster stop changes.
The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed with prior intimation.

The main idea is to define k centroids, one for each cluster. In other words, centroids do not move any more.

PROPOSED ALGORITHM: K-MEANS

The k-means algorithm for partitioning, where each cluster’s center is represented by the mean value of the objects in the cluster [7]

Input:
k: the number of clusters, D: a data set containing n objects
Output:
A set of k clusters

Algorithm
1. Arbitrarily choose k objects from D as the initial cluster centers;

2. repeat
3. (re)assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;
4. update the cluster means, i.e., calculate the mean value of the objects for each cluster;
5. until no change;

The Naive Bayes is one of the most simple and powerful algorithms for classification based on Bayes Theorem. It is very easy to build and particularly useful for very large datasets. There are three types in Naïve Bayes model. They are Gaussian, Multinomial, Bernoulli.

Bernoulli : The binomial model is useful if feature vectors are binary (i.e. zeros and ones). So, it is used to predict a patient who has a disease or not.

Bayes theorem provides a way of calculating posterior probability P (c | x) from P(c), P(x) and P (x | c). The equation,

$$P(c \mid x) = \frac{P(x \mid c).\ P(c)}{P(x)}$$

Above,

- P (c | x) is the posterior probability of class (c, target) given predictor (x, attributes.)
- P(c) is the prior probability of class.
- P (x | c) is the likelihood which is the probability of predictor given class.
- P(x) is the prior probability of predictor.

In simple terms, a Naïve Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature

STEPS:

Step 1: Convert the data set into a frequency table

Step 2: Create Likelihood table by finding the probability.

Step 3: Use, Naïve Bayesian equation to calculate the posterior probability for each class. The class with the highest posterior probability is the outcome of prediction.

RESULTS AND FINDINGS

Clustering the patient's age and sex from dataset, here k=3, three clusters are differentiated by using green, blue, red color. And black Star represents centroid of a cluster.

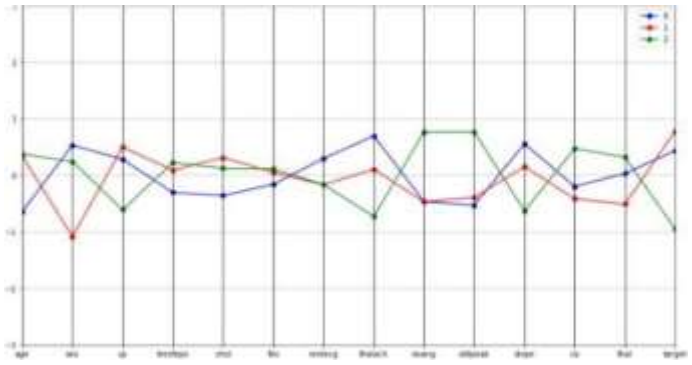

Fig 1. Clustering of Data

Naive Bayes converts a data into following frequency tables,

Table 1. Frequency Table for Gender data

Gender category	+ve Score	-ve Score
M	300	120
F	59	121

Table 2. Frequency Table for Smoking data Categories in Count

Smoking	+ve Rating	-ve Rating
Y	240	180
N	119	61

Similarly, Frequency table is calculated for chest pain, heart rate, Blood pressure, Blood sugar and cholesterol data. Total Frequency is calculated for further use.

Table 3. Frequency Table for Total data

Responses (in Number)	Percentage

Positive	Negative	Total	Positive	Negative
359	**241**	**600**	**0.5983334**	**0.401666**

After submitting user data, the system first forms the cluster and then calculate the ‘yes score’ and ‘no score’ then calculate the probability. Finally provide the result either positive or negative as shown in Fig.2.

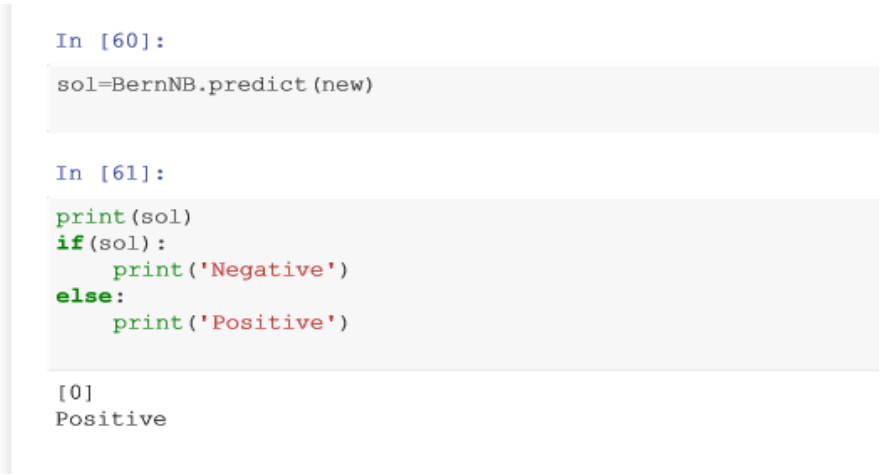

Fig.2 Coding Part

OBSERVED OUTCOME

Heart disease prediction system can serve a training tool to train medical students and nurses to diagnose patients with heart disease. All doctors do not possess expertise in every sub specialty and moreover there is a shortage of resource person at certain places. Therefore, this system helps to them to diagnose patients with heart disease. It can also provide decision support to assist doctors to make clinical decisions or at least provide a “second opinion”.
It only uses two data mining algorithms. Additional data mining techniques can be included to provide better diagnosis. Next constraint is that it only used to diagnose patient with heart disease.

In additional, can improve to show some medical solutions based on their level. Another restriction is that this system based on the 14 attributes. This list may need to be expanded to provide a more comprehensive diagnosis system.

The responses are categorized into positive and negative rating countable. The overall data are fed into frequency table and the performance of the techniques measured. The performance comprised into percentage of frequent sequences with positive and negative ratios. Thus, the tables and the graph predict the response category specifies the clustering centroids with respect to k-means analysis followed by Naïve Bayes axioms towards frequent occurrences for the prediction of heart diseases.

CONCLUSION

Healthcare is one of the most important application areas of Data Mining and Machine Learning. Analyzing a raw healthcare data of heart information will help in the long term saving of human lives and early detection of abnormalities in heart conditions. Clustering and prediction methods help to process raw data and provide a new solution towards heart disease. Heart disease prediction is challenging and very important. However, the mortality rate can be extremely controlled if the disease is detected at the early stages and preventative measures are adopted as soon as possible.

REFERENCES

[1] WHO, World Health Federation, and Global Atlas on Cardiovascular Disease prevention and Control, World Health Organization, 2011.
[2] Chapman, P., Clinton, J., Kerber, R.Khabeza, T., Reinartz, T., Shearer, C., Wirth, R.: "CRISP- DM 1.0: Step by Step Data Mining Guide", SPSS, 1- 78, 2000.
[3] Mrs. Bharati, M. Ramageri: "Data Mining Techniques and Applications", Mill Valley, CA, 2015
[4] Fayyad, U: "Data Mining and Knowledge Discovery in Databases: Implications from scientific databases", Proc.of the 9th Int. Conf. on Scientific and Statistical Database Management, Olympia, Washington, USA, 2-11, 1997.
[5] Mehmed, K.: "Data mining: Concepts, Models, Methods and Algorithms", New Jersey: John Wiley, 2003.
[6] Weiguo, F., Wallace, L., Rich, S., Zhongju, Z.: "Tapping the Power of Text Mining", Communication of the ACM. 49(9), 77-82, 2006.
[7] Jiawei Han and Micheline Kamber: "Data Mining Concepts and Techniques", Second Edition, MK Publishers Elsevier, USA, 2014
[8] Han, J., Kamber, M.: "Data Mining Concepts and Techniques", Morgan Kaufmann Publishers, 2006.

How to cite this article:
G Priyanka, Dr P Revathi, "Performance Analysis on Hybrid Data Mining Techniques for Heart Disease Prediction", International Journal of Intelligent Computing and Technology (IJICT), Vol.5, Iss.1, pp.27-34, 2021