Volume 4 Issue III, March 2016
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Data Mining Techniques to Find Out Heart Diseases: An Overview

Mayuri Takore[1], Prof.R.R. Shelke[2]
[1]ME First Yr. (CSE) , [2]Assistant Professor Computer Science & Engg, Department
H.V.P.M's COET, Amravati

Abstract--Heart disease is a major cause of morbidity and mortality in modern society. Medical diagnosis is extremely important but complicated task that should be performed accurately and efficiently. Although significant progress has been made in the diagnosis and treatment of heart disease, further investigation is still needed. The availability of huge amounts of medical data leads to the need for powerful data analysis tools to extract useful knowledge. There is a huge data available within the healthcare systems. However, there is a task of effective analysis tools to discover hidden relationships and trends in data. Knowledge discovery and data mining have found numerous applications in business and scientific domain. Researchers have long been concerned with applying statistical and data mining tools to improve data analysis on large data sets. Disease diagnosis is one of the applications where data mining tools are proving successful results. This research paper proposed to find out the heart diseases through data mining, Support Vector Machine (SVM), Genetic Algorithm, rough set theory, association rules and Neural Networks.

Keywords--Data Mining, Heart Disease, SVM, rough sets techniques, association rules & clustering.

I. INTRODUCTION

A. Overview of Data Mining

Knowledge discovery in databases is well-defined process consisting of several distinct steps. Data mining is the core step, which results in the discovery of hidden but useful knowledge from massive databases. A formal definition of Knowledge discovery in databases is given as follows:Data mining is the non trivial extraction of implicit previously unknown and potentially useful information about data. Data mining technology provides a user-oriented approach to novel and hidden patterns in the data.The discovered knowledge can be used by the healthcare administrators to improve the quality of service. The discovered knowledge can also be used by the medical practitioners to reduce the number of adverse drug effect, to suggest less expensive therapeutically equivalent alternatives. Anticipating patient's future behavior on the given history is one of the important applications of data mining techniques that can be used in health care management.

A major challenge facing healthcare organizations (hospitals, medical centers) is the provision of quality services at affordable costs. Quality service implies diagnosing patients correctly and administering treatments that are effective. Poor clinical decisions can lead to disastrous consequences which are therefore unacceptable. Hospitals must also minimize the cost of clinical tests.

They can achieve these results by employing appropriate computer-based information and/or decision support systems. Health care data is massive. It includes patient centric data, resource management data and transformed data. Health care organizations must have ability to analyze data. Treatment records of millions of patients can be stored and computerized and data mining techniques may help in answering several important and critical questions related to health care.

Volume 4 Issue III, March 2016
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

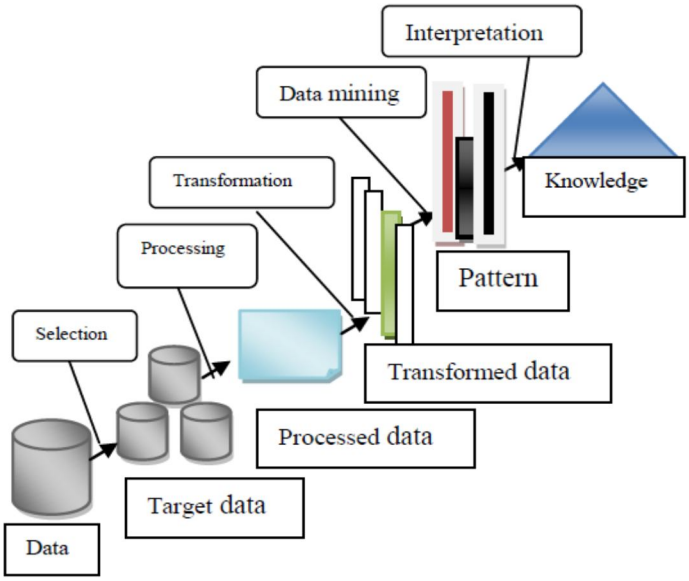

Fig. 1: KDD Process

Once these patterns are found they can further be used to make certain decisions for development of their businesses. Three steps involved are

Exploration

Pattern identification

Deployment

Exploration: In the first step of data exploration data is cleaned and transformed into another form, and important variables and then nature of data based on the problem are determined. Pattern Identification**:** Once data is explored, refined and defined for the specific variables the second step is to form pattern identification. Identify and choose the patterns which make the best prediction. Deployment: Patterns are deployed for desired outcome.

B. Causes and Impact of Heart Diseases

According to WHO report Global atlas on cardiovascular disease prevention and control states that cardiovascular disease (CVDs) are the leading causes of death and disability in the world. Although a large proportion of CVDs is preventable, they continue to rise mainly because preventive measures are inadequate. Over 17.3 million An estimated 17.3 million people died from CVDs in 2008, Over 80% of CVD deaths take place in low- and middle-income countries, 23.6 million By 2030, almost 23.6 million people will die from CVDs.

1) Protect Heart Health: Tobacco use, an unhealthy diet, and physical inactivity increase the risk of heart attacks and strokes. Engaging in physical activity for at least 30 minutes every day of the week will help to prevent heart attacks and strokes.

Eating at least five servings of fruit and vegetables a day, and limiting your salt intake to less than one teaspoon a day, also helps to prevent heart attacks and strokes.

2) Cardiovascular Diseases (Cvds) Key Facts: CVDs are the number one cause of death globally: more people die annually from CVDs than from any other cause.

An estimated 17.3 million people died from CVDs in 2008, representing 30% of all global deaths. Of these deaths, an estimated 7.3 million were due to coronary heart disease and 6.2 million were due to stroke.

Low- and middle-income countries are disproportionally affected: over 80% of CVD deaths take place in low- and middle-income countries and occur almost equally in men and women.

By 2030, almost 23.6 million people will die from CVDs, mainly from heart disease and stroke. These are projected to remain the single leading causes of death.

3) Cardiovascular Diseases: Cardiovascular disease is caused by disorders of the heart and blood vessels, and includes coronary

heart disease (heart attacks), cerebrovascular disease (stroke), raised blood pressure (hypertension), peripheral artery disease, rheumatic heart disease, congenital heart disease and heart failure. The major causes of cardiovascular disease are tobacco use, physical inactivity, an unhealthy diet and harmful use of alcohol. These are the three causes of heart diseases (1) chest pain (2) stroke and (3) heart attack.To prevent and identification of these diseases different techniques of data mining is used through this easily find out heart related diseases and this is the aim of this research studies. Heart disease is the leading cause of death all over the world in the past ten years. Several researchers are using statistical and data mining tools to help health care professionals in the diagnosis of heart disease.

II. DATA MINING ALGORITHMS AND TECHNIQUES

Various algorithms and techniques like Classification, Clustering, Regression, Artificial Intelligence, Neural Networks, Association Rules, Decision Trees, Genetic Algorithm, Nearest Neighbor method etc., are used for knowledge discovery from databases.

A. Classification

Classification is the most commonly applied data mining technique, which employs a set of pre-classified examples to develop a model that can classify the population of records at large. Fraud detection and credit risk applications are particularly well suited to this type of analysis. This approach frequently employs decision tree or neural network-based classification algorithms. The data classification process involves learning and classification. In Learning the training data are analyzed by classification algorithm. In classification test data are used to estimate the accuracy of the classification rules. If the accuracy is acceptable the rules can be applied to the new data tuples. For a fraud detection application, this would include complete records of both fraudulent and valid activities determined on a record-by-record basis. The classifier-training algorithm uses these pre-classified examples to determine the set of parameters required for proper discrimination. The algorithm then encodes these parameters into a model called a classifier. Types of classification models:

Classification by decision tree induction

Bayesian Classification

Neural Networks

Support Vector Machines (SVM)

Classification Based on Associations

B. Clustering

Clustering can be said as identification of similar classes of objects. By using clustering techniques we can further identify dense and sparse regions in object space and can discover overall distribution pattern and correlations among data attributes. Classification approach can also be used for effective means of distinguishing groups or classes of object but it becomes costly so clustering can be used as preprocessing approach for attribute subset selection and classification. For example, to form group of customers based on purchasing patterns,to categories genes with similar functionality. Types of clustering methods

Partitioning Methods

Hierarchical Agglomerative(divisive) methods

Density based methods

Grid-based methods

Model-based methods

C. Predication

Regression technique can be adapted for predication. Regression analysis can be used to model the relationship between one or more independent variables and dependent variables. In data mining independent variables are attributes already known and response variables are what we want to predict. Unfortunately, many real-world problems are not simply prediction. For instance, sales volumes, stock prices, and product failure rates are all very difficult to predict because they may depend on complex interactions of multiple predictor variables. Therefore, more complex techniques (e.g., logistic regression, decision trees, or neural nets) may be necessary to forecast future values. The same model types can often be used for both regression and classification. For example, the CART (Classification and Regression Trees) decision tree algorithm can be used to build both classification trees (to classify categorical response variables) and regression trees (to forecast continuous response variables). Neural networks too can

create both classification and regression models. Types of regression methods

Linear Regression

Multivariate Linear Regression

Nonlinear Regression

Multivariate Nonlinear Regression

D. Association Rule

Association and correlation is usually to find frequent item set findings among large data sets. This type of finding helps businesses to make certain decisions, such as catalogue design, cross marketing and customer shopping behavior analysis. Association Rule algorithms need to be able to generate rules with confidence values less than one. However the number of possible Association Rules for a given dataset is generally very large and a high proportion of the rules are usually of little (if any) value. Types of association rule

Multilevel association rule

Multidimensional association rule

Quantitative association rule

III. SURVEY OF LITERATURE (DIFFERENT DATA MINING TECHNIQUES TO FIND OUT HEART DISEASES)

A. Decision Tree Classification Algorithm

Heart disease or coronary artery disease (CAD) or coronary heart disease (CHD) or ischemic heart disease (IHD) is a broad term that can refer to any condition that affects the heart. For developing clinical decision support systems, literature presents a number of researches that have made use of artificial intelligence and data mining techniques. Till now, several studies have been reported on heart disease diagnosis. These studies have applied different approaches to the given problem and achieved high classification accuracies, of 77% or higher, using the dataset taken from the UCI machine learning repository.

B. UCI Database Description about Decision Tree Classification

The heart disease database from the University of California Irvine. UCI archive is used. This database contains four data sets from the Cleveland Clinic Foundation, Hungarian Institute of Cardiology, V.A. Medical Center and University Hospital of Switzerland. It provides 920 records in total. Originally, the database had 76 raw attributes. However, all of the published experiments only refer to 13 of these: Age, Sex, P, Trstbps, Chol, Fbs, estecg, Thalach, Exang, OldPeak, Slope, Ca, Thal and Num.

C. Clustering D.M. Technique Using K- Means Algorithms

The categorization of objects into various groups or the partitioning of data set into subsets so that the data in each of the subset share a general feature, frequently the proximity with regard to some defined distance measure, is known as Clustering. The clustering problem has been identified in numerous contexts and addressed being proven beneficial in many medical applications. Clustering the medical data into small with meaningful data can aid in the discovery of patterns by supporting the extraction of numerous appropriate features from each of the clusters thereby introducing structure into the data and aiding the application of conventional data mining techniques. Numerous methods are available in the literature for clustering and employed the renowned K-Means clustering algorithm in this approach.

The steps involved in a k-means algorithm are given subsequently: Prediction of heart disease using K – Means clustering technique

K points denoting the data to be clustered are placed into the space. These points denote the primary group centurions.

The data are assigned to the group that is adjacent to the centurion.

The positions of all the K centroids are recalculated as soon as all the data are assigned.

Steps 2 and 3 are reiterated until the centroids stop moving any further. This results in the segregation of data into groups from which the metric to be minimized can be deliberated.

D. Advantages To Using This Technique

The main advantages of this algorithm are its simplicity and speed which allows it to run on large datasets.

With a large number of variables, K-Means may be computationally faster than hierarchical clustering (if K is small).

K-Means may produce tighter clusters than hierarchical clustering, especially if the clusters are globular.

E. Data Mining Through Genetic Algorithms

We start out with a randomly selected first generation. Every string in this generation is evaluated according to its quality, and a fitness value is assigned. Next, a new generation is produced by applying the reproduction operator. Pairs of strings of the new generation are selected and crossover is performed. With a certain probability, genes are mutated before all solutions are evaluated again. This procedure is repeated until a maximum number of generations are reached. While doing this, the all time best solution is stored and returned at the end of the algorithm.

Genetic algorithm have been used in, to reduce the actual data size to get the optimal subset of attributed sufficient for heart disease prediction. Classification is a supervised learning method to extract models describing important data classes or to predict future trends. Three classifiers e.g. Decision Tree, Naïve Bayes and Classification via clustering have been used to diagnose the presence of heart disease in patients. Pairs of strings of the new generation are selected and crossover is performed. With a certain probability, genes are mutated before all solutions are evaluated again. This procedure is repeated until a maximum number of generations are reached.

F. Classification via Clustering

Clustering is the process of grouping similar elements. This technique may be used as a preprocessing step before feeding the data to the classifying model. The attribute values need to be normalized before clustering to avoid high value attributes dominating the low value attributes. Further, classification is performed based on clustering. Experiments were conducted with Weka 3.6.0 tool. Data set of 909 records with 13 attributes. All attributes are made categorical and inconsistencies are resolved for simplicity. To enhance the prediction of classifiers, genetic search is incorporated. Observations exhibit that the Decision Tree data mining technique outperforms other two data mining techniques after incorporating feature subset selection but with high model construction time. Naïve Bayes performs consistently before and after reduction of attributes with the same model construction time.

IV. CONCLUSION

This paper examines the classification techniques in data mining and shows the performance of classification among them. In these classifications accuracy among these data mining techniques has discussed. The result shows the difference in error rates. However there are relatively differences in different techniques. Decision tree and SVM perform classification more accurately than the other methods. Data mining application in heart disease Author name et.al. reported that the major advantage of data mining technique shows the 92.1 % 91.0 % accuracy for the heart disease. We suggest that the age, sex, chest pain, blood pressure, personnel history, previous history, cholesterol, fasting blood sugar, resting ECG, Maximum heart rate, slope, etc. that may be used as reliable indicators to predict presence of heart disease. We also suggest that data should be explored and must be verified from the team of heart disease specialist doctors. In future, we will try to increase the accuracy for the heart disease patient by increasing the various parameters suggested from the doctors by using different data mining techniques.

V. ACKNOWLEDGEMENT

First of all we would especially like to express sincere gratitude to our parents. It gives us great pleasure and satisfaction in presenting the paper on **"DATA MINING TECHNIQUES TO FIND OUT HEART DISEASES: AN OVERVIEW "** Before we get into the depth of the things, we show our sincere gratitude towards respected teachers, guide, colleagues and all who have directly or indirectly helped us in the completion of this paper successfully

REFERENCES

[1] Mrs. Bharati M. Ramageri, :-Data Mining Techniques And Applications‖,Bharati M. Ramageri / Indian Journal of Computer Science and Engineering Vol. 1 No. 4 301-305
[2] Heart disease from http://www.nhlbi.nih.gov/educational/hearttruth/lower-risk/what-is-heart-disease.html
[3] Bala Sundar V, Bharathiar:-Development of a Data Clustering Algorithm for Predicting Heart‖ International Journal of Computer Applications (0975 – 888) Volume 48– No.7, June 2012
[4] R. Gupta, V. P. Gupta, and N. S. Ahluwalia, :-Educational status, coronary heart disease, and coronary risk factor prevalence in a rural population of India‖, BMJ. pp 1332–1336, 19 November 1994.
[5] Panniyammakal Jeemon & K.S. Reddy, :-Social determinants of cardiovascular disease outcomes in Indians‖, pp 617-622, November 2010.

