Data Mining Techniques to Find out Heart Diseases

Lintamol Jose
St.Joseph's College,
Irinjalakuda

***Abstract*--Heart diseases is the major cause of mortality in modern world. Today heart diseases are the number one cause of death. Medical diagnosis is important but it is complicated task to perform. A good investigation is needed to made significant improvements in the diagnosis and treatment of heart diseases. A huge data is available within the health care system is needed to extract useful knowledge from it. Discovery of knowledge and data mining have found numerous applications in business and scientific field. Unfortunately there is an unavailability of effective analysis tools to discover hidden relationships and trends in data. Data mining tools are also providing successful result in the field of disease diagnosis. This research paper proposed to find out the heart diseases through data mining support genetic algorithm, association rules, neural networks and predication. Most effective tools to heart diseases in genetic algorithm and neural networks and support vector machine. There for it is observed that, the data mining could help in the prediction of high or low risk heart diseases.**

Keywords—Genetic algorithm; neural Networks; association rules; predication; Support vector machine.

I. INTRODUCTION

Data mining is the knowledge discovery in databases; it is the process of finding interesting and useful patterns and relationships in large volumes of data. Data mining is widely used in the field of business and scientific field. Data mining technology provides a user-oriented approach to hidden patterns in the data. The resulting knowledge can be used by the healthcare administrators to improve the quality of service. This knowledge can also be used in the medical field to reduce the number of side effects of drug, and to suggest therapeutically equivalent alternatives in low cost. Decisions are often made based on doctors' experience rather than on the
knowledge rich data hidden in the database. This practice leads to errors and excessive medical costs which affects the quality of service provided to patients. data mining can lead to significant changes in the quality of clinical based decisions. Data mining is the process of mining useful knowledge and from large amount of data. It is a is a logical process that is used to search through large amount of data in order to find useful data. The main goal of this technique is to find data that were previously unknown.

Heart disease is the first leading cause of death in high and low income countries and occurs almost equally in men and women. they continue to rise mainly because preventive measures are inadequate. An estimated Over 80% of cardiovascular disease deaths take place in low- and middle-income countries. Tobacco use, an unhealthy diet increases the possibility of heart attacks and strokes. Eating of fruit and vegetables , and limiting the use of salt , also helps to prevent heart attacks and strokes. Heart disease is caused by disorders of the heart and blood vessels. Heart attacks, hypertension, , stroke, peripheral artery disease, rheumatic heart disease and heart failure are also included in this. Use of tobacco and use of alcohol are the major causes of cardio vascular diseases.
Three causes of heart diseases are:
(1) Chest pain
(2) Stroke
(3) Heart attack
The aim of this research studies is to prevent and identification of heart diseases through different techniques of data mining so we can easily find out heart related diseases .

II.DATA MINING ALGORITHMS AND TECHNIQUES

Various algorithms and techniques like Classification, Clustering, Regression, Neural Networks, Association Rules, Genetic Algorithm etc., are used for knowledge discovery from databases.

A.Classification

Classification is the most commonly used data mining technique, which produces a set of pre-classified examples to develop a model that can classify the population of records at largeThis approach results decision tree or neural network-based classification algorithms. The data classification process involves learning and classification. In Learning the data are analyzed by classification algorithm. The data are used to estimate the accuracy of the classification rules. Rules can be applied to the new data set if the accuracy is acceptable. The algorithm uses these pre-classified examples to check the set of parameters required for proper description. The algorithm then encodes these parameters into a model called a classifier. Some of the classification models:

- Neural Networks
- Classification Based on Associations

B. Clustering

Clustering is the process of identification of similar classes of objects. We can
easily identify the differences between groups or classes. In this technique, categories genes with similar functionality. For example, to form group of
customers based on purchasing patterns. Thus we can identify the correlations among data attributes.

Types of clustering methods

- Density based methods
- Grid-based methods
- Model-based methods

C. Predication

predication analysis used to find the relationship between one or more independent variables and dependent variables. From this relationship we can predicts . But sometime it cannot be used in many real-world problems and cannot simply predict. For example, 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.

Types of predication methods

- Multivariate Linear Regression
- Nonlinear Regression
- Multivariate Nonlinear Regression

D. Association rule

Association is usually to find frequent item set 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. The disadvantage of association rule is that 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

E. Neural networks

Neural network is a set of connected input or output units. For each connection a weight present with it. After the learning phase, we can adjusting the weights so as to be able to predict the correct class labels of the input data sets. The main advantage of this technique is that Neural networks have the ability to derive meaning from complicated or undefined data and can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques. It can be applied in many many real world business problems and have already been successfully applied in many
industries. Nowadays neural networks are used in an increasing number of applications.It is used as the best identification pattern or trends in data and is suited for prediction of knowledge.

F. Genetic algorithm

In genetic algorithm we randomly selected first generation. According to its quality, every generation is evaluated and a fitness value is assigned. Then by applying the reproduction operator, a new generation is produced. Pairs of strings of the new generation are selected and crossover is performed. Next step is to perform mutation of genes. After genes are mutated all solutions are evaluated again. This process is repeated until a maximum number of generations are reached. After 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 an optimal subset of attributes and it is sufficient for heart disease prediction. Classification is a supervised learning method to extract models 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.

III. LITERATURE SURVEY

The paper titled *"Data mining: concepts and techniques"* covers the basic information about the concept of data mining and its various techniques applied to various fields so thereby implementing many improving solutions to large problems for implementing data mining and minimizing the causes and leading to good solutions.

The paper titled *"Application of data mining techniques in health care data"* covers the A high-level introduction to data mining as it relates to surveillance of healthcare data is presented. Data mining is compared with traditional statistics, some advantages of automated data systems are identified, and some data mining strategies and algorithms are described. A concrete example illustrates steps involved in the data mining process, and three successful data mining applications in the healthcare arena are described.

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

The categorization of objects into various groups 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 is beneficial in many medical applications. Clustering the medical data extracts small with meaningful data . Numerous methods are available in the literature for clustering and one of this is K-Means clustering algorithm. The k-means algorithm is one of the widely recognized clustering tools that are applied in a variety of scientific and industrial applications. k-means groups the data in accordance with their characteristic values into k distinct clusters. Data categorized into the same cluster have identical feature values. k, the positive integer denoting the number of clusters, needs to be provided in advance. The steps involved in a k-means algorithm are given by:

1. K points denoting the data to be clustered are placed into the space. These points denote the primary group centurions.
2. The data are assigned to the group that is adjacent to the centurion.
3. The positions of all the K centroids are recalculated as soon as all the data are assigned.
4. Steps 2 and 3 are reiterated until the centroids stop moving any further.

B. K-Means and derivatives

The k-Means algorithm assigns each point to the cluster whose center (also called centroid) is nearest. The center is the average of all the points in the cluster — that is, its

coordinates are the arithmetic mean for each dimension separately over all the points in the cluster.

Example: The data set has three dimensions and the cluster has two points X = (x1, x2, x3) and Y = (y1, y2, y3).
Then Z becomes Z = (z1, z2, z3), where z1= x1+y1 , z2= x2+y2 and z3= x3+y3

Advantages

The main advantages of this algorithm are its simplicity and speed which allows it to run on large datasets. b. With a large number of variables, K-Means may be computationally faster than hierarchical clustering (if K is small). c. K-Means may produce tighter clusters than hierarchical clustering, especially if the clusters are globular.[3]

C. 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 [5], 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 diseases.

The paper titled "*Data Mining Techniques to Find out Heart Diseases: an Overview*" covers the different causes and prevention of heart diseases, it briefly explained about the various data mining techniques to find out heart diseases. For example clustering, neural networks, association rules,..etc are the different types of data mining techniques.

The paper titled *"Data Mining Approach to Detect Heart Dieses"* explains the several data mining techniques proposed in recent years for the diagnosis of heart disease. Neural Networks, Bayesian Classification based on clustering, Decision Tree, Genetic Algorithm. Naive Bayes, Decision tree, WAC which are showing accuracy at different levels.

IV. PROPOSED SYSTEM

Support Vector Machines, one of the new techniques for pattern classification, have been widely used in many application areas. Support vector machines (SVMs) have shown superior performance compared to other machine learning techniques, especially in classification problems. SVMs are considered a supervised computer learning method because they exploit prior knowledge of gene function to identify unknown genes of similar function from expression data. SVMs avoid several problems associated with unsupervised clustering methods, such as hierarchical clustering. SVMs have many mathematical features that make them attractive for gene expression analysis, including their flexibility in choosing a similarity function, sparseness of solution when dealing with large data sets, the ability to handle large feature spaces, and the ability to identify outliers. We test several SVMs that use different similarity metrics, as well as some other supervised learning methods, and find that the SVMs best identify sets of genes with a common function using expression data. Finally, we use SVMs to predict functional roles for uncharacterized yeast ORFs based on their expression data[4].

V. CONCLUSION

This paper describes the classification techniques in data mining and shows the performance of classification among them. Accuracy among data mining techniques has explained in these studies. The result shows the different accuracy of data mining methods. There are many differences in the different techniques. Combining with other methods SVM perform classification more accurately. Data mining application in heart disease reported that the major advantage of data mining technique shows the 92.1 % accuracy for the heart disease. Age, sex, chest pain, blood pressure, personnel history, previous history, cholesterol, raising blood pressure, ECG, Maximum heart rate, etc. are the reliable information that can be used to predict presence of heart disease. We also suggest that data should be distributed and must be verified from the team of heart disease specialist doctors. In future, we will try to increase the accuracy of finding heart diseases for the patient by increasing the various parameters suggested from the doctors by using different data mining techniques.

ACKNOWLEDGMENT

I would like to express my sincere gratitude to
Ms.Nisha Peter and all faculty members of Department of computer science. St.Joseph's college, Irinjalakuda.

REFERENCES

[1] Aqueel Ahmed, Shaikh Abdul Hannan, "Data Mining Techniques to Find Out Heart Diseases:An Overview" International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-1, Issue-4, September 2012
[2] 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] http://heart-disease.emedtv.com/ coronary-artery-disease/coronary-artery-disease.html
[3] Vikas Chaurasia,Saurabh Pal, **"**Data Mining Approach to Detect Heart Dieses" , International Journal of Advanced Computer Science and Information Technology (IJACSIT) Vol. 2, No. 4, 2013, Page: 56-66, ISSN:2296-1739© Helvetic Editions LTD, Switzerland www.elvedit.com
[4] The paper titled "support vector machine introductory overview" http://www.statsoft.com/Textbook/Support-Vector-Machines
[5] http://heart-disease.emedtv.com/coronary-artery-disease/coronary-artery-disease.html