Prognosis of Heart Disease Using Knowledge Discovery In Data (KDD)

Mrs. B. Meena Preethi[1], Sasi Revathi. N[2], Divya Dharani. K[3]
[1, 2, 3] Dept of BCA & M.Sc(SS),
[1, 2, 3] Sri Krishna College of Arts & Science

Abstract- *Data mining is a process to analyses a huge amount of data sets and then extracts the meaning of the entire or a particular data. It helps a person to predict the designs and future trends, allowing business to involve in decision making. Data mining applications are able to give the answer to any number of business questions which can take much time to resolve using traditional methods. Huge amount of data that can be generated for the prediction of disease is analyzed using traditional methods and is too complicated along with vast amount to be treated using traditional methods. Data mining provides methodologies and methods for transformation of the data into useful information for decision making. These methods can make process fast and take less time to predict the heart disease with more accuracy. The healthcare sector amass a huge quantity of healthcare data which cannot be mined, to uncover hidden information for conclusive decision making. However, there is a plenty of hidden information in this data which is untapped and not being used appropriately for predictions. It becomes more influential in case of heart disease that is considered as the predominant reason behind death all over the world. In medical field, Data Mining provides many methodologies which are widely used in the medical and clinical decision support systems which should be helpful for diagnosis and predicting of various diseases. These data mining methods can be used in heart diseases prediction involves less amount of time and makes the process much quicker for the predicting diseases with good speed to improve their health. This paper involves different algorithms in which one or more algorithms of data mining used for the prediction of heart disease. By Applying data mining methods to heart disease data which requires to be processed, we can get effective results and achieve reliable performance which will help in decision making in healthcare industry. It will help the medical practitioners to diagnose the disease in less time and predict probable complications well in advance. Identify the major risk factors of Heart Disease categorizing the risk factors in an order which causes damages to the heart such as diabetes, high blood cholesterol, obesity, hyper tension, smoking, poor diet, stress, etc. Data mining methods and functions are used to identify the level of risk factors which helps the patients to take precautions in advance to save their life.*

Keywords- Data mining, heart disease prediction, KNN, ANN, decision tree, SVM

I. INTRODUCTION

Data mining is the analytical process to explore specific data from large volume of data. It is a process that finds previously unknown designs and trends in databases. This information is further used to build predictive models. The main objective is to learn the different data mining methods/algorithms which are used in the prediction of heart diseases using any data mining tool. Heart is the most vital part of the human body as life is dependent on efficient working of heart. A Heart disease is caused due to narrowing or blockage of coronary arteries. This is caused by the deposition of fat on the inner walls of the arteries and also due to build up cholesterol. Heart diseases can be caused due to number of factors:

A. High blood pressure

When the heart pumps blood, the force of the blood pushes against the walls of the arteries causing pressure. If the pressure rises and stays high over the time it is called high blood pressure or hypertension which can harm the body in many ways i.e. increasing the risk of heart stroke or developing heart failure, kidney failure etc.

B. High cholesterol

Cholesterol is a waxy substance found in the fatty deposits in the blood vessels. Increase in the fatty deposits (high cholesterol) does not allow sufficient blood to flow in through the arteries causing heart attacks.

C. Unhealthy diet

Eating too much fast food increases blood pressure and cholesterol level causing the risk of heart attacks.

D. Smoking

It damages the lining of arteries and builds up a fatty material called atheroma which narrows the arteries causing heart attacks.

E. Lack of physical activity

Lack of exercise increases cholesterol level in blood vessels which further increases the risk of heart attacks.

F. Obesity

Obese people are more likely to have high blood pressure, high cholesterol level and diabetes (increase in blood sugar level) which increases the risk of heart strokes in human body. Nowadays, data mining is gaining popularity in health care industry as this industry generates large amount of complex data about hospital resources, medicines, medical devices, patients, disease diagnosis etc. This complex data needs to be processed and analysed for knowledge extraction which will further help in decision making and is also cost effective.

World health organization has estimated 17.5 million people died from cardio vascular diseases in 2012, representing 31 percent of all global deaths. Out of these, an estimated 7.4 million were due to coronary heart disease and 6.7 million were due to stroke. WHO estimated by 2030, almost 23.6 million people will die due to heart disease as written in [1].

Thus, a beneficial way to predict heart diseases in health care industry is an effective and efficient heart disease prediction system. This system will find human interpretable designs and will determine trends in patient records to improve health care.

II. DATA MINING APPLICATIONS

Data mining is used in various fields such as retail industry, telecommunication industry, healthcare industry, financial data analysis, intrusion detection, sports and also in analyzing student's performance.

A. Retail Industry

Data mining is a great application in retail industry as it collects large amount of data which includes transportation, sales and consumption of goods and services. This data expand rapidly due to increase in purchase and sales in business. Data mining helps to identify customer's buying designs and trends that lead to improved quality of customer service and customer's satisfaction.

B. Telecommunication Industry

Telecommunication industry is the most growing industry as it provides various services such as fax, pager, cellular phones and e-mails.

C. Healthcare Industry

Data mining is very useful in healthcare industry in diagnosis of heart diseases, breast cancer and diabetes. It helps in finding patterns and trends in patient's history having the similar risk factor and helps in making decisions.

D. Financial Data Analysis

Financial data in banking is reliable and of high quality which facilitates systematic data analysis in financial industry. It helps in loan payment prediction and customer credit policy analysis. It also helps in clustering of customers for target marketing.

Intrusion Detection: Intrusion is any kind of action that threatens the confidentiality or integrity of network resources from any outside party. With the increased usage of internet and availability of the tools and tricks of intrusion and attacking network, intrusion detection has become an important issue for network administration.

E. Sports

In sports, vast amount of statistics are gathered for each player, team, game and season. Data mining is used in the prediction of performance of players, selection of players and forecast of future events.

F. Student's Performance

Data mining is used to evaluate student's performance using classification method for data classification. Attendance, class test, seminar and assignment marks are collected from the student record to predict the performance of the student at the end of the semester.

III. COMPARATIVE STUDY OF DATA MINING TRENDS FROM PAST TO FUTURE

I. PAST

In the previous years, statistical and Machine learning methods were used on numerical data stored in traditional databases and the computing resources were 4G PL and various related methods.

II. PRESENT

These days, along with the statistical and Machine learning methods, artificial intelligence and design re-organisation methods are also used.

III. FUTURE

In future, for complex data objects which includes high dimensional, high speed data streams, sequence, noise in the time series and for multi instance objects, soft computing methods like fuzzy logic, neural networks and genetic programming is used. Computing resources used would be multi-agent technologies and cloud computing.

IV. DATA MINING METHODOLOGIES

There are various algorithms which can be used to predict the heart disease with accurate data.

A. MAFIA

MAFIA stands for Maximal Frequent Itemsets. Item set of frequent is one of the fundamental data mining problems which has an aim to discover the count of items which include the frequently used itemset in a dataset. The major goal is to find interesting designs from data warehouse in number of data mining tasks like as association rules, cluster classifiers, sequence and many more. The newest method MAFIA exploits an effective algorithm which combines the ideas of old and latest algorithm to configure a realistic algorithm. The algorithm can also use for maximal mining frequent item set for searching with effective pruning algorithm.

B. K-means Clustering

It is an iterative and one of the best unsupervised learning algorithm to divide a given set of data in predefined set of k cluster where k is marked as input variable to solve the conventional clustering problem. The K-means algorithm is a method and frequent to use in medical area and their associated fields. K-means clustering chooses points in multidimensional space to symbolize each k cluster called centroids. A centroid is the point whose ordinates are accepted by evaluating the average of each co-ordinates of samples point that allocated the clusters. The major objective of using k-means clustering to emphasize the overall squared error function or intra-cluster deviation.

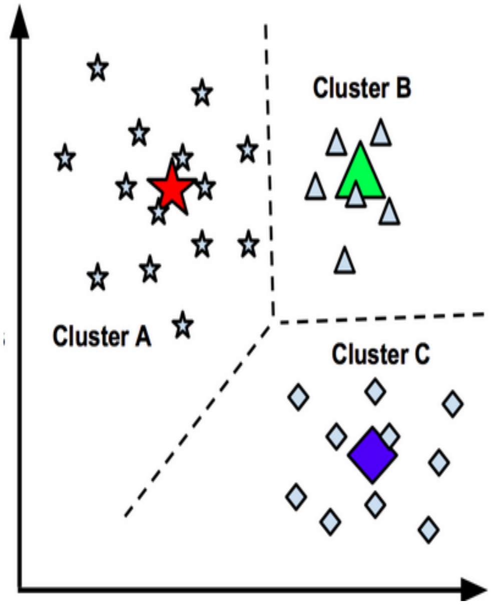

FIG 1: EXAMPLE OF K-MEANS CLUSTERING

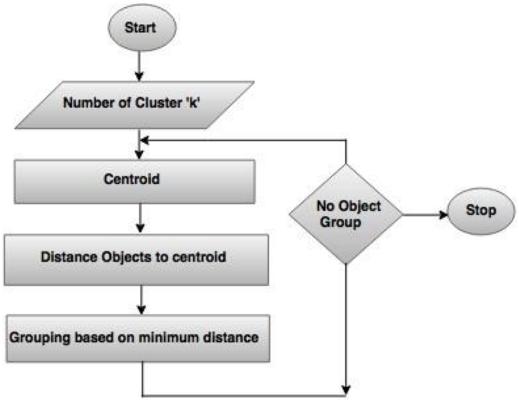
Start
Number of Cluster 'k'
Centroid
Distance Objects to centroid
Grouping based on minimum distance
No Object Group
Stop

FIG 2: K-MEANS CLUSTERING ALGORITHM

C. Artificial neural networks

ANN algorithm is used in computer science and other research disciplines, which is based upon a huge selection of simple neural units (artificial neurons), loosely analogous to the observed behavior of a biological brain's axons. Each neural unit is joined with many others, and links can prevent the energizing state of adjacent neural units. Each individual neural unit computes using outline function. These systems are self-learning and trained, rather than notably programmed, and excel in areas where the solution or feature detection is difficult to express in a traditional computer program.

V. VARIOUS DATA MINING TOOLS USED IN HEART DISEASE PROGNOSIS SYSTEM WITH VERACITY

Abhishek et al in the year 2013 used data mining tool Weka 3.6.4 in heart disease prognosis system using J48 method accomplished 95.56% veracity and using Naïve Bayes accomplished 92.42%.

Rashedur et al in the year 2013 used Neural network method using Weka data mining tool and accomplished 79.19% and to compare various categorizing methods, he used another method fuzzy logic using TANGRA data mining tool and accomplished 83.85% veracity.

Nidhi et al in the year 2012 used data mining tool Weka 3.6.6 in the analysis of heart disease prognosis system and accomplished 99.52% using Naive Bayes. She also used TANGRA data mining tool but could accomplish up to 52.33% only using decision trees. She also tried .NET data mining tool and accomplished up to 96.5% using neural networks.

Resul et al in the year 2009 used SAS base software 9.1.3 achieving 97.4% using neural networks.

VI. DATA MINING METHODS ALSO USED IN DIAGNOSIS OF OTHER DISEASES

Humar et al in 2008 used categorizing, Back propagation, Fuzzy neural network methods for diabetes and heart diseases.

Marcel et al in the year 2007 used Bayesian Categorizing for Characinoid heart disease. Mohammad et al in the year 2012 used C4.5 and C5.0 algorithm for heart disease and breast cancer diagnosis.

VII. DATA MINING AND ITS METHODS

It is main concerned with extracting useful information from large amount of databases. Data mining methods and tools are used to discover unknown designs and trends from the data set. Its main objective is to automatically discover the designs in the dataset with minimal user effort and input. Data mining’s main contribution is in decision making and in forecasting future trends of market. Many organisations use data mining as a tool these days for data analysis as it easily valuates designs and trends of market and produce producting results. Data mining methods are:

A. Association

It is a well known and well researched method for data mining. Association is also known as relation method because designs which are discovered from the dataset depends on the relationship between the items listed.

B. Categorizing

It is a data mining method which is used to classify each item in a data set into one of predefined set of classes or groups. It is a classic data mining method which is based on machine learning.

C. Clustering

It is a data mining method which creates useful cluster of objects that have similar characteristics using automatic method. There is a slight difference between clustering and categorizing.

D. Prognosis

It is a data mining methodology which is used to discover the relationship between independent variables and the relationship between dependent and independent variables.

E. Sequential Designs

It is a data mining methodology that is used to find similar designs or regular events in transaction of data over a business period of time.

VIII. DATA MINING TOOLS

There are various data mining tools and techniques used for data mining are: WEKA, TANAGRA, MATLAB and .NET FRAMEWORK.

A. Weka

It is a data mining tool which was developed in New Zealand by the University of Waikato that makes use of data mining algorithms using JAVA language. WEKA is a collection of machine learning algorithms and their application to the data mining problems. These methods are directly applied to the dataset. WEKA supports data file in ARFF format.

B. Tanagra

It is open source software as researchers can access to the source code and add their own algorithms and compare their performances, if it conforms to the software distribution license.

C. Matlab

It is a data mining tool built in high level language. It gives an interactive environment for visualization, numerical computation and programming. The built in math functions, language and tool explore different approaches and helps a person to reach the solution faster with the spreadsheet of traditional programming languages like C,C++ and JAVA. It is used to analyze data, develop algorithms, and create models and applications.

D. .NET Framework

It is a software framework developed by Microsoft which runs primarily on Microsoft windows. It provides secure communication and regular applications. It provides language interoperability across several programming languages.

IX. PROBLEM FORMULATION

At presently different algorithms are available for clustering the pre-processed data, in the existing work they use K-mean clustering and MAFIA algorithm for Heart disease prognosis system and accomplished the veracity of 89% as we can see that there is a wide scope of improvement, in our proposed system we will use SVM Classifier and GA optimization over the data and will accomplish the veracity more than the current algorithm.

X. OBJECTIVE

1) To study the various algorithm of clustering and classifying data.
2) To implement the K-mean and MAFIA algorithm.
3) To implement the improved algorithm of clustering.
4) Performance analysis of improved system.

XI. CONCLUSION

In medical field, Data Mining provides different methods and had been vastly used in clinical decision support systems that are useful for diagnosis of various kinds of diseases. These data mining methodology are used in heart disease prediction and takes lesser time and make process very fast for the prognosis system to predict the heart diseases with good veracity in order to improve their health. In this work, K-mean clustering and MAFIA algorithm for Heart disease prognosis system and accomplished the veracity of 89% so, there is a vast scope of enrichment. In this proposed system we will implement the improved algorithm of clustering which accomplish the veracity more than the present algorithm.

REFERENCES

[1] Ms. Chaitrali S. Dangare, Dr. Mrs. Sulabha S. Apte, “A data mining approach for prediction of heart disease using neural networks, international journal of computer engineering and technology”,2012

[2] M.A.Nishara Banu and B.Gomathy,” Disease Forecasting System Using Data Mining Methods”, 2014

[3] Aqueel Ahmed, Shaikh Abdul Hannan, “Data Mining Techniques to Find Out Heart Diseases”, International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-1, Issue-4, September 2012.

[4] Ms. Ishtake S.H, Prof. Sanap S.A., “Intelligent Heart Disease Prediction System Using Data Mining Techniques”, International J. of Healthcare & Biomedical Research, 2013

[5] Chitra R and Seenivasagam V, “REVIEW OF HEART DISEASE PREDICTION SYSTEM USING DATA MINING AND HYBRID INTELLIGENT TECHNIQUES”, ISSN: 2229-6956(ONLINE) ICTACT JOURNAL ON SOFT COMPUTING, JULY 2013, VOLUME: 03, ISSUE: 04, 2013

[6] Nidhi Bhatla and Kiran Jyoti, “An Analysis of Heart Disease Prediction using Different Data Mining Techniques”, International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181,Vol. 1 Issue 8, October – 2012

[7] Shadab Adam Pattekari and Asma Parveen, prediction system for heart disease using naïve bayes, International Journal of Advanced Computer and Mathematical Sciences, 2012

[8] Venkatadri.M, Dr. Lokanatha C. Reddy a review on data mining from past to the future. International Journal of Computer Applications, 2011.

[9] Abhishek taneja, Heart Disease Prediction System Using Data Mining Techniques, Oriental Scientific Publishing Co., India, 2013.

[10] Rashedur M. Rahman, Farhana Afroz, Comparison of Various Classification Techniques Using Different Data Mining Tools for Diabetes Diagnosis, Journal of Software Engineering and Applications, 2013.

[11] Nidhi Bhatla Kiran Jyoti, An Analysis of Heart Disease Prediction using Different Data Mining Techniques, International Journal of Engineering Research & Technology (IJERT), 2012.

[12] Humar Kahramanli, Novruz Allahverdi, Design of a hybrid system for the diabetes and heart diseases, Elsevier, 2008.

[13] Marcel A.J. van Gerven, Predicting carcinoid heart disease with the noisy-threshold classifier, Elsevier, 2007.

[14] Mohammad Taha Khan, Dr. Shamimul Qamar and Laurent F. Massin, A Prototype of Cancer/Heart Disease Prediction Model Using Data Mining, International Journal of Applied Engineering Research, 2012

[15] M.Akhil jabbar, Dr.Priti Chandra, Dr.B.L Deekshatulu, Heart Disease Prediction System using Associative Classification and Genetic Algorithm, International Conference on Emerging Trends in Electrical, Electronics and Communication Technologies, 2012