Prediction of Heart Disease using Data Mining Techniques

Era Singh Kajal
Research Scholar
CBS Group of Institution, Jhajjar, Haryana
erasinghkajal@gmail.com

Ms. Nishika
Assistant Professor (CSE Department)
CBS Group of Institution, Jhajjar, Haryana

Abstract—Data mining is process to analyses number of data sets and then extracts the meaning of data. It helps to predict the patterns and future trends, allowing business in decision making. Data mining applications are able to give the answer of business questions which can take much time to resolve traditionally. High amount of data that can be generated for the prediction of disease is analyzed traditionally and is too complicated along with voluminous to be processed. Data mining provides methods and techniques for transformation of the data into useful information for decision making. These techniques can make process fast and take less time to predict the heart disease with more accuracy. The healthcare sector assembles enormous quantity of healthcare data which cannot be mined to uncover hidden information for effectual 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 several methods 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 techniques can be used in heart diseases takes less time and make the process much faster for the prediction system to predict diseases with good accuracy to improve their health. In this paper we survey different papers in which one or more algorithms of data mining used for the prediction of heart disease. By Applying data mining techniques 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 techniques 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, disease prediction, KNN,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 patterns and trends in databases. This information is further used to build predictive models. The main objective is to

learn the different data mining techniques/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:

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.

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.

Unhealthy diet: eating too much fast food increases blood pressure and cholesterol level causing the risk of heart attacks.
Smoking: it damages the lining of arteries and builds up a fatty material called atheroma which narrows the arteries causing heart attacks.
Lack of physical activity: lack of exercise increases cholesterol level in blood vessels which further increases the risk of heart attacks.

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 patterns and will determine trends in patient records to improve health care.

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.

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 patterns and trends that lead to improved quality of customer service and customer's satisfaction.

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

Healthcare Industry: Data mining is very useful in healthcare industry in diagnosis of heart diseases, breast cancer and diabetes. It helps in identifying patterns and trends in patient's records having same risk factor and helps in decision making.

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.

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.

Student's Performance: Data mining is used to evaluate student's performance using classification technique 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.

COMPARATIVE STUDY OF DATA MINING TRENDS FROM PAST TO FUTURE

PAST:- In the previous years, statistical and Machine learning techniques were used on numerical data stored in traditional databases and the computing resources were 4G PL and various related techniques[8].

PRESENT: - These days, along with the statistical and Machine learning techniques, artificial intelligence and pattern reorganisation techniques are also used [8].

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 techniques like fuzzy logic, neural networks and genetic programming is used. Computing resources used would be multi-agent technologies and cloud computing [8].

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

1. **MAFIA – MAFIA** stands for Maximal Frequent Itemsets. Item set of Frequent is one of the fundamental data mining problems which has a goal to find the number of items which include the frequently in dataset. The major goal is to find interesting patterns 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.
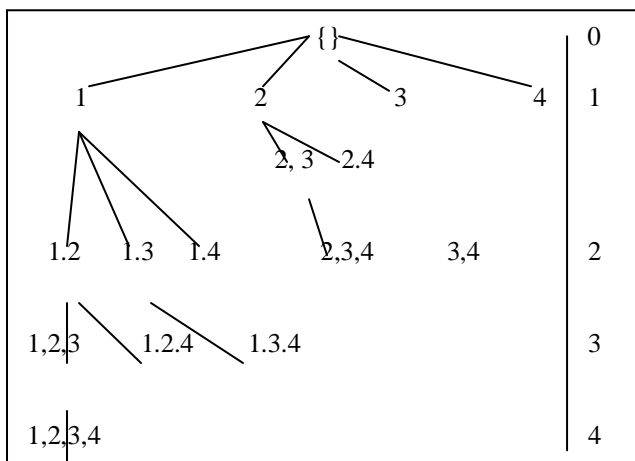
Figure: 1. Relation of Data in MAFIA

2. **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.

DIFFERENT DATA MINING TOOLS USED IN HEART DISEASE PREDICTION SYSTEM WITH ACCURACY

Abhishek et al in the year 2013 used data mining tool Weka 3.6.4 in heart disease prediction system using J48 technique achieved 95.56% accuracy and using Naïve Bayes achieved 92.42%. [9]

Rashedur et al in the year 2013 used Neural network technique using Weka data mining tool and achieved 79.19% and to compare various classification techniques, he used another technique fuzzy logic using TANGRA data mining tool and achieved 83.85% accuracy. [10]

Nidhi et al in the year 2012 used data mining tool Weka 3.6.6 in the analysis of heart disease prediction system and achieved 99.52% using Naive Bayes. She also used TANGRA data mining tool but could achieve up to 52.33% only using decision trees. She also tried .NET data mining tool and achieved up to 96.5% using neural networks. [6]
Resul et al in the year 2009 used SAS base software 9.1.3 achieving 97.4% using neural networks. [11]

DATA MINING TECHNIQUES ALSO USED IN DIAGNOSIS OF OTHER DISEASES

Humar et al in the year 2008 used classification, Back propagation, Fuzzy neural network techniques for diabetes and heart diseases. [12]
Marcel et al in the year 2007 used Bayesian Classification for Characinoid heart disease. [13]

Mohammad et al in the year 2012 used C4.5 and C5.0 algorithm for heart disease and breast cancer diagnosis. [14]
M. Akhil et al in the year 2012 used associative classification and genetic algorithm for the diagnosis of breast cancer, Pima Indian Diabetes and heart disease. [15]

DATA MINING AND ITS TECHNIQUES

Data Mining

It is main concerned with extracting useful information from large amount of databases. Data mining techniques and tools are used to find unknown patterns and trends from the data set. Its main objective is to automatically find the patterns 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 evaluates patterns and trends of market and produce effective results.
Data mining techniques:

Association: it is the best known and well researched method for data mining. Association is also called relation technique because patterns which are discovered from the dataset are based on the relationship between the items.

Classification: it is a data mining technique 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 technique which is based on machine learning.

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

Prediction: It is a data mining technique which discovers relationship between independent variables and relationship between dependent and independent variables.

Sequential Patterns: It is a data mining technique that discovers similar patterns or regular events in transaction data over a business period.

DATA MINING TOOLS:

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

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

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.

MATLAB: It is a data mining tool built in high level language. It provides interactive environment for visualization, numerical computation and programming. The built in math functions, language and tool explore various approaches and helps to reach a solution faster than with the spreadsheet of traditional programming languages like C,C++ and JAVA. It analyse data, develop algorithms, and create models and applications.

.NET FRAMEWORK: It is a software framework developed by Microsoft which runs primarily on Microsoft windows. It provides secure communication and consistent applications. It provides language interoperability (each language can code written in other languages) across several programming languages.

II. LITERATURE SURVEY

Over the years, numerous works have been done related to heart disease prediction system using different data mining algorithms by different authors. They tried to achieve efficient methods and accuracy in finding out diseases related to heart by their work including datasets and different algorithms along with the experimental results and future work that can be done on the system to achieve more efficient results. This paper aims at analyzing different data mining techniques that has been introduced in recent years for heart disease prediction system by different authors.

M. A. Nishara Banu and B. Gomathy [2] used C4.5 algorithm, MAFIA and K-means clustering in the year 2014 using 13 attributes in the dataset achieving 89 percent accuracy.

Aqueel Ahmed et al. [3] show the classification techniques in data mining and show the performance of classification among them. In this classification accuracy among these data mining has discussed. In this decision tree and SVM perform classification more accurately than the other methods and was able to achieve 91% accuracy

Ms. Ishtake et al. [4] developed a prediction system for heart diagnosis using decision tree, Neural Network and Naive Bayes techniques using 15 attributes in the year 2013.

Chitra R.et al. [5] developed the computer aided heart disease prediction system that helps the physician as a tool for heart disease diagnosis. From the analysis it is concluded that neural network with offline training is good for disease prediction in early stage and good performance can be obtained by pre-processed and normalized dataset.

Nidhi Bhatlaet al. [6] projected the study of different data mining techniques that can be employed in automated heart disease prediction systems. The analysis shows that neural network with 15 attributes has shown the highest accuracy. On the other hand, Decision tree has also performed well with 99.62% accuracy by using 15 attributes.

Shadab et al. [7] used Naive Bayes technique in the year 2012 using 15 attributes in the dataset for the heart diagnosis in heart prediction system.

III. PROBLEM FORMULATION

Presently various algorithm is available for clustering the pre-processed data, in the existing work they used K-mean clustering and MAFIA algorithm for Heart disease prediction system and achieved the accuracy of 89% as we can see that there is vast scope of improvement, in our proposed system we will implement SVM Classifier and GA optimization over the data and will achieve the accuracy more than the present algorithm.

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

Conclusion:
In medical field, Data Mining provides various techniques and have been widely used in clinical decision support systems that are useful for predicting and diagnosis of various diseases. These data mining techniques used in heart diseases takes less time and make process fast for the prediction system to predict heart diseases with good accuracy in order to improve their health. In this work, K-mean clustering and MAFIA algorithm for Heart disease prediction system and achieved the accuracy of 89% as we can see that there is vast scope of improvement. In our proposed system we will implement the improved algorithm of clustering which achieve the accuracy 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