Datamining Techniques on Predicting Heart Disease (An Extensive Survey)

[1]A.Menaka, [2] Dr.M.Senbagavalli, [3]D.Sudaroli
[1] Assistant Professor, Department of CSE, Jayam College OF Engineering & Technology, Dharmapuri, Tamilnadu.
[2] Associate Professor, Department of Information Technology, Alliance College of Engineering and Design, Alliance University,Bangalore,Karnataka.
[3]Assistant Professor, Department of Information Technology, Alliance College of Engineering and Design, Alliance University,Bangalore,Karnataka.

Abstract

An enormous amount of information is generated in medical corporation (ameliorate facilities, healing focuses) still the information is not efficiently used. Data mining plays prominent roles for forecast of medical diseases. The health care system is "knowledge poor" but "data rich". The medical practitioner can help the patients by forecasting the heart disease before occurring. Medical data mining consist of great potential for analysing hidden patterns present in the data sets of medical field. Those patterns can help in clinical diagnosis. In this paper, various technologies in data mining for prediction of heart disease are mentioned. Mining is the method of extracting large sets of data to bring out patterns that are previously and hidden unknown relationships and detection of knowledge for better understanding of data in prevention of heart diseases. There are several types of data mining techniques available such as classification techniques through naive Bayes (NB), Decision tree (DT), Neural network (NN), Genetic algorithm (GA), Artificial Intelligence (AI) and Clustering algorithms like Support vector machine(SVM), K-NN. The main advantages of this paper are: early prediction of heart disease and diagnosis correctly on time and providing treatment with minimal cost.

Keywords: *Genetic algorithm (GA), Artificial Intelligence (AI), K Nearest Neighborhood (KNN) Algorithm, Neural network (NN), Decision tree (DT) Algorithm.*

I. INTRODUCTION

Data mining is the process of extracting valuable data from the large database. In current life style health diseases are increasing to a very great extent. Heart is the major organ in human body. It is the centre of the circulatory system. [1]It function as pump that drive blood to whole parts of the body through blood vessels, supplying a constant supply of oxygen as well as nutrients which is required to human body. If the heart ever stops functioning and ceases to pump blood, the body will shut down and within very less time a person will expire. "Cardiovascular disease is the leading cause of illness and death worldwide," said Dr. Stephen Weng, of Nottingham University's National Institute for Health Research School [2]. Change in lifestyle, work related stress and bad food habits contribute to the increase in rate of several heart related diseases.

The usage of present technology in health care industry is increasing day by day to provide information which is more useful for doctors in decision making process. It helps doctors and physicians in disease management, medications and discovery of patterns and relationships among diagnosis data. Current approaches to predict cardiovascular risk fail to identify many people who would benefit from preventive treatment, while others receive unnecessary intervention [1]. Taking a survey of present population it is seen that about sixty percentages are suffering from heart diseases. Early detection of heart diseases can prevent the death rate, people are not aware about the detection of heart. Health care industries are aiming to diagnose the disease at early stages. In most cases it is noticed at the final stages of disease or after death[3].

There are many types of data mining classifiers which can be applied to predict diseases. In this paper we are focusing on heart disease prediction [4].The term heart disease includes several types of disorders which may damage the heart. Most common types of heart diseases are congenital heart disease, heart failure, Hypertensive heart disease, cardiomyopathy, heart murmurs, rheumatic heart disease, pulmonary stenosis and coronary artery disease.

These diseases are the main reasons of death for huge number of people all over the globe. There are several causes of heart disease like smoking, age, gender, high blood pressure, high blood cholesterol, unbalanced diet, lack of exercise, over stress, poor hygiene and family history of heart disease etc. We can use these causes as risk factors to predict heart disease. If the function of heart is not done properly means, it will affect other human body part too[5].

The level of lipids or fats increased in the blood are causes the heart disease. The lipids are in the arteries hence the arteries become narrow and blood flow is also become slow .Age is the non-modifiable risk factor which also a reason for heart disease. Smoking is the reason for 40% of the death of heart diseases. Because it limits the oxygen level in the blood then it damage and tighten the blood vessels.

Data mining is the process of analysing large set of data and summarizing into useful information.
Data mining techniques are:

1. Association
2. Classification
3. Clustering

Associative analysis helps in bringing out hidden relationships among data items in a large data set.
Classification: This is tagging or classifying data items into different user-defined categories.
Clustering: This is partitioning a huge set of data into related sub-classes.

Various datamining techniques such as Naïve Bayes, KNN algorithm, Decision tree, Neural Network, ID3, and Genetic Algorithm are used to predict the risk of heart disease.

The rest of the paper is organized as follows. Literature reviewis explained in Section II. Section III includes methodology.Data mining techniques are described in Section IV. Finally Section V draws the conclusion of our findings.

II. LITERATURE SURVEY

Different types of studies have been done to focus on prediction of heart disease. Various datamining techniques are used for diagnosis and achieved different accuracy level for different methods [6]. Rupali and Patil [7] described an improved decision support system using two data mining classifiers namely Naive Bayes and Jelinek-Mercer smoothing for heart disease prediction.Chitra and Seenivasagam [8] proposed a hybrid intelligent algorithm for improvement of classifiers' accuracy to predict heart disease.

Shamsher and Shukla [9] analysed some classification techniques for prediction of heart disease with decreased number of attributes. Jyoti and Dipesh [10] compared the predictive performance of data mining techniques. Khaled and Das [11] evaluated the different data mining techniques to find frequent pattern based on cost, performance, speed and accuracy.

Hlaudi and Mosima [12] applied J48, Naive Bayes, Bayes Net, REPTREE and CART for predicting heart attacks. In that paper, classification techniques J48, REPTREE and SIMPLE CART showed the best accuracy.

Taneja [13] found that J48 that used selected attributes outperformed the Naive Bayes and Neural Network by achieving the highest accuracy.

P.K Anooj et al. [14] presented a weighted fuzzy rule-based system for the diagnosis of heart disease, the system will automatically retrieve knowledge from the patient's data. The proposed system for the prediction of heart disease consists of two phases: (1) automated approach for the generation of weighted fuzzy rules and (2) developing a fuzzy rule-based decision support system. The weighted fuzzy rules were used to build the system using Mamdani fuzzy inference system.

Aditya Methaila et al. [15] desire to use data mining Classification Modeling Techniques, such as Decision Trees, Naïve Bayes and Neural Network, in addition to weighted association Apriori algorithm and MAFIA algorithm in Heart Disease Prediction.

Shimpy Goyal et al. [16] discussed Data Mining Techniques to Predict Heart Disease based on K-means and apriori algorithm. The researchers also presented the challenges in detecting and diagnose the diseases and analyse results of research.

Deepika N et al. [17] used Pruning Classification Association Rule (PCAR). Pruning Classification Association Rule comes from Apriori algorithm. The proposed method deletes minimum frequency item with minimum frequency item sets and deletes infrequent items from item sets then the frequent item set is discovered.

Valli M S and Arasu G T [18] used An Efficient Feature Selection Technique of Unsupervised Learning Approach. Feature selection (FS) is a procedure which efforts to select more informative features. The main aspect of feature selection is give high accuracy performance with minimal feature subset. They proposed the unsupervised rough set method for clustering text for Web opinion mining and conducted more experiments and had benchmarked with the unsupervised algorithm which gives higher micro accuracy results.

The prediction of heart disease requires a huge size of data which is too complex and massive to process and analyze by conventional techniques. Diverse data mining techniques are being used by experts. Our objective is to find out the suitable data mining technique that is computationally efficient as well as accurate for the prediction of heart disease [4].

III.DATA MINING TOOLS

Data mining tools provide ready to use implementation of the mining algorithms. Most of them are free open source software's so that researchers can easily use them. They have an easy to use interface. Some of the popular data mining tools are WEKA, RapidMiner, TANAGRA, MATLAB etc. Some of them are discussed as follows

1. WEKA

It stands for Waikato Environment for Knowledge Learning.Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes.

2. Rapidminer

Formerly called as YALE (Yet Another Learning Environment) is a readymade, open source, no-coding required software, which gives advanced analytics. Written in Java, it incorporates multifaceted data mining functions such as data pre-processing, visualization, predictive analysis, and

can be easily integrated with WEKA and R-tool to directly give models from scripts written in the former two. Besides the standard data mining features like data cleansing, filtering, clustering, etc, the software also features built-in templates, repeatable work flows, a professional visualisation environment, and seamless integration with languages like Python and R into work flows that aid in rapid prototyping.

3. TANAGRA

It is free open source data mining software for academic and research purposes. It proposes several data mining methods from exploratory data analysis, statistical learning, machine learning and databases area. TANAGRA is more powerful, it contains some supervised learning but also other paradigms such as clustering, factorial analysis, parametric and non-parametric statistics, association rule, feature selection and construction algorithms. The main purpose of Tanagra project is to give researchers and students an easy-to-use data mining software, conforming to the present norms of the software development in this domain (especially in the design of its GUI and the way to use it), and allowing to analyse either real or synthetic data.

4. Apache Mahout

Mahout is primarily a library of machine learning algorithms that can help in clustering, classification and frequent pattern mining. It can be used in a distributed mode that helps easy integration with Hadoop. Mahout is currently being used by some of the giants in the tech industry like Adobe, AOL, Drupal and Twitter, and it has also made an impact in research and academics. It can be a great choice for anyone looking for easy integration with Hadoop and to mine huge volumes of data.

5. MOA

Massive Online Analysis (MOA), as the name suggests, is primarily data stream mining software that is well suited for applications that need to handle volumes of real-time data streams at a high speed. Stream mining algorithms typically require faster computations without storing all of the datasets in the memory and have to get the work done within a limited time. MOA is well suited for these requirements. Weka and MOA can be closely linked to each other and either of the classifiers can be called from the other one. For those looking to analyse and mine information from real-time data, MOA can be the best choice.

6. XL Miner

XLMiner is the only comprehensive data mining add-in for Excel, with neural nets, classification and regression trees, logistic regression, linear regression, Bayes classifier, K-nearest neighbours, discriminant analysis, association rules, clustering, principal components, and more.

7. MATLAB

It is the short form for matrix laboratory. It supports a multi-paradigm numerical computing environment. It is a fourth-generation programming language. MATLAB provides matrix manipulations, plotting of functions and data, algorithm implementations, creation of user interfaces and interfacing with programs written in other languages including C, C++,C#, Java, Fortran and Python .

8. Orange

Python users playing around with data sciences might be familiar with Orange. It is a Python library that powers Python scripts with its rich compilation of mining and machine learning algorithms for data pre-processing, classification, modelling, regression, clustering and other miscellaneous functions.

IV.PROPOSED SYSTEM

In the proposed system data mining techniques are used for early diagnosis of the heart diseases. A large amount of health data which regrettably are not mined to find hidden information for effective decision process.

This research provides a model “prediction of heart disease using data mining techniques” such as

A. Genetic algorithm
B. K-means algorithm
C. MAFIA algorithm
D. Decision tree classification

A. Genetic algorithm

A genetic algorithm (GA) is a search heuristic that imitates the process of natural evolution. This heuristic is routinely used to generate useful solutions to optimization and search problems. Genetic algorithms belong to the larger class of evolutionary algorithms (EA), which generate optimized solutions using techniques inspired by natural evolution, such as inheritance, mutation, selection, and crossover. A typical genetic algorithm requires:

1. A genetic representation of the solution domain,
2. A fitness function to evaluate the solution domain. Here genetic algorithm is used to take out attribute from a huge attribute set.

The extracted attribute are as follows,

a. age: age in years
b. sex: sex (0= male; 1= female)

3. cp: chest pain type

Value 1: typical angina
Value 2: atypical angina
Value 3: non-anginal pain
Value 4: asymptomatic

4. trestbps: resting blood pressure (in mm Hg on admission to the hospital)
5. chol: serum cholestoral in mg/dl
6. fbs: (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
7. restecg: resting electrocardiographic results Value 0: normal

Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)

Value 2: showing probable or definite left ventricular

Hypertrophy by Estes' criteria

8. thalach: maximum heart rate achieved
9. exang: exercise induced angina (1 = yes; 0 = no)
10. Old peak = ST depression induced by exercise relative

to rest

11. Slope: the slope of the peak exercise ST segment Value Upsloping

Value 2: flat

Value 3: downsloping

12. ca: number of major vessels (0-3) colored by flourosopy
13. thal: 3 = normal; 6 = fixed defect; 7 = reversable defect
14. num: diagnosis of heart disease (angiographic disease status)

Value 0: < 50% diameter narrowing

Value 1: > 50% diameter narrowing

B. K-means algorithm

k-means clustering algorithm is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. The method follows a simple way to cluster a given data set through a certain number of clusters fixed apriori. The main idea is to define k centers is one for each cluster. These centers should be arranged in a smart way because of different location causes different result in the clustering.

So, the better choice is to arrange them as much as possible far away from each other. The next stride is to take each point belongs to the given data set and associate it to the nearest center. When no point is pending, the first step is completed and an early group age is done. At this point we need to re-calculate k new centroids as barycentre of the clusters resulting from the previous step. After identification of these k new centroids, a new binding has to be done between the same data set points and the nearest new center. A loop has been generated, as a result of this loop it is notice that the k centers change their location step by step until no more changes are done or in other words centers do not move any more. Finally, this algorithm aims at minimizing an objective function know as squared error function given by:

Where, $||x_i - v_j||$ is the Euclidean distance between xi and v_j. 'c_i' is the number of data points in i^{th} cluster 'c' is the number of cluster canters.

Input: The number of clusters k, and a database containing n objects.

Output: A set of k clusters which minimizes the squared-error criterion.

Method:

1) Arbitrarily choose k objects as the initial cluster centers
2) Repeat the process
3) 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

C. Frequent Pattern mining using MAFIA

Mining frequent item sets is an active area in data mining that aims at searching interesting relationships between items in databases. It can be used to address to a wide variety of problems such as discovering association rules, sequential patterns, correlations and much more. The proposed approach utilizes an efficient algorithm called MAFIA (Maximal Frequent Itemset Algorithm) which combines diverse old and new algorithmic ideas to form a practical algorithm.

The proposed algorithm is employed for the extraction of association rules from the clustered dataset besides performing efficiently when the database consists of very long item sets specifically.

Pseudo code for MAFIA:

```
MAFIA(C, MFI, Boolean IsHUT)
{
        name HUT = C.headC.tail;
        if HUT is in MFI
        stop generation of children and return
        Count all children, use PEP to trim the tail, and
        recorder by increasing support,
For each item i in C, trimmed_tail
{
        IsHUT = whether i is the first item in the tail
        newNode = C I
        MAFIA (newNode, MFI, IsHUT)
}
if (IsHUT and all extensions are frequent)
{
        Stop search and go back up subtree
        If (C is a leaf and C.head is not in MFI)
        Add C.head to MFI
        }
}
```

The cluster that contains data most relevant to heart attack is fed as input to MAFIA algorithm to mine the frequent patterns present in it.

D. Decision tree

Decision tree is a classification technique. Decision tree learning methods are most commonly used in data mining. The goal is create a model to predict value of target variable based on input values. Training dataset is used to create tree and test dataset is used to test accuracy of the decision tree. Each leaf node represents the target attribute's value depend on input variables represented by path by path from root to leaf node. First, an attribute that splits data efficiently is selected as root node in order to create small tree. The attribute with higher information is selected as splitting attribute.

Decision tree algorithm involves three steps:

1. For a given dataset S, select an attribute as target class to split tuples in partitions.
2. Determine a splitting criterion to generate a partition in which all tuples belong to a single class. Choose best split to create a node.
3. Iteratively repeat above steps until complete tree is grown or any stopping criterion is fulfilled.

The J48 algorithm recursively classifies data until it has been categorized as perfectly as possible. This technique gives

maximum accuracy on training data. The overall concept is to build a tree that provides balance of flexibility & accuracy.

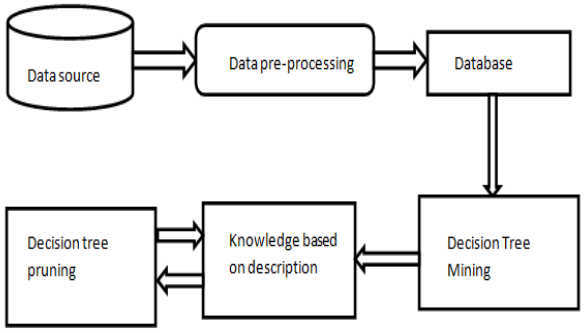

Figure 1: Generation of Decision tree using heart Disease DataSet

The activities of Decision Tree Classifier are as follows:

1. **Data Source**
   - Need to collect necessary information from various resources
   - Eliminate unwanted information
   - Categorize the information and it act as primary source information for data preprocessing.

2. **Data Preprocessing**
   - **Cleansing** - process of removing the corrupt
   - **Integration –** combining data from several disparate sources
   - **Transformation –** process of converting data from one format or structure to other type

3. **Database –** collection of data with organized manner(store and access)

4. **Decision Tree Mining –** Tree used for Regression and Classification

5. **Knowledge based on description –** Retrieving valid information based on description

6. **Decision Tree Pruning –** it is the process of reduces the size of decision trees by removing parts of the tree that do not provide power to classify instances.

V. CONCLUSION

The main motivation of this Research is to provide an intuition about detecting heart disease risk rate using data mining techniques. Decision Tree has tremendous efficiency using fourteen attributes, after applying genetic algorithm to reduce the actual data size to get the optimal subset of attribute acceptable for heart disease prediction. In our future work, we have planned to design and develop an efficient heart attack prediction system with patient prescription support using the web mining and data warehouse techniques which will help the public to get awareness about heart disease and to reduce the death ratio in future.

REFERENCES

1. Tanvi Sharma, SahilVerma, Kavita" Intelligent Heart Disease Prediction System using Machine Learning: A Review" International Journal of Recent Research Aspects ISSN: 2349-7688, Vol. 4, Issue 2, June 2017, pp. 94-97
2. https://www.digitaltrends.com/computing/artificial-intelligence-cardiovascular-disease/
3. SarathBabu, Vivek EM, Famina KP, Fida K, Aswathi P, Shanid M, Hena M" Heart Disease Diagnosis Using Data Mining Technique" International Conference on Electronics, Communication and Aerospace Technology 978-1-5090-5686-6/17/$31.00 ©2017 IEEE
4. Marjia Sultana*, Afrin Haider and Mohammad ShorifUddin" Analysis of Data Mining Techniques for Heart Disease Prediction"978-1-5090-2906-8/16/$31.00 ©2016 IEEE
5. Theresa Princy. R, J. Thomas" Human Heart Disease Prediction System using Data Mining Techniques" 978-1-5090-1277-0/16/$31.00 ©2016 IEEE
6. Miss. Chaitrali S. Dangare, Dr. Mrs. Sulabha S. Apte, "ImprovedStudy of Heart Disease Prediction System using Data MiningClassification Techniques", International Journal of ComputerApplications (0975 – 888), Volume 47– No.10, pp.44-48, June 2012.
7. Rupali, R.Patil, "Heart disease prediction system using Naive Bayes and Jelinek-mercer smothing," International Journal of Advanced Research in Computer and Communication Engineering, vol. 3, no. 5, May 2014.
8. R. Chitra and V.Seenivasagam, "Review of heart disease prediction system using data mining and hybrid intelligent," ICTACT Journal on Soft Computing, vol. 03, no. 04, July 2013.
9. Shamsher Bahadur Patel, Pramod Kumar Yadav and Dr. D. P. Shukla, "Predict the diagnosis of heart disease patients using classification mining Techniques," IOSR Journal of Agriculture and

Veterinary Science (IOSR-JAVS), vol. 4, no. 2, pp. 61-64, Jul.-Aug. 2013.

10. JyotiSoni, Ujma Ansari and Dipesh Sharma, "Prediction data mining for medical diagnosis: An overview of heart disease prediction," International Journal of Computer Applications (0975-8887), vol. 17, March 2011.
11. Mohammed Abdul Khaled, Sateesh Kumar Pradhan and G.N. Dash, "A survey of data mining techniques on medical data for finding locally frequent diseases," International Journal of Advanced Research in Computer Science and Software Engineering, vol. 3, pp. 149-153, August 2013.
12. Hlaudi Daniel Masethe and Mosima Anna Masethe, "Prediction of heart disease using classification algorithms," in Proceeding of the World Congress on Engineering and Computer Science, vol. II, San Francisco, USA, 2014.
13. Abhishek Taneja, "Heart disease prediction system using data mining techniques," Oriental Journal of Computer Science & Technology, vol. 6, pp. 457-466, December 2013.
14. P.K. Anooj, Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules, Journal of King Saud University – Computer and Information Sciences (2012) 24, 27–40.
15. Shimpy Goyal and Dr.Rajender Singh Chhillar , A Literature Survey on Applications of Data Mining Techniques to Predict Heart Diseases, International Journal of Engineering Sciences Paradigms and Researches (IJESPR) (Vol. 20, Issue 01) and (Publishing Month: May 2015)
16. Aditya Methaila, Prince Kansal, Himanshu Arya, Pankaj Kumar, Early Heart Disease Prediction Using Data Mining Echniques, Sundarapandian et al. (Eds) : CCSEIT, DMDB, ICBB, MoWiN, AIAP – 2014 pp. 53–59, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4807.
17. BoshraBahrami, Mirsaeid Hosseini Shirvani, “Prediction and Diagnosis of Heart Disease by Data Mining Techniques”, Journal of Multidisciplinary Engineering Science and Technology (JMEST) ISSN: 3159-0040 Vol. 2 Issue 2, February–2015.
18. M S Valli,G T Arasu,”An Efficient Feature Selection Technique of Unsupervised Learning Approach for Analyzing Web Opinions”, Journal of Scientific and Industrial Research ,Vol. 75, April 2016, pp. 221-224.