A Study of Heart Disease Prediction in Data Mining

S.Vijayarani
Assistant Professor
Department of Computer Science
School of Computer Science and Engineering
Bharathiar University
Coimbatore, Tamil Nadu, India
vijimohan_2000@yahoo.com

S.Sudha
Research scholar
Department of Computer Science
School of Computer Science and Engineering
Bharathiar University
Coimbatore, Tamil Nadu, India
sudhas253@gmail.com

Abstract: **The data mining techniques can extract the hidden information from the large databases. It helps to find the relationships and patterns from the data. Data mining is used for various applications such as business organizations, e-commerce, health care industry, scientific and engineering. In the health care industry the data mining is mainly used for predicting the diseases from the datasets. In this survey paper, we have studied and analyzed how data mining techniques such as classification, clustering, fuzzy system and association rules are used for predicting the heart diseases. This paper also gives the advantages and disadvantages of the existing techniques. It also discusses the future enhancements of the existing works.**

Keywords: Heart disease, Prediction, Association rules, Classification, Clustering, Fuzzy system.

I. INTRODUCTION

Data mining is defined an extraction of useful knowledge from the large data repositories. Compared with other data mining application areas, medical data mining plays an important role and it has some unique characteristics. In medical domain, the medical data mining has the high potential for extracting the hidden patterns in the datasets. These patterns are used for clinical diagnosis. The medical data are widely distributed, voluminous and heterogeneous in nature. The data is collected and then integrated to provide a user oriented approach and to find the novel and hidden patterns from the data. [7]

Based on Heart Disease the WHO (World Health Organization) estimated that twelve million deaths occur worldwide. Half of the deaths occurs in United States and other developed countries based on cardio vascular diseases. Most of the death occurs in adults. Heart disease was the major causes of deaths of different countries include India. In United States every 34 seconds the heart disease kills one person. There are different categories of heart diseases. The important heart diseases are Cardiovascular Disease, Cardiomyopathy and Coronary heart disease.

Medical diagnosis plays vital role and yet complicated task that needs to be executed accurately and efficiently. To reduce cost for achieving clinical tests, an appropriate computer based information and decision support is required. The existing techniques are studied and compared for finding the efficient and accurate systems. This paper aims to analyze the different predictive data mining techniques proposed in recent years for the diagnosis of heart diseases**.**

The rest of this paper is organized as follows. Section 2 describes various types of heart diseases. Section 3 analyses the various heart disease prediction research papers and the future enhancements using various data mining techniques also given. Conclusions are given in Section 4.

II HEART DISEASES

Now a day's many of the people are affected by heart diseases. There are different types of heart diseases. They are discussed as follows.

Coronary Artery Disease: When the combination of fatty material, calcium and scar tissue (plague) builds up in the arteries that supply the heart with blood

through this, the disease should develops. Through these arteries called the coronary arteries, the heart muscle (myocardium) gets the oxygen and other nutrients it needs to pump blood. Coronary artery disease is America's No.1 killer, affecting more than 13 million Americans. [16]

Enlarged Heart (Cardiomegaly): A heart condition that causes the heart to become larger than normal as a result of heart disease. Cardiomegaly is most often linked to high blood pressure, but it can also occur as a result of other heart conditions, such as congestive heart failure, and other non-cardiac causes such as long-term anemia.

Heart Attack: A heart attack is the death of, or damage to, part of the heart muscle because the supply of blood to the heart muscle is severely reduced or stopped.

Heart Valve Disease: Valvular heart disease refers to several disorders and diseases of the heart valves, which are the tissue flaps that regulate the flow of blood through the chambers of the heart.

Congenital Heart Disease: Congenital heart disease refers to a problem with the heart's structure and function due to abnormal heart development before birth. Congenital means present at birth.

Heart Muscle Disease (Cardiomyopathy) : Cardiomyopathy is a chronic disease of the heart muscle (myocardium), in which the muscle is abnormally enlarged, thickened, and/or stiffened. The weakened heart muscle loses the ability to pump blood effectively, resulting in irregular heartbeats (arrhythmias) and possibly even heart failure. [16]

Dilated Cardiomyopathy: The chambers of the heart are dilated (enlarged) because the heart muscle is weakened and cannot pump effectively. There are many causes, the most common being myocardial ischemia (not enough oxygen supplied to the heart muscle) due to coronary artery disease. [16]

Hypertrophic Cardiomyopathy: Cardiomyopathy is an ongoing disease process that damages the muscle wall of the lower chambers of the heart. It is a form of cardiomyopathy in which the walls of the heart's chambers thicken abnormally. Hypertrophic cardiomyopathy is also referred as idiopathic hypertrophic sub aortic stenosis and asymmetrical septal hypertrophy.

Restrictive Cardiomyopathy: Cardiomyopathy is an ongoing disease process that damages the muscle wall of the lower chambers of the heart. Restrictive cardiomyopathy is a form of cardiomyopathy in which the walls of the heart become rigid.

III HEART DISEASE PREDICTION

In [15], the neural network approach is used for analyzing the heart disease dataset. Applying feed forward neural network model and back propagation learning algorithm with variable learning rate and momentum the heart disease database are trained by the neural network. The input layer contains 13 neurons to represent 13 attributes. It consists of 4 class labels namely normal person, first stroke, second stroke and end of life. The output layer consists of two neurons to represent these four classes. Some of the neural networks are constructed with and without hidden layer that is single and multilayer networks are trained. The dataset was collected from Cleveland database [3]. This dataset classifies the person into normal and abnormal person based on heart diseases. The dataset consists of 414 instances, 13 attributes and a class attribute. Both test and training data are used for performance analysis. In a trained network, the test data is given as the input. With the adjusted weights, the output of the net is calculated. From the experimental results, the author concluded that efficiency of the classification process is increased by applying parallel approach which is adopted in the training phase. In future this work will be enhanced by applying genetic algorithm using neural networks.

In [8], association rule mining technique is used for predicting heart attack. In this paper, the author proposed a novel method CBARBSN, for association rule mining based on sequence numbers and clustering the transactional database for predicting heart attack. The two important steps of this process are, first the medical data is transformed into binary and the proposed method is applied to the binary transactional data. The data is collected from Cleveland database [3]. The medical data contains 14 attributes. From the results, the author concluded that the proposed algorithm performs better than the existing ARNBSN ( ) algorithm. The performance of the algorithms is compared based on the execution

time. Further this work will be enhanced by measuring optimum accuracy for clustering algorithm.

In [5], the data mining classification techniques namely RIPPER classifier, decision tree, Artificial Neural Networks and Support vector machine are used for predicting cardiovascular heart disease. The performance factors used for comparing these techniques are sensitivity, accuracy, specificity, error rate, true positive rate and false positive rate. To measure the unbiased estimate of prediction models the author used 10 fold cross validation method. This model was developed by using data mining classification tool weka version 3.6. It contains 14 attributes and 303 instances. From an experiment, the results are compared. Error rates for RIPPER, Artificial Neural Networks, Support vector machine and Decision tree are 02.756, 0.2248, 0.1588 and 0.2755 respectively. The accuracy of RIPPER, Artificial Neural Networks, Support Vector Machine and Decision tree are 81.08%, 80.06%, 84.12% and 79.05% respectively. When compared to four classification models, the Support Vector Machine has given least error rate and highest accuracy. The author concluded that the Support Vector Machine is the best technique for predicting the cardiovascular disease. In future, in order to improve the efficiency of the classification techniques by creating meta models.

In [14], the author proposed enhanced K-means clustering algorithm for predicting coronary heart disease. There are two strategies are used for enhancing K-means clustering algorithm. First the author proposed weighted ranking algorithm to overcome the problem of random selection of initial centroids. Second the attributes associated with weights concerned by the physicians are taken into account in both ranking and the K-means algorithm instead of assigning unit weight to all the attributes. The heart dataset was collected from UCI machine learning repository [2]. Moreover 35 conditions are carried out to assign weights to attributes. From an experiment the author concluded that the proposed algorithm improves the consistency and quality of the final clusters. The unique clusters generates in turns of consistency. In future the accuracy is further enhanced to the dataset from the same region of the physicians by assigning weights and also improves the efficiency of the process by unique cluster.

In [12], the heart disease is predicted by the frequent feature selection method. From the use of fuzzy measure and relevant nonlinear integral the performance of the algorithm becomes good. The feature selection attribute reflects the importance of non additive of the fuzzy measure. The author predicts the likelihood of patients getting a heart disease by using medical profiles. The proposed approach was implemented in Java. The sample combinations of normal and risk level heart attack parameters with their values and weight ages are mentioned. The normal level of prediction comprises the weightage of lesser value (0.1) and other than 0.1 is considered as higher risk levels. Finally the proposed algorithm reduces the computational time and improves accuracy. The proposed work can be further expanded for automation of heart disease prediction. Real data are collected from health care organizations and compared with optimum accuracy with the available techniques.

In [9] the coronary artery disease was effectively diagnosed by rotation forest algorithm in order to support clinical decision-making process. It utilizes the Artificial Neural Networks with Levenberg-Marquardt back propagation algorithm of rotation forest ensemble method as base classifiers. The algorithm is implemented in matlab. From an experiment, the author diagnosed the disease by comparing the performance of base classifiers in terms of sensitivity, accuracy, AUC and specificity on two things i) without rotation forest classifier, the utmost performance of classifiers and ii) with rotation forest algorithm it actually improves the performance of classifiers. As a result it is observed that Levenberg-Marquardt was the best classifier with or without random forest. The accuracy is improved to 91.2% of original classification accuracy which is an improvement of 7%. In future the proposed work may be used to develop efficient expert systems for the diagnosis of heart disease.

In [13] the heart disease is predicted by a decision support using naïve bayes. The author predicts the likelihood of patients getting heart disease by using medical profile information such as

age, gender, blood sugar and blood pressure. This application is implemented in web based questionnaire. By checking the probability the author validates the patient getting heart disease or not. The dataset is collected from Cleveland database [16]. From an analysis the author concluded that this experiment could answer complex queries with its own strength and interpretation. It is accessed for detailed information and accuracy. In future this work is enhanced by using continuous data in medical profiles for heart disease prediction.

In [4] the fuzzy expert system is designed for heart disease diagnosis with reduced number of attributes. The author finds that how genetic algorithm and fuzzy logic combine together for efficient and cost effective diagnosis of heart disease. The genetic algorithm and two models of fuzzy system Mamdani and Takagi-sugeno were used to find the cost. The dataset were taken in Cleveland clinic foundation dataset [16]. The input field is a set of all the selected features and the output of the system is to get a value '1' or '0' that indicates the presence or absence of disease. It is further enhanced by using art classifiers like Decision tree, Naïve bayes, Classification via clustering and SVM classifier.

In [6], the heart attack symptoms are predicted using biomedical data mining techniques. The author used data classification which is based on supervised machine learning algorithms. For data classification the Tanagra tool is used. Using entropy based cross validations and partitioned techniques, the data is evaluated and the results are compared. The algorithms used in these techniques are K-nearest neighbors, K-means and Mean Clustering Algorithm (EMC) is the extension of the K-mean algorithm for clustering process which reduces the number of iterations. As a result the author analyzed that the mean clustering algorithm performs well when compared to other algorithms. To run the data the time taken is very fast and it gives the result of accuracy about 82.90%. Further this work will enhanced by applying unsupervised machine learning algorithm.

In [1] discussed the performance analysis of classification data mining techniques for heart disease prediction. The three algorithms used in this work are naïve bayes, WAC and apriori. The performance evaluation is based on classification matrix, it displays the frequency of correct and incorrect prediction. The analyzed model is viewed in Lift charts, Bar charts and Pie charts. To build and access the model DMX query language and functions are used. This work can be further enhanced and expanded by using other data mining techniques like Time series, Clustering and Association rules. Instead of categorical data, the continuous data can be used.

In [11] the author predicted heart disease using decision support based on probability. It can predict the likelihood of patients getting heart disease based on medical profiles. To train nurses and medical students to diagnose patients with heart disease this serves as a training tool. Normally 15 attributes are listed for predicting heart disease with basic data mining techniques and other approaches example: Clustering, Time series, ANN, Soft Computing approaches etc. can also be implemented. The result of analyzed data finds that Decision tree outperforms and some time Bayesian classification having similar accuracy. In future this work will extended by applying optimistic technique namely fuzzy systems, genetic algorithm etc.

In [10] using data mining techniques the author predict heart disease from horoscope of a person. It finds out the possibilities of suffering a person from heart disease. Horoscope has 12 regions and each region is called 'house'. Based on house the author predicts the heart disease. The analysis is implemented in weka. The algorithms used in this work are Decision table, Multilayer perception, J48 and LWL. From this, the decision table performed well when compared to other algorithms. In future this work will extended by applying clustering algorithm in data mining.

IV CONCLUSION

Around 18 million people 7% of the Indians are affected by heart disease. Heart disease is mostly affected the person under the age of 65. This paper mainly focuses on three different categories of heart diseases namely cardiovascular disease, coronary artery disease and cardiomyopathy. Various heart disease prediction research papers are studied. This

paper also analyzes how data mining techniques namely classification, clustering, fuzzy system and association rules are applied to the health data sets for predicting heart diseases. In future, new algorithms and techniques are to be developed which overcome the drawbacks of the existing system.

REFERENCES

[1] N. Aditya Sundar, P. Pushpa Latha, M. Rama Chandra, "Performance Analysis of Classification Data Mining Techniques over Heart Disease Data base" [IJESAT] international journal of engineering science & advanced technology ISSN: 2250–3676, Volume-2, Issue-3, 470 – 478

[2]. Blake, C.L., Mertz, and C.J.: "UCI Machine Learning Databases", http://mlearn.ics.uci.edu/databases/heartdisease

[3]."Cleveland heart disease dataset" sci2s.ugr.es/keel/dataset.php?cod=57

[4] E.P. Ephzibah, "A Hybrid Genetic-Fuzzy Expert System for Effective Heart Disease Diagnosis" D.C. Wyld et al. (Eds.): ACITY 2011, CCIS 198, pp. 115–121, 2011. © Springer-Verlag Berlin Heidelberg 2011

[5] Esra Mahsereci Karabulut & Turgay İbrikçi "Effective Diagnosis of Coronary Artery Disease Using The Rotation Forest Ensemble Method" June 2011 / Accepted: 30 August 2011 / Published online: 13 September 2011 # Springer Science+Business Media, LLC 2011

[6] V.V.Jaya Rama krishniah, D.V.Chandra Sekar, Dr.K.Ramchand H Rao, "Predicting the Heart Attack Symptoms using Biomedical Data Mining Techniques" Volume 1, No. 3, May 2012 ISSN – 2278-1080 The International Journal of Computer Science & Applications (TIJCSA)

[7]. Han, J., Kamber, M.: "Data Mining Concepts and Techniques", Morgan Kaufmann Publishers, 2006

[8] MA.JABBAR, Dr.PRITI CHANDRA, B.L.DEEKSHATULU "Cluster Based Association Rule Mining For Heart Attack Prediction" JTAIT Vol. 32 No.2 October 2011.

[9] Milan Kumari, Sunila Godara "Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction", IJCST Vol. 2, Issue 2, June 2011

[10] Mr. Pankaj S. Kulkarni, Ms. V. C. Belokar, Dr. S. S. Sane, Prof. N. L. Bhale, "Heart Disease Prediction from Horoscope of a Person Using Data Mining", International Journal of Scientific and Research Publications, Volume 2, Issue 7, July 2012

[11] Dr. D. Raghu. T. Srikanth, Ch. Raja Jacub, "Probability based Heart Disease Prediction using Data Mining Techniques" IJCST Vol. 2, Issue 4, Oct - Dec. 2011, ISSN: 0976-8491 (Online) | ISSN: 2229-4333(Print)

[12] S.Sarumathi, N.S.Nithya, "Effective Heart Disease Prediction System Using Frequent Feature Selection Method" International ournal of Communications and Engineering Volume 01– No.1, Issue: 01 March2012

[13] Mrs.G.Subbalakshmi, Mr. K. Ramesh, Mr. M. Chinna Rao, "Decision Support in Heart Disease Prediction System using Naive Bayes" Indian Journal of Computer Science and Engineering (IJCSE), ISSN: 0976-5166 Vol. 2 No. 2 Apr-May 2011

[14] R. Sumathi, E. Kirubakaran "Enhanced Weighted K-Means Clustering Based Risk Level Prediction for Coronary Heart Disease" European Journal of Scientific Research ISSN 1450-216X Vol.71 No.4 (2012), pp. 490-500 © Euro Journals Publishing, Inc. 2012

[15] Dr. K. Usha Rani "Analysis of Heart Diseases Dataset Using Neural Network Approach" (IJDKP) Vol.1, No.5, September 2011

[16].www.webmd.com/heart-disease/guide/heart-disease-symptoms-types

Mrs. S.Vijayarani has completed MCA and M.Phil in Computer Science. She is working as Assistant Professor in the School of Computer Science and Engineering, Bharathiar University, Coimbatore. Her fields of research interest are data mining, privacy, security issues and data streams. She has published papers in the international journals and presented research papers in international and national conferences.

Ms. S.Sudha has completed M.Sc in Software Systems. She is currently pursuing her M.Phil in Computer Science in the School of Computer Science and Engineering, Bharathiar University, Coimbatore. Her fields of interest are privacy in data mining.