Heart Disease Prediction System Using Data Mining Techniques - A Review

Megha Shahi¹, Er. Rupinder Kaur Gurm²

¹Research Scholar, ²Asst. Professor

1²Dept of computer Science Engineering, RIMT-IET, Mandi Gobindgarh, India

Abstract-Data mining is the computer based process of analyzing large sets of data and then extracting the meaningful patterns. Data mining tools predict future trends which allow business to make knowledge-driven decisions. Data mining tools can answer the business questions which take much time to resolve in real world. The huge amount of data generated for prediction of heart disease are too complex to be processed and analyzed by traditional methods. Data mining provides the methodology to transform these mounds of data into useful information for decision making. It takes less time for the prediction of the disease by using data mining techniques with more accuracy. In this paper we survey different papers in which one or more algorithms of data mining are used for the prediction of heart disease. In some papers it is shown that a SVM provide effective and efficient accuracy as compared to other data mining techniques in the prediction of heart disease. Applying data mining techniques in case of heart disease treatment data can provide as reliable performance as that achieved in diagnosing heart disease.

Keywords:- Data mining, Heart disease, Data mining techniques.

I. INTRODUCTION

Data mining is the computer based process of extracting useful data from large sets of databases. The data mining is most useful in an exploratory analysis because it allow extraction of nontrivial information from the large volumes of data. The data mining tools are useful for answering the business questions for predicting the various diseases in the healthcare field. The disease prediction plays an important role in data mining. Multiple type of diseases can be predicted by using data mining such as heart diseases, breast cancer, lung cancer etc.

This paper explores the different data mining techniques which are used in health care field for the prediction of heart diseases. These techniques provide invisible patterns which can be used for health diagnosis in healthcare data. Data mining techniques are effective approach to the latest and indefinite patterns in the data. The information which is obtained can be used by the healthcare administrators to get better results. The heart disease was most crucial reason for heart attack victims in the countries like India, United States etc. Data mining techniques like association rule mining, clustering, classification algorithms such as SVM, Decision tree, C4.5 algorithm, Neural Network, Naïve Bayes are used to explore the different kind of heart based problems.

A. Heart Disease

One of the important organ of our body is heart. If the operation of a heart is not proper then it will affect the other body parts of human such as brain, Kidney etc. It is nothing more than a pump, which pumps blood throughout the body. If the circulation of blood in the body is inefficient then the organs like brain suffers and if the heart stops working altogether then death may occur within minutes. Life is completely dependent upon the efficient working of heart. The term heart disease refers to the disease of heart and the blood vessel system within it.

Coronary artery disease (CAD) is also known as atherosclerotic cardiovascular disease, coronary heart disease, atherosclerotic heart disease or ischemic heart disease (IHD), is the most common type of heart disease which cause heart attacks.

Fig 1.2.1: Coronary heart disease

The heart disease is caused by plaque building up along the inner walls of the arteries of heart and which narrows the arteries and reduces the blood flow to heart. While the symptoms and signs of coronary artery disease are noted in advanced state of heart disease and the most individuals with the coronary artery disease show no evidence of heart disease for decades as the disease progresses before the first symptom occur and often a “sudden” heart attack finally arises. The heart disease is main cause of heart attack according to our research work.

a. Risk factors for heart disease are:-

Some main risk factors of heart disease according to our survey are given below:-

Family history of heart disease: - The most of people know that the heart disease can run in families. If anybody has a family history of heart disease then he/she may be at greater risk for heart attack, stroke and some other heart diseases.

Smoking: - The smoking is major cause of heart attack, stroke and other arterial disease. Nearly 40% of people die from smoking tobacco due to heart and blood vessel diseases. A smoker’s risk of heart attack reduces only after one year of not smoking.

Cholesterol: - The abnormal levels of lipids in the blood are risk factor of heart diseases. Cholesterol is a substance found among the lipids in the bloodstream and in all the body cells. The high level of triglyceride (the most common type of fat in body) combined with the high levels of LDL...
The lack of exercise is one of the health problems such as heart disease, high blood pressure, healthy weight. Obesity puts anybody at higher risk for condition of any person significantly above his or her ideal weight. It increases the risk of having the heart attack or stroke and developing a heart failure, kidney failure and other peripheral vascular disease.

Obesity: - The term obesity is used to describe the health condition of any person significantly above his or her ideal weight. Obesity puts anybody at higher risk for health problems such as heart disease, high blood pressure, stroke, diabetes etc.

Lack of physical exercise: - The lack of exercise is one of the risk factors for developing coronary artery disease (CAD). Lack of physical exercise increases the risk of CAD because it also increases the risk for high blood pressure and diabetes.

II. LITERATURE SURVEY

The paper aims at reviewing the various data mining techniques introduced in the recent years for heart disease prediction. Different data mining techniques have been used in the diagnosis of different heart disease datasets. Some of the papers used only one technique for diagnosis of heart disease and other researchers use more than one data mining techniques for diagnosis of the heart disease.

- **Parisa Narvaei et.al [1]** “Application of Multilayer Perceptron Neural networks and Support Vector machines in Classification of Healthcare Data” In this paper the researcher compare accuracy of multilayer perceptron neural networks and support vector machine on heart disease dataset. It also analyzed the effectiveness of support vector machine using a dataset of 303 patients in classification. In this research work the result show support vector machine is able to classify more accurately then multilayer perceptron Neural Networks.

- **DeepaliChanda et.al. [2]** “Diagnosis of Heart Disease Using Data Mining Algorithm” In health concern business the data mining plays a significant task for predicting the various diseases. The numeral number of tests must be requisite from patient data for detecting a disease. However using the data mining techniques can reduce the number of tests that are required. Cardiovascular disease is principle source of deaths widely and the prediction of the heart disease is significant at untimely phase. In order to reduce the number of deaths because of heart diseases there has to be a quick detection technique.

- **M.Akhiljabbar, et.al.[4]** “Heart Disease Prediction System Using Associative Classification and Genetic Algorithm” In this researcher want to proposed that heart disease is the single largest cause of death in the developed countries and one main contributors to the disease burden in the developing countries. The data from registrar general of India and Andhra Pradesh coronary heart disease cause 30% of deaths in the rural areas. So there is very much requirement to develop a decision support system for the prediction of heart disease patients.

- **Ashish Kumar Sen1 [10]** “A Data Mining Technique for Prediction of Coronary Heart Disease Using Neuro-Fuzzy Integrated Approach Two Level” The healthcare institutes enrich the repository of the patients. The author proposed that he implemented the association rules mining based novel idea for finding the co-occurrences of heart disease carried by a patient using the healthcare repository.

- **Asghar, Set al[3]** “Automated Data Mining Techniques: A Critical Literature Review “ In this paper researcher want to propose that the data mining techniques has emerged as one of major research domain in the recent years in order to extract the implicit and useful information. This information can be comprehended by humans easily. This information extraction was computed and evaluated manually using the statistical techniques.

- **Niti Guru et al. [5]** “Decision Support System for Heart Disease Diagnosis Using Neural Network” In this the researcher uses neural networks for prediction of heart disease among patient’s. The experiments were carried out on a sampled data set of the patient’s records. The supervised network has been advised for diagnosis of the heart disease. The training was carried out with the help of the back propagation algorithm.

- **K. Srinivas et al. [6]** “Application of Data Mining Techniques in Healthcare and Predication of Heart Attacks” In this paper classification based data mining algorithms such as Rule based, Decision tree, Naive Bayes and Artificial Neural Network are used on massive volume of healthcare data. Tanagra, the data mining tool was used for exploratory data analysis, statistical learning algorithms and machine learning. The performance of different classifiers is evaluated and their results are analyzed. The results of comparison are based on the 10 tenfold cross validations.

- **N. Deepika et al. [7]** “Association rule for classification of Heart Attack Patients” In this paper extraction of significant patterns from the heart disease data warehouse was presented. The heart disease data warehouse contains the screening clinical data of the heart patient’s. Initially, the data warehouse is preprocessed to make mining process more effective and then association rule was used on dataset. Later applied equal interval binning with the approximate values based on the medical expert advice on heart attack data. The result is compared on 10 tenfold cross validation.

- **D. Shanthi, et al. [8]** “Designing an Artificial Neural Network Model for the Predication of Thromboembolic Stroke” In this paper functional model of ANN was proposed to aid existing diagnosis methods. The Back propagation algorithms was used to train ANN
architecture and test various categories of stroke disease. The data was standardized so that error free results are generated. The data were analyzed in the dataset to define data anomalies and column parameters.

- **Dangare et al.[9]** “Improved Study of Heart Disease Prediction System using the Data Mining Classification Techniques” In this paper the researcher propose a prediction system for heart disease that contain huge amount of data which is used to extract hidden information for making intelligent medical diagnosis. The main objective of this paper was to build intelligent heart prediction system that provide diagnosis of heart disease using historical heart dataset. In this Multi-layer Perceptron Neural Networks (MLPNN) is used to maps a set of input data onto a set of appropriate layer.

A. **Data Source**

A total 76 medical attributes (factors) which are numeric-valued have been obtained from different Heart Disease database websites. While the database have 76 raw attributes but only 14 of them are used actually which are listed below.

1. **Age**- It take age in years as the input.
2. **Gender**- It take two values as the input (value 1=Male and value 0=Female).
3. **Pain Type**- It take four values as the input which show the type of chest pain as value 1= typical type-1 angina, value 2= typical type angina, value 3= non-angina pain, value 4= asymptomatic.
4. **Trest Blood Pressure**- The resting blood pressure (in mm Hg) on admission to the hospital.
5. **Chol**- The serum cholesterol which is in mg/dl.
6. **Fasting Blood Sugar**- It take two values as the input (value 1= FBS>120 mg/dl and value 0= FBS<120 mg/dl).
7. **Restecg**- The resting electrographic results take three values as the input (value 0= normal, value 1= having ST-T wave abnormality, value 2= showing definite left ventricular hypertrophy).
8. **Thalach**- The maximum heart rate achieved by the patient.
9. **Exang**- The exercise induced angina which take two values as the input (value 1= Yes and value 0= No).
10. **Old peak**- The ST depression induced by exercise relative to rest.
11. **Slope**- The slope of peak exercise ST segment take three values as the input (value 1= unsloping, value 2= flat and value 3= down sloping).
12. **CA**- The number of major vessels colored by fluoroscopy which take three values as the input (value= 0-3).
13. **Thal**- It take three input values (value 3= normal, value 6= fixed defect and value 7= reversible defect).
14. **Num**- This is diagnosing attribute which have two input values (value 0= < 50% diameter narrowing i.e. no heart disease) and value 1= >50% diameter narrowing i.e. has heart disease).

### III. DATA MINING TECHNIQUES USED

**K-nearest-neighbor (KNN)**: The classification is one of most fundamental and simple classification methods and it is one of the first choices for the classification study when there is little or no prior knowledge about the distribution of data. The K-nearest-neighbor classification was developed from the need to perform different analysis when parametric estimates of the probability densities are unknown or difficult to find. In 1951, the unpublished US Air Force School of Aviation Medicine report (Fix and Hodges) introduced a non-parametric method for pattern classification that also known as the k-nearest neighbor rule. In 1967, some of formal properties of k-nearest-neighbor rule was worked out for instances it show that for k=1 and n is infinite k-nearest-neighbor classification is bounded by twice the Bayes error rate.

**Naïve Bayes**: - It is a simple technique for constructing classifiers, modals that assign the class labels to the problem instances and represented as vectors of feature values where class labels are drawn from finite set. Naïve Bayes is not a single algorithm for training such classifiers but it is a family of algorithms based on some common principle- All Naïve Bayes classifiers assume that the value of particular feature is independent of the value of some other feature when class variable is given.

**SVM**: - The support vector machine training algorithm build a model that assigns new examples into one category or other which make it a non-probabilistic binary linear classifier. The SVM model is a representation of examples as the points in space which are mapped so that the examples of separate categories are divided by clear gap that is as wide as possible. The new examples are then mapped into predicted and same space to belong to a category based on which gap side they fall.

**Decision Tables**: - The decision tables are precise yet compact way to model the complex rule sets and their corresponding actions. Decision tables like flowcharts, switch-case statements, if-then-else and associate conditions contain actions which are perform for analysis.

**Decision Tree**: - Decision trees are non-parametric supervised learning method which is used for classification of dataset. The main aim of decision tree is to build a model that predicts the value of a target variable by learning simple decision rules from dataset. Decision tree classify instances by starting from the root of tree and moving through it until a leaf node.

**Neural Network**: - An artificial neural network (ANN) is often called as “neural network (NN)” which is a mathematical model or computational model based on biological neural network. In other words it is an emulation of biological neural system and it also maps a set of input data onto the set of appropriate output data. It consists of three layers input layer, hidden layer and output layer. There is connection between each layer and weights are assigned to each of connection.

IV. OPEN SOURCE TOOLS FOR DATA MINING

**WEKA Tool**: WEKA is a data mining system that was developed by the University of Waikato in New Zealand that implements the data mining algorithms using JAVA language. WEKA is a state of the art facility for developing machine learning techniques for real world data mining problems. It is used to implement algorithms for classification, data preprocessing, and clustering and association rules.
TANAGRA: Tanagra is a free data mining software for research and academic purposes. It proposes several data mining method from statistical learning, exploratory data analysis, machine learning and database area. It is an open source project as every researcher can access source code and add own algorithms as it conforms to the software distribution license.

MATLAB: MATLAB is an interactive environment for numerical computation, programming and visualization. Using MATLAB we can develop algorithms, analyze data, and create applications and models. The tool and built-in math functions enable to explore multiple approaches.

Orange: Orange is an open data analysis for experts and novice. Data Mining used through components for machine learning and visual programming of python scripting. The add-ons used for text mining and bio-informatics.

.NET Framework: .NET framework is a software framework that runs primarily Microsoft windows and provide languages interoperability across the several programming languages. .NET Framework provide a comprehensive and consistent application to developers that has visually stunning user experiences and secure communication.

V. CONCLUSION

The main objective of our work is to provide a study of different data mining techniques that can be used in automated heart disease prediction systems. Various data mining classifiers are defined in this work which has emerged in recent year for effective and efficient heart disease diagnosis. The analysis shows that different technologies are used in all the papers by using different number of attributes. So different technologies used show different accuracy to each other. In some papers it is shown that SVM provide effective and efficient accuracy about 85% as compared to other data mining techniques in prediction of heart disease. So applying data mining techniques help health care professionals in the diagnosis of heart disease is having success, the use of data mining techniques to identify a suitable treatment for the heart disease patients has received less attention.

REFERENCES


[21] Carlos Ordonez, Edward Omiechinski, “Mining Constrained Association Rules to Predict Heart


