Predictive Data Mining Procedures for the Prediction of Coronary Artery Disease

Swathi P¹, Yogish H K², Sreeraj R S³

¹Student, ²PG Coordinator, East West Institute of Technology, Bangalore, Karnataka, India
³Senior Analyst at Accenture Services Pvt. Ltd., Bangalore, Karnataka, India

Abstract—Data mining can be defined as the process of discovering actionable information from huge volumes of data. In this paper, we deal with the prediction of heart disease, specifically coronary artery disease, by taking into account the wealth of clinical data available in the healthcare systems. Certain mathematical analysis to derive appropriate patterns and trends that exist in data supports in predictive data mining. Typically, these patterns cannot be derived by traditional data exploration because the relationships are either too complex or there prevails too much data to be dealt with.

Keywords—coronary artery disease, decision tree induction, Naïve Bayes, neural network, predictive data mining

I. INTRODUCTION

Data mining is the procedure of exploring large stores of data to discover patterns and trends that go beyond simple analysis. This process of extracting the trends and patterns from large volumes of data leads to efficient findings in various perspectives. It plays an important role in the field of Information Technology as it facilitates the knowledge discovery from various domains.

Data mining uses advanced mathematical algorithms to segregate the data and get the measure of the probability of future events. A data mining algorithm is a set of assessments that creates a data mining model from the voluminous data. The algorithm uses the results for creating the mining model by defining the favoring parameters. The parameters so obtained are then applied across the entire data set to extricate the patterns that have practical value and circumstantial statistics.

Data mining, also known as Knowledge Discovery, plays an important role in the field of Information Technology as it facilitates in the knowledge discovery from various domains.

A. Medical data

Data mining can also be used to mine the medical data. The Healthcare domain produces huge amount of data about patients, diseases, diagnosis, medicine etc.

It contributes a lot to the process of identification and prediction of various sort of metabolic syndromes. Thereby, various sorts of diseases can be predicted and diagnosed at the early stages. One of the major provocations being faced by the healthcare organizations, such as hospitals, medical centers and so on is the distribution of excellent quality services at reasonable price. Quality service indicates identifying the nature of illness by examining the symptoms in patients accurately and regulating the treatments that are productive.

B. Heart diseases

In this paper we consider the concept of heart diseases. We shall go through some factors that can cause the occurrence of certain heart conditions such as those that affect one’s heart's muscle, valves or rhythm and so on. The conditions related to heart require immediate attention, since every passing second has the value of a life in it as it passes by. Heart diseases can be caused by various factors out of which certain common factors can be mentioned as, age, sex, family history, high blood pressure levels, high cholesterol levels, obesity, stress, diabetes[7] etc. By considering these concepts and their respective measures we can predict the occurrence of heart disease. In this paper we take into consideration certain other measures as liver fatigue, triglycerides, blood urea nitrogen and enzymes. Several predictive data mining techniques are applied on these factors and an optimal and accurate result is obtained with high performance features. There are numerous types of heart diseases out of which we deal with coronary artery disease.

C. Coronary artery disease

Also called as coronary heart disease, it is the most prevalent type of heart disease across the world. It is a circumstance in which normal flow of blood and oxygen to the heart is restricted. The supply is restricted due to the plaque deposits in the coronary blood vessels. A plaque is a thick, wax-like coating that forms on the inner walls of vessels contributing to the heart or other areas in the body. Plaques are typically made up of cholesterol and other fats. They build up over time, restricting blood flow.

339
II. PROBLEM STATEMENT

The predictions of coronary artery disease are based only on certain fixed set of attributes. The patient is made to undergo various tests but only a few parameters from within the test reports are considered leaving very minute findings that would contribute towards prediction of heart disease more effectively. The algorithms so far used either concentrate only on accuracy or performance. A hybrid methodology is needed which can assure both accuracy and performance efficiently.

III. LITERATURE SURVEY

A. Prediction of heart disease using genetic algorithm for selection of optimal reduced set of attributes

Shruti Ratnakar et al. [2] proposed a Heart Disease Prediction System which is implemented as web-based questionnaire application using Naïve Bayes. They compared two papers and found that in the paper based on classification model, the intensity of risk level of heart disease is not predictable. Improvement is required to increase its consistency and efficiency. The paper based on Bayes’ Rule is quite effective but further enhancement can be done in terms of numbers of attributes used.

B. Intelligent Heart Disease Prediction System Using Data Mining Techniques

In this paper [3], Sellappan Palaniappan et al. proposed the development of An Intelligent Heart Disease Prediction System (IHDPS) using data mining techniques as Naïve Bayes, Neural Network, and Decision Trees which is based on the limited number of attributes. The attribute list may need to be expanded to provide a more comprehensive diagnosis system. Another limitation is that it only uses categorical data.

C. A Data mining Model for predicting the Coronary Heart Disease using Random Forest Classifier

A. Sheik Abdullah et al, in their paper [4], developed a mining model with the random forest classification algorithm that provides the functionality of predicting the occurrence of various events related to each patient record which should have at least one of the following in their history: Angina, Acute Myocardial Infarction (AMI), Percutaneous Coronary Intervention (PCI), and Coronary Artery Bypass Graft (CABG). Thereof the prevention of risk factors with its associated cost metrics and an improvement in overall prediction accuracy is observed.

D. Comparison of ANNs, Fuzzy Logic and Neuro-Fuzzy Integrated Approach for Diagnosis of Coronary Heart Disease: A Survey

In the paper [5], Nitin Kumari et al. presented a survey on diagnosis of coronary heart disease using ANN, fuzzy logic and neuro-fuzzy integrated approach. All these techniques were compared to find out which one was better among the three. It had been found that although ANN and fuzzy logic have a lot of advantages but these techniques have some disadvantages too. Neuro-fuzzy approach is the combination of ANN and fuzzy logic. It comprises of advantages of both ANN and fuzzy logic. Advantages of artificial neural networks include massive parallelism, robustness, and learning in data-rich environment. They presented a survey on diagnosis of coronary heart disease diagnosis using three soft computing techniques.

E. Decision Support in Heart Disease Prediction System using Naïve Bayes

Mrs.G.Subbalakshmi et al have, in the paper [6], developed a Decision Support in Heart Disease Prediction System providing decision support using Naïve Bayesian Classification technique. The system extracts hidden knowledge from a historical heart disease database. The model proposed could answer complex queries, each with its own strength with respect to ease of model interpretation, access to detailed information and accuracy.
Data mining techniques maybe applied to construct decision model or system for clinical procedures such as prediction, diagnosis and treatment plans. By applying the above said data mining techniques on the parameters furnished above, an efficient coronary artery disease (CAD) prediction system can be obtained. The study on how to bring the Random Forest classifiers into picture for achieving the same is going on. Once the better solution is obtained, the work will progress in that direction.

V. CONCLUSION

Data mining is a process of scrutinizing data from different outlooks and assembling the knowledge from it. Data which is a great asset to healthcare organizations have to be transformed into information first. Predictive data mining is the process of automatically creating a classification model from a set of examples, called the training set, which belongs to a set of classes. Predictive data mining techniques maybe applied to construct decision model or system for clinical procedures such as prediction, diagnosis and treatment plans. By applying the above said data mining techniques on the parameters furnished above, an efficient coronary artery disease (CAD) prediction system can be obtained. The study on how to bring the Random Forest classifiers into picture for achieving the same is going on. Once the better solution is obtained, the work will progress in that direction.

REFERENCES


The data mining techniques that are used are Naïve Bayes, decision trees and neural networks. The Naïve Bayes classifiers have the minimum error rate in comparison to all other classifiers. Bayesian classifiers have exhibited high accuracy and speed when applied to large databases. Decision tree builds classification or regression models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. Decision trees can handle both categorical and numerical data. A neural network usually involves a large number of processors operating in parallel, each with its own small sphere of knowledge and access to data in its local memory. Typically, a neural network is initially "trained" or fed large amounts of data and rules about data relationships.

By applying these algorithms to the newer set of attributes, a more efficient system is developed. Both, in terms of, accuracy and performance are obtained in a fraction of some milliseconds. The system that is proposed helps the medical practitioners to improve their practice as well as provide quality service.

IV. PROPOSED SYSTEM

In heart disease prediction system, the input attributes play a major role for efficient prediction. Certain factors that are considered are as follows [8]:

- Age: age in years
- Sex: sex (1 = male; 0 = female)
- Chest pain type: cp
  - Value 1: typical angina
  - Value 2: atypical angina
  - Value 3: non-anginal pain
  - Value 4: asymptomatic
- Cholesterol (chol): serum cholesterol in mg/dl
- Lipid levels: Low high-density-lipoprotein (HDL) levels (<0.91 mmol/L [<35 mg/dL]) and high low-density-lipoprotein (LDL) levels are independently associated with CAD (HDL especially in women).
- Diabetes mellitus: Increased risk is related to hyperglycemia and hyperinsulinemia, both of which are atherogenic.
- Hypertension: Systolic and diastolic blood pressures are independent risk factors, but the systolic is the preferred marker.
- Smoking: Promotes atherogenesis, ischemia, and thrombogenesis.
- Family history: Especially premature disease (parent with MI before age 60), but the independent effect is difficult to quantify.
- Left ventricular hypertrophy: A powerful independent risk factor for CAD (more so than diabetes mellitus or smoking).
- Homocysteine: Higher levels are associated with a 20% to 40% increased risk of cardiovascular events.
- C-reactive protein: The prevalence of CAD is increased by 50% for each doubling of the C-reactive protein level. The pathophysiological significance of this relationship remains unclear.

Other risk factors: These include obesity, high levels of uric acid, triglycerides, lipoprotein Lp, tissue plasminogen activator antigen, fibrinogen, and leukocytes.


