Intelligent Risk Analysis and Prediction of Heart Disease Using Data Mining Techniques

Ms.A.Thamizhiniyal M.C.A., M.Phil., Assistant Professor, Department of Computer Science, Bon Secours College for Women -Vilar Bypass,Thanjavur-613006, Tamilnadu, India.

A.Vidhya, M.Sc., Computer Science, Department of Computer Science, Bon Secours College for Women -Vilar Bypass,Thanjavur-613006, Tamilnadu, India.

Abstract—The World Health Organization estimates that by 2030 there will be approximately 350 million young people (below 30 to 40 years) with heart diseases associated with renal complications, stroke and peripheral vascular disease. Heart disease is most common in present era. The treatment cost of heart disease is not affordable by most of the patients. So we can reduce this problem by a Heart Disease Prediction System (HDPS). It is helpful for earlier diagnosis of heart disease. Data mining techniques are used for the construction of HDPS. In health care field some systems use large healthcare data in varied forms such as images, texts, charts and numbers. Our aim is to analyze the risk factors and system conditions to detect heart disease early. Using effective methods to identify and extract key information that describes aspects of developing a prediction model, sample size and number of events, risk predictor selection. Using the new algorithm called Intelligent Support Vector [ISV], we can easily identify the heart disease with various attributes and risk factor specifications. Based on these parameters, the analysis of high risk factors of developing heart disease is identified using mining principles. Use of data mining algorithms will result in quick prediction of disease with high accuracy.

Index Terms — Data Mining, Health Care Mining, Risk Attributes, Risk Analysis.

I. INTRODUCTION

The main aim is to predict the Heart Diseases based on the parameters, the analysis of high risk factors of developing heart diseases are identified using Intelligent Support Vector [ISV] Algorithm with rule mining techniques. We propose a clinical application of mining techniques to identify sets of abnormal conditions (and the patient subpopulations who suffer from these conditions) that imply significantly increased risk of heart diseases. Intelligent Rule mining principles and Intelligent Support Vector [ISV] Algorithm on this approach to produce extensive results based on risk factor analysis and symptoms specified. The computer aided systems help the doctor as a tool for predicting and diagnosing heart disease. The medical field is dealing with huge amount of data regularly as well as handling that large data by traditional way may affect the results.

Our main contribution is a comparative evaluation of these extended summarization techniques that provides guidance to practitioners in selecting an appropriate algorithm for a similar problem. Heart is one of the important organs in blood circulatory system of all living organism. There are many elements which make problems to heart. They are smoking, poor eating methodology, high pulse, cholesterol and high blood pressure etc. The diagnosis of heart disease in earlier stage is a challenging problem for the medical industry. Data mining based heart disease prediction system can help in determining the heart disease during early stages. The prediction system helps to reduce the high risk of heart disease. Prediction is done based on the current data given to the system. For building Heart Disease Prediction System we use Microsoft Visual Studio tool. Diagnosis of heart disease is a complicated task in medical field. So it is needed to develop an efficient disease prediction system for the earlier detection of disease. One of the earliest systems for heart disease detection was proposed by Meghna Sharma et al. and they propose a hybrid technique in data mining for heart disease prediction. Here a prototype which can extract unknown data related with heart disease from a past heart disease database record is developed.

They put an idea of hybrid technique methodology which can be implemented in future to have accuracy of almost 99% or with least error. In past researches, authors analyses various papers on heart on using different data mining technologies. Also they make comparative study on the performance of three classifiers like Naïve Bayesian classifier, Decision trees and Probabilistic Neural Network (PNN).

The analysis showed that artificial neural networks gave accuracy of 94.6 percentages in heart disease prediction. In earlier systems, author developed an intelligent Heart Disease Prediction System. This system is builds using Naive Bayes algorithm and it also uses a smoothing technique (Jelinek mercer smoothing) to improves the performance. Here the
system is proposed using Cleveland heart disease database as the input dataset. Each attribute of the dataset were fed to the Naive Bayesian classifier and it produces the prediction results based on the classification process. We can conclude that efficiency can be improved with the use of smoothing technique. This model could answer complex queries which traditional decision support systems cannot. In earlier systems, Bhuvaneswari Amma proposes a system using genetic algorithm and neural network, which is helpful for cardiovascular disease prediction. Observed demerits of this system are: less accuracy and no extraction of hidden data.

Fig.1 Data Flow Modeling

It consists of training dataset and user input as the test dataset. Weka data mining tool with api is used to implement the heart disease prediction system. The source code of Weka is in java. The system is designed with java swing and use Weka api to call the different methods of Weka. The components used are instances, different classifiers and methods for evaluation. Supervised learning method is used here. A supervised learning algorithm analyses the training data and deduces a function from the labelled training set. It can be used for mapping new examples. The training data obtained from cleveland heart disease database is the training example. This training data consist of the class label and its corresponding value. Naive Bayesian, J48 and Random Forest classifiers are supervised learning algorithms. They learn from the provided training examples. When a new instance with same attributes as in training data with different values other than those in the training example comes, these algorithms correctly classify the new instance based on the generalization created from the training set. Naive Bayesian, J48 and Random Forest classifiers are classify the new observation into two categories on the basis of training dataset. The training dataset is in the ARFF format. The training set consists of 14 attributes including the class attribute. Heart disease prediction system accepts input from the user through a graphical user interface. All the attributes needed for classification is received from a text field. The graphical user interface is built using swing. The next process is to transfer the user input obtained from graphical user interface into a file of CSV (Comma separated Value) extension. Then the CSV file is converted into ARFF file. Weka api provide native methods for converting from CSV to ARFF. The converted user input is treated as test data. The test data set will contain all the attributes of training dataset. If the user did not enter an attribute value a "?" will be assigned at the value of that corresponding attribute. Weka will handle this missing value. This test data is run on Naive Bayes algorithm, Random Forest algorithm, and J48 algorithm.

II. SYSTEM ANALYSIS

A. Existing System Summary

Heart Disease index is in essence a predictive model that assigns a score to a patient based on their estimated risks of suffering. Collins et al. conducted an extensive survey of Heart Disease, which indices describing the risk factors and the modeling technique that these indices utilized. Lots of past classification approaches such as SVM, Neural networks, Bayesian Classification and so on are derived but all are facing struggles in certain limits. However all these methods which is developed in past systems fails to maintain the risk attributes and visualizing the summary to patients. Association Rules are used to indexing the disease summary, but the resulting does not constitute a Heart Disease index because the study does not designate a particular outcome of interest. And they do not assess or predict the risk of Heart Disease in patients, but they discovered some significant associations between diagnosis codes.

Disadvantages

- Comparatively Low in Performance because of classical Map Reduce and Classification techniques.
- Because of traditional classification approaches system requires large storage space to maintain the risk factor and patient records.
- Difficult to Search the medicinal records in a single server.
- Maintenance cost is high.
- Low level of efficiency.

III. RELATED WORKS

A. Prediction of Heart Disease Using a Hybrid Technique in Data Mining Classification - Ankita Dewan and Meghna Sharma - 2010

Heart disease prediction is treated as most complicated task in the field of medical sciences. Thus there arises a need to develop a decision support system for detecting heart disease.
of a patient. In this paper, we propose efficient genetic algorithm hybrid with the back propagation technique approach for heart disease prediction. Today medical field have come a long way to treat patients with various kind of diseases. Among the most threatening one is the Heart disease which cannot be observed with a naked eye and comes instantly when its limitations are reached. Bad clinical decisions would cause death of a patient which cannot be afforded by any hospital. To achieve a correct and cost effective treatment computer-based and support Systems can be developed to make good decision. Many hospitals use hospital information systems to manage their healthcare or patient data.

These systems produce huge amounts of data in the form of images, text, charts and numbers. Sadly, this data is rarely used to support the medical decision making. There is a bulk of hidden information in this data that is not yet explored which give rise to an important query of how to make useful information out of the data. So there is necessity of creating an excellent project which will help practitioners predict the heart disease before it occurs. The main objective of this paper is to develop a prototype which can determine and extract unknown knowledge (patterns and relations) related with heart disease from a past heart disease database record. It can solve complicated queries for detecting heart disease and thus assist medical practitioners to make smart clinical decisions which traditional decision support systems were not able to. By providing efficient treatments, it can help to reduce costs of treatment.

B. Heart Disease Prediction System using Naïve Bayes and Jelinek-mercer smoothing - Ms.Rupali R.Patil - 2014

Heart disease is most common in present era. The treatment cost of heart disease is not affordable by most of the patients. So we can reduce this problem by a Heart Disease Prediction System (HDPS).It is helpful for earlier diagnosis of heart disease. Data mining techniques are used for the construction of HDPS. In health care field some systems use large healthcare data in varied forms such as images, texts, charts and numbers. But this data is hardly visited and are not mined. This problem can be avoided by introducing HDPS. This system would enhance medical care and it can also reduce the costs. The system can handle complex queries for detection of heart disease and thus help to make intelligent medical decisions. This paper proposes a HDPS based on three different data mining techniques. The various data mining methods used are Naive Bayes, Decision tree (J48), Random Forest and WEKA API. The system can predict the likelihood of patients getting a Heart disease by using medical profiles such as age, sex, blood pressure, cholesterol and blood sugar. Also, the performance will be compared by calculation of confusion matrix. This can help to calculate accuracy, precision, and recall. The overall system provides high performance and better accuracy.

C. Predictions in Heart Disease Using Techniques of Data Mining - Monika Gandhi and Dr. Shailendra Narayan Singh - 2015

As huge amount of information is produced in medical associations (healing facilities, therapeutic focuses) yet this information is not properly utilized. The health care system is "data rich" however "knowledge poor ". There is an absence of successful analysis methods to find connections and patterns in health care data. Data mining methods can help as remedy in this circumstance. For this reason, different data mining techniques can be utilized. The paper intends to give details about various techniques of knowledge abstraction by using data mining methods that are being used in today's research for prediction of heart disease. In this paper, data mining methods namely, Naive Bayes, Neural network, Decision tree algorithm are analyzed on medical data sets using algorithms.

D. Cardiovascular Disease Prediction System using Genetic Algorithm and Neural Network - Bhuvaneswari Amma N.G. - 2014

Medical Diagnosis Systems play a vital role in medical practice and are used by medical practitioners for diagnosis and treatment. In this paper, a medical diagnosis system is presented for predicting the risk of cardiovascular disease. This system is built by combining the relative advantages of genetic algorithm and neural network. Multilayered feed forward neural networks are particularly suited to complex classification problems. The weights of the neural network are determined using genetic algorithm because it finds acceptably good set of weights in less number of iterations. The dataset provided by University of California, Irvine (UCI) machine learning repository is used for training and testing. It consists of 303 instances of heart disease data each having 14 attributes including the class label. First, the dataset is preprocessed in order to make them suitable for training. Genetic based neural network is used for training the system. The final weights of the neural network are stored in the weight base and are used for predicting the risk of cardiovascular disease. The classification accuracy obtained using this approach is 94.17%.

E. Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques - Chaitrali S. Dangare - 2012

Healthcare industry is generally "information rich", but unfortunately not all the data are mined which is required for discovering hidden patterns & effective decision making. Advanced data mining techniques are used to discover knowledge in database and for medical research, particularly in Heart disease prediction. This paper has analysed prediction systems for Heart disease using more number of input attributes. The system uses medical terms such as sex, blood pressure, cholesterol like 13 attributes to predict the likelihood of patient getting a Heart disease. Until now, 13 attributes are used for prediction. This research paper added two more attributes i.e. obesity and smoking. The data mining classification techniques, namely Decision Trees, Naive Bayes, and Neural Networks are analyzed on Heart disease database. The performance of these techniques is compared,
based on accuracy. As per our results accuracy of Neural Networks, Decision Trees, and Naive Bayes are 100%, 99.62%, and 90.74% respectively. Our analysis shows that out of these three classification models Neural Networks predicts Heart disease with highest accuracy.

F. An Efficient Classification Tree Technique for Heart Disease Prediction - S. Vijiyarani et. al. - 2013

The data mining can be defined as discovery of relationships in large databases automatically and in some cases it is used for predicting relationships based on the results discovered. Data mining plays a vital role in 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. Various data mining techniques are available for predicting diseases namely Classification, Clustering, Association rules and Regressions. This paper analyzes the classification tree techniques in data mining. The aim of this paper is to investigate the experimental results of the performance of different classification techniques for a heart disease dataset. The classification tree algorithms used and tested in this work are Decision Stump, Random Forest, and LMT Tree algorithm. Comparative analysis is done by using Waikato Environment for Knowledge Analysis or in short, WEKA. It is open source software which consists of a collection of machine learning algorithms for data mining tasks.

A. Proposed System Summary

In our proposed system, we extend our previous study by incorporating a wide variety of predictors alongside the diagnosis codes. We use the advanced methodology of combining survival analysis and distributional rule mining techniques. In this approach, we review the basic concepts underlying this methodology: survival analysis, rule mining and Intelligent Support Vector [ISV]. Disease Investigation takes place when we obtain full information about a patient and system dynamically classifies and generates the summary regarding patient health. For example, if a patient drops out of the weight level, we may not know that he/she is suffered, but with our approach we can easily identify and get back soon from the disease with the help of earlier predictions based on risk factor analysis and symptom attributes. The ability to use such partial information and the ability to take time into account are the key characteristics of survival analysis making it a mainstay technique in clinical research.

Advantages
- Timing consumption is low for order processing.
- Maintenance cost is comparatively low.
- High accuracy to retrieve medicinal details from the hospital server.
- Quicker order processing.
- Provide the efficient authentication and authorization process to achieve privacy and security to the system.

IV. System Implementation

A. Identify Patients Risk Attributes

In response to the pressing need to identify patients at high risk of heart disease early, numerous diabetes risk indices (risk scores) have been developed. Some of these indices gained acceptance in clinical practice and are used as guidance in treatment: patients presenting high risk scores are treated more aggressively. These scores only provide a quantification of the risk, they are not suggestive of the factors that may have caused the elevation of the risk. Moreover, these scores utilize individual risk factors in an additive fashion without taking interactions among them into account. Heart Disease is the major physical syndrome, which is a constellation of diseases including hyperlipidemia (elevated triglyceride and low HDL levels), hypertension (high blood pressure) and central obesity (with body mass index exceeding 30 kg/m2) and so on. These diseases interact with each other, with cardiac and vascular diseases and thus understanding and modeling these interactions is important.
B. Applying Rule Mining Techniques

Rule Mining Techniques such as Association rules, Apriori and so on are implications that associate a set of potentially interacting conditions with elevated risk. The use of association rules is particularly beneficial, because in addition to quantifying the Heart Disease risk, they also readily provide the physician with a “justification”, namely the associated set of conditions. This set of conditions can be used to guide treatment towards a more personalized and targeted preventive care or Heart Disease management. While association rules themselves can be easily interpreted, the resulting rule sets can sometimes be very large, eroding the interpretability of the rule set as a whole. Especially, in this work, we consider a rich set of risk factors, namely co-morbid diseases, laboratory results, medications and demographic information that are commonly available in Electronic Medical Record (EMR) systems.

C. Symptoms and Attribute Summarization

A number of successful rule mining set summarization techniques have been proposed but no clear guidance exists regarding the applicability, strengths and weaknesses of these techniques. The focus of this manuscript is to review and characterize four existing association rule summarization techniques and provide guidance to practitioners in choosing the most suitable one. A common shortcoming of these techniques is their inability to take Heart Disease risk—a continuous outcome into account. In order to make these techniques more appropriate, we had to minimally modify them: we extend them to incorporate information about continuous outcome variables.

D. Disease Follow-Up Strategies

In this module we are concerned with two types of events: progression to Heart Disease and last follow-up. We define a status variable indicating whether a patient progressed to Heart Disease at any time during the study and we also define follow-up, which is the time from the beginning of the study to progression to Heart Diseases or last follow-up, whichever occurred earlier. Last follow-up denotes the last encounter when data is available about the patient. Note that patients who progressed to Heart Disease are no longer followed as far as survival models are concerned.

V. RESULTS AND DISCUSSION

The following figure, Fig.3 illustrates the Administrator Home Page view of the proposed system, in which this home page allows the administrator to navigate into the features in the proposed approach.

Fig.3 Administrator Home Page

The following figure, Fig.4 illustrates the Registration view of the proposed system, in which this registration page allows the user to register their identity into the system.

Fig.4 Registration Page

The following figure, Fig.5 illustrates the Patient Authentication Page view of the proposed system, in which it allows the patient to navigate into the main page, if the credentials are valid.

Fig.5 Administrator Authentication

The following figure, Fig.6 illustrates the Risk Factor Updation view of the proposed system.
Heart disease is the leading cause of death for both men and women. Know the warning signs and symptoms of a heart attack so that you can act fast if you or someone you know might be having a heart attack. The chances of survival are greater when emergency treatment begins quickly. This system mainly focuses on the experimental analysis of various approaches of heart attack disease prediction researches and proves the efficiency via proposed results of ISV.

VI. CONCLUSION

Heart disease is the leading cause of death for both men and women. Know the warning signs and symptoms of a heart attack so that you can act fast if you or someone you know might be having a heart attack. The chances of survival are greater when emergency treatment begins quickly. This system mainly focuses on the experimental analysis of various approaches of heart attack disease prediction researches and proves the efficiency via proposed results of ISV.

References