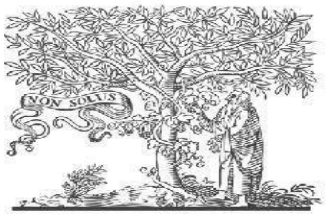
International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

COPY RIGHT



ELSEVIER
SSRN

IJIEMR Transactions, online available on 4th Jan 2014. Link :http://www.ijiemr.org/downloads.php?vol=Volume-03&issue=ISSUE-01

Title **DATA MINING TOOLS AND TECHNIQUES ANALYSIS ON HEART DISEASE PREDICTION**

Volume 03, Issue 01, Pages: 47–54.

Paper Authors

K. SRINIVAS, B. KAVITHA RANI

Jyothishmathi Institute of Technology & Science Karimnagar, Telangana, India

USE THIS BARCODE TO ACCESS YOUR ONLINE PAPER

To Secure Your Paper As Per UGC Guidelines We Are Providing A Electronic Bar Code

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

DATA MINING TOOLS AND TECHNIQUES ANALYSIS ON HEART DISEASE PREDICTION

[1]K. SRINIVAS,[2] B. KAVITHA RANI

[1.2]Associate Professor, Department of CSE, Jyothishmathi Institute of Technology & Science

Karimnagar, Telangana, India

ABSTRACT

The World Health Organization (WHO) estimated that cardiovascular diseases (CVD) are the major cause of mortality globally, as well as in India. They are caused by disorders of the heart and blood vessels, and includes coronary heart disease (heart attacks). Data mining acts as a major role in the construction of an intellectual prediction model for healthcare systems to detect Heart Disease (HD) using patient data sets, which support doctors in diminishing mortality rate due to heart disease. We have investigated three data mining techniques: Naïve Bayes, Artificial neural network, and J48 decision tree algorithms. Our Analysis Shows that these three classification models Naïve Bayes predicts heart disease with higher Accuracy. The successful application of data mining in highly visible fields like e-business, marketing and retail has led to its application in other industries and sectors. Among these sectors just discovering is healthcare. The Healthcare industry is generally "information rich", but unfortunately not all the data are mined which is required for discovering hidden patterns & effective decision making. Discovery of hidden patterns and relationships often goes unexploited. Advanced data mining modeling techniques can help remedy for this situation.

INTRODUCTION

Data mining is process of extracting useful information from large amount of databases. Data mining is most useful in an exploratory analysis because of nontrivial information in large volumes of data. Data mining is the process of extracting data for finding buried patterns which can be transformed into significant format. Data mining knowledge afford a user-oriented approach to new and concealed patterns in the data. The knowledge which is exposed can be used by the healthcare practitioners to get better quality of service and to reduce the extent of adverse medicine effect. Hospitals have to reduce the charge of medical tests. They can attain these consequences by employing suitable decision support systems. Health care data is enormous. It consists of patient centric data, resource organization data and altered data. Medical care organizations must have capability to explore data. Treatment records of millions of patients can be hoarded and data mining techniques will aid in answering numerous essential and decisive questions interrelated to health care. Data mining techniques has been performed

in healthcare domain. This realization is in the arouse of explosion of difficult medical data. Medicinal data mining can utilize the veiled patterns present in huge medical data which otherwise is left undiscovered. Data mining techniques which are useful to medical data include association rule mining for finding frequent patterns, prediction, classification and clustering. Data mining techniques are more useful in predicting heart diseases, breast cancer, lung cancer, diabetes and etc.

The electronic health record accumulates large amount of health information which are necessary to be mined for finding unknown information for effective decision making. Due to the daunting disease such as heart disease the mortality rate is increased every year. As the data is very huge, researchers feel tough to extract data. The data mining techniques are applied to sort this problem. With the help of patient's Electro cardiogram (EKG or ECG), Echocardiography (ECHO) test reports and doctor's practice, Diagnosis is being done. Medical diagnosis is yet challenging and complicated task that needs to be done efficiently and accurately in addition with patient's Electrocardiogram (EKG or ECG), Echocardiography (ECHO) test reports. A suitable computer based information and decision support should be supported for decreasing the cost while doing the process of clinical test reports.

2. LITERATURE REVIEW

Kim, Jae-Kwon, et al., [9] this paper proposes the Fuzzy Rule-based Adaptive Coronary Heart Disease Prediction Support Model (FbACHD_PSM), which stretches comfortable reference to coronary heart disease patients. The projected model uses a mining method approved by medical experts to deliver approvals.

Seera, Manjeevan, and CheePeng Lim [10] in this paper, a hybrid intelligent system that comprises of the Fuzzy Min-Max neural network, the Organization and Random Forest model, and Regression Tree is proposed, and its value as a decision support device for medical data classification is observed. The hybrid intelligent system targets to abuse the benefits of the constituent models and, at the same time, ease their impediments.

Bashir, Saba, Usman Qamar, and M. Younus Javed [11] the goal of the proposed research is to envisage the heart disease in a patient more precisely. The proposed structure customs common vote based novel classifier collaborated to amalgamate different data mining classifiers. UCI heart disease dataset is practiced for assessment and results.

Shabana, ASMI P., and S. Justin Samuel [12] different data mining techniques such as Decision Tree, Naive Bayes, Association Rule and Linear Regression are practiced to envisage the heart disease. Data mining techniques in overall diagnosis realistic over all disease treatment. Data sets explore if hybrid data mining techniques can attain comparable (or better) results in classifying appropriate actions as that attained in the diagnosis. In this paper, the proposed work is to more precisely predict the existence of heart disease with new attributes of the disease and using association rules.

Aljaaf, A. J., et al., [13] in this study, a multi-level risk assessment of developing

heart failure has been proposed, in which five levels of risks in heart failure can be predicted using C4.5 decision tree classifier. In contrast, we are enhancing the primary prediction of heart failure over concerning three core risk factors with the heart failure data set.

Bashir, Saba, UsmanQamar, and Farhan Hassan Khan [14] this research paper presents a novel classifier collaborative structure based on improved bagging approach with multi-objective weighted voting structure for analysis and prediction of heart disease. The proposed structure overwhelms the boundaries of orthodox performance by exploiting a collective of five heterogeneous classifiers: linear regression, Naive Bayes, instance-based learner, quadratic discriminant analysis and support vector machines.

Kim, Jaekwon, Jongsik Lee, and Youngho Lee [15] established model for CHD prediction must be aimed bestowing to rule-based procedures. In this study, a fuzzy logic and decision tree (classification and regression tree [CART])-driven CHD prediction model was advanced for Koreans. Datasets derived from the Nutrition Examination Survey VI (KNHANES-VI) and Korean National Health was exploited to produce the proposed method.

Joshi, Sujata, and Mydhili K. Nair [16] In this research, the classification-based data mining techniques are practiced to healthcare data. This research emphasis on the forecast of heart disease using three classification techniques namely Naïve Bayes, Decision Trees, and K Nearest Neighbor.

Chadha, Ritika, et al., [17] The objective of the research is to accumulate, organize and analyse the numerous data mining techniques that have been implicit and instigated in the latest years for Heart Disease Prediction. This paper attempts to highlight blatant assessments and focus to the boons and banes of each technique.

Choi, Edward, et al., [18] explored whether the practice of deep learning to typical temporal relations among proceedings in electronic health records (EHRs) would advance model performance in forecasting preliminary diagnosis of heart failure (HF) associated to conservative techniques that snub momentarily.

Saxena, Kanak, and Richa Sharma [19] In this study, we have structured a framework that can competently discover the doctrines to forecast the risk level of patients considering the given constraint about their health. The main influence of this research is to assist a non-specialized doctor to brand accurate decision about the heart disease risk level. The guidelines engendered by the projected system are arranged as Pruned Rules, Original Rules, rules without duplicates, Sorted Rules, Classified Rules and Polish. The implementation of the structure is evaluated as far as arrangement exactness and the results reveal that the structure has astonishing prospective in expecting the coronary illness risk level all more accurately.

3. RELATED WORKS

Many experiments are being carried out for evaluating the performance of Naïve Bayes and Decision Tree algorithm. The results observed so far indicated that Naïve Bayes outperforms and sometimes Decision Tree.

In addition to that an optimization process using genetic algorithm is also being planned in order to reduce the number of attributes without sacrificing accuracy and efficiency for diagnosing the heart disease. There are many possible algorithms for the diagnosis of heart disease which are:

A. Naïve Bayes

A Naive Bayes classifier predicts that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature [8]. This classifier is very simple, efficient and is having a good performance. Sometimes it often outperforms more sophisticated classifiers even when the assumption of independent predictors is far. This advantage is especially pronounced when the number of predictors is very large. One of the most important disadvantages of Naive Bayes is that it has strong feature independence assumptions.

B. Decision Trees

Decision Trees (DTs) are a non-parametric supervised learning method used for classification. The main aim is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. The structure of decision tree is in the form of a tree. Decision trees classify instances by starting at the root of the tree and moving through it until a leaf node. Decision trees are commonly used in operations research, mainly in decision analysis. Some of the advantages are they can be easily understanding and interpret, robust, perform well with large datasets. Also they are able to handle both numerical and categorical data. Decision-tree learners can create over-complex trees that do not generalise well from the training data is one limitation.

C. Clustering

Clustering is a process of partitioning a set of data (or objects) into a set of meaningful sub-classes, called clusters. It helps users to understand the natural grouping or structure in a data set. Clustering is an unsupervised classification and has no predefined classes. They are used either as a stand-alone tool to get insight into data distribution or as a pre-processing step for other algorithms. Moreover, they are used for data compression, outlier detection, understand human concept formation. Some of the applications are Image processing, spatial data analysis and pattern recognition. Classification via Clustering is not performing well when compared to other two algorithms. All these algorithms are implemented with the help of WEKA tool for the diagnosis of heart diseases.

Data set of 294 records with 13 attributes. These algorithms have been used for analyzing the heart disease dataset. The Classification Accuracy should be compared for this algorithm. After the comparison attributes are to be reduced for further purpose.

4. 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, Rapid Miner, TANAGRA, MATLAB etc. Some of them are discussed as follows.

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

1. WEKA

It stands for Waikato Environment for Knowledge Learning. It is a computer program that was developed at the University of Waikato in New Zealand for the purpose of identifying information from raw data. WEKA supports different standard data mining tasks such as data pre-processing, classification, clustering, regression, visualization and feature selection. The basic premise of this application is to utilize a computer application that can be trained to perform machine learning capabilities and derive useful information in the form of trends and patterns. Originally written in C, the WEKA application was then completely rewritten in Java and is now compatible with almost every computing platform. Its user friendly graphical interface allows for quick set up and operation.

2. Rapid Miner

Formerly called as YALE (Yet Another Learning Environment), is an environment for providing data mining and machine learning procedures including data loading and transformation (ETL), data preprocessing and visualization, modeling, evaluation and deployment. Rapid Miner is written in the Java programming language. Also, it can be used for text mining, multimedia mining, feature engineering, data stream mining etc.

3. TANAGRA

It is a free data mining software designed for academic and research purposes. It proposes several data mining methods such as exploratory data analysis, statistical learning and machine learning. TANAGRA comprises some paradigms and algorithms such as clustering, association rule, parametric and nonparametric statistics, factorial analysis, feature selection and construction algorithms.

4. Apache Mahout

It is a project of the Apache Software Foundation designed for free implementations of distributed or otherwise scalable machine learning algorithms that focus primarily in the areas of collaborative filtering, clustering and classification. Apache Hadoop is another open source, Java-based programming framework which supports the processing and storage of extremely large data sets in a distributed computing environment. It is a part of the Apache project which is sponsored by the Apache Software Foundation.

5. 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.

6. Java

Java is a high level programming language developed by Sun Microsystems and now owned by Oracle Inc. It is widely used for developing and delivering content on the web. Java has numerous object oriented programming features much like C++, but is simplified to eliminate language features that cause common programming errors. Java language is well suited for use in World

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

Wide Web. Java applets (small Java applications) can be downloaded from a web server and run on a computer by a Java-compatible web browser.

7. C

C was developed by Dennis M. Ritchie at Bell Labs for the Unix Operating System in the early 1970s.It was originally intended for writing system software's. C is a high-level, general-purpose programming language which is ideal for developing firmware and portable applications.

8. Orange

It is a toolkit for data visualization, machine learning and data mining. It is interactive and can be used as a Python library.

5. DATA MINING TECHNIQUES

Data Mining is the process of extracting valid, authentic, and actionable information from large databases. Data Mining is an analytic process designed to explore data in search of consistent patterns and/or systematic relationships between variables, and then to validate the findings by applying the detected patterns to new subsets of data. The ultimate goal of data mining is prediction - and predictive data mining is the most common type of data mining and one that has the most direct business applications. Data mining strategies fall into two broad categories namely Supervised Learning and Unsupervised Learning.

Supervised Learning methods are deployed when there exists a field or variable (target) with known values and about which predictions will be made by using the values of other fields or variables (inputs). Unsupervised Learning methods tend to be deployed on data for which there do not exist a field or variable with known values, while fields or variables do exist for other fields or variables.

A. Feature Selection

Feature selection is a process used in machine learning in which a subset of the features available from the data is selected for application of a learning algorithm. It is necessary because it is computationally not feasible to use all available features or because of problems of estimation when limited data samples are present. Feature selection from the available data is vital to the effectiveness of the methods employed. Extracted features can be ranked with respect to their contribution and utilized accordingly. Existing feature selection methods for machine learning typically fall into two broad categories; those which evaluate the worth of features using the learning algorithm that is to be ultimately applied to the data, and those which evaluate the worth of features by using heuristics based on general characteristics of the data. The former is referred to as wrappers and the latter filters.

B. Classification Techniques

The classification task in machine learning is to take each instance of a dataset and assign it to a particular class. A classification-based system attempts to classify all the patient either having heart disease or not. The challenge in this is to minimize the number of false positives and false negatives. Classification maps a data item into one of several predefined categories. These algorithms normally output "classifiers", for example, in the form of decision trees or rules. An ideal application in intrusion detection will be to gather sufficient "normal" and "abnormal"

audit data for a user or a program, and then apply a classification algorithm to learn a classifier that will determine (future) audit data as belonging to the normal class or the abnormal class.

C. Clustering Techniques

Data clustering is a common technique for statistical data analysis which is used in many fields, including machine learning, data mining, pattern recognition, image analysis and bioinformatics. It is the classification of similar objects into different groups, or more precisely, the partitioning of a data set into subsets or clusters, so that the data in each subset share some common trait which is often proximity according to some defined distance measure.

Machine learning typically regards data clustering as a form of unsupervised learning. Clustering is useful in intrusion detection as malicious activity should cluster together, separating itself from non-malicious activity [4]. Clustering is one of the most useful tasks in data mining process for discovering groups and identifying interesting distributions and patterns in the underlying data. Clustering problem is about partitioning a given data set into groups (clusters) such that the data points in a cluster are more similar to each other than points in different clusters. Clustering provides some significant advantages over the classification techniques which already discussed, in that it does not require the use of a labelled data set for training.

D. Association Rule Mining

Association rules are if/then statements [1] that help to uncover relationships between unrelated data in a database, relational database or other information repository. Association rules are used to find the relationships between the objects which are frequently used together. Applications of association rules are basket data analysis, classification, cross-marketing, clustering, catalogue design, and loss-leader analysis etc.

CONCLUSIONS

In this paper we have discussed some of effective techniques that can be used for heart diseases. classification and the accuracy of classification techniques is evaluated based on the selected classifier algorithm. An important challenge in data mining and machine learning areas is to build precise and computationally efficient classifiers for Medical applications. The performance of Naive Bayes shows high level comparison with other classifiers and also the accuracy of the Decision Tree and Bayesian Classification further improves after applying genetic algorithm to reduce the actual data size to get the optimal subset of attribute sufficient for heart disease prediction.

REFERENCES

[1] Carlos Ordonez, “Improving Heart Disease Prediction using Constrained Association Rules”, Technical Seminar Presentation, University of Tokyo, 2004.

[2] Franck Le Duff, CristianMunteanb, Marc Cuggiaa and Philippe Mabob, “Predicting Survival Causes After Out of Hospital Cardiac Arrest using Data Mining Method”, Studies in Health Technology and Informatics, Vol. 107, No. 2, pp. 1256-1259, 2004.

[3] W.J. Frawley and G. Piatetsky-Shapiro, “Knowledge Discovery in Databases:

An Overview", AI Magazine, Vol. 13, No. 3, pp. 57-70, 1996.

[4] Heon Gyu Lee, Ki Yong Noh and Keun Ho Ryu, "Mining Bio Signal Data: Coronary Artery Disease Diagnosis using Linear and Nonlinear Features of HRV", Proceedings of International Conference on Emerging Technologies in Knowledge Discovery and Data Mining, pp. 56-66, 2007.

[5] Kiyong Noh, HeonGyu Lee, Ho-Sun Shon, Bum Ju Lee and Keun Ho Ryu, "Associative Classification Approach for Diagnosing Cardiovascular Disease", Intelligent Computing in Signal Processing and Pattern Recognition, Vol. 345, pp. 721-727, 2006.

[6] Latha Parthiban and R. Subramanian, "Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm", International Journal of Biological, Biomedical and Medical Sciences, Vol. 3, No. 3, pp. 1-8, 2008.

[7] Niti Guru, Anil Dahiya and Navin Rajpal, "Decision Support System for Heart Disease Diagnosis using Neural Network", Delhi Business Review, Vol. 8, No. 1, pp. 1-6, 2007.

[8] Sellappan Palaniappan and Rafiah Awang, "Intelligent Heart Disease Prediction System using Data Mining Techniques", International Journal of Computer Science and Network Security, Vol. 8, No. 8, pp. 1-6, 2008.