FEATURE SELECTION METHODS FOR HEART DISEASE PREDICTION WITH DATA MINING TECHNIQUES

Uma K* and Dr. M. Hanumanthappa**

* Department of Computer Science and Applications, Bangalore University, Bengaluru, India.
** Department of Computer Science and Applications, Bangalore University, Bengaluru, India.

ABSTRACT: Heart disease is now turned out most deadly disease throughout the world. To aid early and correct diagnose of heart disease, data mining techniques are used widely. This paper presents, classification techniques applied for prediction of heart disease with different feature selection methods. And correlates the accuracy of each data mining techniques for heart disease prediction.

KEYWORDS: Classification techniques, Data mining, Feature selection method, Heart disease.

INTRODUCTION

Heart disease is one of the leading disease causes death in worldwide since 1999 according to estimation of WHO factsheet (WHO, 2016). It is increasing and has become a true deadly disease that has no limit. A lot of people have died from heart disease compared to any other diseases. The death rates of heart disease are also influenced by nature of different countries in the major risk factors, particularly blood pressure, blood cholesterol, smoking, physical activity and diet. Even as genetic factors play a part, 80% to 90% of people are losing their lives from heart disease have one or more major risk factors that are influenced by lifestyle. Correct diagnosis at an early stage followed by right successive treatment can result in significant lifesaving. Age groups are more at risk of getting heart disease. A new study says that heart disease can be controlled efficiently if it is identified at an early stage. But it is not easy to do perfect analysis because of many difficult factors of heart diseases. For example, other than the heart and very often heart diseases may exhibit various syndromes. Due to this complication, there is a requirement for automate the process of medical diagnosis which can assist medical practitioners in the diagnostic process.

Data Mining

Data mining is a process of discovering the previously unknown pattern from large dataset. According to Jaiwei Han and Micheline Kamber, Data mining is the most powerful and motivating concept of discovering the hidden pattern from the voluminous data (Han & Kamber). Presently, Data mining is significantly used in the healthcare industry to transfer the more complex data into useful information. Due to existence of huge data in healthcare, there is a need of strong method to handle the data and extracts the useful information from healthcare. With the growth and maintenance of large data repositories of structured and unstructured data, health organizations are increasingly using data analytics. It also includes data mining to analyse and utilize the patterns and relationships found in the data to make improved clinical and other health-related decisions.

Heart Disease

Heart Disease is the number one fatal-disease causing death worldwide by the factsheet of World Health Organization-2016. Heart disease is a word used to describe the different conditions affecting the heart. Heart disease could be related to any one of the following disease such as cardiovascular disease, coronary artery disease, angina, arrhythmia, congenital heart disease, myocardial infarction, etc. More deaths occurs by heart related disease than any other diseases. According to World Health Organization (WHO) more than 31% global death occurs due to heart disease during 2012.

Classification Techniques

Classification is the process of finding a model that describes and distinguishes data classes and finds the target class label. The resulting model is constructed based on the analysis of a set of training data (Han & Kamber). Classification techniques are applied popularly in the process of disease diagnosis.

Decision Tree

A decision tree is a supervised classification method which generates a tree and a set of procedures, representing the model of different classes from a given data set. A tree is too easy to represent and recognize and opposing to noise in data (Han & Kamber). The decision tree is one of the most frequently used classification techniques of data analysis. The healthcare decision making systems, literature reveals a number of researches that have made use of and data mining techniques.

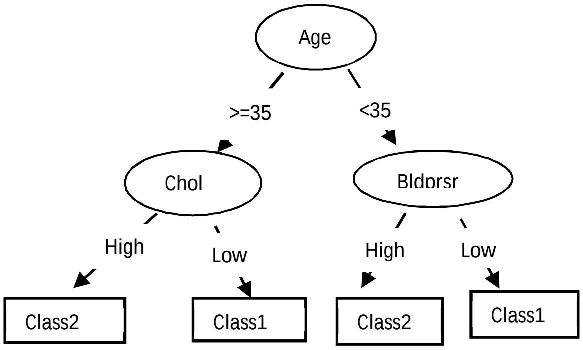

Fig.1. Decision tree

Bayesian classifier

It is a statistical classification method to predict the class membership probabilities. Bayesian classifier works based on Bayes theorem and estimates the posterior probability. Naïve Bayesian classifier is a simple classifier that assumes attribute independence it is high speed when applied to large databases and comparable in performance to decision trees. Let *X* be an unknown class label of data sample, Let H_i be the hypothesis that *X* belongs to a particular class C_i, P (H_i) is the probability class that tells *X* belongs to a particular class C_i. It can be estimated by n_i/n from training data samples, *n* is the total number of training sample data and n_i is the number of training data samples of class C_i

$$P(H_i \mid X) = \frac{P(X \mid H_i)P(H_i)}{P(X)}$$

Support Vector Machine

Support Vector Machine (SVM) is a supervised machine learning algorithm for classification and regression problems [7]. The SVM classifies the data based on support vectors. To differentiate the classes the SVM finds the hyper-plane. Support vectors are simply the co-ordinates of individual observation. Support Vector Machine is a frontier which best segregates the two classes by hyper-plane. The SVM classifier objective is to maximize the margin.

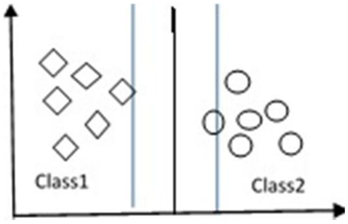

Fig.2. SVM

Bagging

Bagging is a “bootstrap” collective method that randomly creates sample dataset group from original dataset by training each classifier on a random redistribution of the training set.

Classification via Regression

Classification via regression method comprises the class for doing classification using regression approach. For every class value in classification one regression model will built. Classifies the dataset based on the relationship of variables estimation.

Feature Selection Methods

Feature selection method plays a very significant role in medical data mining to remove irrelevant or redundant features present in the data. Feature selection is a procedure to extract the feature subset to reduce the large data volume(R Suganya et al,). Feature selection is a key point for prediction since the prediction evaluates the attributes and identifies the correct data. Feature selection can be done in two ways: i) Filter approach-a statistical measure is applied to identify the scoring of each attribute and assign value. ii) Wrapper approach- A predictive model used to evaluate a combination of features and assign a score based on model accuracy. In WEKA data mining tool, the attribute selection process has two parts: Attribute evaluator- selects the attribute subset and Search method- searches the best subset of feature. For example, CfsSubsetEval, CorrelationAttributeEval, GainRatioAttribute, InfoGainAttribute, Wrapper Method, etc.

RELATED WORK

A lot of work has been conducted for diagnosing and prediction of heart disease by applying different Data mining techniques.

In 2013, Vikas Chaurasi and Saurabh Pal, worked on prediction of heart disease using data mining techniques. To predict the heart disease the researchers taken three popular data mining algorithms such as CART, ID3 and decision Table and 10 fold cross validation methods to measure the unbiased estimate. The dataset contains 303 instances with 11 attributes. The data is analysed and implemented in Weka tool. The performance of the three classifiers are evaluated, it results 83.49% accuracy for CART, 72.93% for ID3 and 82.50% for Decision Table.

In 2014, Deepali Chandran, proposed a work for identifying feature selection techniques and developed new machine learning algorithms. That are designed for providing automatic computer aided analysis and decision support system for heart disease diagnosis. The researcher collected the online available heart disease dataset from University of California, Irvin (UCI). 14 attributes are selected. The authors used the information gain and Adaptive Neuro Fuzzy Inference System (ANFIS) for finding heart disease. Information gain is used to select the attributes which is have the more information.

In 2015, Umair Shafique et al, have been worked on heart disease to find out the interesting patterns from data of heart patients. The selection of only 14 attributes out of 76 attributes present 597 records by applying an attribute selection method using a WEKA tool for different data. Authors had taken classification which is a data mining technique and implemented with the following algorithms, Neural Network, Decision tree and Naïve Bayes and conducted 4 experiments and achieved 82.914% accuracy of result with a Naïve Bayes classification algorithm.

Motivation

Today, the main challenge facing by healthcare industry includes hospitals, healthcare centres are delivery of eminence service in affordable cost for the society. Due to misdiagnosis of heart disease more people losing their lives. The facility may include the correct diagnosis of patients and effective treatments at lower cost. Poor clinical diagnosis might cause deaths especially in case of heart disease. The hospitals must also minimize the cost of medical test by achieving results using the appropriate decision support system and techniques employed. Ultimately, an existing abundant medical data (patient records) raises a query “How to explore the data to get potential information and make use of this information for predicting absence or presence of heart disease”.

Problem Domain

Every hospitals, healthcare centres, diagnostic centres generates a huge amount of data such as patient particulars, test reports, etc. The Heart patient details contains many features which help to predict the heart disease. This high volume of medical data offered data mining techniques to discover the hidden pattern and diagnose the patients correctly. The historical medical data needs analytical methods to analyse the data and to extract the potential information from it. Data mining is the one such type of process to extract the hidden pattern and used as an analytical method to analyse the historical data. The problem domain contains to predict the heart disease presence and provides treatment in early stage.

Problem Definition

Many researchers worked on different disease management, diagnosis and prognosis, heart disease prediction system using different data mining techniques and other analytical methods. All the related work done using only 14 attributes of the dataset. The proposed work contains prediction of heart disease dataset with class label presence or absence by applying various classification techniques on trained data in two different scenarios. One scenario is applying classification techniques with 18 attributes and another is to apply same classification techniques with feature selection method. The early prediction of disease will help to physicians for better treatment.

Problem Statement

The performance of classification techniques for predicting heart disease with 18 attributes and selected attributes by applying feature selection methods.

PROBLEM DESIGN AND SOLUTION METHODOLGY

The classification of heart disease method with and without attribute selection consists following steps.

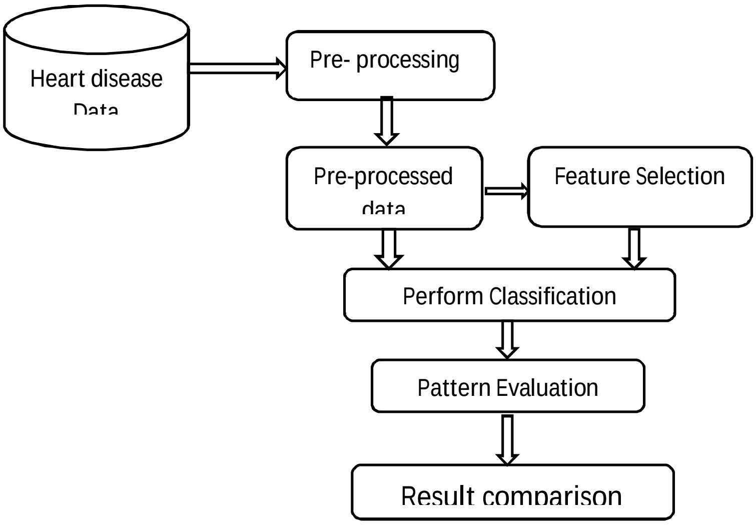

Fig. 3. Proposed method

Data Collection and Pre-processing

For conducting an experiment the heart disease dataset collected through online from UCI (University of California, Irvin), Machine Learning Repository. The collected heart disease dataset consists 689 instances and 18 attributes with missing records. The collected heart disease dataset contain noisy, inconsistency and missing values. While pre-processing, those missing records were find and replaced with appropriate value using ReplaceMissingValues filter using Weka tool. This filter scans all the records and replaces the missing values by mean mode method. The attributes are, age, sex, high cholesterol, hypertension, blood sugar. Smoking, family-history, diabetes, etc.

Experiment: Scenario-1

The experiment conducted using Weka machine learning tool. This dataset contains 689 unique records with 18 attributes. In this scenario, five different classification techniques such as SVM, Bagging, Naïve Bayesian, Classification via regression and J48 decision tree are applied on pre-processed data. Each classification algorithms performs differently on same data. Individually, chosen algorithms are classifies the heart disease dataset with the class label presence and absence.

Experiment: Scenario-2

Prediction of dataset by applying feature selection method experiment was conducted with many attribute selection methods using WEKA tool. They are, CfsSubsetEval, Information Gain, Gain Ratio and Wrapper method. Each feature selection techniques calculates the value of every attributes and assigns the ranking using ranker method. And some feature selection techniques use the search method to search the best attribute from dataset.

RESULT ANALYSIS

The attention of this research work is to identify the suitable classification techniques and features for heart disease prediction. To discover the best classification algorithm to predict an experiment was conducted on the heart disease dataset by applying different classification techniques. The two experiment results the different computation time and accuracy for five different classifier techniques. The accuracy of each experiment is as shown in the below table.

Table.1. Comparison of Classification algorithms with and without feature selection

Classification Techniques	Accuracy (%)					
	Scenario -1	Scenario-2				
		Cfs (6)	Info Gain (11)	Gain Ratio (9)	Correlation (9)	Wrapper (10)
SVM	99.7	86.4	99.7	98.4	98.0	99.4
Bagging	91.7	89.6	91.9	93.8	91.7	91.6
Naïve Bayes	92.7	87.1	92.6	92.2	91.4	92.6
Cls. regression	99.7	89.6	99.7	99.4	97.0	99.4
J48	92.0	89.3	79.0	76.8	98.8	85.6

For the first approach of experiment, the selected SVM, Bagging, Naïve Bayes, Classification via Regression and J48 algorithms achieves the accuracy of 99.7%, 91.7%, 92.7%, 99.7% and 92.0 % respectively to classify the heart disease presence and absence data. In this scenario, the SVM and classification via regression gives same and highest accuracy meanwhile the bagging provides less accuracy i.e. 91.7%. It shows that the classification using regression method and support vector method are efficient. Similarly, the second experiment using feature selection approach realizes the attribute values with achieving variant accuracy for different algorithms. The selected five attribute selection methods performs differently on five classification algorithms. As shown in the table.1 each method chooses the various attribute pairs and classifies the dataset. Among five methods, CfsSubsetEval method gives the lowest accuracy and remaining methods are achieves nearly identical accuracy as previous experiment. The proposed method focuses on the significant of attributes in the classification. The accurate classification of data requires the most relevant and genuine attributes.

CONCLUSION

This study was focused on the use of data mining techniques in healthcare specifically in Heart Diseases. Heart disease is a fatal disease which cause death globally. The online available heart patient’s data from UCI repository used for conducting an experiment. There were 689 unique instances. The classification techniques and feature selection methods are applied differently to classify the data accurately. Some important point were considered to choose suitable tool for mining, on the basis of them Weka machine learning software were used for experiments. Accuracy is taken to evaluate the performance of the algorithms. The experiments results that SVM and regression algorithms have the highest accuracy among all that is 99.7%. This study shows that the data mining can be used to predict about heart disease efficiently and effectively.

REFERENCES

[1] Deepali Chandan (2014), *Diagnosis of Heart disease using Data Mining Algorithm*, International Journal of Computer Science and Information Technologies, Vol(2), 1678-1680.

[2] Eman AbuKhousa & Piers Campbell (2012), *Predictive Data Mining to Support Clinical Decisions: An Overview of Heart Disease Prediction System*, International Conference on Innovations in Information Technology (IIT) 978-1-4673-1101-4/1.

[3] Hlaudi Daniel Masethe, Mosima Anna Masethe (2014), *Prediction of Heart Disease Using Classification Algorithms*, Proceedings of the world Congress on Engineering and computer Science Vol II., WCECS 2014, 22-24 October, 2014, San Francisco, USA

[4] Jaiwei Han and Micheline Kamber, Data Mining Concepts and Techniques”, Second Edition, ISBN 13:978-1-55860-901-3.

[5] R. Suganya et al., (2016), *A Novel Feature Selection Method For Predicting Heart Disease With Data Mining Techniques*, Asian Journal of Information Technology 15 (8): 1314 1321, ISSN: 1682-3915.

[6] Umair Shafique et al., 2015, *Data Mining in Healthcare for Heart Diseases*, International Journal of Innovation and Applied Studies, ISSN 2028-9324 Vol. 10 No. 4 march 2015, pp. 1312-1322.

[7] Vikas Chaurasia,, Carib.j. (2013), *Early Prediction of Heart Diseases Using Data Mining Techniques*, Caribbean Journal of Science and Technology, SciTech, Vol.1,208-217, ISSN : 0799-3757.

[8] World Health Organization - 2016 [online] Fact Sheet available: http://www.who.int/cardiovascular_diseases/en