HEART DISEASE PREDICTION USING CLASSIFICATION TECHNIQUES WITH FEATURE SELECTION METHOD

***Uma K**
Assistant Professor, Bangalore University, Bangalore

****M. Hanumathappa**
Professor, Dept. of Computer Science & Application, Bangalore University, Bangalore

ABSTRACT

Heart disease is now turned out most deadly disease throughout the world. Due to misdiagnosis of heart disease more people losing their lives. Hence, there is a need of automate the system for correct diagnose the heart disease based on the historical data. To aid early and correct diagnose of heart disease, many data mining techniques are used to predicting the disease. The high volume of medical data offered data mining techniques to discover the hidden pattern. Classification technique is one among the data mining techniques predict the heart disease. This paper presents, classification techniques applied for prediction of heart disease in two scenario such as dataset with all 13 attributes and 6 attributes selected by attribute selection method. For achieving the results, the selected classification techniques are Support vector machine, Neural Network (Multilayer perception), Bagging, Classification via regression and Simple logistic. And correlates the accuracy and time taken to build the prediction model for all used classification techniques in two different scenario.

Keywords: *Data mining, Classification techniques, Heart disease, Attribute selection method.*

I. INTRODUCTION

Medical data mining takes the great potential for discovering the hidden patterns in the data sets of the medical field. These patterns can be used for clinical diagnosis [3], however this diagnosis is to disclose the presence or absence of the disease. Diagnosis of disease is one of the promise application for many purpose; an abundant of patient data are available and in the same time, the need for accurate diagnosis. So this has encouraged to attract many researchers to work in this field.

Data mining is a process of discovering the meaningful information from huge amount of data [8]. The data mining techniques are very beneficial to predicting the various diseases in the healthcare industry. Disease prediction plays most important role in the data mining [6]. Data Mining is a process of discovering interesting patterns and knowledge from huge amount of data. It refers to extracting or mining knowledge from large amount of data. Extracting knowledge it is also called knowledge mining from data or knowledge extraction or Knowledge Discovery from Data (KDD). The knowledge discovery process typically involves datacleaning, data integration, data selection, data transformation, pattern discovery, pattern evaluation and knowledge presentation. Nowadays, healthcare organization generates a voluminous data that results lack of information to make the right decision. Data mining techniques can be used to extract the needful information from healthcare organizations. Data mining is also widely used in healthcare for various applications such as detection of fraud

and abuse in healthcare, customer relationship management, detection of diseases and treatment effectiveness, and availability of healthcare services at lower cost. Heart disease is the major cause of deaths.

According to World Health Organization (WHO) estimation, every year 12 million deaths are happening globally because of heart disease. And 80% of humans are losing their lives due to heart disease. Based on the present assessment WHO estimated that 23.4 million people will die because of heart disease by 2030 [9]. Prediction by using data mining techniques gives us accurate result of disease.

Heart diseases remain the biggest cause of deaths for the last two decades. Recently computer technology and machine learning techniques to develop software to assist doctors in making decision of heart disease in the early stage. The diagnosis of heart disease depends on clinical and pathological data. Heart disease prediction system can assist medical professionals in predicting heart disease status based on the clinical data of patients. In biomedical field data mining plays an essential role for prediction of diseases with interrelated symptoms and signs especially when the patients suffer from more than one type of disease of the same category. The physicians may not able to diagnose it correctly.

II. CLASSIFICATION TECHNIQUES

Data mining is a process of finding the potential information from a large amount of data. In present era, Data mining techniques are significantly applied in medical domain for decision making, disease diagnosis etc. Classification technique is more adopted in the field of disease diagnosis to classify the dataset into number of classes.

Classification model can be built through learning process. It is describing a predetermined set of data classes or concepts [11]. The model is constructed from available data i.e. classified examples. These examples consist of set of features. This model is induced through supervised learning process, the classified examples would be processed as training data. The training example is presented to learning algorithm with its class and let to extract specific knowledge gradually. After, classifier was built, this model would be evaluated according to the answers accuracy that model will give through testing.

A. Support Vector Machine:

Support Vector Machine (SVM) is a most popularly used supervised machine learning algorithm for both regression and classification challenges [7]. It perform classification by finding the hyper-plane that differentiate the two classes very well. 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. In the below shown figure, H1 does not separate the classes. H2 does, but only with a small margin. H3 separates them with the maximum margin. The margin is the distance between separating hyper-plane and the training samples that are closest to this hyper-plane.

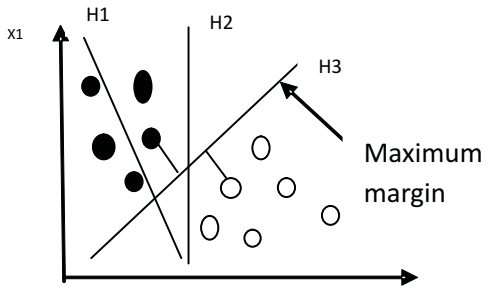


Fig.1.Support Vector Machine

B. Bagging:

Bagging [Breiman, 1996] is a "bootstrap" ensemble method that creates individuals for its ensemble by training each classifier on a

random redistribution of the training set [10]. Each classifier's training set is generated by randomly drawing, with replacement, N examples - where N is the size of the original training set; many of the original examples may be repeated in the resulting training set while others may be left out. Each individual classifier in the ensemble is generated with a different random sampling of the training set.

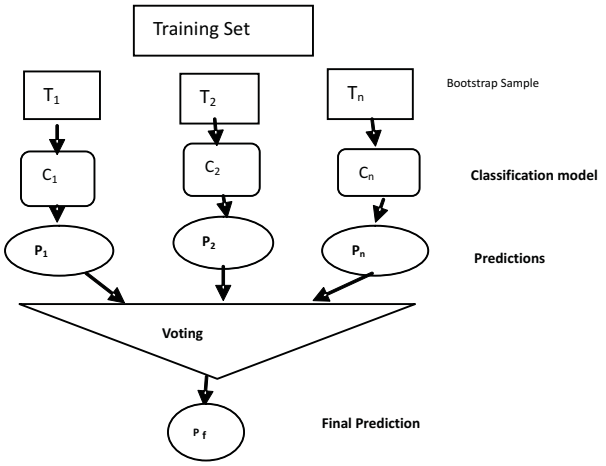
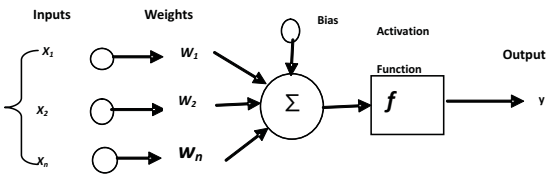


Fig.2. Bagging

C. Neural network:

It is a mathematical model, based on biological neural networks [4]. It consists of an interconnected group of artificial neurons and processes information using a connectionist approach to computation. In most cases an ANN is an adaptive system that changes its structure based on external or internal information that flows through the network during the learning phase.



D. Classification via Regression:

Classification is a supervised method to classify the data items into binary or multi class variables [10]. This will helps to build the prediction model. Classification via regression method includes class for doing classification using regression methods. Class is binary valued and one regression model is built for each class value. Classification is done with the process of estimating the relationship of variables.

E. Simple Logistics:

It is a classifier for building linear logistic regression models. It works on LogitBoost algorithm. LogitBoost with simple regression functions as base learners is used for fitting the logistic models. The optimal number of LogitBoost iterations to perform is cross-validated, which leads to automatic attribute selection. The advantage of Simple Logistic is that it has built-in attribute selection, it stops adding Simple Linear Regression models when the cross-validated classification error no longer decreases.

III. FEATURE SELECTION METHOD

Feature selection is the technique of choosing the appropriate attributes or related attributes for classification [12]. Feature selection techniques can be applied in two stages. Firstly, in the preprocessing stage for selecting the related attributes from raw data. Secondly, in the process of dimensionality reduction stage. Feature selection method can be mainly classified into two types, filter and wrapper approach. Filter approach is independent of the data mining algorithm and it is to be applied to choose the attributes meanwhile evaluate the relevance of feature by considering the only at key points of the data. In the wrapper approach of feature selection uses the result of the data mining algorithm to measure the attribute subset. The feature subset is taken and applied wrapper approach to measure the performance generated by the data mining algorithms.

IV. RELATED WORK

Abhishek Taneja et al. [1] in the year 2013 used data mining tool WEKA 3.6.4 in heart disease

prediction system using J48 technique achieved 95.56% accuracy and using Naïve Bayes technique achieved 92.42% and using Neural Network achieved 94.85%. In this paper, various data mining techniques have been studied for the diagnosis of heart disease. The data for the study collected from PGI, Chandigarh which has a total of 15 attributes. A total of 7339 instances were trained. Of the 15, only 8 attributes were selected.

G. Subbalakshmi et al. [2] in the year 2011 has developed a Decision Support in Heart disease Prediction System (DSHDPS) using Naïve Bayes data mining modeling technique to discover the relationship between variables in database in the healthcare industry. This model could answer the complex queries. It is resulted out as the most effective model in the prediction of heart disease.

Mr Akhil Jabbar et al. [3] in the year 2012 proposed evolutionary algorithm for heart disease prediction. They used the genetic algorithm to predict the heart diseases in Andhra Pradesh population. They used the Association Rule Mining based on the sequence number and clustering for heart attack prediction.

V. PROPOSED WORK

The proposed work contains the knowledge extraction process from two different scenarios with same Hungarian heart disease dataset. In first scenario, an experiment is conducted with Hungarian heart disease dataset with 13 attributes. And in second scenario same experiment is conducted with 6 attributes which are selected by 'CfsSubsetEval' attribute selection method. Finally, comparison of the obtained result for two different scenario is done. The procedure for the process involves following steps.

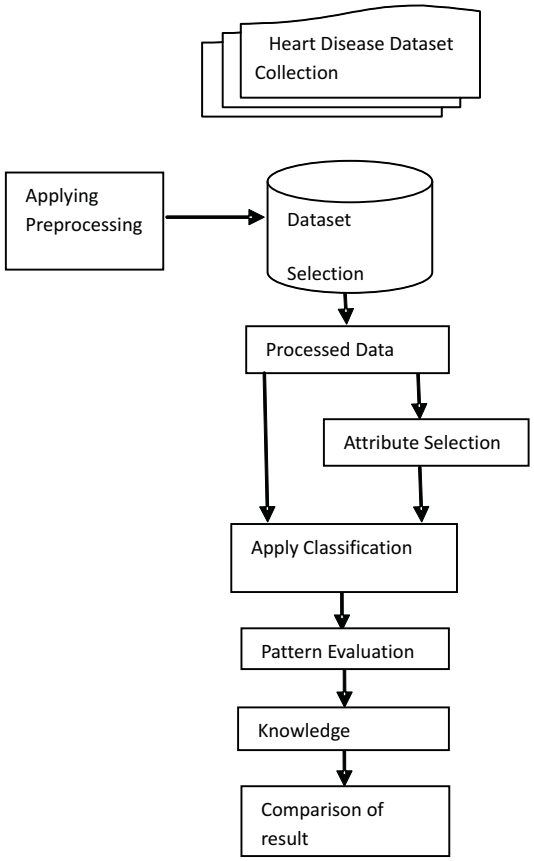
i) Data collection

ii) Target data selection

iii) Preprocessing

iv) Apply Classification techniques with and without attribute selection method

v) Result Analysis



The steps are shown in Fig.4.

Proposed Method

The Hungarian Heart disease dataset is drawn from University of California, Irvin (UCI) machine learning repository through online. The dataset contains 293 instances with 76 attributes, however only 13 attributes were taken to conduct an experiment since 13 attributes are plays vital role to predict the heart disease. The experiment carried using Weka data mining tool. Weka is a machine learning tool in which all the knowledge discovery process can be performed. In the next context, pre- processing of data is done with the help of Weka filters to remove noisy value and replace the missing values. The

dataset for experiments had some missing records, those missing records were find and replaced with appropriate value using 'ReplaceMissingValues' filter from Weka tool. This filter scans all the records and replaces the missing values by mean mode method. After pre-processing, the test is directed in mode. In the first mode, the selected fie different classification techniques such as SVM, Bagging, Multilayer Perceptron, Classification via regression and simple logistic are applied on pre-processed data with 13 attributes and one class attribute for prediction. The attributes are as follows.

Sl No.	Attribute Name	Value	Description
1	Age	29-62(Numerical value)	Age of patient in years
2	Gender	0-Male , 1-Female	Sex of Patient
3	Cpt	1 - typical angina, 2 - atypical angina 3 - non-anginal pain, 4 – asymptomatic	Chest pain type
4	Restbp	Numerical value In mm/ Hg(140mm/Hg)	Resting blood pressure
5	Chol	Numerical value in mg/dl(289mg/dl)	Cholesterol in mg/ dl
6	Fbs	0 – false , 1 – true	Fasting blood sugar > 120 mg/dl (1 = true; 0 = false)
7	Restecg	0 – normal, 1 - having ST-T, 2 – hypertrophy	Resting electrocardiographic results
8	Thalach	Numerical value (140, 173)	Maximum heart rate achieved
9	Exang	0 – no ,1 – yes	Exercise induced angina (1 = yes; 0 = no)
10	oldpeak	Numerical value ()	ST depression induced by exercise relative to rest
11	Slope	1 – upsloping, 2 – flat, 3 – downsloping	The slope of the peak exercise ST segment
12	Ca	0 – 3 vessels (numeric value)	Number of major vessels (0 -3) colored by fluoroscopy
13	Thal	3 – normal, 6 - fixed defect, 7 - reversable defect	Thalassemia

Table.1. Attribute description.

In the second mode of experiment, attribute selection method is apply to select the attributes based on 'CfsSubsetEval' method. This method selects the 6 attributes out of 13 attributes. The selected attributes are gender, chest pain type, cholesterol, thal, exang and oldpeak. For this reduced dataset, the earlier selected five classification techniques are applied. Each techniques are performed and produces the result in terms of confusion matrix and time for building classifier model. Based on the accuracy of classifier, computation time and attribute selection the results of two experiment is discussed.

V. RESULT AND DISCUSSION

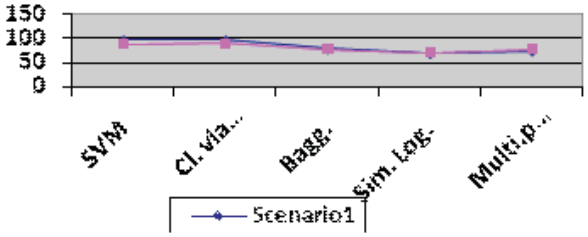
The main focus of this research work is to identify the suitable data mining techniques for heart disease prediction. To discover the best prediction algorithm an experiment was

conducted on the heart disease dataset. By applying different classification techniques try to find out which techniques are giving more accurate results for heart disease prediction. The two experiment results the different computation time and accuracy for five different classifier techniques. For the first approach of experiment the selected SVM, Classification via Regression Bagging, Multilayer Perceptron and Simple logistic classification techniques are achieves the accuracy of 97.9%, 97.9%, 78.4%, 74.3% and 69.2% respectively. And time taken to build each classifier is SVM – 0.5 sec, bagging – 0.0.5, Classification via regression – 1.13 sec, Simple logistic – 3.06 sec and multilayer perceptron – 1.17 sec.

For the second type of experiment i.e. for 6 attributes dataset, the selected classification techniques achieves the accuracy and computation time like Classification via regression - 91.1% and 0.66s, SVM – 89.4% and 0.49s, Bagging – 79.1% and 0.05s, Multilayer perceptron – 79.1% and 1.16s and Simple logistic – 71.6% and 1.58s respectively. The following table describes the summarized result of the experiment.

Sl. No	Algorithm	Scenario1 (13 Attributes)		Scenario2 (6 Attributes)	
		Accuracy	Time (in sec)	Accuracy	Time (in sec)
1	SVM	97.9%	0.5s	89.4%	0.49s
2	Classification via regression	97.9%	1.13s	91.1%	0.66s
3	Bagging	78.4%	0.05s	79.1%	0.05s
4	Simple logistic	69.2%	0.42 s	71.6%	1.58s
5	Multilayer perceptron	74.3%	1.17 s	79.1%	1.19s

Table.2. Summary of the result.



Graph.1. Comparison of two scenario

Comparison of algorithm in different scenario:

The two experiments were conducted with same heart disease dataset and with different scenario. First experiment setup is with all 13 attributes and second one is with 6 attributes of reduced dataset by applying attribute selection method. Five selected classification techniques are performed and achieves the result as mentioned in the above table.2 and Graph.1.

From the description of above table, the SVM (97.9%, 89.4%) and classification via regression (97.9%, 91.1%), Bagging (78.4%, 79.1%), Simple logistic (69.2%, 71.6%) and Multilayer perceptron (74.3%, 79.1%) techniques are achieved different accuracy in two scenario. With this result, it shows that the attributes selection approach gives less accuracy compared to general method of classification for SVM and Classification via regression techniques. In the other side, the three classification techniques such as Bagging, Simple logistic and Multilayer perceptron achieved more accuracy with attribute selection approach compared to general approach. Similarly, the time taken to build each classifiers depends on the accuracy obtained from each

techniques. It indicates that if the accuracy of attribute selection approach is greater than the general approach then the time taken to build the classifier also long. Based on the obtained result and comparison the SVM and Classification via regression techniques are having greater accuracy with general approach method.

VI. CONCLUSION

This research work was focused on the practice of data mining techniques in healthcare specifically in Heart Diseases. Heart disease is a serious disease which may cause death. The online available Hungarian heart patient's data is drawn from UCI repository and used for conducting an experiment. There were 293 unique instances in dataset with 13 attributes. The classification which is a data mining technique was implemented with following algorithms, SVM, Bagging, Multilayer perceptron, classification via regression and simple logistic. Some important points were considered to choose suitable tool for mining, on the basis of them Weka machine learning software were used for experiments. To evaluate the performance of the algorithms different performance metrics were considered that are accuracy and time to build model.

The experiments show that both SVM and classification via regression classification algorithms have the highest accuracy among all that is 97.9% with 13 attributes. This study shows that the data mining can be used to predict about heart disease efficiently and effectively. The results or the outcomes of this experiments may be used as assistant tool to help in making more consistent diagnosis of heart diseases.

Individual techniques are not sufficient to achieve the desired result. In future, ensemble techniques are applied to get more accuracy.

REFERENCE

[1], Abhishek Taneja et al., "Heart Disease Prediction System Using Data Mining Techniques", Oriental Journal of Computer Science & Technology ISSN: 0974671, December 2013, Vol.6, No. (4).

[2] Akhil Jabbar et al.," Heart Disease Prediction System Using Associative Classification and Genetic Algorithm", International Conference on Emerging trends on Electrical, Electronics and Communication Technologies, ICIECT, 2012.

[3], Boris Milovic et al., "" Prediction and Decision Making in Health Care Using Data Mining", International journal of public health science (IJPHS), Vol. 1, No. 2, December 2012, ISSN:2252-8806.

[4], Fausett, Laurene (1994), "Fundamentals of Neural Networks: Architectures, Algorithms and Applications", Prentice-Hall, New Jersey.

[5] , G. Subbalakshmi et al., "Decision support in heart disease prediction system using naive bayes", Indian journal of Computer Science and Engineering, ISSN: 0976-5166. Vol. 2 No. 2 Apr-May 2011

[6], Hian Cbye Kob et al., "Data mining Applications in Healthcare" Journal of Healthcare Information management-Vol.19, No.2.

[7], Hloudi Daniel Masethe et al., "Prediction of Heart Disease Using Classification Algorithms", Proceedings of the world

Congress on Engineering and computer Science 2014 Vol II.

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

[9], Umair Shafique et al., "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.

[10] WWW.wikipedia.org/wiki/Bagging

[11] WWW.wikipedia.org/wiki/Machine learning

[12] Sunita Beniwal et al., "Classification and Feature Selection Techniques in Data Mining", International Journal of Engineering Research & Technology (IJERT) Vol. 1 Issue 6, August – 2012, ISSN: 2278-0181.