IJARIIT
INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 4.295
(Volume 4, Issue 3)
Available online at: www.ijariit.com

Best feature selection for heart disease prediction using data mining

Anu Mehra
anumehra14@gmail.com
Amity School of Engineering and Technology,
Noida, Uttar Pradesh

Dr. Shailendra Narayan Singh
Snsingh36@amity.edu
Amity School of Engineering and Technology,
Noida, Uttar Pradesh

ABSTRACT

Data mining is way for extracting the valuable knowledge patterns from huge amount of data .Various data mining tools and techniques are used in medical world for predicting the diseases. Heart disease is one of the common disease now days .This paper presents different feature selection attribute Evaluator models Working with all attributes in data is not always useful so for this study we filtered selected attributes that will give maximum accuracy .This study compares various classification techniques for predicting heart disease with different set of attributes selected using evaluators .The data set used for the study is Cleveland heart dataset it has 14 attributes and 303 instances . For achieving the results, the different classification algorithms are applied on selected attributes taken after feature selection. The classifiers used in study are J48, SMO, Multilevel perception, Bagging Naive Bayes Random Tree. And Comparative study is done on the basis of accuracy achieved from different classifier

Keywords: *Heart Disease, Data Mining, Feature Selection of Attributes, WEKA.*

1. INTRODUCTION

Heart is a very important organ in human body There are various factors which may be considered when monitoring heart like patient family history of heart disease , high pulse, cholesterol level , obesity , smoking habits[8] .There are various types of heart illness like Coronary illness, Angina pectoris .The challenging problem with medical industry is to identify heart problem .Thus by applying data mining concepts and algorithm can help in predicting the heart disease and reduces the risk factor attached with heart disease For Heart disease prediction we use WEKA tool which is a data mining tool developed using Java. For designing prediction system, various data mining algorithms are used. Heart disease dataset with 14 attributes and 303 instances is the training dataset used for analysis the data set used in study consists of attributes like patient like gender, age, level of blood sugar, cholesterol. As per WHO, Cardiovascular disease (CVDs) takes life of 17.7 million people every year, 31 % of all global death [2].

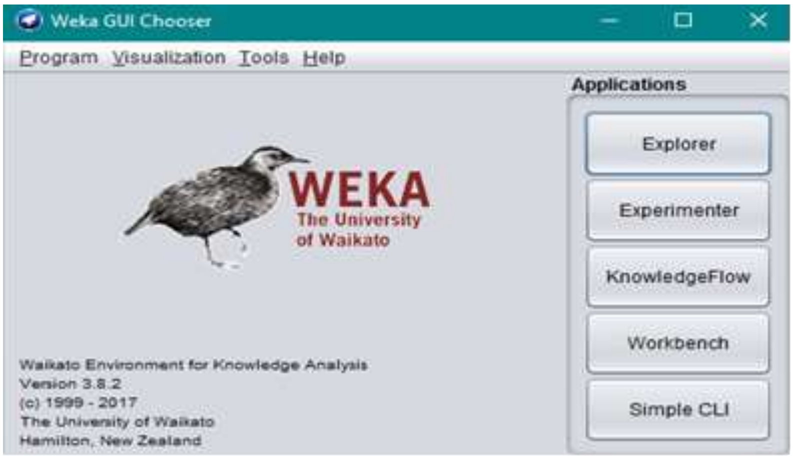
Data Mining is used in diverse areas of market research, fraud detection to health care sector. Data mining help in various sector of our society in making a prediction based on past result collected Data mining is a process of finding accurate and meaningful information from a very large set of data. The purpose of Data Mining is to be able to extract information from and make sense of large amounts of information [12].

This paper presents various attribute selector with search models to filter the attributes that will give a higher percentage of accuracy. The selected attributes are then classified using different classifiers. The output of applying classifiers on selected attributes are compared on the basis of correctly classified, incorrect classified, and accuracy of the classification model. For selecting attributes four attribute selectors are used Wrapper subset Evaluator, CFS subset Evaluator, Infogain Evaluator, Classifier subset Evaluator

2. METHODOLOGY

Weka tool is used in this study for the purpose of comparative study on different attribute subset evaluator Weka is Graphical user interface tool developed under the University of Waikato in New Zealand

It has a huge collection of machine learning algorithms for data mining. WEKA has different tabs for data pre-processing, classification, regression, clustering, association rules and visualization.



Weka GUI Chooser

Waikato Environment for Knowledge Analysis
Version 3.8.2
(c) 1999 - 2017
The University of Waikato
Hamilton, New Zealand

Fig I WEKA Interface

3. ANALYSIS OF CLASSIFICATION ALGORITHM

Firstly the study on complete data set is done by applying different classifiers on it like Naive Bayes, SMO, Bagging, J48, Random Forest, Random Tree, Multilayered perception

The table below compares the data set on the basis of the correctly classified instance, incorrect classified instance and accuracy in the percentage that is achieved by applying different classify on the heart dataset. First, we have applied the different classifiers on the complete data set.

Table I Different accuracy percentage with the complete data set

S.NO	Classifier's Name	Correctly classified Instances	Incorrect classified Instance	Accuracy
1	Naive Bayes	253	50	83.5
2	SMO	255	48	84.2
3	Bagging	246	57	81.2
4	J48	235	68	77.5
5	Random Forest	252	51	83.2
6	Multilayered perception	245	58	80.9
7	Random Tree	225	78	74.3

4. FEATURE SELECTION METHOD

FSA	Feature Selection Method	No of Attribute Selected
FSA1	WrapperSubset Evaluator	5 (3,9,11,12,13)
FSA2	CFS Subset Evaluator Best first forward	7(3,7,8,9,10,12,13)
FSA3	InfoGain Attribute Evaluator Ranker	13(13,3,12,10,9,8,11,1,2,7,6,4,5)

FSA	**Feature Selection Method**	**No of Attribute Selected**
FSA4	Classifier subset Evaluation	6(3,5,9,11,12,13)

Working with all the attribute sometimes gives less accuracy .because all the attribute may or may not give exact classification and accuracy. So feature selection is one the important method to achieve good accuracy for classification.

Attribute Selection for the purpose of data mining will improve prediction performance and classification accuracy. This can be applied in both supervised and unsupervised learning

Table II. Attribute Selection by Each Method

FSA	**Feature Selection Method**	**No of Attribute Selected**
FSA1	WrapperSubset Evaluator	5 (3,9,11,12,13)
FSA2	CFS Subset Evaluator Best first forward	7(3,7,8,9,10,12,13)
FSA3	InfoGain Attribute Evaluator Ranker	13(13,3,12,10,9, 8,11,1,2,7,6,4,5)
FSA4	Classifier subset Evaluation	6(3,5,9,11,12,13)

5. RESULT

After applying various attribute selector the accuracy of the result has increased. Wrapper subset evaluator gives 82.5 % accuracy, CFS attribute selector gives 74.3 % accuracy with different classifier Infogain Attribute selector gives 77.2 % accuracy and Classifier subset Evaluator gives 77.2 % accuracy

Table III Result of Different Classification techniques

Classifiers	**FSA1**	**FSA2**	**FSA3**	**FSA4**
J48	80.5	77.2	77.5	80.2
Naive Bayes	84.8	84.5	84.5	84.8
Multilayer perception	80.8	82.5	84.5	82.8
Random forest	83.2	81.2	81.8	80.5
SMO	82.5	83.5	84.2	84.8
Bagging	83.1	81.5	81.2	82.5
Random Tree	82.5	74.3	76.9	77.2
Average correctly classified	82.5	80.7	81.5	81.7

6. RELATED WORK

Keerthana T K proposed a heart disease prediction system which predicts the likelihood of a patient having heart disease by using his medical profile. The system uses three classification algorithm for predicting the risk of heart disease.[1].

Monika Gandhi et al. [3]. Designed a system for predicting heart disease using various classification algorithm the paper describes the various type of heart illness and distinguishes various classification algorithms with their advantages and disadvantages

Ranganatha S et al. [4] proposed a system that predicts the heart disease of patients admitted in the hospital by taking the input of patients. The output is human understandable words system implements ID3 and Naïve Bayesian algorithms.

Ajay Patel et al.[15] proposed a system that predicts the presence of high risk for heart disease or not. The proposed model uses heart data set for prediction and applied Naive Bayes and WAC classifier.

AH, Chen et.al. [16] In his study uses the artificial neural network for classification and prediction of heart dataset. The application was developed using C and C# language the result of the system is 80% accuracy.

T.Georgeena.S et al. [17] find critical attributes using of Apriori algorithm. The Efficiency of the system is measured by factors like recall, F measure, and ROC space. The proposed model is developed using ASP.Net

Durairaj M,et al.[18] proposed TrainBr algorithm of Multilayer Perceptron a type of artificial neural network gives highest accuracy . In his study trainer Algorithm gives 92.3 % accuracy .

7. CONCLUSION

The study in this paper is carried to find the best feature selection model that can provide higher accuracy . After comparing the result with different classifier the output of wrapper subset evaluator as compared to other evaluators is higher in term of accuracy gained after applying classifier on selected attributes. Hence after analyzing the values of different parameters and accuracy achieved with different classifiers, it has been observed that all the evaluators gives higher accuracy when different classification algorithm is applied to selected attributes in data set rather than complete data set. The wrapper subset evaluator method gives 82.5% of accuracy on heart dataset, which is higher among all the attribute evaluator method In the presence of large volume of data working on selected attributes may give higher accuracy in comparison to working with complete large data set.

8. REFERENCES

[1] Keerthana T K "Heart Disease Prediction System using Data Mining Method" International Journal of Engineering Trends and Technology volume 47 2017.

[2] www.who.int/cardiovascular_diseases/en/

[3] Monika Gandhi, Dr. Shailendra Narayan Singh "Predictions in Heart Disease Using Techniques of Data Mining" 2015 IEEE

[4] Ranganatha s., pooja raj h.r., Anusha c, Vinay s.k "medical data mining and analysis for heart disease dataset using classification techniques" IEEE 2014.

[5] K.srinivas, Dr.g.raghavendra Rao, dr. A.govardhan "analysis of coronary heart disease and prediction of a heart attack in coal mining region using data mining techniques" IEEE 2010.

[6]https://docs.oracle.com/cd/b19306_01/datamine.102/b14339 /3predictive.htm

[7]https://docs.oracle.com/cd/b28359_01/datamine.111/b28129/ classify.htm#i1005746

[8] s.sharmila, dr.m.p.indra Gandhi, " analysis of heart disease prediction using datamining techniques" international journal of advanced networking & applications (ijana) pages: 93-95 (2017)

[9]https://www.cdc.gov/dhdsp/data_statistics/fact_sheets /fs_heart_disease.htm

[10] Mrs. Bharati m. Ramageri "data mining techniques and applications" ijcsc vol. 1 no. 4 301-305

[11] http://archive.ics.uci.edu/ml/datasets

[12] nature of biotechnology. Data mining. Nature Biotechnology, 18:237–238, 2000.

[13]https://docs.oracle.com/cd/b28359_01 /data mine.111/b28129/algo_nb.htm#babiidde

[14] Gaganjot Kaur, Amit Chhabra "improved j48 classification algorithm for the prediction of diabetes" international journal of computer applications (0975 – 8887 volume 98 – no.22, July 2014

[15] Ajad Patel, Sonali Gandhi, Swetha Shetty, prof. Bhanu tekwani " heart disease prediction using data mining" international research journal of engineering and technology (irjet) volume: 04 issue: 01 | Jan -2017

[16] Ah chen, sy huang, ps hong, ch cheng, ej lin "hdps: heart disease prediction system" issn 0276-6574 IEEE 2011.

[17]T.georgeena.s. Thomas, siddhesh.s. Budhkar, siddhesh.k. Cheulkar, akshay.b.choudhary, rohan singh: heart disease diagnosis system using apriori algorithm: international journal of advanced research in computer science and software engineering, volume 5, issue 2, february (2015)

[18] Durairaj m, revathi v prediction of heart disease using back propagation mlp algorithm "international journal of scientific & technology research volume 4, issue 08, august 2015 "