RESEARCH REVIEW JOURNALS
Volume-04
Issue-03
March -2019

ISSN: 2455-3085 (Online)
RESEARCH REVIEW International Journal of Multidisciplinary
www.rrjournals.com [UGC Listed Journal]

Development and Evaluation of a collection of Heart Disease Prediction Models using Decision Tree Algorithm based on the various Feature Selection Methods

[1]Qureshi Mujtaba Ashraf and [2]Srivastava Azad Kumar

[1]Research Scholar, Department of Information Technology, Mewar University, Chittorgarh(Raj), India.
[2]Professor, Department of Computer Science, Mewar University, Chittorgarh (Raj), India

ARTICLE DETAILS

Article History
Published Online: 13 March 2019

Keywords
Heart Disease Prediction System, Data Mining, Neural Networks, Feature selection, WEKA tool

***Corresponding Author**
Email:mujtaba170@gmail.com

ABSTRACT

Cardiovascular disease is considered one of the largest fatal diseases in the world. Heart diseases have surpassed more than 41% contribution to the mortality rate of human race. Grave concern is needed by the world community predominantly the medical science domain, to curb these fatal diseases before to enter into the unrestrained zone. So well planned and vigilant steps are needed to shrink the effects of these cardiovascular lethal diseases. In this research paper various heart disease prediction models are developed using decision tree algorithm based upon the four feature selection methods. Feature selection methods selects and ranks the high profile attributes to form effective prediction models. A comparative study is performed to analyze and select the high performance model based on the particular feature selection method.

1. Introduction

Data mining refers to the extraction of valuable and unknown information from the dissimilar and distant databases. The hidden and unseen information of diverse databases will be of no use if data mining technology has not been come into existence. Applicability of data mining technology has been extensively captivated in the diverse and significant fields. One of the most important fields where data mining techniques have become imperative is the healthcare industry. Data mining techniques plays an important role to make predictions regarding diverse fatal diseases and thus to take remedial steps to control those diseases well on time. Cardiovascular diseases cause approximately 41% deaths alone in the world, as per WHO study. However data mining techniques came to rescue to the human community involved in the various kinds of heart diseases. Data mining technology uses various techniques as per the choice and requirement to predict dissimilar heart diseases well on time.

In this research paper decision tree technique is used to develop various heart disease prediction models based on the various feature selection techniques available used to enhance the purity and selection of ranking of applied datasets. Finally a model which shows an acceptable performance measures is selected and adopted by means of comparative study analysis.

2. Literature Review

Palaniappan et al. [1] developed an intelligent heart disease prediction model using naïve bayes, decision tree and neural network.In [2] an intelligent prediction system is developed for cardiovascular diseases. This is implemented using the .net platform and answer is obtained by query called “what if query”.

Carlos Ordonez [3] used association rules for the prediction of cardiac diseases. Simple mapping algorithm is used which continuously treats features as categorical or numerical values. A developed algorithm is put to use to obtain constrained association rules. Decision tree is preferred as it breaks numerical values in automatic manner.

Milan Kumari et al. [4] proposed cardiovascular diseases prediction system using various data mining algorithms like ANN, SVM, Decision Tree and RIPPER classifier. Author used various performance measures to analyze and compare the used techniques. Out of these four data mining techniques SVM presents better results in comparison to other models for the prediction of heart diseases.Moloud Abder et al. conducted a comparative study of various data mining techniques to analyze the best classifier having better performance results to predict CVDs. The techniques used are Decision Tree (C5.0), Artificial Neural Network, Support Vector Machine (SVM), K-Nearest Neighborhood (KNN) and Logistic Regression. The model developed using decision tree presents higher accuracy i.e. 93.02% in comparison to other techniques/models. Decision rules are mostly simple and easy to understand for experts and other professionals.Boshra. Et al analysed various data mining techniques for the prediction of heart diseases like KNN, Decision Tree (J48), Naive Bayes and SMO. A comparative study is made based on various performance measures like sensitivity, specificity, accuracy etc of the said techniques. Results show that Decision tree (J48) is an acceptable and best classifier for the prediction of cardiovascular diseases.

Xing et al. [5] presents technique to find the probability of continued existence of heart patients by hybridizing of various renowned techniques. These techniques are neural networks, support vector machine and decision tree.

The researchers [6] applied the various data mining techniques like decision trees, naïve bayes, neural networks, association classification and genetic algorithm for diagnosis the CVD using heart datasets.

Nidhi Bhatla et. al. [7] in 2012 conducted an experimental work, “A Novel Approach for Heart Disease Diagnosis using Data Mining and Fuzzy Logic”. The main focus of this study was to decrease the number of attributes utilized for the prediction of heart diseases. Also to develop the CVD predictive model

having good predictive results. Various data mining techniques are put to use. However results show that Naive Bayes and Decision Tree has over powered to other techniques.

M. Anbarasi et al. [8] presented acceptable heart diseases prediction system using GA with feature subset selection. Also three classifiers namely naive bayes, decision tree and classification using clustering was employed. From this experimental work decision tree has shown good results for heart diseases prediction.

Tu et al. [9] develops a sharp and acceptable cardiovascular prediction system. The author used the decision tree C4.5 and naive bayes.

Su et al. 2011 [10] investigated three data mining techniques which are Decision Tree (C4.5), BPN and Baysian Network to mine medical datasets.

3. Framework

The framework of the present experimental work and programming work flow is depicted in lucid approach in Figure 1;

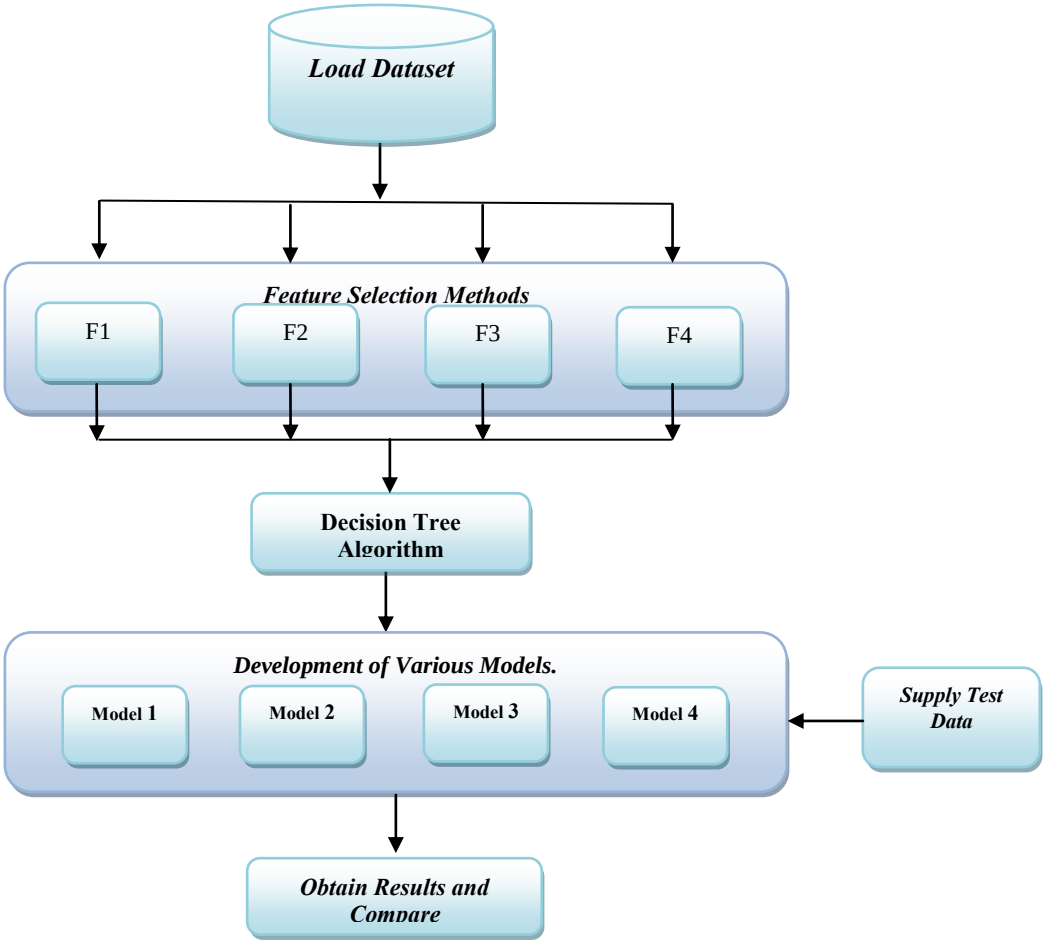

Figure 1: Framework

The abbreviations applied in the framework are defined below;
F1 stands for ReliefAttributreEval with Ranker-T.
F2 stands for InfoGainAttributeEval with Ranker-T.
F3 stands for CfsSubsetEval with Greedy Stepwise.
F4 stands for Correlation Attribute Eval with Ranker-T

4. Decision Tree Algorithm

Decision Tree is considered one of the effective data mining techniques, introduced in 1960's. The working mechanism of decision tree is easy to understand, doubt free and ease of application. The decision tree develops a model in the form of tree like structure, so along with other unfussiness features debugging process is trouble-free. The structure of tree is divided into root node, internal node and the leaf node. Root node acts as a parent node and leaf node is the final output of the prediction value in the decision tree structure. Actually decision tree works on the Information gain method of the attributes used in the applied datasets. Information gain method splits the attributes in the tree like structure as per their value of information provided using the following equation 1 which decides the attribute to act as a root node, internal node/nodes and leaf node/nodes.

$$E(S) = -P(P)\log_2 P(P) - P(N)\log_2 P(N) \quad \text{eq. (1)}$$

The algorithm for the decision tree is given as:

Step 1: *Obtain the information gain of all the attributes used in the dataset.*
Step 2: *Arrange the attributes on the basis of information gain in decreasing order.*
Step 3: *The attribute with highest value of information gain is selected as root node.*
Step 4: *After that obtain the information gain by applying same formula.*
Step 5: *Further divide the nodes using the highest information gain method.*
Step 6: *Process is repeated until each attribute is not set as leaf nodes in the branches of the decision tree formed.*

5. Methodology

Diverse feature selection methods are used to select the attributes having strong relation with the predictive power of the developed models. Also these methods play vital role to remove the redundant and irrelevant features from the loaded datasets. Correspondingly we loaded the obtained dataset to the four selected feature selection methods namely Relief AttributreEvaluation with Ranker-T, InfoGainAttributeEval with Ranker-T, CfsSubsetEval with GreedyStepwise and CorrelationAttributreEval with Ranker-T in separate manner to obtain the refined datasets having attributes arranged as per their weightage to develop the predictive models. The various models are developed using these obtained datasets but only first 8 attributes arranged in order, are selected with 720 instances. All the heart disease predictive models are supplied with the test datasets to evaluate the performance measures. Finally the results shown by these models are compared and the acceptable model is selected and adopted for the prediction of heart diseases based on the used feature selection methods. WEKA simulation tool is used to imitate this research work. Models are developed using 60% training data and 40% as test data. *Model 1* is developed using *ReliefAttributeEval with Ranker-T. Model 2 is developed using InfoGainAttributeEval with Ranker-T. Model 3 is developed by means ofCfsSubsetEval with Greedy Stepwise. Model 4 is developed using CorrelationAttributeEval with Ranker-T.*

A. Dataset Description

A wide variety of attributes related to heart diseases is collected by means of various methods. The choice of attributes is made from the 14 attributes shown below in table 1. We used 720 instances with only 08 attributes on the basisof their ranking system by means of various feature selection methods to develop various prediction models.

Table 1
Collection of Attributes.

S.No	Name	Description	Range
1	Age	Age in Years	1 -100
2	Sex	Male/Female	female=0, male=1
3	cp	Chest pain type	Typical angina=1,atypical angina=2,non-anginal pain=3,asymptomatic pain=4
4	trestbp	Resting blood pressure	120/80- 140/100
5	Serumcho	Serum cholesterol in mg/dl	126-564
6	fbs	Fasting blood sugar level	Yes=1, No=0
7	restecg	Resting electrographic results	Normal=0,ST_T_wave_abnormality=1, Ventricular hypertrophy=2
8	thalach	Maximum heart rate achieved	82-125
9	exang	Exercise induced angina	Yes=1,No=0
10	oldpeak	ST depression induced by exercise	71-202
11	peakSlope	the slope of the peak exercise ST segment	1-3
12	numVessels	number of major vessels colored by fluoroscopy	0-3
13	thal	the defect type of heart	normal=3,fixed_defect=6,rever-defect=7
14	Diseases	Yes/no	Yes=1,N0=0

B. Feature selection methods

Feature selection methods play a vital role to form a set of highly refined and distinguished attributes and thus to contribute good enough to develop accurate prediction models. In actual, feature selection method is composed of attribute evaluator and search method. In this research work we selected four feature selection methods to develop four respective models for comparative analysis of their performance measures. Feature selection methods arrange the supplied attributes as per their significance and ranking for prediction. So in this research work we selected the first 8 attributes in order after the every feature selection method is applied and finally use 720 instances of these 8 attributes to decision tree to develop models. The used feature selection methods are given as;

B.1: ReliefAttributeEval with Ranker-T:-Subsequent to the application of attributes shown in table 1 by means of ReliefAttributeEval with Ranker-T, the order of attributes obtained based as per their significance for prediction is as 2,6,4,1,8,7,3,5,10,12,11,9,13 Thus here we opted attributes of s no's in table 1 as *2,6,4,1,8,7,3 and 5* i.e., first 8 selected attributes.

B.2: InfoGainAttributeEval with Ranker-T:-Subsequent to the application of attributes shown in table 1 by means of *InfoGainAttributeEval with Ranker-T*, the order of attributes obtained based as per their significance for the prediction is as 2,6,8,5,4,1,7,3,9,11,13,10,12. Thus here we opted attributes of s no's in table 1 as *2,6,8,5,4,1,7 and 3* i.e., first 8 selected attributes.

B.3: CfsSubsetEval with GreedyStepwise:-Subsequent to the application of attributes shown in table 1 by means of *CfsSubsetEval with GreedyStepwise*, the order of attributes obtained based as per their significance for prediction is as 2,6,8,5,4,1,7,9,3,10,12,11,13. Thus here we opted attributes of s no's in table 1 as *2,6,8,5,4,1,7 and 3* i.e., first 8 selected attributes.

B.4: CorrelationAttributeEval with Ranker-T:-Subsequent to the application of attributes shown in table1 by means of *CorrelationAttributeEval with Ranker-T*, the order of attributes obtained based as per their significance for prediction is as 2,6,8,1,7,5,4,3,9,12,10,13,11. Thus here we opted attributes of s no's in table 1 as *2,6,8,1,7,5,4 and 3* i.e., first 8 selected attributes.

C. Performance measures

The model evaluation is performed by using various performance measures presented subsequently. Accuracy, precision, recall (sensitivity) and ROC curve are taken into consideration for the comparison analysis between the developed models. Some of performance measures taken into consideration are defined below as;

Accuracy = (TP+TN)/TP+TN+FP+FN)
Precision = (TP/TP+FP)
Recall (Sensitivity) = (TP/TP+FN)
Specificity= (TN)/ (TN + FP)
Here,

TP is defined as the number of true positive: Patients of heart disease correctly accepted/recognized as heart diseases.

TN is defined as the number of true negative: Healthy people correctly accepted/recognized as healthy.

FP is defined as the number of false positive: Healthy people wrongly accepted/recognized as Heart Disease.

FN is defined as the number of false negative: Patients of heart incorrectly accepted/recognized as well and healthy.

6. Experimental Results

Processed datasets are obtained by means of four feature selection methods i.e. F1, F2, F3 & F4 to develop four predictive models namely M1, M2, M3 & M4 respectively using decision tree data mining technique. WEKA simulation tool is employed to perform the experimental work. Following performance measures are revealed by the four developed predictive models, shown in the tabular form in table 2:

Table 1: Performance Measures shown by various Models.

Model Name	Accuracy	T.P. Rate	F.P. Rate	Precision	Recall	ROC Area
M1	84.05 %	84.1%	16.7%	84.0%	84.1%	85.8%
M2	84.54%	84.3%	15.7%	84.6%	84.5%	87.2%
M3	86.15%	86.2%	14.6%	86.20%	86.2%	90.7%
M4	83.57%	83.6%	16.4%	83.7%	83.6%	87.0%

The results obtained using the four models are displayed below in graphical form in figure 2;

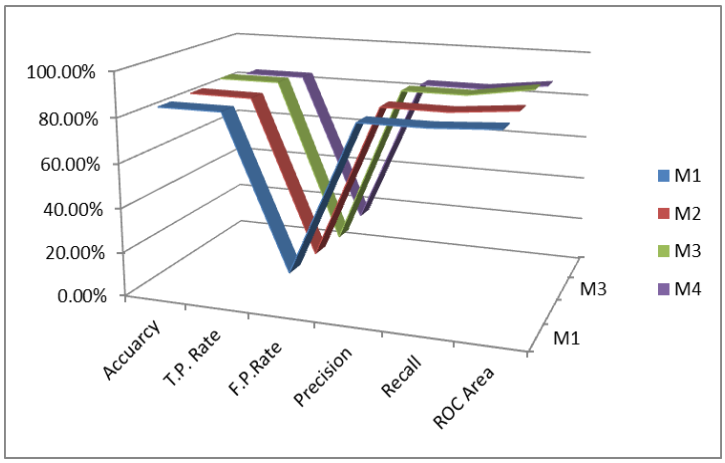

Figure 2: Results shown in graphical form.

7. Conclusion and Future Work

Decision tree algorithm is used to develop various cardio-vascular prediction models based on the dissimilar feature selection methods. Feature selection methods have great impact upon the development and efficiency of the prediction models. In this research paper four feature selection methods are selected and adopted to obtain the highly refined and high rank variables, which are used to develop heart disease prediction models by the application of decision tree algorithm. By analyzing the models using comparative study methods,

model 3 shows higher performance results in comparison to other three models. The model 3 uses the Cfs Subset Evaluation with Greedy Stepwise method to refine and select higher rank attributes. Here an important conclusion came up i.e. the Cfs Subset Evaluation method with Greedy Stepwise search method developed model 3 with high performance results by supplying the more meaningful and highly refined attributes in comparison to other three feature selection methods.

This research work can be enhanced by the application of more than the four selected feature selection methods for ranking and selection of high profile attributes to develop prediction models. Also more instances of data with more number of attributes can be used to develop and test the prediction models.

Acknowledgement

I am very thankful to my supervisor Professor Dr. Azad Kumar Srivastava of his support and assistance to make this research work successful.

References

1. X. Yanwei et al., "Combination Data Mining Models with New Medical Data to Predict Outcome of Coronary Heart Disease", Proceedings of International Conference on Convergence Information Technology, pp. 868-872, 2007.
2. Sellappan Palaniappan, Rafiah Awang "Intelligent Heart Disease Prediction System Using Data Mining Techniques"Department of Information Technology Malaysia University of Science and Technology Block C, Kelana Square, Jalan SS7/26 Kelana Jaya, 47301 Petaling Jaya, Selangor, Malaysia .
3. Carlos Ordonez, EdwardOmincenski and Levien de Braal ,"Mining Constraint Association Rules to Predict Heart Disease", Proceeding of 2001, IEEE International Conference of Data Mining, IEEE Computer Society, ISBN-0-7695-1119-8, 2001, pp: 433-440.
4. Milan Kumari, Sunila Godara, "Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction", IJCST Vol. (2), Issue (2), June 2011.
5. Yanwei Xing, Jie Wang and Zhihong Zhao (2007). Combination data mining methods with new medical data to predicting outcome of Coronary Heart Disease. IEEE. p1-5.
6. K. Sudhakar, "Study of Heart Disease Prediction using Data Mining," vol. 4, no. 1, pp. 1157–1160, 2014.
7. Nidhi Bhatla, Kiran Jyoti, "A Novel Approach for Heart Disease Diagnosis using Data Mining and Fuzzy Logic", International Journal of Computer Applications, Volume 54–No.17, (pp 16-21), September 2012, ISSN 0975 – 8887.
8. M. Anbarasi, E. Anupriya, N.Ch.S.N.Iyengar, "Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm"; International Journal of Engineering Science and Technology, Vol. 2(10), 2010.
9. My Chau Tu AND Dongil Shin (2009). A Comparative Study of Medical Data Classification Methods Based on Decision Tree and Bagging Algorithms. IEEE. P1-5.
10. Jenn-Lung Su, Guo-Zhen Wu, I-Pin Chao (2001). The Approach Of Data Mining Methods For Medical Database. *IEEE*. p1-3.

