ISSN (Print) : 0974-6846
Indian Journal of Science and Technology, Vol 8(35), DOI: 10.17485/ijst/2015/v8i35/87458, December 2015 ISSN (Online) : 0974-5645

Performance Analysis of Various Data Mining Techniques in the Prediction of Heart Disease

Kodali Lohita, Adusumilli Amitha Sree, Doreti Poojitha, T. Renuga Devi* and A. Umamakeswari

School of Computing, SASTRA University, Thanjavur – 613401, Tamil Nadu, India;
renugadevi@cse.sastra.edu

Abstract

Objective: The main objective of the work is to compare the heart disease prediction accuracy of different data mining classification technique and to find the best technique with minimum incorrectly classified instances. Different classification techniques are used to predict heart disease based on the factors that cause these diseases which include family history, age, obesity and some other factors. **Method:** This work is carried out in three phases. The First Phase is preprocessing of data set. The attributes like trestbps, cholesterol, tpeakbps and age are normalized and missing values are handled appropriately. The second phase is feature selection. The greedy hill climbing best first attribute evaluator is used to identify the subset of attributes based on its individual prediction ability. The third phase is comparison of prediction accuracy of different techniques in literature. **Findings:** The work has been evaluated using the performance metrics like accuracy, specificity, sensitivity, confusion matrix to prove the efficiency of different techniques. It was concluded that the Bagging algorithm achieved highest accuracy compared with other algorithms.

Keywords: Classification, Data Mining, Heart Disease Prediction

1. Introduction

Heart Disease is one among the major life threatening diseases. Statistics indicates that the mortality rate in India is about 45%, United Kingdom is about 38% and Australia is about 26.30%. WHO identified the significance and usefulness of data mining techniques in diagnosis of Heart disease.

Coronary Heart Disease (CHD) occurs due to the accumulation of plaque along the walls of the arteries in the heart and makes the arteries harder, this condition is called atherosclerosis. The lumen of arteries becomes narrow and results in the formation of blood clots so blood flow to the heart is reduced. The risk of CHD may increase due to factors like age, high cholesterol, diabetes, obesity, smoking and high blood pressure. The people having CHD feel healthy for long years before they start experiencing symptoms like chest pain.

In[1] Logistic Regression (LR), Radial Basis Function (RBF), CART, self-organizing feature maps (SOFM) and Multi-layer Perceptron (MLP) are used to predict the probability of presence of coronary heart disease which resulted in 0.783 area under ROC curve. In[2] Naive Bayes, k-Nearest Neighbor (k-NN) and Decision list is used for the classification of data and resulted in 52.33% of accuracy by naive bayes algorithm.

In[3] naive bayes, decision tree, K-Nearest Neighbor (KNN), neural networks along with CFS subset, Chi-squared, Consistency subset, filtered attribute, filtered subset and Gain ratio attribute selection methods are used to predict the presence of heart disease which resulted in 85.5% accuracy by Naive Bayes with CFS subset attribute selection method.

In[4] a dataset consisting of 303 patient records with 54 features called Z- Alizadeh Sani dataset is used for classification using Bagging, Naive Bayes, Sequential Minimal Optimization (SMO), and Neural Network algorithms in association with feature selection and feature creation algorithms. The analysis resulted in 94.08% accuracy for SMO algorithm with created features.

* Author for correspondence

In[5] a K-Nearest Neighbor (KNN) and genetic algorithm based hybrid method is proposed. The hybrid method resulted in greater than 95% accuracy.

In[6] a hybrid approach is proposed which is a combination of Feature Selection (FS), fuzzy weighted preprocessing and Artificial Immune Recognition System (AIRS) for medical decision support systems. Feature selection is done using C4.5 decision tree algorithm, normalization done in the range of [0, 1] fuzzy weighted preprocessing and AIRS classifier is used classification. This hybrid method resulted in 92.59% accuracy. In[7] a Traditional Risk Factors (TRF) plus plaque plus Carotid Intima - Media Thickness (CIMT) proved best for the prediction of CHD than individual factors. In[8] HER is used to accurately predict heart related hospitalization. The patients with chest pain and RBBB have higher risk of CAD. Chest X-ray, ECG and 2D echocardiography were the tests considered. For effective prediction of CAD Angiography was suggested[9].

In[10] important health issues of young women for the assessment of cardiovascular risk factors based on Risk Assessment Index (RAI) is considered. The RAI scores were used to categorize women into low, moderate and high risk groups. These RAI scores were compared to the subjects lipid profiles to assess the risk of cardiovascular disease. In[15] genetic algorithm and fuzzy inference system is used for effective prediction of heart disease inpatients. Fuzzy Gaussian membership function and defuzzification using centroid method improves the performance of the system. In[16] adaptive Neuro fuzzy inference system with adaptive group based K-Nearest Neighbour algorithm has been used to classify the data in prediction of heart disease and cancer in diabetic patients. The survey of different data mining techniques involved in risk prediction of heart disease[17] prove the hybrid approach is best compared to single model approach.

2. The Dataset

The Heart disease dataset from UCI Machine Learning repository[11] with 75 attributes and 902 instances is considered. The attribute, "num" represents the class attribute. If the patient is likely to get CHD, the value of class attribute is zero (True), else the value is one (False). The dataset is given as input to seven classification models.

2.1 Data Pre Processing

Data Preprocessing is the process in which data cleaning, data transformation and data reduction are done. It is significant step in data mining. The data cleaning process include filling the missing values, smoothing noisy data, removal of outliers and clearing up inconsistencies. Data Transformation includes normalization and aggregation techniques and data reduction consists of reducing the amount of data but producing the same results. For Heart disease dataset the missing values are handled by replacing the missing values with a constant - 9. Next the values of the attributes like trestbps, cholesterol, tpeakbps and age are normalized. The normalization of values is done by taking their mean values into consideration. Table 1 represents the sample Data Set and the way how the missing values are handled is shown in Table 2.

Table 1. Raw Data Set

smoke	cigs	dig	Famhist	Painloc
1	-	1	0	1
-	20	-	0	0
0	40	0	-	1
1	-	-	1	-
-	0	1	0	-
0	-	1	1	1
1	10	-	-	1
1	-	0	-	-
0	15	-	0	1

Table 2. Preprocessed Data Set

Smoke	cigs	dig	Famhist	Painloc
1	-9	1	0	1
-9	20	-9	0	0
0	40	0	-9	1
1	-9	-9	1	-9
-9	0	1	0	-9
0	-9	1	1	1
1	10	-9	-9	1
1	-9	0	-9	-9
0	15	-9	0	1

Figure 1 Gives the overall description of the system in which first the raw data is preprocessed and the processed data is given to different algorithms and the output's of each algorithm are compared against each other.

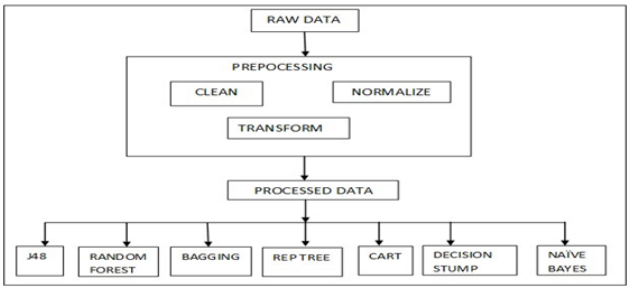

Figure 1. System overview.

3. Implementation Methodology

The classification models considered for implementation are J48, Random forest, Bagging, REP Tree, CART, Naïve Bayes and Decision Stump[12-14].

3.1 J48

It is a simple decision tree learning technique. The main aim of this technique is to build a single decision tree in top-down manner. Greedy search technique is employed at every node in order to test each attribute. The selection of the best attribute done based on a metric named information gain is used. For accurate definition of information gain entropy value is calculated.

Given a set S contains true and false.

$$Entropy(S) = - [P(true)\log_2 P(true) - P(false)\log_2 P(false)]$$

Where,

P (true): Proportion of true in S.

P (false): Proportion of false in S.

The metric information gain is calculated as follows:

$$Gain(S, A) = [Entropy(S) \ \Sigma ((|S_v|/ |S|) * Entropy (S_v))]$$

Where:

Σ is each value v of all possible values of attribute A

S_v = subset of S for which attribute A has value v

$|S_v|$ = number of elements in S_v

$|S|$ = number of elements in S

The preprocessed training data is used to build the classifier. While building the classifier, information gain and entropy are calculated for each attribute. The attribute with highest entropy value is chosen as the root node as shown in Figure 2. The built classifier is used to evaluate the test data. The results of the evaluation accuracy is 94.78%, sensitivity of 90.85%, ROC area of 0.962, mean absolute error of 0.0555, specificity of 98.70%, correctly and incorrectly classified instances are 291 and 16 respectively.

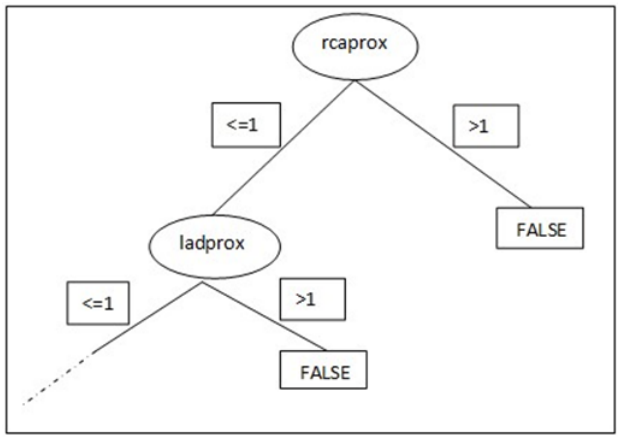

Figure 2. J48 Tree representation.

The advantages of J48 are that it can handle missing values, able to process erroneous datasets, can handle numerical, nominal and textual data. However this algorithm runs slow for large and noisy datasets and space complexity is high.

3.2 Random Forest

It is an Ensemble model *i.e* collection of decision trees. A different subset of the training data are selected and given to each classifier. Individual classifier is a "weak learner", while all the classifiers taken together are the "strong learner". When building the decision trees, each time a split is considered, a random sample of n predictors is chosen as split candidates from the full set of m predictors.

The split is allowed to use only one of those n predictors. A fresh sample of n predictors is taken at each split and typically n is approximately the square root of m. The subset should be carefully selected and should be at least 66% of the total data. Smaller subset produces lower error rate but lower predictive power.

While building a decision tree for each subset, approximately 9 (approximate square root of 75) predictors are considered at a time without replacement i.e. there is no redundancy among the subsets. For each subset of training data, randomly 9 predictors are considered using which decision trees are built. While building each classifier, information gain and entropy are calculated for each attribute. Using a voting mechanism, best classifier is chosen among all the classifiers. The voted classifier is used to evaluate the test data. The results of the evaluation are accuracy of 95.77%, sensitivity of 94.44%, ROC area

of 0.990, mean absolute error of 0.1747, specificity of 96.93%, correctly and incorrectly classified instances are 294 and 13 respectively.

The advantages of random forest includes handling missing values automatically, not very sensitive to outliers in training data and works efficiently for large databases. But it is observed to over fit for some datasets with noisy classification tasks and difficult for interpretation.

3.3 Bagging

It is also an Ensemble model but the difference between random forest and bagging lies in the selection of predictor subset size n. If n is equal to m, then it is bagging. Bagging yields best results when the predictors are non-correlated. The attribute with highest entropy value is chosen as the root node. Voting mechanism is done to choose classifiers. The results of the evaluation are accuracy of 97.39%, sensitivity of 93.94%, ROC area of 0.992, mean absolute error of 0.1452, specificity of 100%, correctly and incorrectly classified instances are 299 and 8 respectively.

The advantages are that it is fast and scalable, produces lower error rates and gives higher accuracy. The drawback is its computational complexity

3.4 Rep Tree

REP Tree is a fast decision tree learning technique. It constructs a decision tree using metric named information gain or by reducing the variance. Trees are pruned back fitting (reduced-error pruning). Numeric attributes values are sorted only one time. By dividing the corresponding instances into samples, missing values are dealt.

REP Tree considers all 75 attributes. At each step it does reduced error pruning. While traversing from top to the bottom of a tree over the internal nodes, checks if replacing the internal node with most repeated class that will not reduce the accuracy of the tree. It is done until further pruning decreases accuracy.

3.5 Decision Stump

Decision stump is generally a single-level decision tree. It means that the decision tree has one root connected to its leaves. The decision is based only on a single feature. Decision stumps are generally used as base learners in Ensemble learning methods.

3.6 CART

A CART Tree is a binary tree. It splits a node into two child nodes repeatedly, beginning with the root node which has the entire learning sample. CART methodology consists of three parts - creation of maximum tree, choosing the correct tree size, classifying the recent data by constructed tree.

In CART for the given dataset, it finds the predictor to be considered for splitting which minimizes the mean impurity i.e. the impurity of the descendant subsets is less than that of parent. This is done recursively until it can't be split further. Then prune the tree back to where it best matches the held out data.

3.7 Naive Bayes Classifier

Naive Bayes classifier is a simple probabilistic model based on Bayes' theorem. In this model, classifiers assume that the value of a particular attribute is independent of the value of any other attribute, in the class variable.

The probability of data record *X* having the class label C*i* is:

$P(Ci|X) = P(X|Ci) * P(Ci) / P(X)$

The class label Ci with largest conditional probability value determines the category of the data record. Naive bayes considers 2 classes true and false. Using the formula it decides whether the given instance belongs to class true or false.

3.8 Performance Measure

Accuracy, sensitivity and specificity are some of the common metric used for performance measure.

3.9 Classification Matrix

A classification matrix is a tabular representation that allows visualizing the performance of a technique. It specifies the number of True Positives (TP), False Positives (FP), True Negatives (TN) and False Negatives (FN) in a two class problem. Where TP represents the number of correctly classified instances in class 1, TN represents the number of correctly instances in class 2, FP represents the number of correctly classified instances in class1 and FN represents the number of correctly instances in class 2.

3.9.1 Sensitivity and Specificity

Sensitivity and specificity are defined as follows

$$Specificity = \frac{TN}{(TN + FP)}$$

$$Spensitivity = \frac{TP}{(TP + FN)}$$

3.9.2 Accuracy

Accuracy is defined as the ratio of the number of correctly classified instances to the total number of instances of the test data.

$$Accuracy = \frac{TP + TN}{(TP + TN + FP + FN)}$$

4. Experimental Results

4.1 Performance Comparison of Algorithms

In order to employ the classification techniques, Weka tool[12] is used. The performance measures for J48, Bagging, Random Forest, REP Tree, Naive Bayes, CART and Decision Stump Algorithms executed on the whole dataset are represented in Table 3.

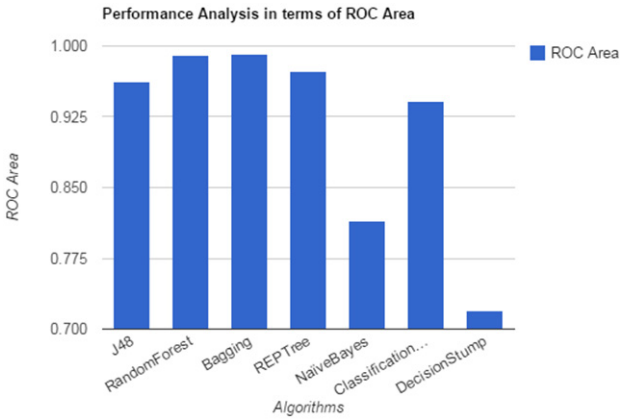

Figure 3. Accuracy comparison.

Table 4. The Confusion Matrix for Algorithms

Algorithm		Actual class1	Actual class2
J48	Predicted class1	139	2
	Predicted class2	14	152
Random Forest	Predicted class1	136	5
	Predicted class2	8	158
Bagging	Predicted class1	124	0
	Predicted class2	8	175
REP Tree	Predicted class1	141	0
	Predicted class2	9	157
Naive Bayes	Predicted class1	125	16
	Predicted class2	70	96
CART	Predicted class1	125	16
	Predicted class2	27	139
Decision Stump	Predicted class1	141	0
	Predicted class2	93	73

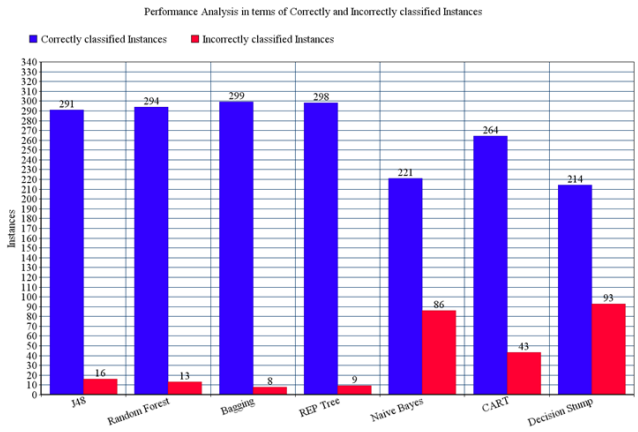

Figure 4. Correctly and Incorrectly classified instances.

Table 3. Performance comparison of Algorithms

Algorithms	Accuracy	Sensitivity	Specificity	F -Measure	Precision	Recall	ROC Area
J48	94.78%	90.85%	98.70%	0.948	0.948	0.951	0.962
Random Forest	95.77%	94.44%	96.93%	0.958	0.958	0.958	0.990
Bagging	97.39%	93.94%	100%	0.974	0.976	0.974	0.992
REP Tree	97.07%	94%	100%	0.971	0.972	0.971	0.973
Naïve Bayes	71.99%	64.10%	85.71%	0.715	0.758	0.720	0.815
CART	85.99%	82.23%	89.68%	0.860	0.863	0.860	0.941
Decision Stump	69.71%	60.26%	100%	0.676	0.817	0.697	0.720

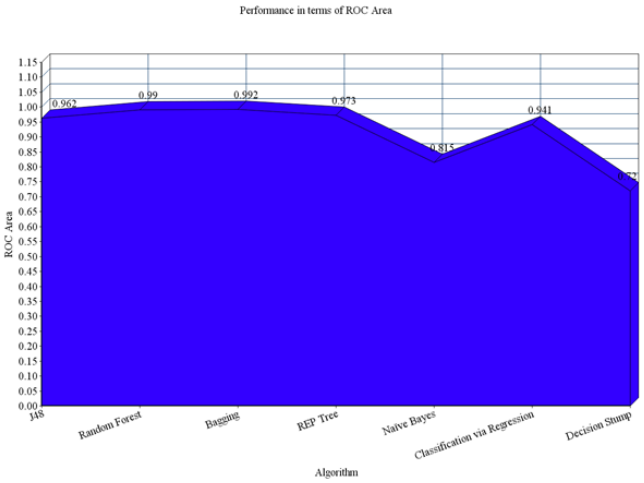

Figure 5. Area under ROC curve.

The different performance measures that are being compared are accuracy, sensitivity, specificity, F - Measure, Precision, Recall and Area under ROC Curve. The Confusion matrix for each algorithm is mentioned in Table 4. Figure 3 shows the accuracy comparison of all the algorithms and Figure 4 represents the number of correctly and incorrectly classified instances. Figure 5 shows the Area under ROC curve performance measure. The results in Table 3 suggest that Bagging algorithm has highest accuracy and ROC area followed by REP Tree, Random Forest, J48, CART, Naive Bayes and Decision Stump algorithms. While Random Forest has high sensitivity value but Bagging has high specificity, F - Measure, Precision and Recall values. So as shown in Table 3 Bagging algorithm is the best algorithm with highest accuracy at 97.39% when compared to other algorithms.

5. Conclusion and Future Work

In this study, various machine learning techniques are applied on the dataset and the results are obtained. The attributes mentioned in the dataset are some possible signs of Coronary Heart Disease. As per the study Bagging algorithm achieved highest accuracy of 97.39% when compared other algorithms like J48 and Random Forest. The achieved accuracy is higher when compared to the methods mentioned in the literature. Future work of this study is selection of algorithms which is a major issue to achieve better performance. Identification of significant attributes from all the available attributes is also a challenging task. Furthermore this work can be extended for various types of heart disease like coronary artery disease, heart failure and so on. Selecting preferential attributes which provides more accuracy is a major issue.

6. References

1. Kurt I, Ture M, Kurum AT. Comparing performances of logistic regression, classification and regression tree and neural networks for predicting coronary artery disease. Journal of Expert Systems with Applications. 2008; 34(1):366-74.
2. Raj kumar A, Reena G. Sophia. Diagnosis of heart disease using data mining algorithm. Global Journal of Computer Science and Technology. 2010; 38(10):38-44.
3. Peter TJ, Somasundaram K. An empirical study on prediction of heart disease using classification data mining techniques. IEEE - International Conference on Advances in Engineering, Science and Management. 2012; 514-8.
4. Alizadehsani R, Habibi J, Hosseini MJ, Mashayekhi H, Boghrati R, Ghandeharioun A, Bahadorian B, San ZA. A data mining approach for diagnosis of coronary artery disease. Computer methods and programs in biomedicine. 2013; 3:52-61.
5. Jabbar MA, Deekshatulu BL, Chandra P. Classification of Heart disease Using K- Nearest Neighbor and Genetic Algorithm. First International Conference on Computational Intelligence: Modeling Techniques and Applications. 2013; 10:85-94.
6. Polat K, Gunes S. A hybrid approach to medical decision support systems: Combining feature selection, fuzzy weighted pre-processing and AIRS. Computer methods and programs in biomedicine. 2007; 88:164-74.
7. Nambi V, Chambless L, Aaron R. Folsom, Hu Y, Mosley T, Volcik K, Boerwinkle E, Christie MB. Carotid Intima-Media Thickness and Presence or Absence of Plaque Improves Prediction of Coronary Heart Disease Risk. Journal of the American College of Cardiology. 2010; 55(15):1600-7.
8. Dai W, Theodora SB, William GA, Mela T, Saligrama V, Ioannis CP. Prediction of hospitalization due to heart diseases by supervised learning methods. International journal of Medical Informatics. 2015; 84(3):189-97.
9. Solbannavar R, Patted SV, Halkatti P, Aurora R. Clinical and Angiographic Correlation of Chest Pain with Right Bundle Branch Block. Indian Journal of Science and Technology. 2015; 8(3):208-15.
10. Thilagamani S, Uma MS. Risk appraisal for cardiovascular disease among selected young adult women in Coimbatore, India. Indian Journal of Science and Technology. 2010; 3(6):672-5
11. UCI Machine Learning Repository. Available from: http://archive.ics.uci.edu/ml/datasets/Heart+Disease, 2015.
12. Weka tool: Available from: http://www.cs.waikato.ac.nz/ml/weka/downloading.html,2015.
13. Breiman L. Bagging predictors: Machine Learning. 1996; 24:123-40.

14. Chi CM, Wu CC, Ching-Huang Lai, Hans-Bernd Bludau, Huei-Jane Tschai, Lu Pai, Shih-Ming Hsieh, Nian-Fong Chu, Angus Klar, Reinhold Haux, and Thomas Wetter. A Bayesian expert system for clinical detecting coronary artery disease. Journal of Medical Sciences. 2009; 29(4):187-94.
15. Santhanam T, Ephzibah EP. Heart disease prediction using hybrid genetic fuzzy model. Indian Journal of Science and Technology. May 2015; 8(9):797-803.
16. Kalaiselvi C, Nasira GM. Prediction of heart diseases and cancer in diabetic patients using data mining techniques. Indian Journal of Science and Technology. July 2015; 8(14).
17. Purusothaman G, Krishnakumari P. A survey of data mining techniques on risk prediction: Heart Disease. Indian Journal of Science and Technology. June 2015; 8(12):1-5.