International Journal of Computing Academic Research (IJCAR)
ISSN 2305-9184, Volume 5, Number 1 (February 2016), pp.46-55

http://www.meacse.org/ijcar

Proficiency Comparison of Random Forest and J48 Classifiers for Heart Disease Prediction

Lakshmi Devasena C

Department of Operations & IT, ISB Hyderabad, IFHE University, Hyderabad, India

Abstract

Healthcare industries collect capacious clinical data which can be used to diagnose the patients in intelligent way. Advancements in Information & Technology played a energetic role in storing and processing such huge collection of clinical data. Decision support systems in Medical field are intended to support doctors in their diagnosis. This offers effort to widen knowledge and understanding of frequent specialists and facilitates the diagnosis process, using patients' data from clinical databases. These systems help in foreseeing the most serious cardio vascular diseases like 'Heart Attack'. The term Heart disease covers the various diseases that affect the heart. The exposure of heart disease from various symptom or factors is an issue which is not free from false presumptions often accompanied by unpredictable effects. Identification of heart disease is a momentous and tedious task in medicine. It is requisite to find the best fit classification algorithm that has superior accuracy on classification in heart disease prediction. This research work compares the efficiency of Random forest and J48 classifiers for prediction of heart diseases using different measures.

Keywords: Heart Disease Prediction, J48 Classifier, Proficiency Comparison, Random Forest Classifier

Introduction

Medical data mining is an investigating field of data mining, where diverse data mining and classification techniques are used to foresee the diseases based on the available clinical data. Health care industries accumulate huge amount of data of patients which can be used for this drive. Even the severe diseases like 'Heart Attack' have some common symptoms which are used to foresee the disease. Based on the past obtainable data if a classification model could be prepared, and then it is easy for the medical practitioner to forecast the disease using basic clinical data and starts the treatment without waiting for other medical modality results. Medical data mining approach used in Medical decision support systems to support diagnosing process. Classification algorithms play a significant role for this purpose. The accuracy of the classification will be based on the exact and sufficient training data available. Varieties of classification algorithms are available and Computer Science and Engineering Researchers have an opportunity to study the algorithms and propose the best performing algorithm. This research work investigates and compares the performance of Random forest and J48 classifiers in predicting Heart Disease.

Literature Review

There are significant works available in literature to predict heart disease. They normally compared the neural network based and nearest neighborhood algorithms for heart disease prediction. In [1], a neuro-fuzzy integrated approach of two levels is implemented to predict the Heart Disease. In in [2] and [11], a combined technique of C4.5, K-means and Maximal Frequent Item set Algorithm is used to extract and predict Heart Disease is presented. A combined approach of Feature Subset Selection with Principal component Analysis and Artificial Neural Network is used to predict Heart Disease in [3]. SPAM algorithm using Nearest Neighbor Classifier is proposed to predict Heart Diseases in [4]. Extrapolation of Heart Disease which

utilizes Genetic Algorithm for grouping of Optimal Reduced Set of Attributes and then Decision Tree and Naive Bayes classifiers is described in [5]. Ability Comparison of C5.0 and the C4.5 decision tree algorithms is described in [6] and how the rules can be used in evidence based medicine is explained. Heart attack forecast using Association Rule Mining using clusters with sequence number is presented in [7] and [8]. Literature survey on Heart Disease prediction is condensed in [9], [13], [16] and [27]. Evaluation of SMO, Multilayer Perceptron and Logistic Function on Heart Disease prediction is elaborated in [10]. In [12], Heart Disease prediction using K- Nearest Neighbors is presented. The advantages, uses and possibilities of Data Mining in Health care to predict diseases is detailed in [14] and [22]. In [15], an adaptive Neuro-Fuzzy Inference system with Hybrid Learning algorithm for Heart Disease prediction is described. Heart Disease prediction and classification using Artificial Neural Network with Multilayer Perceptron which uses Back Propagation algorithm is detailed in [17] and [30]. In [18], Heart disease prediction using Classification and Regression Tree Model is explained and the results are compared with existing research papers. In [19], Heart Disease estimation using Cascaded Neural Network Classifier is described and the same is compared with the ability of Support vector machine algorithm. Data Mining Techniques like Naive Bayes, Neural Networks and Decision tree for prediction of Heart Diseases in advance is explained in [20] and [28] and the same using ID3, CART and Decision Tree classifiers are proposed in [23]. In [21], Nine Voting Equal Frequency Discretization Gain Ratio Decision Tree is described for Heart disease prediction and it is compared with Bagging algorithm and Decision Tree classifier. In [24], proficiency analysis of Support Vector Machine, Neural Network and K-Means Clustering are explained. Web-based application named Decision Support in Heart Disease Prediction System is detailed using data mining technique [25]. Relative study of Decision Table, Naive Bayes and J48 algorithms for heart disease prediction is given in [26]. In [29], Heart disease prediction using Decision Tree with K-Means, Naive Bayes, and Weighted Associative Classifier with Apriori Algorithm is presented. Performance of Naive Bayes classifier and Support Vector machine are compared in [31]. Prediction of Heart Disease using Naive Bayes and Jelinek-mercer smoothing is explained in [32]. Proficiency Comparison of Memory Based Classifiers for Heart Disease Prediction is done in [33]. Expertise Comparison of RIDOR, ZeroR and PART Classifiers for Intelligent Heart Disease Prediction is carried out in [34]. This work investigates the performance comparison of Random forest and J48 classifiers for prediction Heart Disease.

Dataset Used

This work uses the Statlog Heart Disease database from UCI machine learning repository [35] with a total of 270 instances which has 13 medical attributes. It contains 150 patient details without heart disease and 120 patient details with heart disease. The diagnosis class value“1” is used to designate the absence of heart disease and value “2” is used to designate the presence of heart disease. The attributes are used here are: age, sex, chol, cp, trestbps, fbs, restecg, oldpeak, slope, thalach, exang, ca, and thal.

Methodology Used

In this research work, Random Forest and J48 Classifiers are evaluated for adeptness valuation of heart disease prediction.

J48 Classifier.J48 classifier is a direct C4.5 decision tree for classification, which creates a binary tree. It is supreme beneficial decision tree approach for classification problems. This method constructs a tree to model the classification process. After the tree is erected, the algorithm is used in each tuple in the database [36].

Algorithm J48 [36]:
INPUT:
P’ //Training data
OUTPUT
JT //Decision tree
DTBUILD (*P’)

{
JT=φ;
JT= Create root node and label with splitting attribute;
JT= Add arc to root node for each split base and label;
For each arc do
P'= Database created by applying splitting predicate to P';
If stopping point reached for this path, then
JT'= create leaf node and label with appropriate class;
Else
JT'= DTBUILD(P');
JT= add JT' to arc;
}

While building a decision tree, J48 neglects the missing values i.e. the value for that item can be foreseen based on what is identified about the attribute values for the other records. The main idea here is to split the data into range based on the attribute values for that item that are identified in the training sample [36].

RandomForest Classifier. Random Forests [37] are broadly believed to be the finest "off-the-shelf" classifiers envisaging high-dimensional data. It is an assortment of tree predictors such that each tree relies on the values of a random vector sampled autonomously and distributed equally for all trees in the forest. The generalization error for forest touches to a limit as the number of trees in the forest becomes hefty. The generalization error of a forest of tree classifiers relies on the strength of the individual trees and the association between them in the forest. Different subset of training data is selected, with replacement, to train each tree. Remaining training data are used to estimate variable of importance and errors. Class assignment is made by the number of votes from all the trees and for deteriorating the average of the results is used. It is similar to bagged decision trees with barely some difference as stated below:

1. For each split point, the search is not over all 'p' variables but just over 'm-try' variables
2. No pruning necessary. Trees can be grown until each node contains just very few observations (1 or 5).

Merits of Random Forest over bagged decision trees include:

1. better prediction.
2. almost no parameter tuning necessary with Random Forest.

Performance Measures Used

Variousscales are used to scale the performance of the classifiers compared.

Classification Accuracy. Any classification method could have an error rate and it may fail to classify appropriately. Classification accuracy is deliberated as correctly classified instances divided by Total number of instances and then multiplied by 100.

Root Mean Square Error. Root mean squared error (RMSE) is used to measure dissimilarities between values predicted and actually obtained by the classifier model. It is calculated by taking the square root of the mean square error.

Mean Absolute Error. Mean absolute error (MAE) is the average of the variance between predicted and obtained value in all test cases. It is a good measure to estimate the performance.

Confusion Matrix.A confusion matrix includes information about obtained and predicted groupings done by a classification model.

Results and Discussion

Open source machine learning tool is used to investigate the performance of Random Forest and J48 Classifiers. The performance is tested out using the entire Training set as well as using different Cross Validation methods. The class is predicted by considering all 13 attributes of the dataset.

Performance of Random ForestClassifier. The overall evaluation summary of Random Forest Classifierusing entire training set and different cross validation methods is given in Table I. The performance of Random Forest Classifierin terms of Correctly Classified Instances and Classification Accuracy is depicted in Fig. 1 and Fig. 2. The confusion matrix for different test mode and the specificity

and sensitivity is given in Table II. Random Forest Classifiergives 99.63% accuracy for the training data set. Various cross validation methods are used to check its actual performance. Random Forest gives an average of 79.33% accuracy for heart disease prediction.

Table 1. Random Forest Classifier Overall Evaluation Summary

Test Mode	No. of instances	Correctly Classified Instances	Incorrectly Classified Instances	Accuracy (%)	Kappa Statistics	Mean Absolute Error	Root Mean Squared Error	Time Taken to Build Model (Sec)
Training Set	270	269	1	99.6296	0.9925	0.0844	0.1425	0.13
5 Fold CV	270	207	63	76.667	0.5215	0.2863	0.3998	0.05
10 Fold CV	270	219	51	81.111	0.614	0.2719	0.3736	0.05
15 Fold CV	270	216	54	80	0.5888	0.2556	0.37	0.05
20 Fold CV	270	213	57	78.8889	0.5664	0.2656	0.3729	0.06
50 Fold CV	270	216	54	80	0.5895	0.2733	0.3728	0.05

Table 2. Random Forest Classifier Confusion MatrixSummary

Test Mode	Absent	Present	TP	FN	FP	TN	Sensitivity	Specificity	Precision
Training Set	150	120	150	0	1	119	1	0.99166	0.99629
5 Fold CV	150	120	126	24	39	81	0.84	0.675	0.76667
10 Fold CV	150	120	130	20	31	89	0.86667	0.74166	0.81111
15 Fold CV	150	120	132	18	36	84	0.88	0.7	0.8
20 Fold CV	150	120	130	20	37	83	0.86667	0.69166	0.78889
50 Fold CV	150	120	131	19	35	85	0.87333	0.70833	0.8

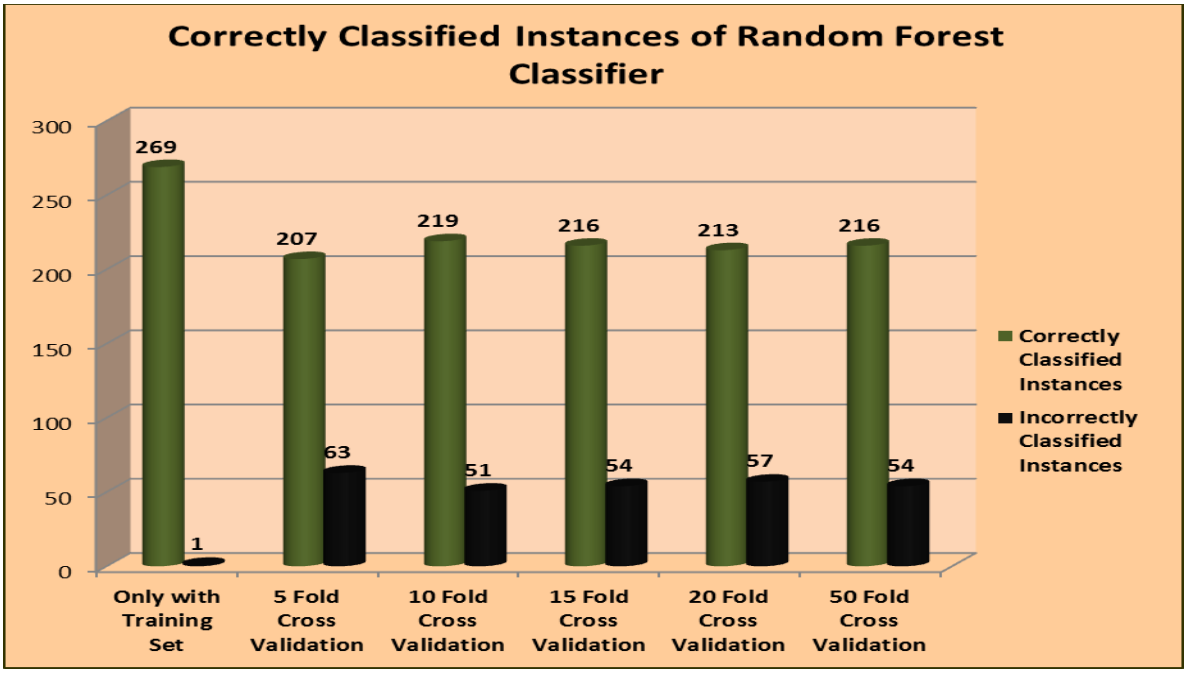

Fig1. Correctly Classified instances of Random Forest Classifier

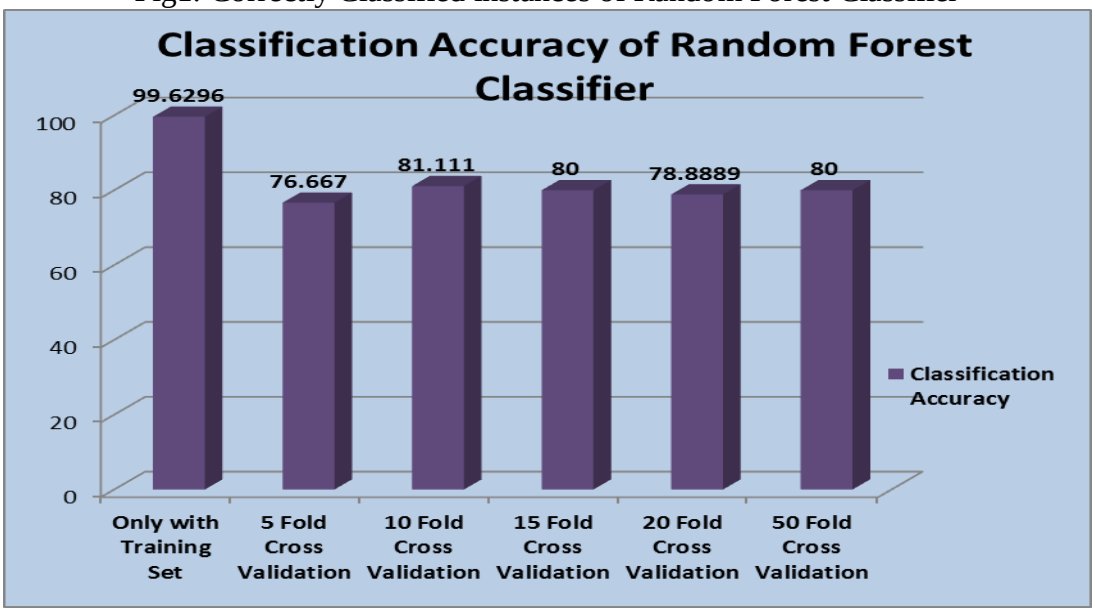

Fig 2. Classification Accuracy of Random Forest Classifier

Performance of J48 Classifier. The overall evaluation summary of J48 Classifierusing training set and different cross validation methods is given in Table III. The performance of J48 Classifierin terms of Correctly Classified Instances and Classification Accuracy is shown in Fig. 3 and Fig. 4. The confusion matrix for different test mode is given in Table. J48 Classifiergives 91.4815% accuracy for the training data set. Various cross validation methods are used to validate its actual performance. J48 gives anaverage of 77.26% accuracy for heart disease prediction.

Table 3. J48Classifier Overall Evaluation Summary

Test Mode	No. of instances	Correctly Classified Instances	Incorrectly Classified Instances	Accuracy (%)	Kappa Statistics	Mean Absolute Error	Root Mean Squared Error	Time Taken to Build Model (Sec)
Training Set	270	269	1	99.6296	0.9925	0.0844	0.1425	0.13
5 Fold CV	270	207	63	76.667	0.5215	0.2863	0.3998	0.05
10 Fold CV	270	219	51	81.111	0.614	0.2719	0.3736	0.05
15 Fold CV	270	216	54	80	0.5888	0.2556	0.37	0.05
20 Fold CV	270	213	57	78.8889	0.5664	0.2656	0.3729	0.06
50 Fold CV	270	216	54	80	0.5895	0.2733	0.3728	0.05

Table 4. Random Forest Classifier Confusion MatrixSummary

Test Mode	Absent	Present	TP	FN	FP	TN	Sensitivity	Specificity	Precision
Training Set	150	120	150	0	1	119	1	0.99166	0.99629
5 Fold CV	150	120	126	24	39	81	0.84	0.675	0.76667
10 Fold CV	150	120	130	20	31	89	0.86667	0.74166	0.81111
15 Fold CV	150	120	132	18	36	84	0.88	0.7	0.8
20 Fold CV	150	120	130	20	37	83	0.86667	0.69166	0.78889
50 Fold CV	150	120	131	19	35	85	0.87333	0.70833	0.8

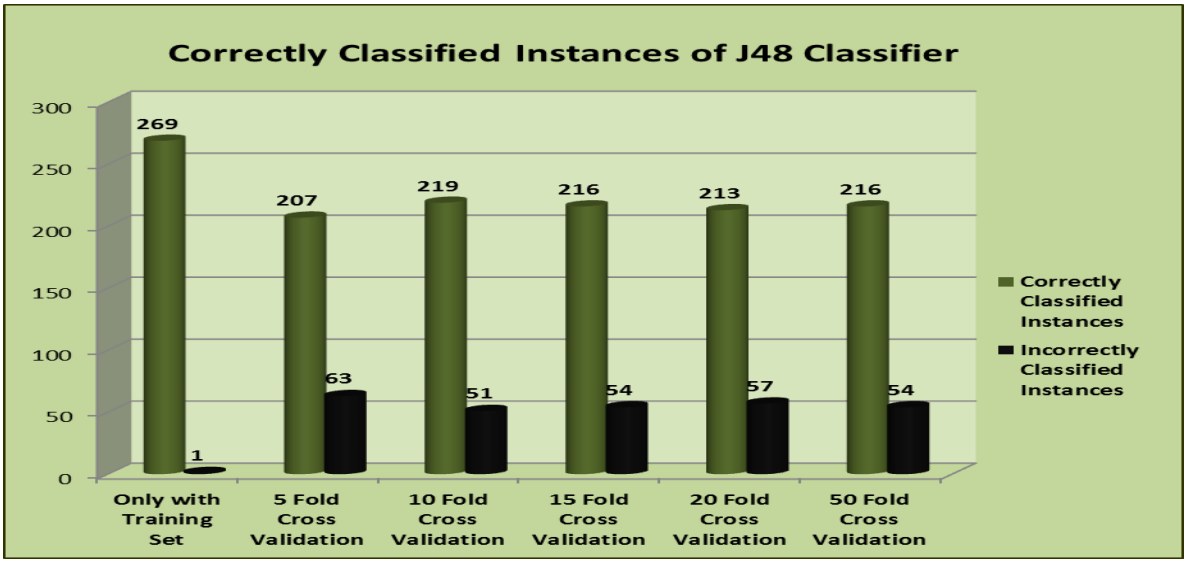

Fig 3.Correctly Classified instances of J48 Classifier

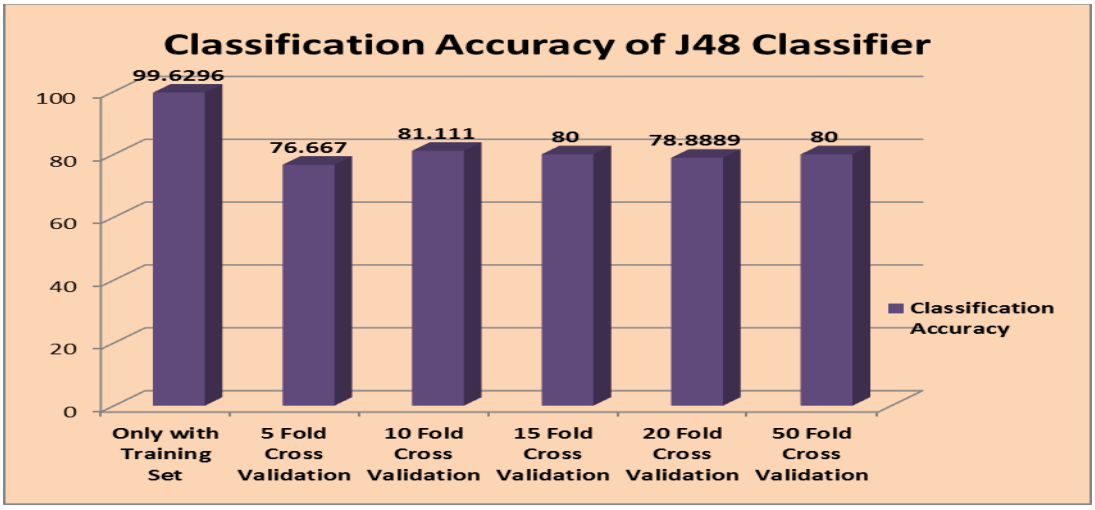

Fig 4. Classification Accuracy of J48 Classifier

Comparison of Random Forest and J48Classifiers. The comparison of performance between Random Forest and J48 Classifieris depicted in Fig 5 and Fig. 6 for heart disease predictionusing Correctly Classified Instances and Classification Accuracy. The complete ranking is done by comparing different measures like correctly classified instances, MAE, classification accuracy and RMSE values and other statistics using various testing modes. Consequently, it is perceived that Random Forest classifier outpaces the J48 Classifier.

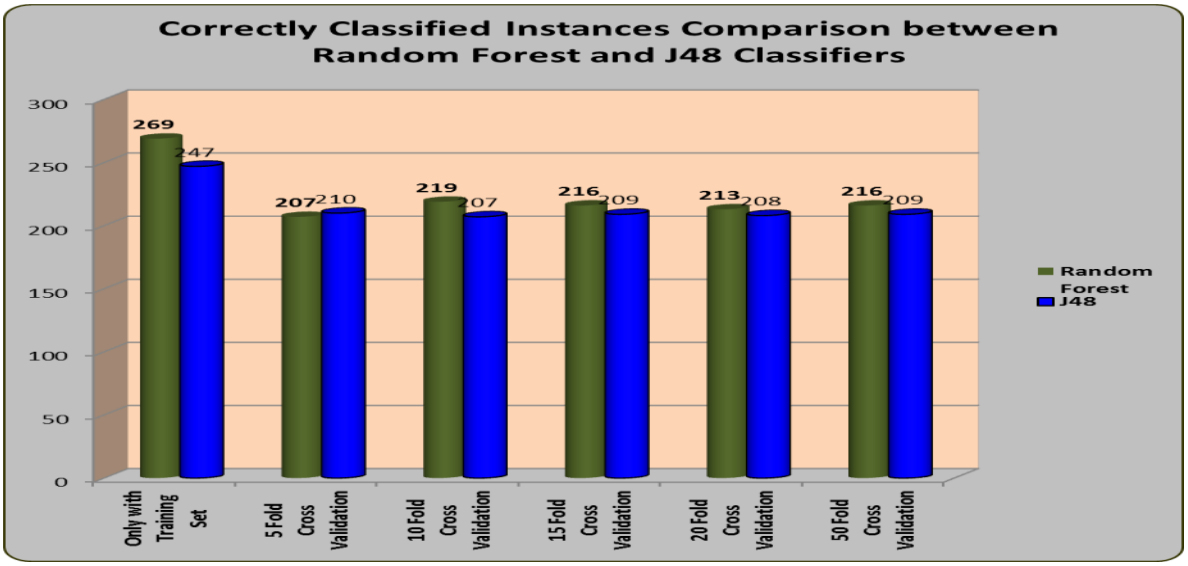

Fig 5. Comparison of Correctly Classified Instances between Random Forest and J48 Classifiers

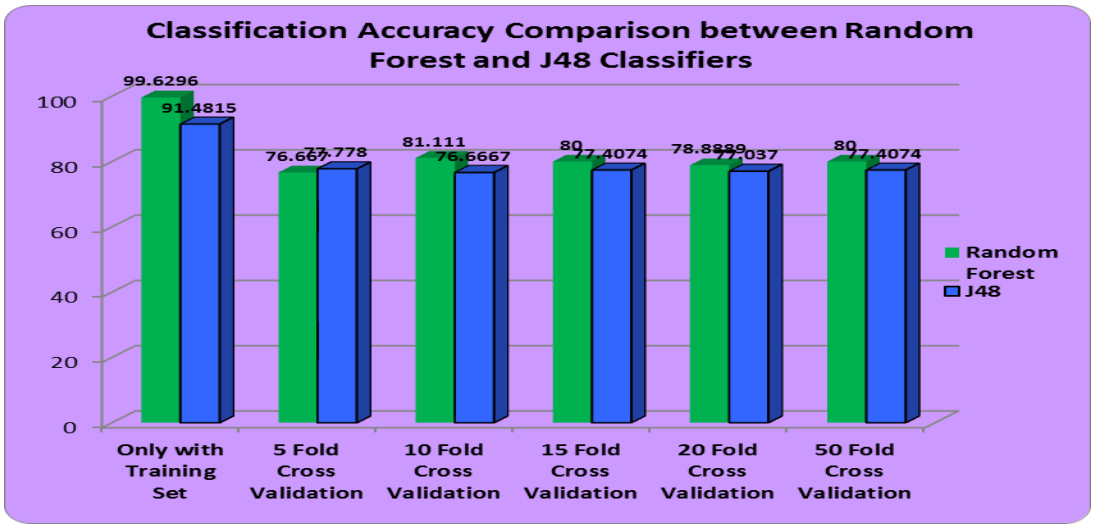

Fig 6.Comparison of Classification Accuracy between Random Forest and J48 Classifiers

Conclusion

This work investigated the efficiency of Random Forest and J48 Classifiers for heart disease prediction. Experimentation is accomplished using the open source machine learning tool. Effectiveness comparison of both the classifiers has been done using different scales of performance evaluation measures. Eventually, it is perceived that Random Forest Classifier performs better than J48 Classifier for heart disease prediction by taking measures including Classification accuracy, Errors and Time taken to build the model.

References

The author expresses her deep gratitude to the Management of IBS Hyderabad, IFHE University and Operations & IT Department of IBS Hyderabad for constant support and motivation.

References

[1] Ashish Kumar Sen, Shamsher Bahadur Patel, and D. P. Shukla, "Data Mining Technique for Prediction of Coronary Heart Disease Using Neuro-Fuzzy Integrated Approach Two Level,", Int. Journal of Engg. and Comp. Science, vol. 2 Issue 9, pp. 2663-2671, Sept 2013.
[2] V. Manikantan and S. Latha, "Predicting the Analysis of Heart Disease Symptoms Using Medicinal Data Mining Methods," International Journal on Advanced Computer Theory and Engineering, vol. 2, no. 2, pp. 5-10, 2013.
[3] M. Akhil Jabbar, B.L Deekshatulu, and Priti Chandra, "Classification of Heart Disease using Artificial Neural Network and Feature Subset Selection," Global Journal of Computer Science and Technology Neural & Artificial Intelligence, vol. 13, no. 3, 2013.
[4] S. Sandhiya, P. Pavithra, A. Vidhya, S. Jegan and S. Saranya, "Novel Approach for Heart Disease verdict Using Data Mining Technique," International Journal of Modern Engineering Research, pp. 10-14.
[5] Shruti Ratnakar, K. Rajeswari, and Rose Jacob, "Prediction of Heart Disease Using Genetic Algorithm For Selection of Optimal Reduced Set of Attributes," International Journal of Advanced Computational Engineering and Networking, vol. 1, no. 2, pp. 51 – 55, April-2013.
[6] Mohammad Taha Khan, Shamimul Qamar and Laurent F. Massin, "A Prototype of Cancer/Heart Disease Prediction Model Using Data Mining," International Journal of Applied Engineering Research, vol.7 no. 11, 2012.
[7] M.Akhil jabbar, Priti Chandra, and B.L Deekshatulu, "Heart Disease Prediction System using Associative Classification and Genetic Algorithm," International Conference on Emerging Trends in Electrical, Electronics and Communication Technologies, 2012.
[8] MA. Jabbar, Priti Chandra, and B.L. Deekshatulu, "Cluster Based Association Rule Mining for Heart Attack Prediction," Journal of Theoretical and Applied Information Technology, vol. 32, no.2, pp. 196-201, October 2011.
[9] S.Vijayarani and S.Sudha, "A Study of Heart Disease Prediction in Data Mining," International Journal of Computer Science and Information Technology & Security, vol. 2, no.5, pp. 1041-1045, October 2012.
[10] S.Vijayarani and S.Sudha, "Comparative Analysis of Classification Function Techniques for Heart Disease Prediction," International Journal of Innovative Research in Computer and Communication Engineering, vol. 1, Issue 3, pp. 735-741, May 2013.
[11] G.Karthiga, C.Preethi, R.Delshi, and Howsalya Devi, "Heart Disease Analysis System Using Data Mining Techniques," International Journal of Innovative Research in Science, Engineering and Technology, vol. 3, Special Issue 3, pp. 3101 – 3105, March 2014.
[12] Mai Shouman, Tim Turner, and Rob Stocker, "Applying k-Nearest Neighbour in Diagnosing Heart Disease Patients," International Journal of Information and Education Technology, vol. 2, no. 3, pp. 220 – 223, June 2012.
[13] K.Srinivas, G.Raghavendra Rao, and A.Govardhan, "Survey on Prediction of Heart Morbidity Using Data Mining Techniques," International Journal of Data Mining & Knowledge Management Process (IJDKP), vol.1, no.3, pp. 14-34, May 2011.
[14] Boris Milovic and Milan Milovic, "Prediction and Decision Making in Health Care using Data Mining," International Journal of Public Health Science, vol. 1, no. 2, pp. 69 – 78, December 2012.

[15] Negar Ziasabounchi and ImanAskerzade, "ANFIS Based Classification Model for Heart Disease Prediction," International Journal of Electrical & Computer Sciences, vol.14, no. 2, pp. 7-12, April 2014.
[16] S.J.Gnanasoundhari, G.Visalatchi, andM.Balamurugan, "A Survey on Heart Disease Prediction System Using Data Mining Techniques," International Journal of Computer Science and Mobile Applications, vol. 2, Issue. 2, pp. 72-77, February 2014.
[17] Chaitrali S. Dangare, and Sulabha S. Apte, "A Data Mining Approach for Prediction of Heart Disease using Neural Networks," International Journal of Computer Engineering and Technology, vol. 3, Issue 3, pp. 30-40, October - December 2012.
[18] Mohammad Subhi Al-batah, "Testing the Probability of Heart Disease Using Classification and Regression Tree Model," Annual Research & Review in Biology, vol. 4, no. 11, pp. 1713-1725, 2014.
[19] R. Chitra and V. Seenivasagam, "Heart Disease Prediction System Using Supervised Learning Classifier," Bonfring International Journal of Software Engineering and Soft Computing, vol. 3, no. 1, pp. 1-7, March 2013.
[20] Vikas Chaurasia, et al, "Early Prediction of Heart Diseases Using Data Mining Techniques," Carib.j.SciTech, vol. 1, pp. 208-217, 2013.
[21] Mai Shouman, Tim Turner, and Rob Stocker, "Using Decision Tree for Diagnosing Heart Disease Patients," Proceedings of the 9-th Australasian Data Mining Conference 2011 (AusDM'11), Ballarat, Australia, pp. 23 – 29, 2011.
[22] Priti V. Wadal and S. R. Gupta, "Predictive Data Mining For Medical Diagnosis: An Overview Of Heart Disease Prediction," International Journal of Engineering Research and Applications, pp. 1 -4, April 2014.
[23] Aditya Methaila, Prince Kansal, Himanshu Arya, and Pankaj Kumar, "Early Heart Disease Prediction using Data Mining Techniques," Sundarapandian et al. (Eds) : CCSEIT, DMDB, ICBB, MoWiN, AIAP - 2014, pp. 53-59.
[24] Aqueel Ahmed and Shaikh Abdul Hannan, "Data Mining Techniques to Find Out Heart Diseases: An Overview," International Journal of Innovative Technology and Exploring Engineering, vol. 1, Issue 4, pp. 18-23, September 2012.
[25] D. Raghu, T. Srikanth, and Ch. Raja Jacub, "Probability based Heart Disease Prediction using Data Mining Techniques," IJCST, vol. 2, Issue 4, pp. 66-68, October - December 2011.
[26] Hari Ganesh S and Gajenthiran M, "Comparative study of Data Mining Approaches for prediction Heart Diseases," International organization of Scientific Research IOSR Journal of Engineering, vol. 04, Issue 07, pp. 36-39, July 2014.
[27] Hariganesh S and Gajenthiran M, "A Survey: Data Mining Approaches for Prediction of Heart Disease," International Journal of Engineering Science Invention, vol. 3, Issue 4, pp. 44-46, April 2014.
[28] K. Thenmozhi, P. Deepika, and M.Meiyappasamy, "Different Data Mining Techniques Involved in Heart Disease Prediction: A Survey," International Journal of Scientific Research, vol. 3, Issue 9, pp. 67-68, September 2014.
[29] Aswathy Wilson, Jismi Simon, Liya Thomas, and Soniya Joseph, "Data Mining Techniques For Heart Disease Prediction," International Journal of Advances in Computer Science and Technology, vol. 3, no.2, pp. 113- 116, February 2014.
[30] Manjusha B. Wadhonkar, P. A. Tijare, and S. N. Sawalkar, "Classification of Heart Disease Dataset using Multilayer Feed forward back propagation Algorithm," International Journal Application or Innovation in Engineering & Management, vol. 2, Issue 4, pp. 213-220, April 2013.
[31] G. Parthiban and S.K.Srivatsa, "Applying Machine Learning Methods in Diagnosing Heart Disease for Diabetic Patients," International Journal of Applied Information Systems, vol. 3, no.7, pp. 25-30, August 2012.
[32] Rupali R.Patil, "Heart Disease Prediction System using Naive Bayes and Jelinek-mercer smoothing," International Journal of Advanced Research in Computer and Communication Engineering, vol. 3, Issue 5, May 2014.
[33] Lakshmi Devasena, C. 2015. Comparative Analysis of Memory Based Classifiers for Intelligent Heart Disease Prediction. International Journal of Applied Engineering Research (IJAER), Volume 10, Number 81, pp. 109-113.

[34] Lakshmi Devasena, C. 2014. Proficiency Comparison of ZeroR, RIDOR and PART Classifiers for Intelligent Heart Disease Prediction. International Journal of Advances in Computer Science and Technology (IJACST), Vol.3, No.11 Special Issue, Pages: 12-18.
[35] UCI Machine Learning Data Repository – http://archive.ics.uci.edu/ml/datasets.
[36] Tina R. Patil, Mrs. S. S. Sherekar, " Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification", International Journal Of Computer Science And Applications Vol. 6, No.2, Apr 2013, pp. 256 - 261.
[37] Leo Breiman (2001). Random Forests. Machine Learning. 45(1),pp.5-32.