ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 10, October 2015

Influence of Cross Validation Parameter for the Classification Algorithms by using Yeast Dataset

S. Kalaivani[1], S. Gandhimathi[2]

Department of Computer Science, PGP Arts and Science College, Namakkal[1]

Head of the Department, PGP Arts and Science College, Namakkal[2]

Abstract: Data mining is a process which finds useful patterns from large amount of data. Data mining is the core part of the Knowledge Discovery in Database (KDD). It is also called as knowledge discovery process, knowledge mining from data, knowledge extraction or data /pattern analysis. Data mining techniques can be classified into summarization, classification, clustering, association rules and trend analysis. Classification aims to discover a small set of rules in the database that forms an accurate classifier. There are different classification methods such as decision tree, Rule Induction, Fuzzy rule , Neural Networks etc., In this paper we are analyzing the performance of 3 classification algorithms namely J48 Decision Tree, Decision Table, RBFNetwork. We use the Yeast datasets for calculating the performance of classification algorithms by using the training set parameter. And finally a comparative analysis based on the performance factors such as the Classification and execution time is performed on all the algorithms.

Keywords: Classification, J48 Decision Tree, Decision Table, RBFNetwork, Yeast dataset.

1. INTRODUCTION

Classification is a derivation of function or model which determines the class of an object based on its attributes. A set of object is given as the training set in which ever object is represented by a vector of attributes along with its class. A classification function or model is derived based on the analysis of the set of training data. Such a classification model is used to predict class label of objects for which class is unknown. [1].

In this paper comparison is made to find out which test option is the best for classification algorithms. We use the training set parameter to calculate the data set values. This paper uses the yeast dataset for comparison of those algorithms. And our paper is structured as follows. Section 2 describes the literature review, Section 3 describes the methodology for the Yeast dataset and Section 4 describes our experimental result. And finally Section 5 gives the Conclusion and Future work.

2. LITERATURE REVIEW

B. Kavitha, et al., presented the classification methods such as ID3, J48, Naive Bayes and One R. Their result shows that ID3 and J48 method carry the highest accuracy and sensitivity with 7 and 14 attributes. The Naive Bayes holds the highest degree of specification for all three dimensionalities [2].

Tina R. Patil, et al., compared the Performance Analysis of Naive Bayes and J48. Classification Algorithm for Data Classification .The results on the dataset shows that the efficiency and accuracy of J48 is better than that of Naïve bayes [3].

C.K. Chan, et al., compared numerically to the conventional preprocessing approaches such as data elimination, averaging, imputation to treat missing values. The efficiencies were confirmed by the classification accuracies through BayesNet, Lazy Kstar, Decision table and Part method classifiers [4].

Hongjun Lu,et al., build an efficient scalable classifiers in the form of decision tables by exploring capabilities of modern relational database management systems. They implemented the unique features of the approach include its high training speed, linear scalability and simplicity [5].

Hyontai Sug, et al., suggests a better sampling technique based on branching information of decision tree for radial basis function networks when target data set is very large like census data. Their experiment with census income data in UCI machine learning repository shows a promising result [6].

Kenneth J McGarry, et al., compared the Knowledge Extraction from Radial Basis Function Networks and Multilayer Perceptrons RBF networks are localist types of learning technique Local learning systems generally contain elements that are responsive to only a limited section of the input space [7].

3. METHODOLOGY

Using the trees classification we find the best algorithm for the Yeast dataset. The flow diagram for the comparative analysis is shown in Fig 1.

A. Data Set

The yeast dataset has been collected from the keel repository. The data mining tool weka is used for analyzing the performance of classification algorithms.

B. Classification

In the Data mining, the classification technique can be used to predict group membership for data instances. The classification is similar to the clustering technique, and in that it also sectors the customer records into distinct sector called classes. In order to predict the outcome of the datasets, the algorithm processes a training set containing a set of attributes and the respective outcome, usually called target or prediction attribute. In this paper we have

DOI 10.17148/IJARCCE.2015.41053

analysis the classification algorithms to predict which algorithm is not suitable for the yeast dataset. In the classification we compare three algorithms for J48 Decision Tree, Decision table, RBFNetwork find out which one fits the effectively for the yeast dataset.

The classification algorithms are listed below.
1. J48
2. Decision Table
3. RBFNetwork

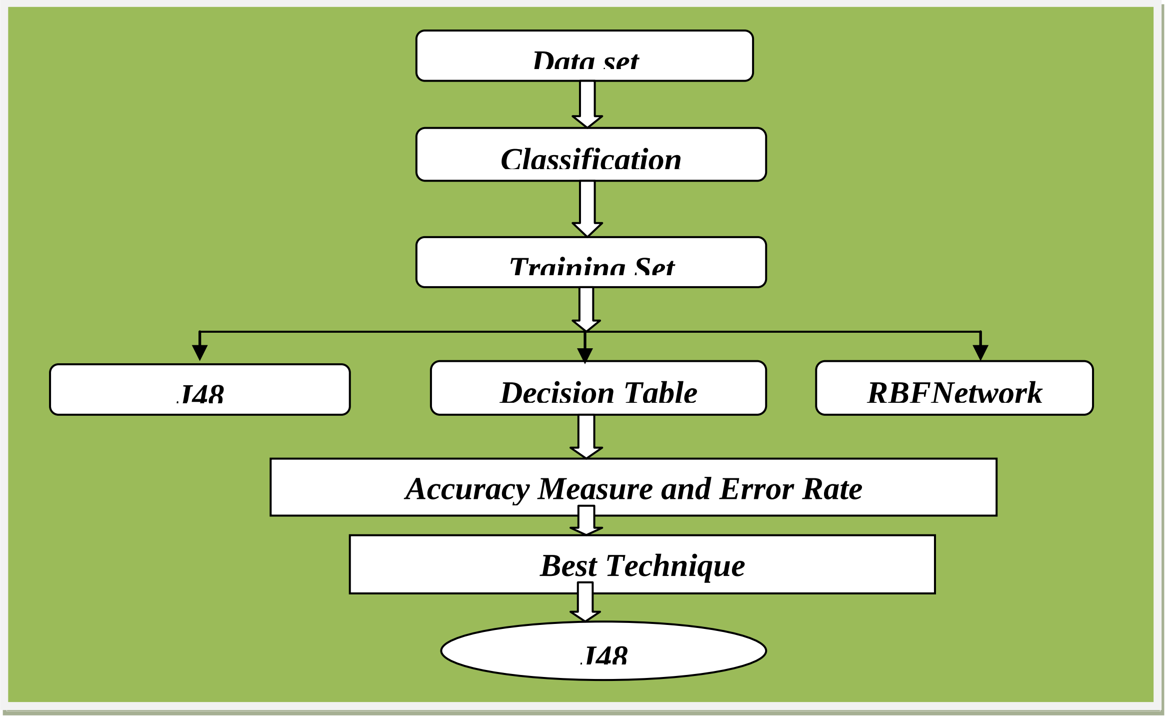

Figure 1. Comparative analysis

1.) J48
Decision Tree Algorithm is to find out the way the attributes-vector behaves for a number of instances. Also on the bases of the training instances the classes for the newly generated instances are being found. J48 is an extension of ID3. The additional features of J48 are accounting for missing values, decision trees pruning, continuous attribute value ranges, derivation of rules, etc. In other algorithms the classification is performed recursively till every single leaf is pure, that is the classification of the data should be as perfect as possible [8].

2.) Decision Table
Decision Table algorithm classifier summarizes the dataset with a decision table' which contains the same number of attributes as the original dataset. Then, a new data item is assigned a category by finding the line in the decision table that matches the non-class values of the data item. By eliminating attributes that contribute little or nothing to a test model of the dataset, the algorithm reduces the likelihood of over-fitting and creates a smaller and condensed decision table [9].

3.) RBFNetwork
A Radial Basis Function neural network has an input layer, a hidden layer and an output layer. The neurons in the secreted layer contain Gaussian transfer functions whose outputs are inversely proportional to the expanse from the center of the neuron. RBF networks are similar to K-Means clustering and PNN/GRNN networks.

The main difference is that PNN/GRNN networks have one neuron for each point in the training file, whereas RBF networks have a variable number of neurons that is usually much less than the number of training points. For problems with small to medium size training sets, PNN/GRNN networks are usually more accurate than RBF networks, but PNN/GRNN networks are impractical for large training sets. Although the implementation is very different, RBF neural networks are conceptually similar to Nearest Neighbor (k-NN) models. The basic idea is that a predicted target value of an item is likely to be about the same as other items that have close values of the predictor variables.

4. EXPERIMENTAL MEASURE

In this paper we calculate the experimental measures by using the performance factors such as the classification accuracy and execution time. And also we find out the accuracy measure and error rate to determine the best algorithm for the Yeast dataset. The performance factor for the classifiers is shown in Table 1. And the accuracy measure for classification algorithms is shown in Table 2.
From the experimental results, it is inferred that for the training set parameter, the J48 algorithm provides better Precision, TP rate, F-measure, Kappa and the ROC values for the Yeast dataset.
And also the J48 algorithm provides low false predictive rates than the other algorithms.

DOI 10.17148/IJARCCE.2015.41053

TABLE 1. PERFORMANCE FACTORS FOR THE CLASSIFICATION ALGORITHMS

Algorithm	TP Value	FP Value	Precision	F Measure	ROC Value	Kappa Statistics
J48	0.831	0.833	0.825	0.825	0.957	0.781
Decision Table	0.375	0.236	0.152	0.215	0.627	0.155
RBFNetwork	0.711	0.104	0.712	0.706	0.91	0.623

TABLE 2. ACCURACY MEASURES FOR CLASSIFICATION ALGORITHMS

Algorithm	Correctly classified Instances value (%)	Incorrectly classified Instances value (%)
J48	83.085	16.915
Decision Table	37.81	62.189
RBFNetwork	71.14	28.85

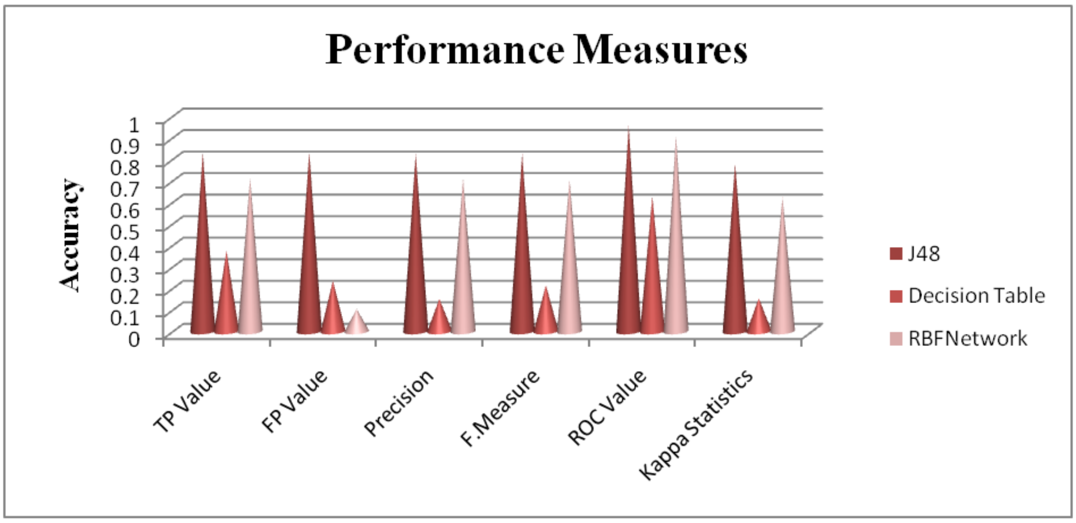

Figure 2. Performance Measures for the Classifier algorithms

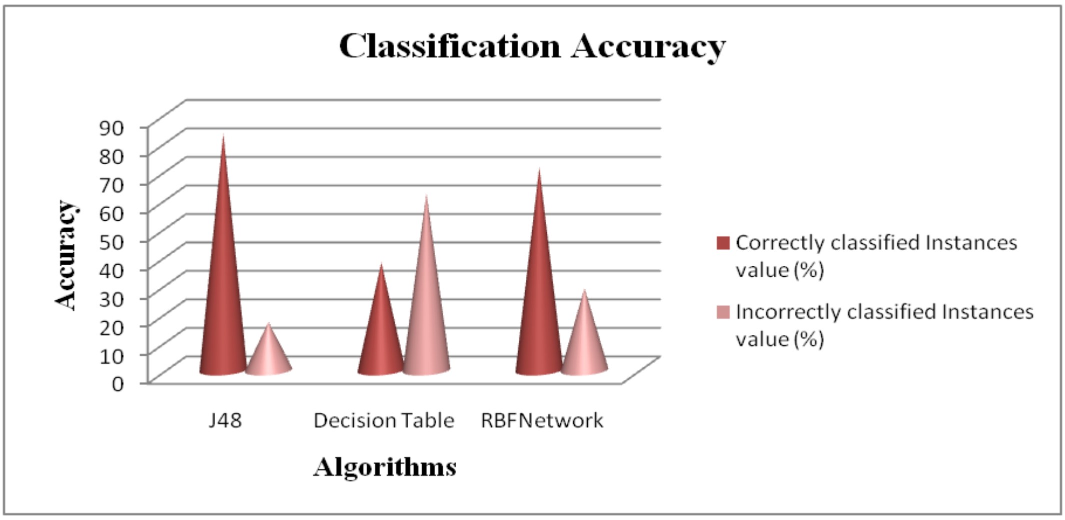

Figure 3: Accuracy Measure for the Classifier algorithms

The performance factors for the classification algorithms are shown in Fig. 2. And the accuracy measure for the classifiers is shown in Fig. 3.

For Decision Table algorithm it is inferred that for the training set parameter, the Precision, ROC, F-Measure, TP Rate and kappa values gives poor results than other algorithms. For RBF Network algorithm it is inferred that for the training set parameter, the Precision, ROC, F-Measure, TP Rate and kappa values gives poor results than J48 and provides better results than Decision Table algorithms. The Error rate measure for the classification is depicted in Table 3 and 4. And also Accuracy error rate measure for the classifier is shown in the Fig. 4 and Fig.5. The experiment was carried out to the yeast datasets by using the cross validation parameter. From the results it is inferred that the J48 algorithm performs well as compare to the Decision Table and RBFNetwork. The J48 algorithm gives more correctly classified instances compare to others.

DOI 10.17148/IJARCCE.2015.41053

TABLE 3. ERROR RATE MEASURE FOR CLASSIFICATION ALGORITHM

Algorithm	Mean Absolute Error	Root Mean Squared Error
J48	0.578	0.169
Decision Table	0.161	0.283
RBFNetwork	0.084	0.208

TABLE 4. ERROR RATE MEASURE FOR CLASSIFICATION ALGORITHMS

Algorithm	Relative Absolute Error	Root Relative Squared Error
J48	33.316	57.89
Decision Table	92.88	95.822
RBFNetwork	48.728	70.818

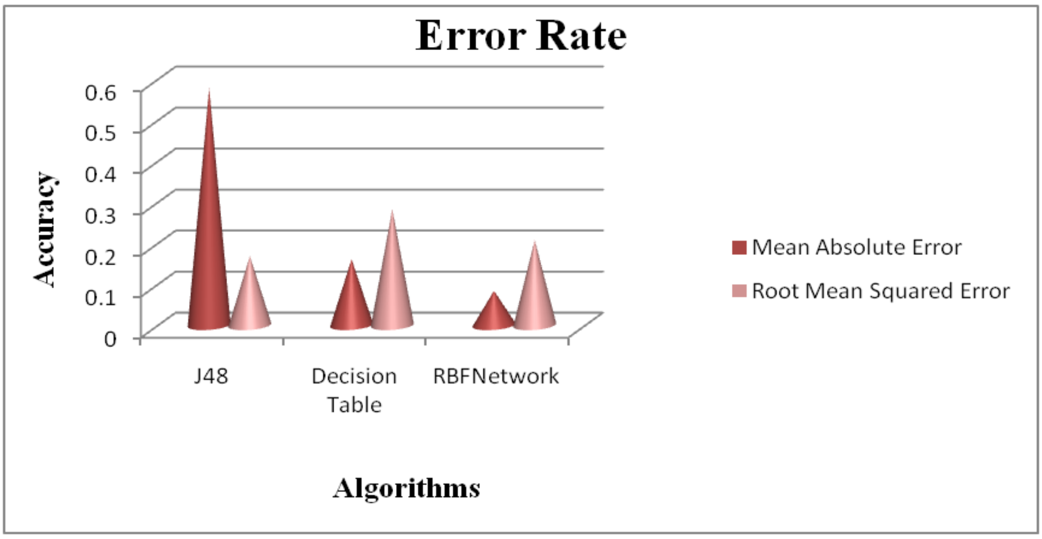

Figure 4. Accuracy error rate measure for classification algorithms

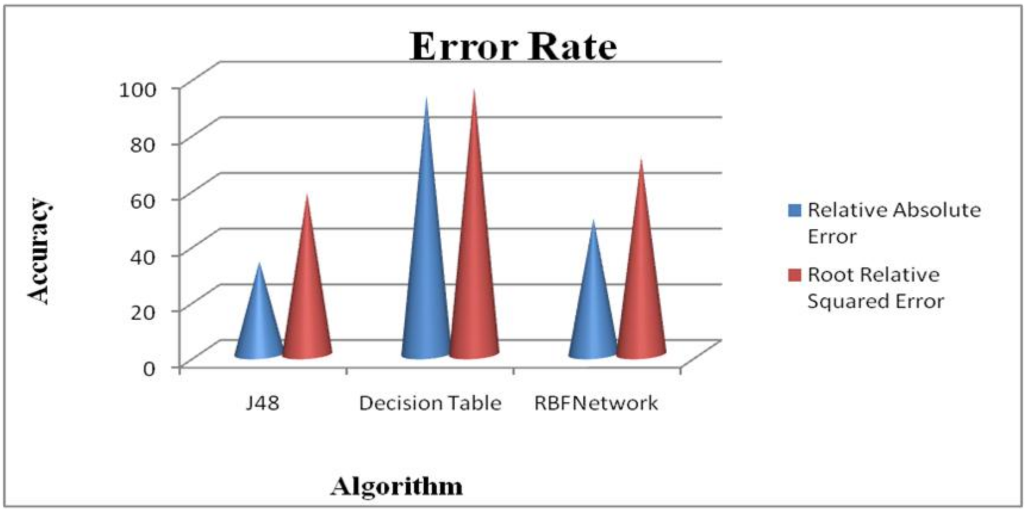

Figure 5. Accuracy error rate measure for classification algorithm

V. CONCLUSION AND FUTURE WORK

In this paper we have analyzed the performance of 3 classifier algorithms namely J48 Decision Tree, Decision Table and RBFNetwork. We used the yeast datasets for calculating the performance by using the training set parameter. And finally we have analyzed the algorithms by using the performance factors such as the classification accuracy and the performance factors.

From the results, it is observed that the J48 algorithm provides better results than the other algorithm.

In Future these classifications can be experimented on other datasets also. And in future we can modify the J48 algorithm to obtain more effective results. And also the classification algorithms can be analyzed using different parameters such as the training set, percentage split, and supplied test set.

DOI 10.17148/IJARCCE.2015.41053

REFERENCES

[1]. Jiawei Han, Micheline Kamber, Jian Pei," Data Mining: Concepts and Techniques: Concepts and Techniques", ISBN 978-0-12-381479-1, 3rd edition.

[2]. B. Kavitha, S. Karthikeyan, B. Chitra, "Efficient Intrusion Detection with Reduced Dimension Using Data Mining Classification Methods and Their Performance Comparison", Information Processing and Management Communications in Computer and Information Science Volume 70, 2010, pp 96-101.

[3]. 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 ISSN: 0974-1011.

[4]. C.K. Chan, W.P. Loh, I. Abd Rahim, "Data Elimination cum Interpolation for Imputation: A Robust Preprocessing Concept for Human Motion Data", Procedia - Social and Behavioral Sciences, Volume 91, 10 October 2013, Pages 140–149, PSU-USM International Conference on Humanities and Social Sciences, http://dx.doi.org/10.1016/j.sbspro.2013.08.411.

[5]. Hongjun Lu and Hongyan Liu," Decision Tables: Scalable Classification Exploring RDBMS Capabilities",Proceedings of the 26th International Conference on Very Large Databases, Cairo, Egypt, 2000.

[6]. Hyontai Sug," Generating Better Radial Basis Function Network for Large Data Set of Census", International Journal of Software Engineering and Its Applications Vol. 4, No. 2, April 2010.

[7].Kenneth J.McGarry,Stefan wermter and John MacIntyre," Knowledge Extraction from Radial Basis Function Networks and Multi_layer Perceptrons", Neural Networks, 1999. IJCNN '99. International Joint Conference on (Volume:4) ISSN :1098-7576 DOI: 10.1109/IJCNN.1999.833464 Publisher:IEEE

[8]. Gaganjot Kaur and Amit Chhabra ," Improved J48 Classification Algorithm for the Prediction of Diabetes", International Journal of Computer Applications (0975 – 8887) Volume 98 – No.22, July 2014.

[9]. Geetali Banerji and Kanak Saxena," An Efficient Classification Algorithm for Real Estate domain ",International Journal of Modern Engineering Research (IJMER) Vol.2, Issue.4, July-Aug. 2012 pp-2424-2430 ISSN: 2249-6645.

[10]. Sai Satyanarayana Reddy,P.Ashok Reddy and V.Krishna Reddy," a perspective of data mining method based on drbf neural networks", Journal of Theoretical and Applied Information Technology © 2005 - 2010 JATIT & LLS. All rights reserved.

DOI 10.17148/IJARCCE.2015.41053