

Random Forest Algorithm Is a Best Classifier Measure Based On Accuracy and Error Rate Using Weka Tool

^[1]R.SatyaRajendra Singh, ^[2]K.Mahesh Kumar, ^[3]G.Arun,
^[1] Head of CSE ^[2] Asst.Professor, Dept. of CSE, ^[3] Student, CSE Dept.
^{[1][2][3]} Sri Venkateswara College of Engineering

Abstract—Random Forest algorithm is an one of the most essemble method for Classification analysis. Classification is a supervised learning approach, which maps a data items into predefined classes. There are various classification algorithms proposed in the literature. In this paper authors have used four classification algorithms such as J48, Random Forest (RF), Reduce Error Pruning (REP) and Logistic Model Tree (LMT) to classify the “CAR SALES NOMINAL” open source Data Set. Waikato Environment for Knowledge Analysis (WEKA) has been used in this paper for the experimental result and they found that Random Forest algorithm classify the given data set, it shows better results than the other algorithms for this specific data set. In this paper, the performance of classifier algorithms is evaluated for 5 fold cross validation test. finally ,Random Forest Algorithm have been proved as a benchmark classification method for getting maximum accuracy value and minimum error rate value.

Keywords—Decision Tree, J48, Random Forest, REP, LMT, Classification, Cross-Validation, Supervised Learning and Performance Measure.

I. INTRODUCTION

Classification is a tree based structure which is a concept of data mining (machine learning) technique. It used to predict data instances through attributes. Classification is a method where one can classify future data into known classes. In general this approach uses a training data set to build a model and test data set to validate it. Popular classification techniques include decision trees, Naïve Bayes, Logistic regression, etc. The accuracy of the supervised classification will be much better than unsupervised classification, but depends on prior knowledge. J48 tree algorithm basically uses the divide-and-conquer algorithm by splitting a root tree into a subset of two partitions of child nodes [13]. Random Forest is a machine learning classifier that works over many iterations of the same technique but with a different approach [12], [8]. Reduced Error Pruning performed as well as most of the other pruning methods in terms of accuracy and better than most in terms of tree size [11].

It is very difficult to select any prediction techniques in practical situation, because prediction depends on many factors like nature of problem, nature of data set, uncertain availability of data. Machine learning algorithms are most significant classifiers to solve a variety of problems in software development and mainly in software fault prediction. Prediction of faulty and non-faulty modules have been done by so many researchers and organizations involved in it but still there is a blackness of

best techniques that always outperforms other methods overall. Hence more reasonable research is required for the assessment of more result. This data set “CAR SALES NOMINAL” has been analyzed and statistical calculation have been done on it to clarify the faulty and non-faulty module. The main focus of this paper is to compare the performance of classification in various machine learning algorithms using open source data set. The solution for this problem may be provided by calculating confusion matrix, accuracy and error rate in taken data set.

Classification is a technique which organizes data of a given class. The proposed model architecture is shown in Figure 1. It describes the applied classification algorithm, which finally classifies the faulty and non-faulty module.

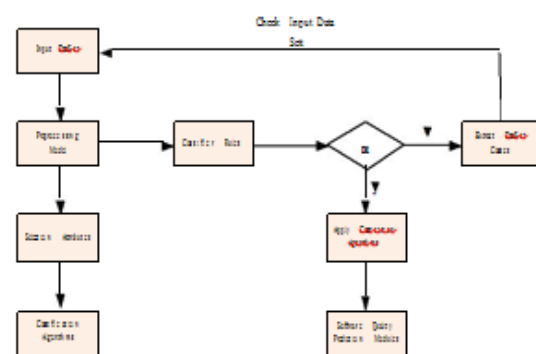


Fig 1: Proposed Model Architecture

Classification may be used in classifying cancer cells as working or damaged, classifying any card transactions as authorized or unauthorized, classifying food items as vitamins, minerals, proteins or carbohydrates, classification of news into sports, weather, stocks etc.

The rest of the paper is organized as follow. In section 2 authors have described the basic classification learning algorithms. Section 3 describes the performance measures for classification. Section 4 explains the experimental result and analysis. Conclusion and future work is shown in section 5.

II. CLASSIFICATION LEARNING ALGORITHMS

Classification techniques can be compared on the basis of predictive accuracy, speed, robustness, scalability and interpretability criteria [10]. In data mining classification tree is a supervised learning algorithm. So one can prepare popular classifiers: J48, Random Forest, Reduce Error Pruning, and Logistic Model Tree. For comparison purpose, authors have also prepared the fault-prone filtering techniques. A classification model is able to identify the fault-prone (fp) module correctly. The algorithm C5.0 is superior to C4.5. J48 is the enhanced version of C4.5 but the working of both algorithms are very similar. The goal of decision tree is to predict to response on a categorical dependent variable to measure a more predictor. The CAR SALES NOMINAL uses a 5 attributes and 14 instances as shown given a Table.1.

Table 1. The data set used in our analysis list

Car Sales Nominal Data Set	
Attributes	5
Instances	14
Sum of Weight	14

2.1 Decision Trees

A decision tree is a flow-chart-like tree structure. The internal node denotes a test on an attribute, each branch represents an outcome of the test, and leaf nodes represent classes or class distribution [5][10]. The top most node in a tree shown by oval is a root node. Further internal nodes are

represented by rectangles, and leaf nodes are denoted by circles which are depicted in figure 3.

2.1.1 J48 Algorithm

J48 is a tree based learning approach. It is developed by Ross Quinlan which is based on iterative dichotomiser (ID3) algorithm [13]. J48 uses divide-and-conquer algorithm to split a root node into a subset of two partitions till leaf node (target node) occur in tree. Given a set T of total instances the following steps are used to construct the tree structure.

Step 1: If all the instances in T belong to the same group class or T is having fewer instances, than the tree is leaf labeled with the most frequent class in T.

Step 2: If step 1 does not occur then select a test based on a single attribute with at least two or greater possible outcomes. Then consider this test as a root node of the tree with one branch of each outcome of the test, partition T into corresponding T1, T2, T3....., according to the result for each respective cases, and the same may be applied in recursive way to each sub node.

Step 3: Information gain and default gain ratio are ranked using two heuristic criteria by algorithm J48.

2.1.2 Random Forest Algorithm

Random Forest algorithm was initially developed by Leo Breiman, a statistician at the University of California [12] Berkeley. Random Forests is a method by which one can calculate accuracy rate in better way. Some attributes of Random Forest is mentioned below [7].

- a) It efficiently works on large data sets (training data sets).
- b) It provides consistent accuracy than the other algorithms.
- c) By this method estimate missing data if any and retains the accuracy rate even if the bulk of the data is missing.
- d) It also provides an estimate of important attributes in the classification.
- e) The strength of RF is mentioned below [4][6] [8].

RF may produce a highly accurate classifier for more data sets.

RF has much simplicity.

RF provides a fast learning approach.

2.1.3 Reduce Error Prune

This method introduced by Quinlan [3]. It is the simplest and most understandable method in decision tree pruning. For every non-leaf sub tree of the original decision tree, the change in misclassification over the test set is examined. The REP incremental pruning developed by Written and Frank in 1999 is a fast regression tree learner that uses information variance reduction in the data set which is splitted into a training set and a prune set.

When any one traverse the tree from bottom to top then he she may apply the procedure which checks for each internal node and replace it with most frequently class, keeping in mind about tree accuracy, which must not reduced. Now the node is pruned. This procedure will continue until any further pruning would decrease the accuracy.

2.1.4 Logistic Model Tree

Logistic Model Tree (LMT) [2] algorithm makes a tree with binary and multiclass target variables, numeric and missing values. So this technique uses logistic regression tree. LMT produces a single outcome in the form of tree containing binary splits on numeric attributes.

2.2 Cross-Validation Test

Cross-validation (CV) method used in order to validate the predicted model. CV test basically divide the training data into a number of partitions or folds. The classifier is evaluated by accuracy on one phase after learned from other one. This process is repeated until all partitions have been used for evaluation [1]. The most common types are 10-fold, n-fold and bootstrap result obtained into a single estimation.

III. PERFORMANCE MEASURES FOR CLASSIFICATION

One can use following performance measures for the classification and prediction of fault prone module according to his/her own need.

3.1 Confusion Matrix

The confusion matrix is used to measure the performance of two class problem for the given data set

Table 2. The right diagonal elements TP (true positive) and TN (true negative) correctly classify Instances as well as FP (false positive) and FN (false negative) incorrectly classify Instances.

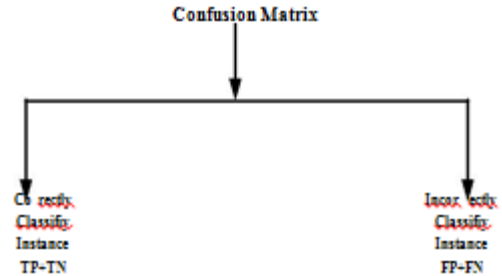


Fig 2: Confusion Matrix

Table 2. Example of Confusion matrix

		Predicted	
		Yes	No
Actual	Yes	TP	FN
	No	FP	TN

Table 3. Actual Vs Predicted Confusion matrix

		Predicted	
		Faulty - Prone	Non-Faulty Prone
Actual	Faulty	True Positive (TP)	False Negative (FN)
	Non-faulty	False Positive (FP)	True Negative (TN)

Total number of instances = Correctly classified instance + Incorrectly classified instance

Correctly classified instance = TP + TN

Incorrectly classified instance = FP + FN

3.2 Cost Matrix

A cost matrix is similar to confusion matrix but minor difference is with finding the value of cost accuracy through misclassification error rate.

Misclassification error rate = 1 – Accuracy

3.3 Calculate Value TPR, TNR, FPR, and FNR

One can calculate the value of true positive rate, true negative rate, false positive rate and false negative rate by methods shown below.

$$\begin{aligned} \text{TPR} &= \frac{\text{TP}}{\text{TP} + \text{FN}} \\ \text{TNR} &= \frac{\text{TN}}{\text{FP} + \text{TN}} \\ \text{FPR} &= \frac{\text{FP}}{\text{TP} + \text{FN}} \\ \text{FNR} &= \frac{\text{FN}}{\text{TP} + \text{FN}} \end{aligned}$$

3.4 Recall

Recall is the ratio of modules correctly classified as fault-prone to the number of entire faulty modules.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

3.5 Precision

Precision is the ratio of modules correctly classified to the number of entire modules classified fault-prone. It is proportion of units correctly predicted as faulty.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

3.6 F-Measure

FM is a combination of recall and precision. It is also defined as harmonic mean of precision and recall

$$\text{F-Measure} = \frac{2 * \text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}}$$

3.7 Accuracy

It is defined as the ratio of correctly classified instances to

Total number of instances.

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

IV EXPERIMENTAL WORK AND ANALYSIS

The repository data contains 5 attributes and 14 instances respectively. WEKA [9] tool have been applied on

“CAR SALES NOMINAL” data set taking 5 fold cross validation for performance evaluation of the different algorithms. Table 4 reveals confusion matrix for mentioned four algorithms, which maps the actual and predicted values for the respective algorithms. In Table 5 authors have calculated Precision, Recall, F-measure and accuracy value and they have found that the accuracy of the RF is 57.14% which is best among J48, REP and LMT methods of classification.

Table 6 depicts instances correctly predicted vs. instance incorrectly predicted with accuracy and total execution time taken by each algorithm. The accuracy of RF is greater than other examined techniques but time taken to make model is greater than J48 and REP. One can also observe that total time taken to make a model is minimum for REP model. Here authors have also prepared error rate of each examined algorithm which is mentioned in Table 7. Hence they observed that RF is showing minimum error rate than the other techniques, and J48 algorithm is showing maximum error rate. Figure 4 depicts the accuracy of the taken classifiers as well as error rate has been shown in figure 5.

Figure 3 depicts for prediction whether the Cars will be purchased or not. Here authors have taken an open source data set named “CAR SALES NOMINAL” which contains 5 attributes as well as 14 instances. Conditions have been shown in this figure.

If income is average and financial status is normal then purchase the car otherwise not-purchase it.

If income is good then car will definitely be purchased.

If income is above average and financial status is bad according then car will be purchased otherwise car may not be purchased.

Fig 3: Decision tree for the car sales Nominal data

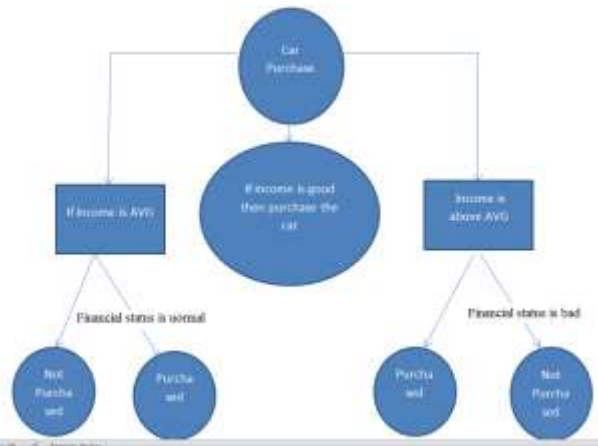


Table 4. Classifiers Confusion Matrix

Actual	Predicted							
	J48		RF		REP		LMT	
	Purch ased	Not-Purch ased	Purch ased	Not-Purch ased	Purch ased	Not-Purch ased	Purch ased	Not-Purch ased
Purch ased	4	5	5	4	7	2	6	3
Not Purch ased	3	2	2	3	5	0	4	1

Table 5. Prediction Performance Measures

Algorithms	TPR	FPR	TNR	FNR	Precision		Recall		F-Measure	Accuracy in %
					PR	NR	PR	NR		
J48	0.444	0.8	0.4	0.556	0.571	0.286	0.444	0.4	0.5	42.85
RF	0.556	0.4	0.6	0.444	0.714	0.429	0.556	0.8	0.825	57.14
REP	0.778	1	0	0.222	0.513	0	0.778	0	0.667	50.00
LMT	0.667	0.8	0.2	0.333	0.8	0.28	0.667	0.2	0.632	50.00

Table 6. Performance Measure about Confusion matrix

Classifier Algorithms	Instances Correctly Predicted	Instances Incorrectly Predicted	Accuracy in %	Total Time Taken to Build Model (in seconds)
J48	6	8	42.85	0.02
RF	8	6	57.14	0.08
REP	7	7	50.00	0.01
LMT	7	7	50.00	0.14

Accuracy of Classifier

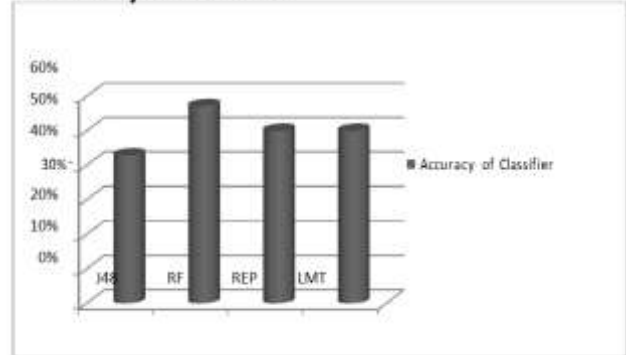


Fig 4: Accuracy of Classifiers

Table 7. Error rate of classifier for 5 fold cross validation

Error Rate in %	J48	RF	REP	LMT
	57.14	42.85	50.00	50.00



Fig 5: Error Rate

V. CONCLUSION AND FUTURE WORK

In this paper authors have examined J48, RF, REP and LMT method of classification and observed that RF is having maximum accuracy and minimum error rate. On the basis of accuracy measures, of the classifiers one can easily provide the guidelines regarding fault-prone prediction issues of any given data set in the respective situations.

Finally, we can conclude that ,we have applied as 5-fold cross validation test on a specific data set. based on a Confusion matrix ,we found out Accuracy and F-Measure values. Accuracy measures affected by using cross validation test, number of tuples and nature of the data set. but ,Random Forest Algorithm can be filled on the missing values,it can overcome the problems of different situations

and also, it regains accuracy values. we felt that ,Random Forest is very flexible and have dynamic nature for measuring accuracy .

In future, we will apply different cross validation tests like 5-fold,6-fold and 10-fold etc on a specific data set ,there may be chances to change the accuracy and error rate values using Random Forest and other classification algorithms.More similar studies on different data sets for machine learning approach is needed to confirm the above finding.

REFERENCES

1. “Multi-dimensional measures function for classifier performance”, 2nd. IEEE International conference on intelligent system, pp.508-513, 2004.
2. N.Landwehr, M. Hall, and E. Frank, “ Logistic modeltrees”. for Machine Learning.,Vol. 59(1-2),pp.161-205, 2005.
3. J.R. Quinlan, “Simplifying decision trees”, Internal Journal of Human Computer Studies,Vol.51, pp. 497-491, 1999.
4. F.Livingston, “Implementation of Breiman’s Random Forest Machine Learning algorithm,” Machine learning Journal, 2008.
5. J.R.Quinlan,”induction of Decision trees : Machine Learning”,vol.1,pp.81-106,1986.
6. G.Biau, L. Devroye, G. Lugosi, “Consisting of Random Forests and other Averaging Classifiers,” Journal of Machine Learning Research, 2000.