Data Analysis of Software Quality Prediction Based on Machine Learning Methods

Satyendra Singh
Deptt. of CSE
KNIT Sultanpur, UP, India
satyendra.cse@gmail.com

Abhay Kumar Agarwal
Deptt.of CSE
KNIT Sultanpur, UP, India
abhay.knit08@gmail.com

Chandrajeet Yadav
Deptt. Of CSE
KNIT Sultanpur , UP, India
chandrajeet86@gmail.com

Abstract: Machine Learning (ML) methods are a field of artificial intelligence system. ML is the scientific field studying how machines can learn, which one of the components of intelligence work. Predictive is a supervised learning approach which is a most important part of machine learning technique. Various machine learning methods are available in a research field. This paper describes only four types of supervise machine learning methods such as J48, Decision Stump (DS), Random Tree (RT) and Logistic Model Tree (LMT). The experimental analysis has been performed on REUSE PREDICTING data of NASA promise open source repository. The results show that the proposed approach achieved better performance in terms of several predictions quality factors for software dataset. The performances of machine learning methods are evaluated for 5 fold cross-validation.

Keywords: Classification, Cross-Validation, Prediction, Software Quality, Machine Learning Methods.

I. INTRODUCTION

A software quality prediction is an important part of software engineering for obtaining producing the best quality product. Software quality prediction is the process of improving a better performance that can identify the faulty and non- faulty modules. In faulty module [1] software suffers from the problem of high development, maintenance, estimation costs, complexity and poor quality. On other hand non-faulty modules have maximum accuracy, recall, precision, F-measure, Kappa Statistic and minimum errors percentage. Several machine learning techniques have been studies to find a relation between quality of defected and non-defected classes. The use of classification for predicting software quality was introduced by Khoshgaftaar et al. [2]. The accuracy of supervised approach is much better than unsupervised approach that depends on prior studies.

J48 decision tree classifier works on the principle of divide-and-conquer approach. In this approach rooted tree splits into two subset of child nodes [3]. The decision stump is a decision tree classifier which is a class of predictive data mining approach that frequently used in machine learning method. Random Tree classifier is very similar to decision tree classifier but it decided the attributes for k-randomly chosen the classes. Logistic model tree is linear regression model for solving classification tasks in statistics and mathematical techniques.

All techniques are implemented by Waikato Environment for knowledge Analysis (WEKA) machine learning tool [10], [11]. The classification techniques used by cross-validation is splitting training and testing dataset. The given dataset used a 5 fold cross-validation. In this dataset four part use for training dataset and remaining one for testing dataset [4].

The rest of the paper is organized as follows: Discuses the machine learning methods, the independent and dependent variables in section 2. Section 3 describe in proposed approach architecture and methods. Section 4 Description of the experimental results can be found in section 5 Discussion the observation result. Section 6 provides to the validity of this work and a conclusion in section 7. Paper references.

II. MACHINE LEARNING METHODS

This concept is used to data mining approach. It is broadly classified into two categories such as Predictive and Descriptive. Predictive is supervised learning approach, which is a classification (categorical) technique. Predictive approach is used to past data and generates conclusions for future prediction. Predictive data mining has its roots in the classical model building process of statistics and medical diagnosis.

Classification technique ensures the risk prediction of software module that depends on dependent and independent variables. The available data item classes are known, and represented by $y = f(x)$. Here y is dependent variable that can varies from $y_1, y_2, y_3, ...y_n$ and f(x) is independent variable that can varies from $f(x_1), f(x_2), f(x_3), ...f(x_n)$. In classification the results are describe in the manner of hierarchical structure. Descriptive is unsupervised learning approach that can be called as continuous classes (regression) technique. This approach used by the users in association rule, clustering and grouping of data items in data mining.

A. **Decision Trees**

Decision tree [5] is one of the simplest algorithms of classification technique that builds a hierarchical model based on non-parametric value. The non-parametric values are independent variables. It is a divide-and-conquer strategy used in Classification and Regression tree (CART). Classification tree are split into three parts; root node, left child and right child.

The various algorithms of supervised learning techniques are described below:

1. **J48 Classifier**

J48 classifier is very similar to C4.5 and C5.0 decision tree classification which is depicted by binary tree classifier. The concept of C4.5 algorithm was introduced by Quinlan [3] in 1999. J48 is easy to understand that derived from C4.5 algorithm. J48 classifier is one of the most popular and powerful decision tree classifiers. J48 classifier is a higher improved version of C4.5 and C5.0 algorithms. WEKA toolkit package has its own version known as J48. J48 is an optimized implementation of C4.5.

2. **Decision Stump**

This method is a simple for binary decision tree classifier consisting of a single node based on one attribute and two branches. All attributes used by the other trees are tested and the one giving the best classifications is chosen to use in the single node. Decision stump algorithm to constructs a decision tree with just one decision node and two classification leaves during training based on a given set of training samples. Decision stump works on both numerical and nominal attributes.

In case of binary features two schemas are identical and missing value is taken as another category [6]. In continuous features, some threshold value is selected, and the stump contains two leaves for values below and above the threshold value.

3. **Random Tree**

Random tree is a tree drawn at random from a set of possible trees. The WEKA Random Tree algorithm builds a tree considering K randomly chosen attributes at each node. Random tree [7] is a decision tree that considers K randomly chosen attributes at each node and allows class probabilities based on back fitting with no pruning.

4. **Logistic Model Tree**

Logistic Model Tree (LMT) uses the concept of logistic regression tree that makes a tree with binary and multiclass target variables, numeric and missing values. LMT produces a single outcome in the form of tree containing binary splits on numeric attributes. The performance of

LMT is evaluated on original datasets taken from the UCI repository in 1998 [8]. The results produced to LMT model are more accurate classifier than other classifier model.

III. PROPOSED APPROACH

In this section we discussed the architecture of proposed model based on supervised learning techniques. This model works on NASA promises datasets, for the prediction of best performance result. The working this model is describe in Figure.1

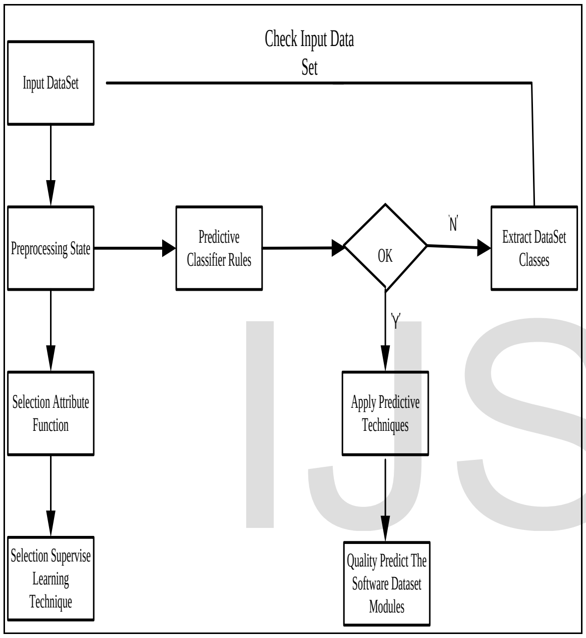
Check Input Data Set
Input DataSet
Preprocessing State
Predictive Classifier Rules
OK
N
Y
Extract DataSet Classes
Selection Attribute Function
Apply Predictive Techniques
Selection Supervise Learning Technique
Quality Predict The Software Dataset Modules

Fig: 1 Proposed Approach

1. Proposed Method:

```
1    Chose an given data set
2    Compute the No. of Instances and No. of Attributes
3    if no. of Instances is not equal to no. of Attributes
4        then find = f(x)
5    Apply Predictive classifier rule
6        if selected algorithm = OK
             then move to Step 7
           otherwise move to Step 9
7    Apply predictive techniques
8    Determined result if no
9    then extract the dataset class
10   Check the given data set Step 1
```

This proposed approach to show the how to evaluated the better performance to any software datasets. So analysis the prediction factors in terms of accuracy, speed, error rate, robustness, interoperability and some others quality factors.

2. Receiver Operating Characteristic

The receiver operating characteristic (ROC) is define as the curve plot of x and y axis's. This curve plotted as an x-axis false positive rate (FPR) or 1-specificity versus y-axis true positive rate (TPR) or sensitivity is an evaluating the performance for assessing the accuracy of predictions. This curve on x-axis and y-axis for varying cut-off points of test values.

This is generally represented in a square box for convenience and it's both axes are from 0 to 1 [9].

The area under the curve (AUC) is combined measure of sensitivity and specificity for assessing valid result. The maximum value of area under the curve (AUC) = 1 and it means cut-off point are tested for prediction.

IV. EXPERIMENTAL RESULT

In this paper, proposed model is evaluated using REUSE PREDICTING taken from NASA database system. This dataset contain a 28 attributes and 24 instances used in machine learning methods as shown table below:

Table 1. Data set

REUSE Dataset	
Attributes	28
Instances	24

The experimental work implement by WEKA (Waikato Environment for Knowledge Analysis) for 5 fold cross validation test [10], [11]. The experiment result decision tree classifier methods are tabulate form in table 2. This table to show the correctly and in-correctly classify instances, size of dataset and time taken to build a model.

Table 2: Performance Measure of REUSE Predicting dataset

ML Methods	Instances Correctly Predicted	Instances Incorrectly Predicted	Size of Data set (KB)	Total Time Taken to Build Model (in seconds)
J48	23	1	14.4	0.01
DS	22	2	14.4	0.03
RT	20	4	14.4	0.02
LMT	22	2	14.4	0.28

In table 3 to find out each values in term of TPR, FPR, TNR, FNR, Recall, F-measure and ROC as shown below.

Table 3: Prediction Performance Measures

ML Methods	TPR	FPR	TNR	FNR	ROC Curve	Precision		Recall		F-Measure	
						PR	NR	PR	NR	PR	NR
J48	1	0.111	0.889	0	0.9	0.938	1	1	0.889	0.968	0.941
DS	1	0.222	0.778	0	0. 822	0.882	1	1	0.778	0.938	0.875
RT	1	0.444	0.556	0	0.867	0.789	1	1	0.556	0.882	0.714
LMT	1	0.222	0.778	0	0. 956	0.882	1	1	0.778	0.938	0.875

In above table 4 the experiment result show the classification algorithms with respect to Accuracy, Error Rate, Root mean square error (RMSE), Mean absolute error (MAE), Kappa Statistic we identify them minimum faulty module and maximum non- faulty module.

Table 4: Performance of Machine Learning Methods

ML Methods	MAE	RMSE	Kappa Statistic	Accuracy (%)	Error Rate (%)
J48	0.0801	0.2112	0.9091	95.833	4.166
DS	0.1174	0.2936	0.814	91.666	8.333
RT	0.2041	0.3536	0.6098	83.333	16.666
LMT	0.1397	0.2752	0.814	91.666	8.333

In table 5 represented as a represented the Classification Confusion Matrix.

Table 5: Classifiers Confusion Matrix

Actual Classify	ML Methods	Predicted Classify							
		J48		DS		RT		LMT	
		D	ND	D	ND	D	ND	D	ND
	D	15	0	15	0	15	0	15	0
	ND	1	8	2	7	4	5	2	7

The ROC values (1-specificity and sensitivity) are lies between 0 to 1 and this visualize threshold curve drawn as below in J48 algorithm.

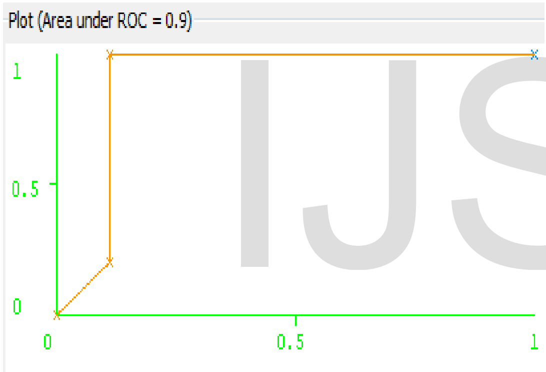

Fig: 2 ROC for Class Success

The area under curve (AUC) maximum value is 1, this curve graphically depicted by snapshoot for attributes success and failure.

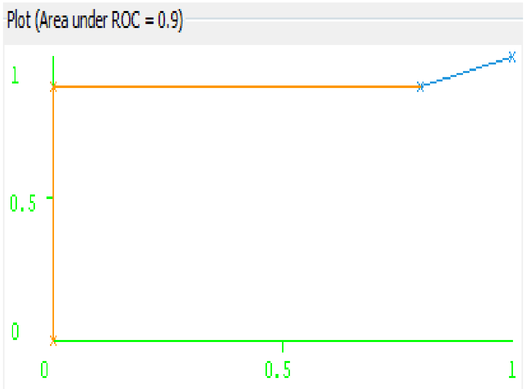

Fig: 3 ROC for Class Failure

Shows the figure 4 and 5 accuracy of machine learning methods and misclassification error rate are respectively for testing 5 fold cross validation.

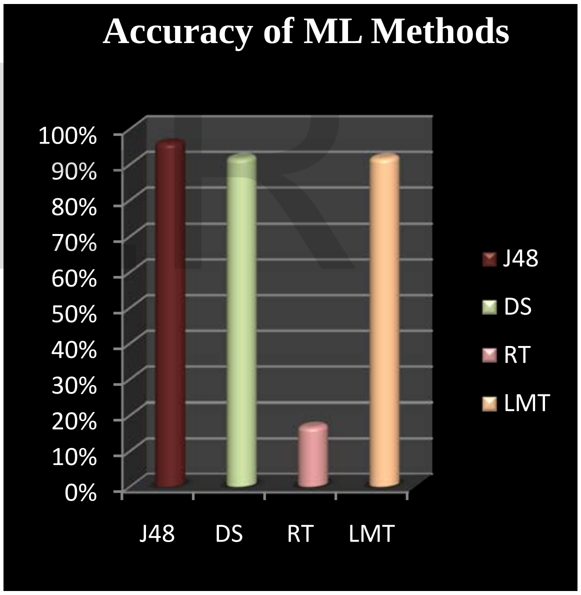

Fig: 4 Accuracy of ML Methods

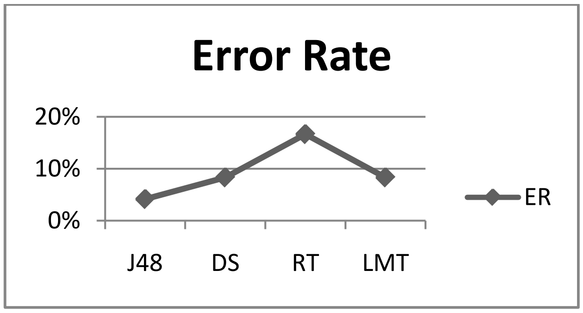

Fig: 5 Misclassification Error Rate

IV. DISCUSSOIN

In this paper we have discussed four topmost machine learning methods to predict faulty or non-faulty modules on open source dataset. Here we have improved the quality and effectiveness of software product was proposed model. Further we have also reduced the estimation cost using various methods.

The result shows that J48 Method provides accuracy of dataset 95.833%. We also found that the model provides minimum misclassification error rate that is 4.166% than the remaining algorithms applied on public REUSE PREDICTING dataset.

Table 2 calculates the multiple values such as TPR, FPR, TNR, FNR, Kappa Statistic, Recall, F-measure ROC and observed that the result i.e. 0.9% obtained for receiver operating characteristic (ROC) (1-specificity and sensitivity) by J48 Method. Figure 2 and 3, represented the area under curve (AUC) plot of x-axis (FPR) versus y-axis (TPR) in measure 0.9% for both case class success and class failure.

Finally we have determined the confusion matrix for all machine learning methods and found that in J48 Method provides 23 correctly classifier instances as well as 1 incorrectly classifier instances where data set size of 14.4 KB.

VI. CONCLUSION AND FUTURE WORK

In this paper authors have examined machine learning methods based on data set REUSE PREDICTING have been used in this paper and increasing accuracy as well as decreasing misclassification error rate is found.

It is also found that the proposed model is quality better accuracy using J48 Method which is 95.833%.

In future, this work may be further used for cyber security aspects of any software system.

We can provide better godliness regarding classification prediction using accuracy, speed, cost, robustness, effectiveness, interoperability, simplicity and bugs (errors) measures and other performance various criteria.

VI. REFERENCES

[1] http/en.wikipedia.org/wiki/Software_bug.

[2] T.M. Khoshgaftaar, E.D. Allen, J.P, Hudepohl, S.J. Aud, Application of neural networks to software quality modeling of a very large telecommunications system," IEEE Transactions on Neural Networks, vol. 8, no. 4, 902-909, 1997.

[3] J. R. Quinlan, "C4.5: Programs for Machine Learning", San Mateo,CA, Morgan Kaufmann Publishers,1993.

[4] N. Laves son and P. Davidson, "Multi-dimensional measures function for classifier performance", 2nd. IEEE International conference on Intelligent system, pp.508-513, 2004.

[5] L. Rokach and O. Maimon. Part C, "Topdown induction of decision trees classifiers - a survey.," Applications and Reviews, IEEE Transactions on Systems, Man, and Cybernetics, vol. 35, pp. 476-487., 2005.

[6] Livingston, Frederick., Implementation of Breiman's Random Forest Machine Learning Algorithm, ECE591Q *Machine Learning Journal Paper*. 2005.

[7] F. Esposito, D. Malerba, and G. Semeraro. A comparative Analysis of Methods for Pruning Decision Trees, IEEE transactions on pattern analysis and machine intelligence, 19(5): pp. 476-491, 1997.

[8] Blake, C. and Merz, C. UCI repository of machine learning

databases.[www.ics.uci.edu/_mlearn/MLRepository.html], 1998.

[9] Fawcett, T.,. ROC Graph, Notes and Pratical Consideration for Data Mining Researchers, 2004.

[10] Gupta, D. L., Malviya, A. K., and Singh, S., Performance Analysis of Classification Tree Learning Algorithms, in International Journal of Computer Applications (IJCA), New York, Volume 55– No.6, October 2012.

[11] WEKA, Available at: http//www.cs.waikato.ac.nz/ml/weka.