Evaluation of Classifiers to Enhance Model Selection

R.Sujatha
Research Scholar
School of Information Technology & Engineering
VIT University, Vellore

D.Ezhilmaran
Guide
School of Advanced Sciences
VIT University, Vellore

***Abstract*— The various tasks like classification, clustering and association rule deriving are performed in the data-mining for the pattern extraction. The performance evaluation measures make each task distinct and meaningful. The plenty of machine learning algorithms helps in the different ways. The classification helps to predict about the future well in advance and make necessary actions thus it otherwise called as actionable data mining. In this paper we plan to give the overview about various classification algorithms by Waikato Environment for Knowledge Analysis otherwise shortly called as WEKA. The measures found in this helps to determine the best model and proposed statistical analysis namely the paired t-test to enhance the model selection. The evaluations make the promising environment for the model selection.**

Keywords- Evaluation; Accuracy; T-Test; Data Mining; Classification; WEKA; Stratified Cross Validation; ROC

I. INTRODUCTION

The classifier helps to make the classification tasks easier. There are many classifiers for a similar data set and the evaluation results are close to each other making it difficult in selecting an ideal model for a data set. In order to find the potential classifier with better accuracy than that of other models, a statistical test called Student paired t test is applied on average accuracy of comparing models. The test concludes with a hypothesis which finds the best classifier for the given data set. The present work examined the performance of different classification methods namely ZeroR, OneR, Naïve Bayes, SMO, Decision Table, PART, J48, Random Tree on the data like Soya beans, Wheat Seeds, Diabetics. The various performance evaluations were taken into consideration and paired t-test is done exclusively for the added advantage for better classification purpose with max accuracy.[14]. The repository of data sets is available in various sites. [4][13]

ZeroR is one of the primitive classifier. According to the WEKA development team, this classifier predicts the majority class in the training data for all rows of test data if the class is categorical. It is not much apt for prediction and minimal performance is given. Still many other algorithms perform worse than this.

OneR is a simple rule-based classifier proposed by Holte's that helps to extract a set of rules based on a single attribute. It is easy to produce reasonable performance on various classification problems by probing only single attribute. According to Holte the minimum number of instances necessary in each class and he suggested it as 6. Returns a rule that finds classification attribute C on the basis of a single predictive attributed A in table T.[2]

The naïve Bayesian (NB) classifier is strong and naturally resistant to noise. It is best approach to supervised classification with extremely high accuracy. The normal distribution is used in the WEKA toolset. In this method a probabilistic summary holding the probability of the class along with the associated probability distribution for each attribute of the training data. The conditional probability is calculated to classify unseen instance assuming that the attributes are autonomous. Based on the type of attributes either the discrete or continuous probability distribution is utilized. [3]

The Sequential Minimal Optimization is popularly called as Platt's SMO algorithm and it is well organized one with excellent computational efficiency. Working sets of size two is selected iteratively with extreme chunking and optimize the target function with respect to them. The optimization subproblems solved analytically by the way of using working sets of size two. The chunking process is iteratively performed until all the training examples fits the optimal conditions. The preliminary thing in the SMO is to choose a pair of indexes, (i_1,i_2) and optimizing the dual objective function in (D) by varying the Lagrange multipliers related to i_1 and i_2 only. The role of threshold parameter β is vital. Since the output error on the ith pattern was defined by Platt as

$$E_i = F_i \sim \beta.$$

Based on the condition **β** is chosen. After several experiments, Platt landed with a good set of heuristics. Employed two-loop approach: the outer loop chooses i_2, the inner loop chooses i_1. The outer loop iterates over all patterns, violating the optimality conditions to make sure that the problem has been solved. He computed the E_i value in parallel. The i_1 is chosen with the target to make a large increase in the objective function. Based on the available E_i , non boundary multiplier indices are used as i_1. Then randomly indexes are chosen from it, if sufficient progress is not met. Thus this affects the running time of SMO.[5][6]

The decision table (DT) has two components schema, list of attributes and body, multiset of labeled instances. The set of instances with the same values for the schema attributes is called a cell. In the WEKA the Decision Table Majority (DTMaj) classifier is used along with the best first search option. From the decision table majority class is returned. [7]

PART is an indirect approach for rule generation. By using the C4.5 statistical classifier the pruned decision tree is generated at iteration. The leaves of the best tree are translated into rules. It is a partial decision tree algorithm. [8]

Random Tree (RT) was introduced by Leo Breiman and Adele Cutler. It can handle both classification and regression problems. The steps involved are it takes the input feature vectors, followed by that it classifies with all the trees in the forest, and final output is the class label which procured the majority of votes. In regression, the average of the responses over the trees in the forest will be the classifier response. In precise it is weak machine learning model that builds tree by taking into consideration K attributes chosen randomly at all node.[9]

J48 is the improved version of C4.5. In this approach it works methodologically. The foremost step is it constructs a very huge tree by taking into account all attribute values and narrow down the decision rule with the help of pruning. Pruning was done using heuristic approach with the aid of statistical significance of splits. The commonly used method is information gain or entropy measure in which the measure that corresponds to level uncertainty in the information. Thus it is like tree structure with root node, intermediate and leaf nodes. Node holds the decision and in turn decision helps to achieve our result.[11][12]

II. MODEL COMPARISON

In this paper the following relations are considered and the various classifiers are applied to check out the model that produces the excellent prediction. [13]

Soy bean relation with 683 instances and 36 attributes

Wheat seed relation with 210 instances and 8 attributes

Diabetes relation with 770 instances and 9 attributes

III. EXPERIMENTAL RESULTS

There are several evaluation techniques such as Cross validation, Holdout, Bootstrap etc. Experiments prove that Cross validation is best equation technique for dataset with less than 1000 instances. In our experiment we use 10 fold CV with all the folds stratified. We intend to apply each fold of a 10 fold CV to two classifiers simultaneously and the accuracy rates fetched from these folds are used to apply for the Paired T test. Three datasets namely soybean, diabetes and wheat seed are used.

The paired t tests are applied to different combination of classifiers and the least significant different classifiers in accuracy are selected for the dataset. For the t test with benchmark of 95% confidence and 9 degrees of freedom(k-1 degrees of freedom for k fold CV), a z value of 2.96 is selected from t table and used in the paired t test. The confidence interval is a measure that speaks about the statistical dispersion of the output. Determining the range of values within which the true parameter value should fall is defined a confidence interval. In case of very small sizes, the confidence interval is wider and larger the size, the confidence interval is larger. In short inverse of the size is the value of confidence interval. The confidence interval for accuracy rate of each of the classifier is computed with a benchmark of 85% confidence that has a z value of 1.96.

A. Soy bean dataset

TABLE I. ACCURACY, CONFIDENCE INTERVAL OF ACCURACY, ROC AREA AND T TEST RESULTS OF 8 CLASSIFIERS FOR SOY BEAN DATASET

Classifiers	ZeroR	OneR	DecisionTable	NaiveBayes	PART	SMO	J48	Random Tree
Accuracy	13.472	39.964	84.027	92.965	91.944	93.70	91.509	80.665
Confidence Interval (+/-)	1.287	6.356	7.085	3.589	4.575	4.301	4.848	7.223
Area under ROC curve	0.484	0.667	0.984	0.996	0.998	0.995	0.998	0.922
Paired T test								
ZeroR vs OneR -> OneR		-30.037						
OneR vs DT-> DT			-35.891					
DT vs NB -> NB				-7.776				
NB vs PART, SMO, J48, RT-> PART, SMO, J48, RT					1.472	-0.893	1.618	8.98
PART vs SMO -> SMO						-2.564		
SMO vs J48, RT -> SMO							3.13	11.535

Note : When comparing Classifier 1 vs Classifier 2, if the t test result is negative and its greater then –z value for 95% confidence interval with 9 degrees of freedom, it indicates Classifier 2 is better than Classifier 1. In the case of ZeroR vs OneR in table I , the t test result -30.037 > -2.26 indicates OneR is better in its accuracy rate than ZeroR .

Whenever a classifier is participating in t test, if it emerges as better classifier among its counterparts, then it is selected as one of the best classifier for the given dataset. With this condition, in the table I we see Naïve Bayes and SMO are best classifiers based on Paired t test results. Further computing the confidence interval for accuracy rate of each of the classifiers will render the true accuracy interval ranges for each of the classifiers. Upon comparing the confidence interval of Naïve Bayes and SMO, we find the interval is narrow and there are no significant differences between the interval ranges of these classifiers.

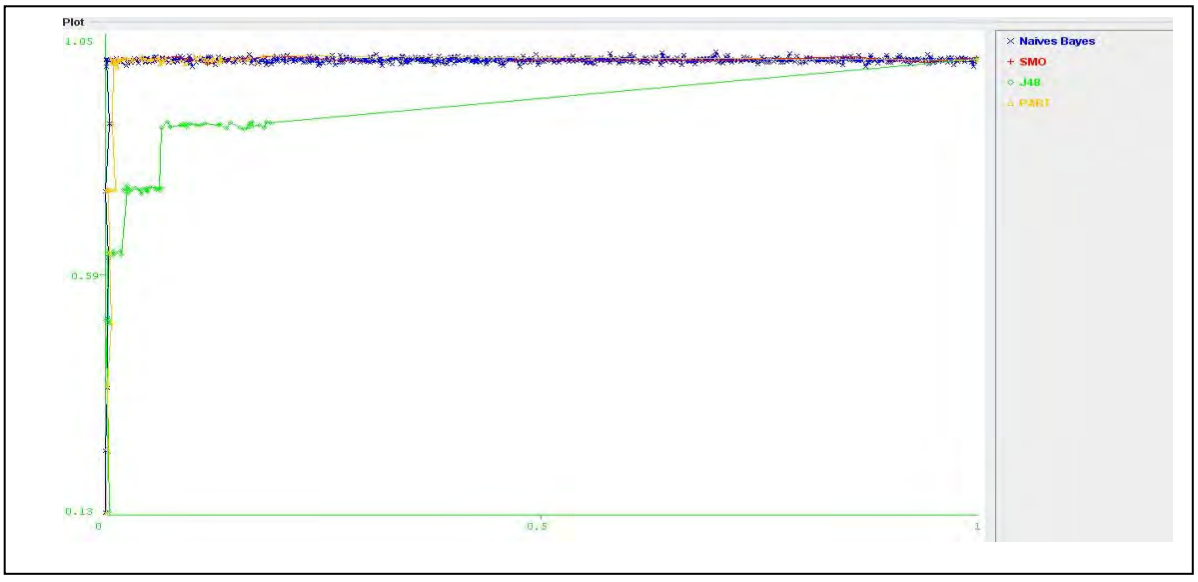

Figure 1. ROC curves for top four classifiers with better accuracy rates. X axis is the True Positive and Y axis is the False Negative

ROC curves are plotted for the top four classifiers with stronger accuracy rates. The Fig. 1 shows SMO and Naïve Bayes are close to each other each having AUC (Area under Curve) as 0.964 and 0.985 respectively. So with combination of Significant Statistical Paired T test, Confidence Interval study on accuracy rate and with ROC curves, we derived confidently that Naïve Bayes and SMO can be equally chosen for soybean data set. The option of giving the user couple or more equally good classifiers will help them to select a single classifier with other better performance attributes without compromising on the accuracy. In this case the user can choose

Naïve Bayes as classifier because the time taken to build and evaluate the classifier is ten times quicker than on SMO classifier without compromising on the accuracy which primarily depicts the quality of a classifier.

B. Wheat seed dataset

TABLE II. ACCURACY, CONFIDENCE INTERVAL OF ACCURACY, ROC AREA AND T TEST RESULTS OF 8 CLASSIFIERS FOR WHEAT SEED DATASET

Classifiers	ZeroR	OneR	DecisionTable	NaiveBayes	PART	SMO	J48	Random Tree
Accuracy	33.334	83.809	87.143	91.429	92.858	93.81	91.905	91.905
Confidence Interval (+/-)	0.002	21.193	12.484	8.577	6.6	8.855	11.683	9.888
Area under ROC curve	0.5	0.879	0.957	0.985	0.944	0.964	0.926	0.911
Paired T test								
ZeroR vs OneR -> OneR		-14.762						
OneR vs DT, NB, PART,SMO-> DT, NB, PART,SMO			-0.886	-1.953	-2.237	-2.552		
NB vs PART,SMO,J48,RT -> NB,PART,SMO,J48,RT					-0.818	-2.236	-0.182	-0.218
PART vs SMO,J48,RT-> PART,SMO,J48,RT						-0.612	0.612	0.557
SMO vs J48,RT-> SMO,J48,RT							0.738	0.768

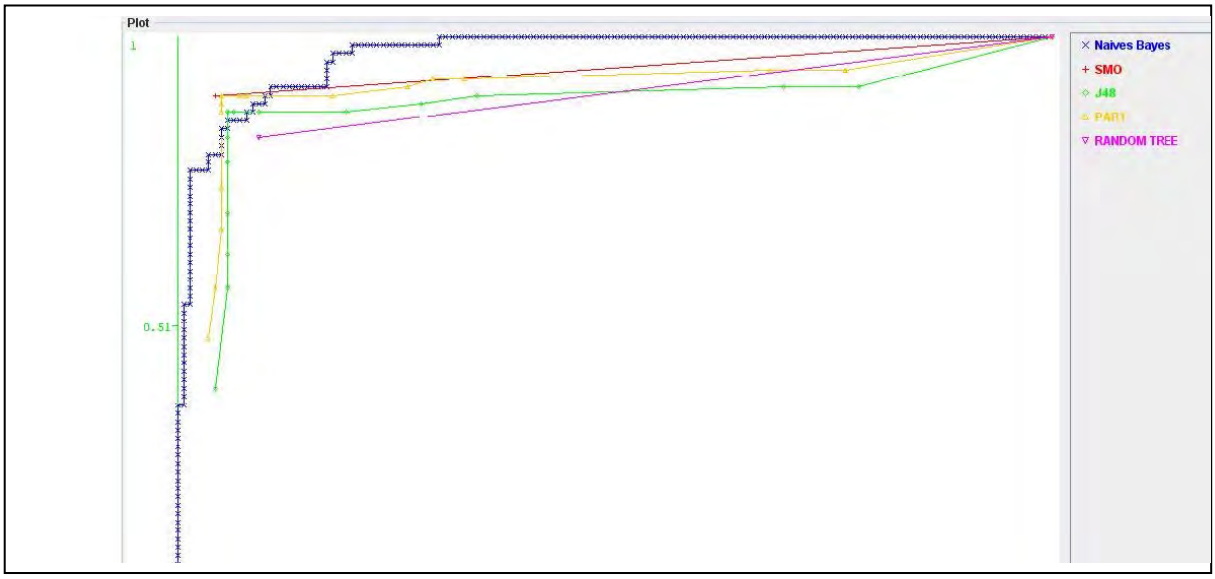

Figure 2. ROC curves for top five classifiers with better accuracy rates. X axis is the True Positive and Y axis is the False Negative

From the above table II and Fig 2 we derive that Naïve Bayes, PART, SMO, J48 and Random Tree are equally good classifiers for the seed dataset.

C. Diabetes dataset

TABLE III. ACCURACY, CONFIDENCE INTERVAL OF ACCURACY, ROC AREA AND T TEST RESULTS OF 8 CLASSIFIERS FOR DIABETES DATASET

Classifiers	ZeroR	OneR	DecisionTable	NaiveBayes	PART	SMO	J48	Random Tree
Accuracy	65.106	72.137	73.31	76.306	74.488	77.345	73.835	69.016
Confidence Interval (+/-)	0.707	9.46	7.946	10.822	9.975	7.968	11.096	13.939
Area under ROC curve	0.497	0.667	0.782	0.819	0.794	0.72	0.751	0.652
Paired T test								
ZeroR vs OneR -> OneR	-4.602							
OneR vs DT vs NB -> NB		-0.627		-3.048				
NB vs PART,SMO,j48,RT -> NB,PART,SMO,j48,RT				1.075		-1.017	1.542	2.702
PART vs SMO,j48,RT ->PART,SMO,j48,RT					-2.576		0.623	2.112
SMO vs j48,RT -> SMO,j48						2.643		3.299

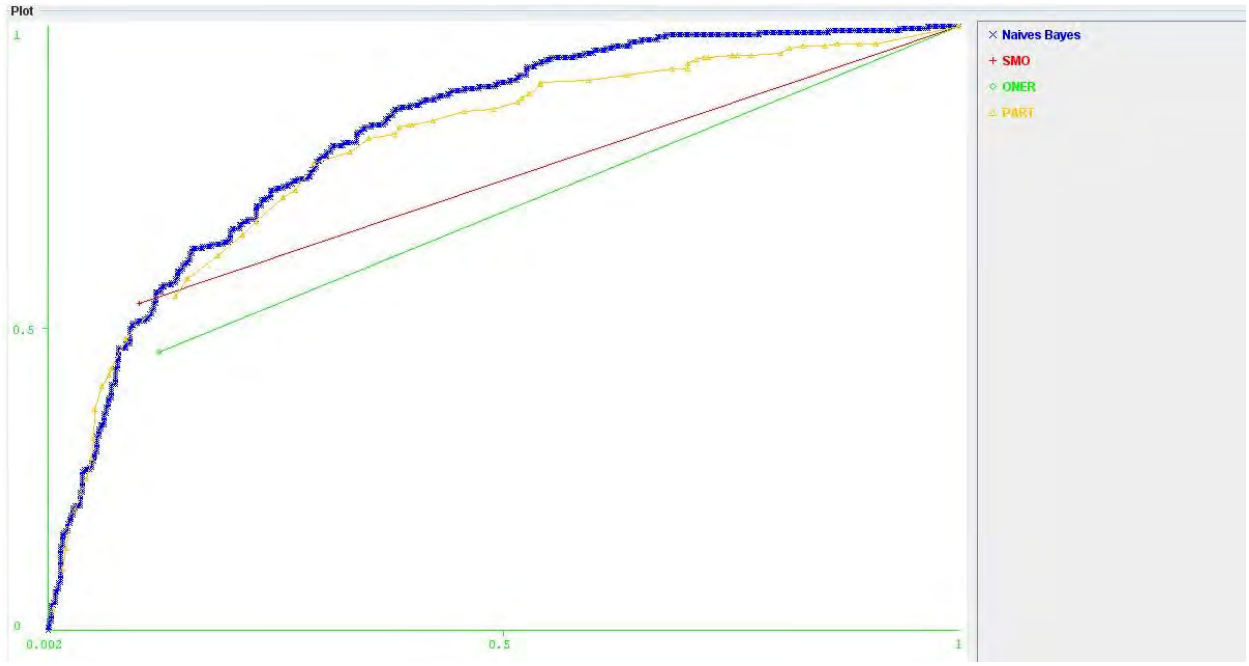

Figure 3. ROC curves for top five classifiers with better accuracy rates. X axis is the True Positive and Y axis is the False Negative

From the above table III and Fig 3 we derive that Naïve Bayes, PART, SMO and J48 are equally good classifiers for the diabetes dataset.

From the above discussion the following classifiers serves good for the data set.

TABLE IV. SELECTED CLASSIFIERS FOR THE RESPECTIVE DATASETS BASED ON THE ABOVE STATISTICAL RESULTS

S No	Dataset	Classifiers Selected
1	Soy bean	Naïve Bayes, SMO
2	Wheat seed	Naïve Bayes, SMO,PART, J48, Random Tree
3	Diabetes	Naïve Bayes, SMO,PART, J48

IV. CONCLUSION

In this paper each of the three nominal dataset is applied to 8 different classifiers and results are evaluated based on the proposed statistical methods. The results confirm that for Soybean dataset- Naïve Byes and SMO, for

Seed dataset - Naïve Bayes, SMO, PART, RandomTree and J48and for Diabetes dataset - Naïve Bayes, SMO, PART and J48 are good classifiers respectively. The proposed method of statistical evaluation can be applied to different datasets and classifiers and better model can be selected.

V. REFERENCES

[1] http://chem-eng.utoronto.ca/~datamining/dmc/zeror.htm
[2] Holt, R.C .1993, "Very simple classification rules perform well on most commonly used datasets" , Machine Learning 11, 63—90.
[3] Pat Langley and Stephanie Sage, "Induction of Selective Bayesian Classifiers", Proceedings of the Tenth Conference on Uncertainty in Artificial Intellignece , 1994, Seattle WA,USA. Morgan Kaufmann Publishers Inc, San Mateo, California.
[4] Data Mining Community's Top Resource , http://www.kdnuggets.com/
[5] J. C. Platt, "Advances in Kernel Methods: Support Vector Machines", B. Schölkopf, C. Burges, and A. Smola, Eds. Cambridge, MA: MIT Press, Dec. 1998. Fast training of support vector machines using sequential minimal optimization.
[6] S.K.Shevade et. al , " Improvements to the SMO Algorithm for SVM Regression" ,IEEE Transactions on Neural Networks, Vol 11, No. 5 September 2000
[7] Ron Kohavi and Daniel Sommerfield, " Targeting Business Users with Decision Table Classifiers" , KDD – 98.
[8] Data Mining : Methods and Techniques, A B M Shawkat Ali, Saleh A. Wasimi, Cengage Learning, 2009
[9] "Bagging Random Tree for Analyzing Breast Cancer Survival", KKU Res. J. 2012; 17(1): 1- 13.
[10] WEKA: Data Mining Software in Java http://www.cs.waikato.ac.nz/ml/weka/
[11] Ian H.Witten, Eibe Frank ,2000 , Data mining : Practical Machine Learning Tools and Techniques with Java Implementations, 2000 Morgan Kaufmann Publishers .
[12] Yugal Kumar et. al., "Analysis of Bayes, Neural Network and Tree Classifier of classification Technique in Data Mining using WEKA", Computer Science & Information Technology , CSCP 2012
[13] UCI Machine Learning Repository: Data Sets, *archive.ics.uci.edu/ml/datasets.html*
[14] Data Mining Concepts and Techniques, Jiawei Han, Micheline Kamber, Morgan Kaufman Publishers, 2003.

R.Sujatha received her B.E. degree in computer science from Madras Uninversity, in 2001, the M.E. degree in computer science from Anna University in 2009 and currently pursuing the Ph.D. degree in Vellore Institute of Technology, Vellore. She was a lecturer and currently assistant professor, with School of Information Technology and Engineering in Vellore Institute of Technology, Vellore. Her area of research interest includes Data mining, Image Processing and Management of Information systems.

Dr. D.Ezhilmaran, received his B.Sc., degree in Mathematics from Madras University, in 1997, the M.Sc., degree in Mathematics from Bharathidasan University, in 1999, the M.Sc., degree in Information Technology from Alagappa University, in 2003, the M.Phil degree in Mathematics from Periyar University, in 2004 and Ph.D. in Mathematics from Alagappa University, in 2011.He worked as Lecturer, Sr. Lecturer and Asst. Professor in KSR College of Technology from 1999 to 2011.Presently he has been working as Asst. Professor of School of Advanced Sciences. His area of interest Fuzzy set theory, Soft set theory and Data mining, Image Processing and Networks using fuzzy concept.