University of Ibadan
Journal of Science and Logics in ICT Research

Error-Rate Evaluation of Classification Data Mining Algorithms in Multidisciplinary Educational Data

***Olamiti A. O.**
aolamiti@yahoo.com

Osofisan A. O.
nikeosofisan@gmail.com

***Corresponding Author**
Department of Computer Science,
University of Ibadan, Ibadan, Nigeria

Abstract

Comparisons tests on Knowledge Discovery in Data (KDD) methods, techniques and tools are carried out to improve on them and also to come up with those that are believed to be "best" for specific domains. Comparing the absolute difference of the error-rates of the algorithms is not enough because the difference should also be tested statistically. Various statistical tests are thus used to determine models/classifiers performances. This study evaluated the performance of the two mostly adopted educational data mining algorithms namely Classification And Regression Trees (CART) and C4.5 with Educational Data (ED) which has the specific characteristic of normal class distribution. The CART and C4.5 were used independently to build ten models for ten ratios. The CART and C4.5 error-rate averages were calculated and their classification performances were compared using two-tailed t-test at $\alpha_{0.05}$. The difference in the error-rates of CART and C4.5 is shown to be statistically significant.

Keywords- Error-rate, Educational data, Performance evaluation, Normal distribution, CART, C4.5.

I INTRODUCTION

Data mining algorithms performance comparison usually involves comparing model error-rate (or the accuracy), model complexity, and model training time. Error-rates are used in measuring models performance and it is the primary index of model performance comparison [1]. The error-rate is an appropriate measure in this study since each training instance of Educational Data (ED) can belong to only one Cumulative Grade Point Average (CGPA) class, that is, all instances are uniquely classifiable [2]. Comparing the absolute difference of the error-rates of the algorithms is not enough because the difference should also be tested statistically [1, 3]. Statistics offers many tests that are designed to measure the significance of any difference between two or more "treatments" [1]. These tests can be adapted for use in determining the better of two learning algorithms but these adaptations as stated by [4], must be with caution. For example the t-test assumes that the test sets for each "treatment" (algorithm) are independent. However both algorithms of this work are compared on the same educational data (ED). As such the test sets are not independent. In view of the situation where tests sets are not independent, various statistical tests have therefore been adapted in the literature for determining the better of two learning algorithms. Two examples are t-test discussed in [5] and two-tailed test reported in [6]. The educational data (ED) used in this study consists of enrolment information of all students in all existing departments of the subject university. It is the same data used in the study reported in [7]. The data is partitioned into six classes A, B, C, D, E and F. The ED exhibits the normal distribution characteristic also known as Gaussian distribution by physicists and bell curve by social scientists [8], which is symmetric with relatively more values at the center (class C) of the distribution and relatively few at the tails (class A and class F).

This study evaluates the performance of CART and C4.5 classification algorithms and determines if there is no significant difference between the models performance.

II. LITERATURE REVIEW

The related works discussed are those where researchers compared performance of one algorithm on more than one datasets; more than one algorithm on one dataset; more than one algorithm on more than one datasets; and those in which the algorithms and/or models comparison performance measure criteria included at least the error-rate/accuracy. These algorithms belong to different data mining techniques which include decision trees/rules, neural networks, logistic regression, Bayesian networks, and genetic

Olamiti, A. O. and Osofisan, A. O. (2018). Error-rate Evaluation of Classification Data Mining Algorithms in Multidisciplinary Educational Data, *University of Ibadan Journal of Science and Logics in ICT Research (UIJSLICTR)*, Vol. 2 No. 1, pp. 8 - 13
©U IJSLICTR Vol. 2 No. 1 June, 2018

algorithms. The datasets could be real world or artificial data from different domains such as education, medicine, business, and science. Also discussed is the two-tailed test which is one of the statistical tests from the literature that have been adapted for determining the better of any two learning algorithms.

Asif et al., [9] analysed the performance of ten classifiers/classification models which included Decision Tree with Gini Index (DT-GI), Naïve Bayes, Neural Networks, and Random Forest with Information Gain on one dataset of students of Information Technology of a public university in Pakistan. They combined the classifier's accuracy and Kappa statistics to arrive at their final conclusion which was that the DT-GI performed the best.

Danjuma and Osofisan [10] compared three data mining algorithms namely, Naïve Bayes, Multilayer Perceptron and J48 decision tree induction on one medical dataset from the University of California at Irvine (UCI) machine learning repository. Their comparative analysis showed that the Naïve Bayes performed best, followed by Multilayer Perceptron while J48 gave the least accuracy. Other performance metrics used included the Receiver Operating Characteristic (ROC) and Area Under the Curve (AUC).

Lavanya and Rani [11] used three algorithms (ID3, CART and C4.5) and five medical datasets from the UCI machine learning repository. The CART models for each of the five datasets produced the highest prediction accuracy rate. They concluded that CART was the best algorithm for classification of data from the medical domain because according to them, accuracy is more important for the classification of medical data. The other performance criterion used in their study was time complexity.

Moshkovich et al., [12] used only one algorithm (C5.0) and four datasets selected from UCI machine learning repository. Each dataset was run twice: setting the attribute types to discrete (categorical) at first run and then to "ordered" at the second run. They reported that a smaller number of induced rules with lower level of errors were achieved with the introduction of ordinal dependencies among categorical attributes. Their conclusion was that the introduction of ordinal dependencies produced essential improvement to classification.

Rahman et al., [13] applied six classification algorithms – JRIP (RIPPER), Naïve Bayes, IBK, Sequential Minimal Optimization (SMO), Multilayer Perceptron, and PART on one dataset obtained from a life insurance company in Bangladesh. The dataset being imbalanced was balanced using Synthetic Minority Over Sampling (SMOTE) and Random Under Sampling (RUS) techniques. Thus there existed two copies of the imbalanced dataset –SMOTE-balanced and RUS-balanced. Each of the six algorithms was then run on either copy. Their findings using the combination of error-rate/accuracy and False Positive (FP) rate was that PART algorithm was best since it gave the least error-rate and least FP-rate with SMOTE-balanced dataset.

Romero et al., [14] compared the performance and usefulness of twenty-five algorithms which included CART, C4.5, AdaBoost, Kernel, and Radial Basis Function Network on one dataset of students of Cordoba University in seven Moodle courses. Their findings showed that the best algorithms with numerical data were CART, Grammar-based genetic programming / genetic algorithm (GAP), Grammar-based genetic programming algorithm (GGP), and Neural Network Evolutionary Programming (NNEP), while CART and C4.5 were best with categorical data. They concluded that some algorithms improved their classification performance when preprocessing task such as discretization was applied to data, but others did not, and that a good classifier model has to be both accurate and comprehensible for the instructors.

Zheng [6] compared the performance of CART and C4.5 algorithms on seven datasets from the UCI machine learning repository and also compared the relationship of training data size to error-rate for both algorithms. The work concluded that since the performance of CART and C4.5 on small data sets was similar, but differs on large data sets; therefore large data sets were more suitable for comparing different algorithms.

Zurada and Lonial [15] examined and compared the effectiveness of four data mining techniques (neural networks, decision trees, logistic regression, and memory-based reasoning) and ensemble model on (a fairly large unbalanced debt recovery) dataset provided by a healthcare company in which the instances/cases with recovered bad debts were under-represented. Their result showed that neural networks, logistic regression, and the combined (logistic regression, neural networks, and decision tree) model produced the best classification accuracy and that the decision tree was the best in predicting "good" customers.

A. Two-tailed Statistical Test (t)

One of the tests from the literature that have been adapted for determining the better of two learning algorithms is the two-tailed test t as reported in [6]. The equations for use in error-rate comparison of two algorithms (A, B) are the following:

$$t = (|E_1 - E_2|) \Big/ \sqrt{\left(q * (1-q) * (1/n_1 + 1/n_2)\right)} \quad (1)$$

$$q = (E_1 + E_2) / 2 \quad (2)$$

where:
E_1 is the error-rate for model M_1 of algorithm A;

E_2 is the error-rate for model M_1 of algorithm B;
n_1 is the number of samples in test set of algorithm A;
n_2 is the number of samples in test set of algorithm B;

With a single test set of size n, equation (1) simplifies to:

$$t = (|E_1 - E_2|) / \sqrt{(q * (1 - q) * (2/n))} \qquad (3)$$

The calculated t is then compared with the tabulated t to be able to determine whether to reject or not reject the null hypothesis, namely that there is no significant difference between algorithms A and B. If calculated t is less than tabulated t then the null hypothesis is accepted otherwise it is rejected. The most commonly used level of significance (α) is 0.05, but for sample size which is usually greater than 120 and since the test is two-tailed, α = 0.05 / 2 (0.025) where degrees of freedom (df) is ∞, is used.

III METHODOLOGY

The educational data (ED) used in this study comprises of records of students that enrolled for undergraduate programmes at the subject university. The data was preprocessed, that is, merged, cleaned, filtered, coded/aggregated, incorporated with appropriate prior knowledge [7]. The preprocessed ED were standardised into formats suitable as input to the CART and C4.5 algorithms respectively. In view of these differences in the formats that are suitable as input for CART and C4.5 there exists two copies of ED, that is, – ED_{CART} and $ED_{C4.5}$. Each copy was thereafter split into pairs of training and testing disjoint sets in ten ratios [16]. Each of the ten train/test pairs induced ten models using CART and C4.5 software respectively [17, 18]. The ten error-rates generated were then averaged as the actual error-rate for that particular splitting ratio. Equation 3 was then used to compute t value to measure if the difference between two error-rates generated from the CART and the C4.5 algorithms at each splitting ratio was not statistically significant. The calculated t value was then compared with the tabulated t at $t_{0.025, \infty}$.

.

IV Results and Discussions

Table 1 shows the number of instances (sizes) in each train-test pair with the test sizes enclosed in parentheses. The test sizes were substituted into Equation 3 as the n values.

The average of the ten error-rates per ratio for CART and C4.5 algorithms are as shown in Table 2.

By substituting average error-rates values for CART and C4.5 from Table 2 as E1, E2 respectively and the n values (test sizes in parentheses) of Table 1 into equation 3 for each of the ten ratios, t values were obtained. The calculated t value for each of the ten ratios is as shown in Table 3.

The calculated t values ranged between 2.10 and 7.37. Since none of the ten calculated t values is less than the tabulated t which is $t_{0.025, \infty} = 1.960$ obtained from standard t-tables; they are each greater than 1.960, thus the null hypothesis, that there is no significant difference at 95% confidence level between each of the two models (per ratio) is rejected. The result implied that there was significant difference in performance between CART and C4.5 algorithms at α = 0.05 / 2 (0.025) where degrees of freedom (df) is ∞ in all the ratios. Figure 2 shows combined test sizes (n) and calculated t values.

This work differs mainly from the related studies in that the researchers carried out performance comparison of algorithms/models but did not divide the data into different ratios. However [6] divided each of the seven datasets into different ratios but none of these datasets was from the educational domain. From Figure 2 it can be seen that the calculated t value has its lowest value when the test size also has its lowest value and also that the calculated t value (except at ratios 70:30; 75:25) decreases as the test size decreases. This observation agrees with that of [6] that when test size is small, the denominator of equation 3 becomes large and therefore the t value is small even with a large value of $|E_{C4.5} - E_{CART}|$. However, the conclusion that there was no significant difference between CART and C4.5 for "small" datasets among the seven datasets used was contrary to our finding where though the ED could be classified as "small" (5202 instances) but the calculated t showed that there was significant difference between CART and C4.5 algorithms.

Conclusion

The performance evaluation using two-tailed test showed that there was significant difference between CART and C4.5 algorithms at α = 0.05 / 2 (0.025) where degree of freedom (df) is ∞ in all the ten ratios.

In the future the performance evaluation will include time complexity and the cost of making wrong classification.

Table 1: The $ED_{CART/C4.5}$ train-test sizes (with test sizes in parentheses)

Class / Ratio	A	B	C	D	E	F	Total
50:50	65(64)	510(510)	1329(1328)	509(508)	167(166)	23(23)	2603(2599)
60:40	77(52)	612(408)	1594(1063)	610(407)	200(133)	28(18)	3121(2081)
66:34	85(44)	673(347)	1754(903)	671(346)	220(113)	30(16)	3433(1769)
66.7:33.3	86(43)	680(340)	1772(885)	678(339)	222(111)	31(15)	3469(1733)
67:33	86(43)	683(337)	1780(877)	681(336)	223(110)	31(15)	3484(1718)
70:30	90(39)	714(306)	1860(797)	712(305)	233(100)	32(14)	3641(1561)
75:25	97(32)	765(255)	1993(664)	763(254)	250(83)	35(11)	3903(1299)
80:20	103(26)	816(204)	2126(531)	814(203)	266(67)	37(9)	4162(1040)
90:10	116(13)	918(102)	2391(266)	915(102)	300(33)	41(5)	4681(521)
95:5	123(6)	969(51)	2524(133)	966(51)	316(17)	44(2)	4942(260)
Total	129	1020	2657	1017	333	46	5202

Figure 1 shows the error-rates of ten runs for each ratio of CART and C4.5 algorithms respectively.

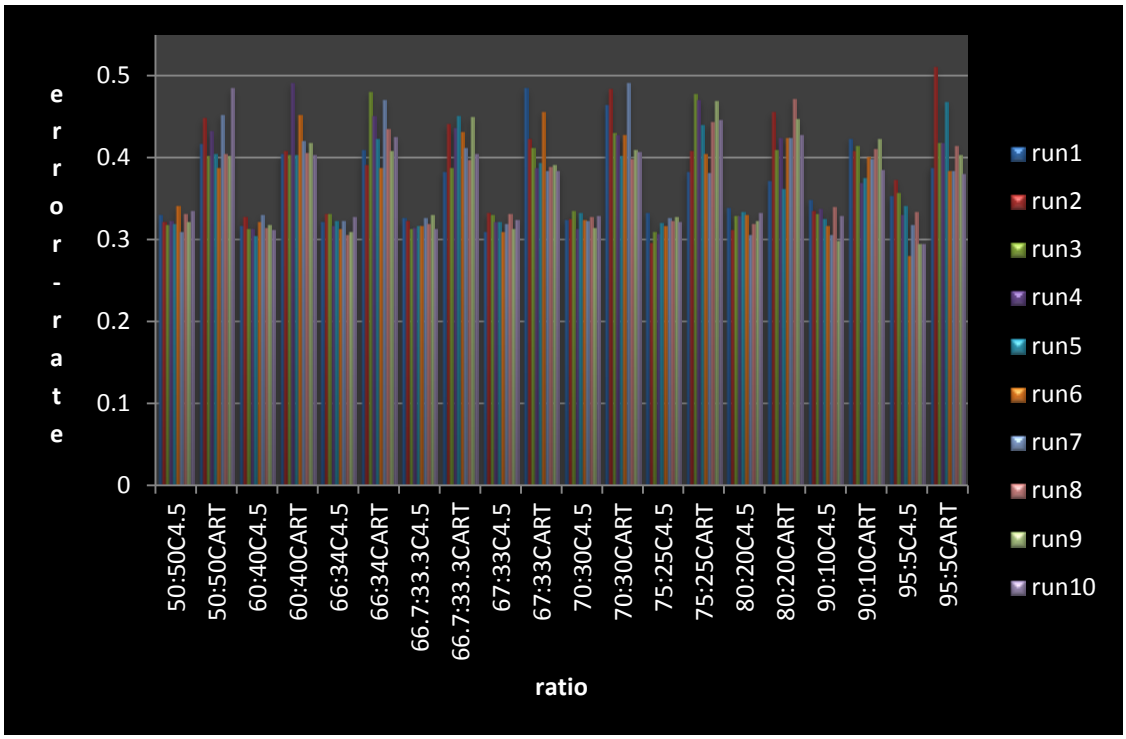

Table 2: The Average Error-Rates of CART and C4.5

Ratio	CART	C4.5
50:50	0.425	0.326
60:40	0.422	0.318
66:34	0.429	0.321
66.7:33.3	0.420	0.321
67:33	0.411	0.322
70:30	0.435	0.326
75:25	0.433	0.319
80:20	0.422	0.326
90:10	0.402	0.327
95:5	0.418	0.329

Table 3: (Calculated) t value

Ratio	t
50:50	7.37
60:40	6.95
66:34	6.63
66.7:33.3	6.03
67:33	5.41
70:30	6.27
75:25	6.00
80:20	4.52
90:10	2.52
95:5	2.10

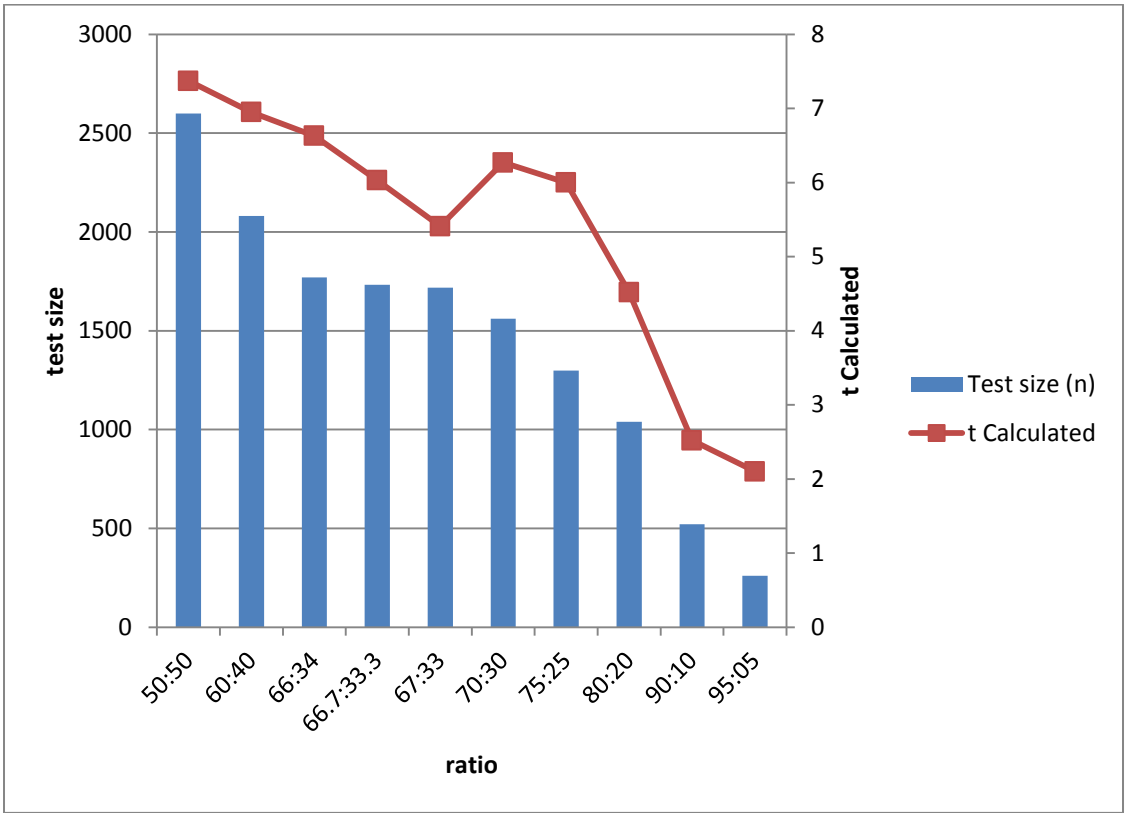

Figure 2: Test size (n) and Calculated t values per ratio

REFERENCES

[1] Witten, I. H., Frank, E. and Hall, M. (2011). *Data Mining Practical Machine Learning Tools and Techniques with Java Implementations*. 3rd edition. Morgan Kaufmann, San Francisco.

[2] Han, J., and Kamber, M. (2006). *Data mining: Concepts and techniques*. 2nd Edition. Morgan Kaufmann, San Francisco.

[3] Olamiti, A. and Osofisan, A. (2009). Academic Background of Students and Performance in a Computer Science Programme in a Nigerian University, *European Journal of Social Sciences*, Vol. 9, No. 4, pp. 564 – 572.

[4] Salzberg, S. (1997) Methodological Note On Comparing Classifiers: Pitfalls to Avoid and a Recommended Approach, *Data Mining and Knowledge Discovery*, Vol. 1, No. 3, pp. 317 - 328.

[5] Bouckaert, R. R. (2003). Choosing between two learning algorithms based on calibrated tests, In T. Fawcett and N. Mishra, editors, Machine Learning, Proceedings of the Twentieth International Conference on Machine Learning, Washington DC, USA, August 21 – 24, pp. 51 – 58.

[6] Zheng, J. (2004). Study on the relationship of training data size to error rate and the performance comparison for two decision tree algorithms. http://etd.lib.ttu.edu/theses/available/etd-06272008-31295019380293/unrestricted/31295019380293.pdf downloaded on 27 September, 2009.

[7] Olamiti, A. O. and Osofisan, A. O. (2016). Investigating The Relationship Between The Size of Training Data and Error-Rate. *The Journal of Computer Science and Its Applications*, Vol. 23, No. 2, pp. 84 – 89.

[8] Lyon, A. (2013). Why are Normal Distributions Normal? http://aidanlyon.com/media/publications/Lyon-normal_distributions.pdf downloaded on 2 November, 2014.

[9] Asif, R., Merceron, A., Ali, A. and Haider, N. (2017). Analyzing undergraduate students' performance using educational data mining. Computer and Education 113 pp 177 - 194

[10] Danjuma, K. and Osofisan, A. O. (2014). Evaluation of Predictive Data Mining Algorithms in Erythemato-Squamous Disease Diagnosis. International Journal of Computer Science Issues. 11: (6): 85 – 94.

[11] Lavanya, D. and Rani, K. U. (2011). Performance Evaluation of Decision Tree Classifiers on Medical Datasets. http://www.ijcaonline.org/volume26/number4/pxc3874247.pdf downloaded on 14 March, 2013.

[12[Moshkovich, H. M., Mechitov, A. I. and Olson, D. L. (2002). Rule induction in data mining: effect of ordinal scales, Expert Systems with application 22, pp. 303 – 311.

[13] Rahman, S., Arefin, K., Masud, S. and Rahman, R. (2017). Analyzing Life Insurance Data with Different Classification Techniques for Customers' Behavior Analysis. Krol, D. et al. (eds.), Advanced Topics in Intelligent Information and Database Systems, Studies in Computational Intelligence 701, DOI 10.1007/978-3-319-56660-3_2.

[14] Romero, C., Ventura, S., Espejo, P. G. and Hervas, C. (2008) Data Mining Algorithms to Classify Students. Proceedings of the First International Conference on Educational Data Mining, Montreal, Canada. June 20 – 21, pp. 8 – 17.

[15] Zurada, J. and Lonial, S. (2005). Comparison Of The Performance Of Several Data Mining Methods For Bad Debt Recovery In The Healthcare Industry, *The Journal of Applied Business Research* Vol. 21, No. 2, pp. 37 – 53.

[16] Olamiti, A. O. and Osofisan, A. (2017). SCHRD: Stratified and Constant Holdout Ratio Divider for Higher Educational Data. *Journal of Computer Science and Its Applications*, Vol. 24, No. 1, pp.

[17] CART Manual (2008), http://www.ifpri.org/pubs/microcom/micro.pdf downloaded on 20 April, 2014.

[18] Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo California.