An Evaluation of Classification Algorithms Using Mc Nemar's Test

Betul Bostanci and Erkan Bostanci

Abstract Five classification algorithms namely J48, Naive Bayes, Multilayer Perceptron, IBK and Bayes Net are evaluated using Mc Nemar's test over datasets including both nominal and numeric attributes. It was found that Multilayer Perceptron performed better than the two other classification methods for both nominal and numerical datasets. Furthermore, it was observed that the results of our evaluation concur with Kappa statistic and Root Mean Squared Error, two well-known metrics used for evaluating machine learning algorithms.

Key words: Classifier Evaluation, Classification algorithms, Mc Nemar's test

1 INTRODUCTION

Evaluating the performance of machine learning methods is as crucial as the algorithm itself since this identifies the strengths and weaknesses of each learning algorithm. This paper investigates the usage of Mc Nemar's test as an evaluation method for machine learning methods.

Mc Nemar's test has been used in different studies in previous research. Dietterich [1] examined 5 different statistical tests including Mc Nemar's test to identify how these tests differ in assessing the performances of classification algorithms. A similar evaluation was performed on a large database by Bouckaert [2]. Demsar [3] has evaluated decision tree, naive bayes and k-nearest neighbours methods

Betul Bostanci
School of Computer Science and Electronic Engineering, University of Essex, Colchester, UK
e-mail: bbosta@essex.ac.uk

Erkan Bostanci
School of Computer Science and Electronic Engineering, University of Essex, Colchester, UK
e-mail: gebost@essex.ac.uk

J. C. Bansal et al. (eds.), *Proceedings of Seventh International Conference on Bio-Inspired Computing: Theories and Applications (BIC-TA 2012)*, Advances in Intelligent Systems and Computing 201, DOI: 10.1007/978-81-322-1038-2_2, © Springer India 2013

using other non-parametric tests including ANOVA (ANalysis Of VAriance) [4] and Friedman test [5, 6].

Other studies have evaluated classifiers using this test over a large set but our method differs in that we use a different criterion that compares how the individual instances are classified and how this is reflected in the whole dataset.

Five different machine learning methods namely J48 (Decision Tree), Naive Bayes [7], Multilayer Perceptron [7] IBK [8] and Bayes Net [9] were used in the experiments. WEKA [10] was used to obtain the classification results of these algorithms. These classification methods are used to classify samples from different datasets. Later, the classification results are analyzed using a non-parametric test in order to identify how a pair of learning methods differ from each other and which of the two performs better.

The rest of the paper is structured as follows: Section 2 presents the nominal and numeric datasets used in the experiments. Section 3 introduces Mc Nemar's test which is the main evaluation method proposed in this study followed by Section 4 where the experimental design is presented. Section 5 presents Mc Nemar's test results and compares them with two conventional evaluation criteria. Finally, the paper is drawn to a conclusion in Section 6.

2 DATASETS

In order to perform a fair evaluation, a relatively large number of datasets obtained from UCI Machine Learning Repository [11] are used. The datasets are selected from the ones including nominal (Table 1) and numeric data (Table 2).

Table 1 Nominal Datasets

Dataset	Number of Instances	Number of Attributes	Number of Classes
Car	1728	7	4
Nursery	12960	9	5
Tic-Tac-Toe	958	10	2
Zoo	101	18	7

Table 2 Numeric Datasets

Dataset	Number of Instances	Number of Attributes	Number of Classes
Diabetes	768	9	2
Glass	214	10	7
Ionosphere	351	35	2
Iris	150	5	3
Segment-Challenge	1500	20	7
Waveform-5000	5000	41	3

3 Mc NEMAR'S TEST

Mc Nemar's test [12, 13] is a variant of χ^2 test and is a non-parametric test used to analyse matched pairs of data. According to Mc Nemar's test, two algorithms can have 4 possible outcomes arranged in a 2×2 contingency table [14] as shown in Table 3.

Table 3 Possible results of two algorithms [13]

	Algorithm A failed	Algorithm A succeeded
Algorithm B failed	N_{ff}	N_{sf}
Algorithm B succeeded	N_{fs}	N_{ss}

N_{ff} denotes the number of times (instances) when both algorithms failed and N_{ss} denotes success for both algorithms. These two cases do not give much information about the algorithms' performances as they do not indicate how their performances differ. However, the other two parameters (N_{fs} and N_{sf}) show cases where one of the algorithms failed and the other succeeded indicating the performance discrepancies.

In order to quantify these differences Mc Nemar's test employs z score (Equation 1).

$$z = \frac{(|N_{sf} - N_{fs}| - 1)}{\sqrt{N_{sf} + N_{fs}}} \tag{1}$$

z scores are interpreted as follows: When $z = 0$, the two algorithms are said to show similar performance. As this value diverges from 0 in positive direction, this indicates that their performance differs significantly. Furthermore, z scores can also be translated into confidence levels as shown in Table 4.

Table 4 Confidence levels corresponding to z scores for one-tailed and two-tailed predictions [13]

z score	One-tailed Prediction	Two-tailed Prediction
1.645	95%	90%
1.960	97.5%	95%
2.326	99%	98%
2.576	99.5%	99%

Following the table, it is worth mentioning that *One-tailed Prediction* is used to determine when one algorithm is better than the other where *Two-tailed Prediction* shows how much the two algorithms differ.

Mc Nemar's test is known to have a low *Type-I* error which occurs when an evaluation method detects a difference between two learning algorithms when there is no difference [1].

4 EVALUATION CRITERION

By adopting the Mc Nemar's test to evaluate classification algorithms, the following criterion is defined: An algorithm is regarded as "successful" if it can identify the class of an instance correctly. Conversely, it is regarded as "failed" when it performs an incorrect classification for an instance.

Using this criterion, the z scores are calculated using Mc Nemar's test for the five classification algorithms. All the algorithms were used with their default parameters as parameter tuning may favor one algorithm to produce better results.

The null hypothesis (H_0) for this experimental design suggests that different classifiers perform similarly whereas the alternative hypothesis (H_1) claims otherwise suggesting that at least one of the classifiers performs differently as shown in Equation 2.

$$\begin{aligned} &H_0 : C_1 = C_2 = C_3 = C_4 = C_5 \\ &H_1 : \exists C_i : C_i \neq C_j, (i, j) \in (1,2,3,4,5), i \neq j \end{aligned} \tag{2}$$

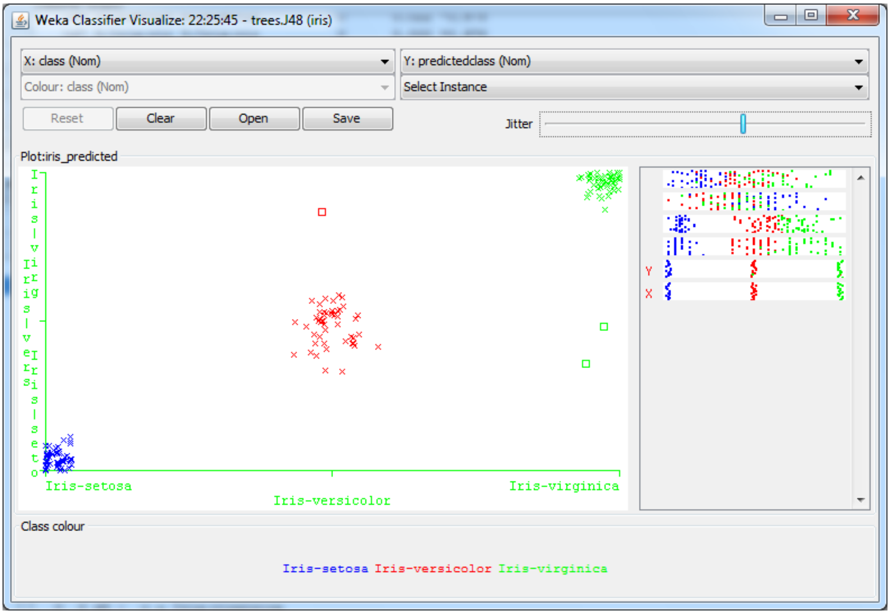
At the end of the experiment, the z scores will indicate whether we should accept H_0 and reject H_1 or vice versa. In order to calculate the z scores, the classification results of the three classifiers must be identified for each individual instance.

This operation is performed for all instances in the given datasets. In WEKA, there are two options to see whether an instance is correctly classified or not. The first option is the graphical one (shown in Figure 1 with the squares while crosses denote correct classifications). The second option to show the incorrect classifications is via the "Output predictions" option of the classifier which displays a "+" in the output next to the instance which has been incorrectly classified.

10-fold cross-validation is used in the evaluation which works as folllows: First the data is separated into 10 sets each having $n/10$ instances. Then, the training is performed using 9 of these sets and testing is performed on the remaining 1 set. This process is repeated 10 times to consider all of the subsets created and the final result for the accuracy is obtained by taking the average of these iterations.

The first option is quite useful to see the result graphically, however in order to calculate the number of correct and incorrect classifications by the classifiers, one needs to export these results into a spreadsheet (e.g. Excel). For this reason, the second method was used to calculate number of instances where the classifiers succeeded and failed. Using these figures, the z scores were calculated using Equation 1.

In order to decide which classifier performed better, N_{sf} and N_{fs} values for two classifiers are examined. For example, classifier A is said to perform better than classifier B if N_{sf} is larger than N_{fs} according to Table 3.



Weka Classifier Visualize: 22:25:45 - trees.J48 (iris)

X: class (Nom)

Y: predictedclass (Nom)

Colour: class (Nom)

Fig. 1 Visualization of Classification Errors in WEKA

5 RESULTS

This section presents the results of the experiment. Results for the Mc Nemar's test will be given first and then these results will be compared with two other evaluation criteria namely Kappa statistic and Root Mean Squared Error (RMSE).

5.1 McNemar's Test Results

In Tables 5 and 6, the arrowheads (←, ↑) denote which classifier performed better in the given datasets. z scores are given next to the arrowheads as a measure of how statistically significant the results are.

By looking at the Mc Nemar's test results for the nominal datasets (Table 5), one can deduce that Multilayer Perceptron has produced significantly better results than J48 and Naive Bayes classifiers (H_1 is accepted with a confidence level of more than 99.5%). J48 classifier performed better than the Naive Bayes for *Nursery* and *Tic-Tac-Toe* datasets. For the *Zoo* dataset, Naive Bayes performed better than J48 and equally to the Multilayer perceptron (H_0 is not rejected.). The performance differences between IBK and all other classifiers were not found to be statistically significant for the *Zoo* dataset but for the rest of the nominal datasets, there were significant differences. Bayes Net shows a poor performance overall except for the *Zoo*

Table 5 Mc Nemar's Test Results for Nominal Datasets

Car

	Naive Bayes	Multilayer Perceptron	IBK	Bayes Net
J48	0	↑ 10.63	↑ 1.62	← 6.93
Naive Bayes		↑ 10.63	↑ 1.62	← 6.93
Multilayer Perceptron			← 9.82	← 15.08
IBK				← 9.75

Nursery

	Naive Bayes	Multilayer Perceptron	IBK	Bayes Net
J48	← 24.66	↑ 17.32	↑ 34.89	← 24.64
Naive Bayes		↑ 34.89	↑ 31.68	0
Multilayer Perceptron			← 12.09	← 34.87
IBK				← 31.66

Tic-tac-toe

	Naive Bayes	Multilayer Perceptron	IBK	Bayes Net
J48	← 8.44	↑ 10.06	↑ 15.73	← 8.56
Naive Bayes		↑ 15.73	← 0.70	← 0.70
Multilayer Perceptron			← 15.90	← 15.90
IBK				0

Zoo

	Naive Bayes	Multilayer Perceptron	IBK	Bayes Net
J48	← 0.67	↑ 1.23	0	↑ 0.5
Naive Bayes		0	0	0
Multilayer Perceptron			0	0
IBK				0

dataset where it performed better than J48 although the result was not statistically significant.

Many differences in the classification performance are noticeable in the numeric dataset results (Table 6). For the *Glass* and *Segment-Challenge* datasets J48 has given better classification performance than Naive Bayes. For the former dataset, the Multilayer Perceptron performed equally with J48 and Naive Bayes produced a poorer classification result than these two. IBK and Bayes Net shows better performance over J48, Naive Bayes and Multilayer Perceptron, however there was no statistically significant performance difference between these two classification methods.

It is interesting to see that the first three (J48, Naive Bayes and Multilayer Perceptron) classifiers performed similarly on the *Ionosphere* dataset (H_0 is not rejected for all pairs.). Some differences can noticeable between these classifiers and Bayes Net however the results are not significant ($z = 0.75$ for Naive Bayes and Multilayer Perceptron) A similar result is also visible when the *Iris* dataset is consided since the values are quite close to zero. For the *Diabetes* dataset, Naive Bayes showed better performance over J48 yet the difference was not very significant for the latter (with a confidence level less than 95%) whereas Naive Bayes performs significantly better than J48 for the *Waveform-5000* dataset.

We can also see that the Multilayer Perceptron did not produce good results for the *Ionosphere* dataset where a relatively large number of attributes are present. This

Table 6 McNemar's Test Results for Numeric Datasets

Diabetes

	Naive Bayes	Multilayer Perceptron	IBK	Bayes Net
J48	↑ 1.61	↑ 0.96	← 0.56	↑ 0.26
Naive Bayes		← 0.56	← 3.40	← 1.29
Multilayer Perceptron			← 2.97	← 0.59
IBK				↑ 2.16

Glass

	Naive Bayes	Multilayer Perceptron	IBK	Bayes Net
J48	← 4.07	0	↑ 4.07	↑ 0.97
Naive Bayes		↑ 4.07	↑ 5.05	↑ 5.24
Multilayer Perceptron			↑ 0.95	↑ 0.97
IBK				0

Ionosphere

	Naive Bayes	Multilayer Perceptron	IBK	Bayes Net
J48	0	0	0	← 1.05
Naive Bayes		0	← 2.71	← 0.75
Multilayer Perceptron			← 2.71	← 0.75
IBK				↑ 1.37

Iris

	Naive Bayes	Multilayer Perceptron	IBK	Bayes Net
J48	0	↑ 0.41	↑ 0.5	← 1.51
Naive Bayes		↑ 0.5	0	← 1.51
Multilayer Perceptron			← 1.16	← 2.00
IBK				← 1.23

Segment

	Naive Bayes	Multilayer Perceptron	IBK	Bayes Net
J48	← 12.95	↑ 1.69	↑ 14.22	← 6.58
Naive Bayes		↑ 14.22	↑ 13.48	↑ 8.53
Multilayer Perceptron			← 0.86	← 7.92
IBK				← 7.05

Waveform-5000

	Naive Bayes	Multilayer Perceptron	IBK	Bayes Net
J48	↑ 6.90	↑ 12.40	← 1.84	↑ 6.71
Naive Bayes		↑ 5.78	← 8.77	← 0.54
Multilayer Perceptron			← 13.89	← 6.05
IBK				↑ 8.59

lower performance can be due to an underfitting problem as the default parameters were used without any parameter tuning.

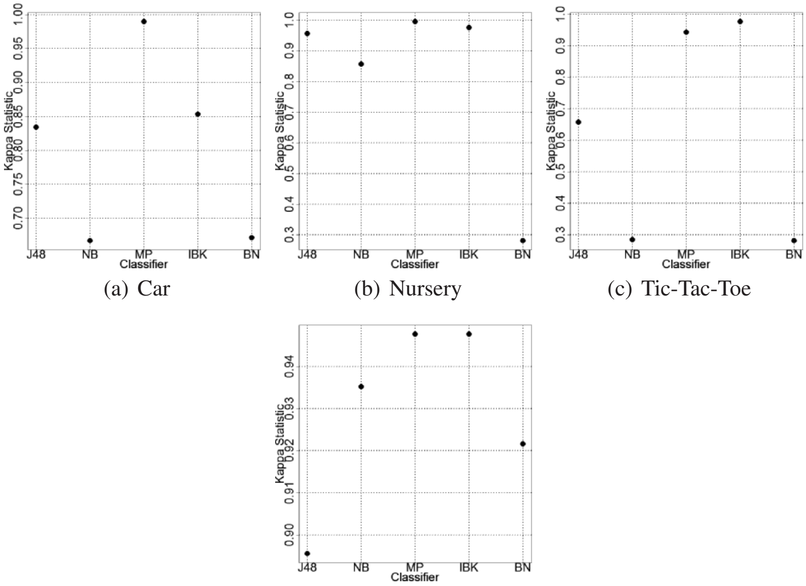
5.2 Comparison with Other Evaluation Criteria

Mc Nemar's test result showed that there are significant discrepancies in the performances of the classifiers. Additional experiments were carried out to see how the results for Mc Nemar's test conform with other evaluation criteria namely Kappa Statistic and Root Mean Squared Error.

5.2.1 Kappa Statistic

Kappa Statistic is a measure of the agreement between the predicted and the actual classifications in a dataset [15]. For this reason, we expect a higher value for a classifier which has more overlapping predictions and observations.

By looking at the nominal datasets in Figure 2, we see that Multilayer Perceptron has the highest value in 3 out of 4 datasets. J48 is better than Naive Bayes except for the *Zoo* dataset (Figure 2(d)). IBK shows good performance in all nominal datasets, although the poorest performance can be seen in the *Car* dataset.



(a) Car

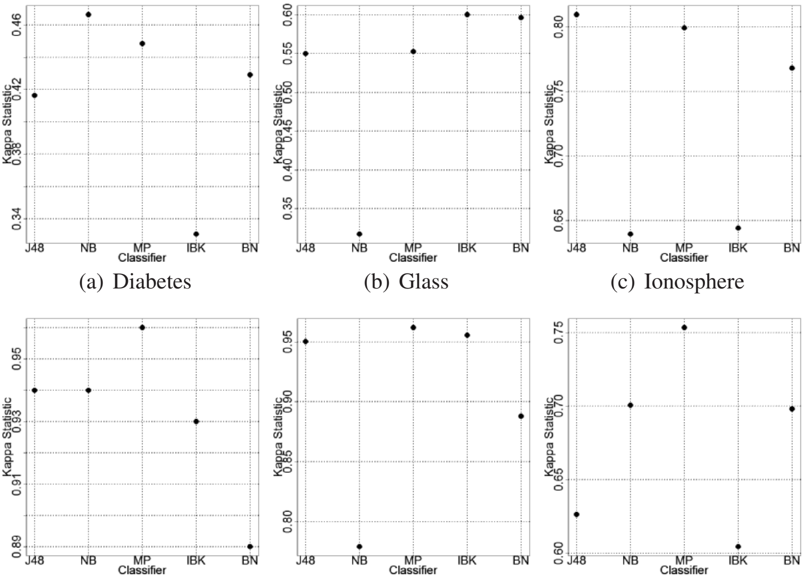
(b) Nursery

(c) Tic-Tac-Toe

(d) Zoo

Fig. 2 Kappa Statistics for Nominal Datasets. NB: Naive Bayes, MP: Multilayer Perceptron, BN: Bayes Net

The ranking between J48 and Multilayer Perceptron changes significantly for the *Glass* and *Segment-Challenge* datasets for the numeric datasets in Figure 3. IBK has a good performance in these two datasets ($Kappa = 0.60$ and $Kappa = 0.95$ respectively). Naive Bayes produced good results only for the *Diabetes* dataset in this group. We can also say the Bayes Net achieves higher classification performance for the numeric datasets than the nominal datasets.



(a) Diabetes (b) Glass (c) Ionosphere

(d) Iris (e) Segment-Challenge (f) Waveform-5000

Fig. 3 Kappa Statistics for Numeric Datasets. NB: Naive Bayes, MP: Multilayer Perceptron, BN: Bayes Net

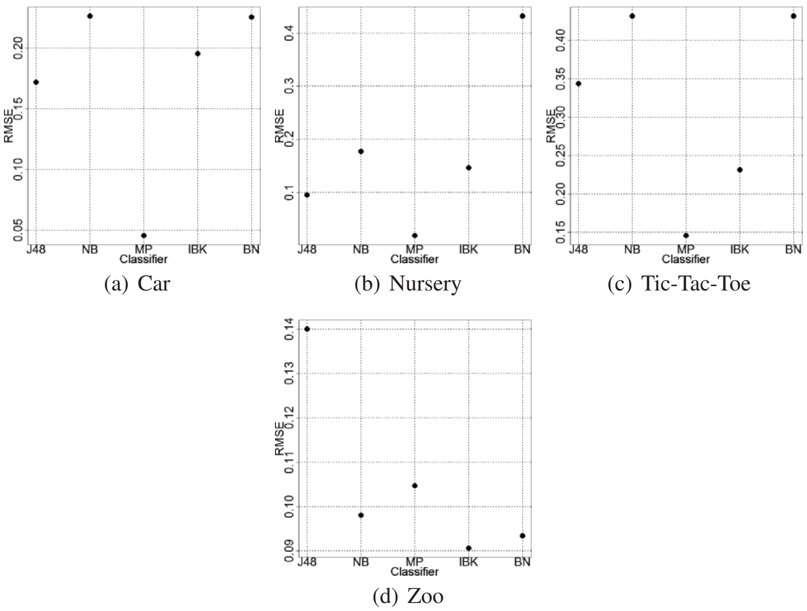
5.2.2 Root Mean Squared Error

Root Mean Squared Error (RMSE) [15] shows the error in the predicted and actual classes which the instances in a dataset belong to. RMSE should have lower values for more accurate classification results.

In nominal dataset results (Figure 4), Multilayer Perceptron had the lowest RMSE values for *Car*, *Nursery* and *Tic-Tac-Toe* datasets. J48 performed better than the Naive Bayes for the these datasets as well, while the ranking changed between them in the *Zoo* dataset shown in Figure 4(d). IBK shows the worst performance on the *Diabetes* and the best performance on the *Zoo* dataset. Bayes Net has poor performance in *Car* and *Tic-Tac-Toe* datasets.

A first look on the results in Figure 5 reveals that the Multilayer Perceptron results in lowest RMSE values for 4 out of 6 numeric datasets. Naive Bayes has a poor performance in *Glass*, *Ionosphere* and *Segment-Challenge* datasets. Naive Bayes showed the lowest performance in all datasets of the numeric dataset results except for the *Diabetes* dataset.

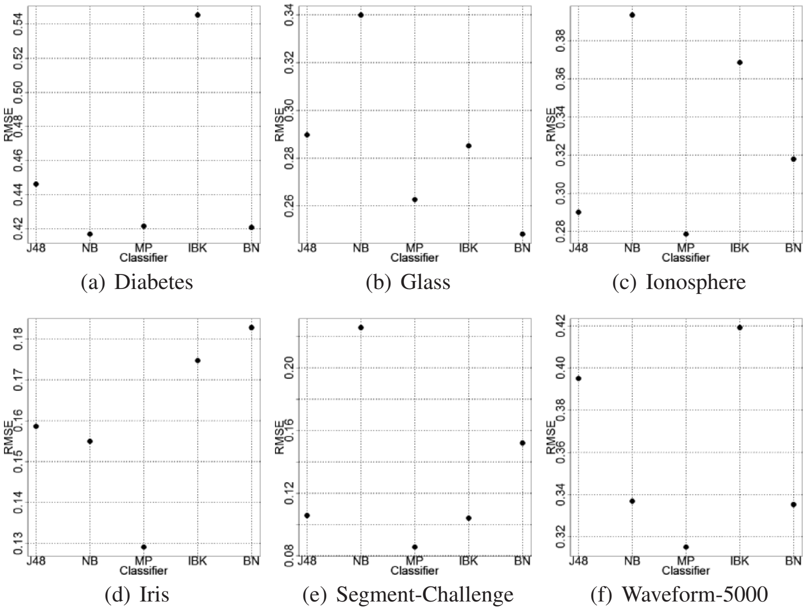
Table 7 show the mean results for all classifiers for the nominal and numeric datasets. Multilayer Perceptron has the highest values for Kappa statistic and lowest values for RMSE showing that the classification results using this classifier are accurate. IBK also shows a good classification performance for nominal and numeric data. Poor results are visible for Naive Bayes and Bayes Net.



(a) Car (b) Nursery (c) Tic-Tac-Toe

(d) Zoo

Fig. 4 RMSE for Nominal Datasets. NB: Naive Bayes, MP: Multilayer Perceptron, BN: Bayes Net



(a) Diabetes (b) Glass (c) Ionosphere

(d) Iris (e) Segment-Challenge (f) Waveform-5000

Fig. 5 RMSE for Numeric Datasets. NB: Naive Bayes, MP: Multilayer Perceptron, BN: Bayes Net

Table 7 Mean Kappa statistic and RMSE values for nominal and numeric datasets

	Nominal		Numeric	
	Kappa	RMSE	Kappa	RMSE
J48	0.82	0.19	0.72	0.36
Naive Bayes	0.69	0.23	0.64	0.31
Multilayer Perceptron	0.97	0.08	0.75	0.25
IBK	0.94	0.17	0.68	0.32
Bayes Net	0.54	0.30	0.71	0.28

From the three evaluation criteria (Mc Nemar's test, Kappa statistic and RMSE), Table 8 can be used to summarize the performance difference over the nominal and numeric datasets where + indicates performance grade. By looking at this summary table, it is evident that the Mc Nemar's test agrees with other evaluation criteria as an important result of the experiments. An exception can be seen for the comparison of Naive Bayes and J48 which is due to insignificance of the differences in Mc Nemar's test.

Table 8 Summary of the performances for nominal and numeric datasets

Mc Nemar's test					
	J48	Naive Bayes	Multilayer Perceptron	IBK	Bayes Net
Nominal	+++	++	+++++	++++	+
Numeric	+	++++	+++++	++	+++
Kappa statistic					
Nominal	+++	++	+++++	++++	+
Numeric	++++	+	+++++	++	+++
RMSE					
Nominal	+++	++	+++++	++++	+
Numeric	+	+++	+++++	++	++++

6 CONCLUSION

This study employed Mc Nemar's test in order to evaluate machine learning algorithms namely J48, Naive Bayes and Multilayer Perceptron, IBK and Bayes Net. By defining the success and failure criteria of Mc Nemar's test as correctly or incorrectly identifying the class of an instance in a dataset, the experiments presented the usage of a non-parametric test as a new method to evaluate classification algorithms.

The results showed that Multilayer Perceptron produced better results than the other methods for both nominal and numerical data. Bayes Net was placed in the lowest ranks for both types of data. Another interesting finding of the experiment is that the results of the Mc Nemar's test mostly conformed with Kappa statistic and RMSE as a justification of method's integrity.

The effect of parameter tuning is considered as future research. In this case, the classifiers will be tuned to achieve the optimal results and then the same tests can be applied to see whether there will be any changes in the rankings.

Acknowledgements The authors would like to thank Nadia Kanwal for discussions on Mc Nemar's test.

References

1. T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural Computation*, vol. 10, pp. 1895–1923, 1998.
2. R. R. Bouckaert and E. Frank, "Evaluating the replicability of significance tests for comparing learning algorithms," in *Proceedings 8th Pacific-Asia Conference*, pp. 3–12, 2004.
3. J. Demsar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine Learning Research*, pp. 1–6, 2006.
4. D. A. Berry, "Logarithmic transformations in ANOVA," *Biometrics*, vol. 43, no. 2, pp. 439–456, 1987.
5. M. Friedman, "The use of ranks to avoid the assumption of normality implicit in the analysis of variance," *Journal of the American Statistical Association*, vol. 32, no. 200, pp. 675–701, 1937.
6. M. Friedman, "A comparison of alternative tests of significance for the problem of m rankings," *The Annals of Mathematical Statistics*, vol. 1, no. 1, pp. 86–92, 1940.
7. P. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*. Pearson, 2006.
8. D. Aha and D. Kibler, "Instance-based learning algorithms," *Machine Learning*, vol. 6, pp. 37–66, 1991.
9. N. Friedman, D. Geiger, and Goldszmidt, "Bayesian network classifiers," *Machine Learning*, vol. 29, pp. 131–163, 1997.
10. M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: An update," *SIGKDD Explorations*, vol. 11, 2009.
11. "UCI Machine Learning Repository." http://archive.ics.uci.edu/ml/, 2012.
12. Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika*, no. 12, pp. 153–157, 1947.
13. A. F. Clark and C. Clark, "Performance Characterization in Computer Vision: A Tutorial."
14. D. Liddell, "Practical tests of 2 2 contingency tables," *Journal of the Royal Statistical Society*, vol. 25, no. 4, pp. 295–304, 1976.
15. I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 2011.