An Evaluation of the Value Added by Informative Metrics

Colin Bellinger, Jeffrey Lalonde, Michael W. Floyd, Vikram Mallur CBELLING@SCS.CARLETON.CA
Carleton University, 1125 Colonel By Drive, Ottawa Ontario, Canada

Elkanzi Elkanzi, Diman Ghazi, Jennifer He, Alain Mouttham
Martin Scaiano, Elaine Wehbe, Nathalie Japkowicz NAT@SITE.UOTTAWA.CA
University Of Ottawa, 800 King Edward Ave. Ottawa Ontario, Canada

Abstract

We present an empirical study of the value added by four commonly employed classifier evaluation metrics and consider whether the informedness of a metric is correlated with its ability to predict the classifier's performance on future datasets, in terms of the true and false positive rates. In doing so, a variety of UCI datasets are examined in order to test the metrics, in general and under specific domain conditions. Through the experiments, we show there is no direct correlation between the informedness of a metric and its ability to predict future performance.

1. Introduction

Classifier evaluation is typically undertaken by employing one of several classes of performance metrics (Caruana & Niculescu-Mizil, 2004): threshold, ranking and probabilistic. Additionally, ensembles of metrics are occasionally used in order to use metrics from multiple classes.

Accuracy (ACC) is an example of a threshold metric; it relays the number of correct classification decisions made by the classifier. It has been noted (Kononenko & Bratko, 1991) that the simplicity of this metric can produce unexpected results under certain domain conditions, such as imbalance.

For this reason, alternative metrics, from the ranking or probabilistic classes, are often preferred. Such metrics are often described as being more informative, as they base their judgements on a wider variety of information. The *area under the ROC curve (AUC)* (Fawcett, 2006), for example, can take true positives and false positives or true and false negatives, at various thresholds, into consideration. Some other metrics commonly considered more informative, are *root mean square error (RMSE)* and the *K&B information score (K&B)* (Kononenko & Bratko, 1991). While the ranking and probabilistic classes of metrics base their decisions on more information, it remains to be seen whether this extra knowledge produces a more accurate prediction of future performance.

The remainder of this paper is formatted as follows: Section 2 provides a motivating scenario. Section 3 describes the experimental methodology and Section 4 provides the results. In Section 5, a discussion of the results is undertaken and Section 6 contains our concluding thoughts.

2. Motivation

Studies, such as (Rosset, 2004), suggest that the use of different metrics for model selection and testing may be beneficial. Related studies on model optimization (Huang & Ling, 2007), suggested that the more informative metrics are indeed beneficial to the optimization process. However, later work (Huang et al., 2008) contradicts the original findings and suggests that when proper statistical tests are applied, model optimization can achieve better results if it is conducted with the metric that will be of interest during future applications.

In this paper, we carry these ideas forward to model selection. In particular, we consider the scenario in which model selection aims to maximize the *true positive rate (TPR)* or minimize the *false positive rate (FPR)*, on a particular class, during the future applications of the classifier. Within this scenario, we examine whether the more informative metrics are more acute predictors of future performance. The *TPR* and *FPR* have been selected to gage the performance of the individual models because their optimization represents two fundamental, albeit often opposing, goals in classification. Furthermore, this process enables us to evaluate the training metrics based on a single test metric, which in

the future could be altered to represent different goals.

3. Experiments

In this study, a series of experiments will be performed in order to test the hypothesis that the more informative metrics, namely *AUC*, *RMSE* and *K&B*, are more proficient than *ACC* at predicting the future performance of classifiers. The following subsections describe the datasets' classification techniques and evaluation approaches applied in the experiments.

3.1. Data

Datasets of the following four characteristics are of particular interest in this paper: binary, multi-classed, balanced and imbalanced, as it is conceivable that these variations may produce different results. Initially, seven binary and seven multi-classed datasets were selected from the UCI repository.

As a means to increase the number of the balanced sets, three multi-class datasets were converted to binary, which resulted in ten binary classification problems and seven mutli-class problems.

In order to determine if class distribution has an effect on the various metrics, the balanced datasets were manipulated to produce imbalanced sets. This approach ensures that any variation in the results obtained were truly a result of the domain conditions, and not other factors that may exist within the datasets. The manipulation process was carried out by randomly removing instances from the minority class, or minority classes, until the majority class represented 75 percent of the instances. This resulted in ten balanced binary datasets, ten imbalanced binary datasets, seven balanced multi-class datasets and seven imbalanced multi-class datasets.

As a final step, the datasets were randomly divided into training (66 percent) and testing (33 percent) sets.

3.2. Classifiers

In this experiment, *Naïve Bayes (NB)*, *Multilayer Perceptron (MLP)*, *J48 decision tree* and *k-nearest neighbour (IBK)* are applied to the classification tasks. As the performance of individual classifiers is not of interest, no steps are taken to optimize their performances. Instead, the classifiers are trained with ten-fold cross validation, and tested in Weka with the default settings.

Table 1. The correlation between four training metrics (rows) and the three test metrics (columns) with respect to the first dataset

	TPR	FPR
ACC	-0.316	0.800
K&B	0.853	-0.135
RMSE	0.211	0.600
AUC	-0.316	0.898

3.3. Evaluation

As a means to evaluate the predictive capabilities of the four metrics, a series of steps were performed. Initially, the performance of each classifier on each of the thirty-four datasets during the training and testing phases, according to the evaluation metrics, was recorded. Following this, a series of ranked lists were compiled in which the classifiers were sorted according to each of the metrics.

Once sorted, the degree of correlation between each training metric and the test metrics were calculated. Table 1 corresponds to the first dataset, and displays the correlation between each of the training metrics and the test metrics. These correlation scores are indicative of the metrics' ability to predict the future performance of the classifiers in terms of *TPR* and *FPR*. For example, *ACC* has a score of *-0.32*. Therefore, it has a slightly negative correlation with *TPR*. This indicates that during training on the first dataset, *ACC* did not accurately predict how the classifiers would perform on the test data. Alternatively, the correlation between *K&B* and *TPR* is quite close to one. Therefore, it was much more acute in its predictions. By ranking these correlation tables and compiling them into one, we begin to paint a picture that indicates the training metric or metrics most correlated with the test metrics. For example, Table 2 shows the first four ranked correlations between each training metric and *TPR* on the ten balanced binary datasets. In addition, the final row displays the summations over the ten domains with respect to each training metric. These summations illustrate the overall predictive power of each training metric. Because they represent the summation of ranks, the lower totals identify the metrics that are more correlated with the test results. In this case, *RMSE* has the smallest summation, therefore, we might conclude that it is most correlated with *TPR*.

This process was repeated for each of the four categories discussed in Section 3.1, and is augmented with the combined group of the thirty-four datasets. The results of each of these experiments are disseminated

Table 2. The training metrics ranked in terms of their correlation to *TPR* on the first four balanced, binary datasets. The final row gives the sum of each training metrics' rank on the ten balanced, binary datasets

Dataset	ACC	K&B	RMSE	AUC
d_1	3.500	1.000	2.000	3.500
d_2	2.500	1.000	2.500	4.000
d_3	2.500	2.500	2.500	2.500
d_4	3.000	3.000	1.000	3.000
	.	.	.	.
total	27.000	25.000	23.500	24.500

Table 3. The rank correlation totals for the training metrics ranked with respect to each of the test metrics over all thirty-four datasets.

	TPR	Rank(TPR)	FPR	Rank(FPR)
ACC	83	2	78	1
KnB	88	4	81	2
RMSE	82	1	94.5	4
AUC	87	3	82.5	3

in the next section.

4. Results

In this section, the results produced in the four domain categories are reported. Prior to that, the results, which ensued from the combining of the thirty-four datasets into an all inclusive group, are revealed. These results are visualized in a series of tables, which depict the measured correlations between each training metric and the two test metrics with respect to the datasets in the particular category. The correlation scores contained in the tables were produced precisely as those in the final row of Table 2 were constructed. In addition, the training metrics are ranked according to the degree of correlation, in order to highlight their relative strengths and weaknesses.

4.1. Overall predictive power

The correlation scores displayed in Table 3 specify the relationship between the training metrics and each test metric over all thirty-four datasets. When the *TPR* is identified as the property for which the model is to be optimized, *RMSE* and *ACC* produce the best correlation score. Alternatively, *ACC* and *k&B* are most correlated with the *FPR*.

Table 4. The rank correlation totals for the training metrics ranked with respect to each of the test metrics over the binary datasets.

	TPR	Rank(TPR)	FPR	Rank(FPR)
ACC	49.5	2	47.5	2
KnB	52	4	50.5	3
RMSE	48	1	57.5	4
AUC	50.5	3	44.5	1

Table 5. The rank correlation totals for the training metrics ranked with respect to each of the test metrics over the multi-class datasets.

	TPR	Rank(TPR)	FPR	Rank(FPR)
ACC	33.5	1	30.5	1.5
KnB	36	3	30.5	1.5
RMSE	34	2	37	3
AUC	36.5	4	38	4

4.2. Cardinality

The results in this subsection highlight the effect of moving from a binary classification problem to one of a multi-class persuasion. An examination of the ranking of the metrics, with respect to *TPR*, in Table 4 and Table 5, reveals that the top two metrics remain the same regardless of the datasets' cardinality. However, their relative position is reversed. In terms of the *FPR*, *ACC* is consistently one of the most correlated training metrics. Alternatively, *AUC* has the strongest correlation with respect to the binary datasets. It is least correlated in the multi-class problem, where *k&B* has a strong correlation with *FPR*.

4.3. Class Balance

Table 6 and Table 7 display the correlation scores produced on the balanced and imbalanced datasets, respectively. Similar to Section 4.2, in which the goal is to optimize the *TPR*, *RMSE* and *ACC* have the strongest correlation. Also reminiscent of this, is the fact what when the nature of the datasets in question change, the relative ranking of the two primary metrics reverse. These properties are also consistent in the cases where the *FPR* is the target for optimization. However, in this case *ACC* and *K&B* are the training metrics most correlated to the test metric of interest.

5. Discussion

In order to determine how meaningful the results reported in Section 4 were, Friedman's rank test (Fried-

Table 6. The rank correlation totals for the training metrics ranked with respect to each of the test metrics over the balanced datasets.

	TPR	Rank(TPR)	FPR	Rank(FPR)
ACC	41.5	2	36	1
KnB	44	3	41.5	2
RMSE	40	1	48.5	4
AUC	44.5	4	44.5	3

Table 7. The rank correlation totals for the training metrics ranked with respect to each of the test metrics over the imbalanced datasets.

	TPR	Rank(TPR)	FPR	Rank(FPR)
ACC	41.5	1	42	2
KnB	44	4	39.5	1
RMSE	42	2	46	4
AUC	42.5	3	42.5	3

man, 1937) was applied. This test judges the significance of the rankings by testing a null hypothesis, which states that the rankings are the results of pure chance. This null hypothesis could not be rejected in any of the five categories discussed, at either the $\alpha_{0.05}$ or $\alpha_{0.1}$ levels. This is the case, both when the model is to be optimized for *TPR* and for *FPR*. Therefore, when these thirty-four datasets are considered collectively, or when they are segregated based on domain conditions, we cannot in fact assume with confidence that any one of these metrics is superior in model selection.

Also worth noting, is that by selecting *TPR* and *FPR* as the test metrics, we are simulating a situation where there is a particular class of interest. For example, when *TPR* is applied, one class is singled out and the true positive rate produced on that class is recorded. A consequence of this is that *AUC*, which also focuses on an individual class of particular interest, can be directly applied in the multi-class domain, and a more comprehensive approach, such as the Volume Under the ROC Curve (VUS) (Mossman, 1999) can be forgone.

6. Conclusions

In this paper, we tested the hypothesis that the more informative metrics are superior to *ACC*, in the task of model selection. In particular, we examined whether *K&B* , *RMSE* and *AUC* were more proficient at selecting the model that would produce optimal results in terms of the *TPR* and *FPR* on future classification tasks. The results indicate that regardless of domain conditions, none of these metrics are significantly more predictive than the others, with respect to model selection of this nature.

This paper presents our initial work on the predictive power of evaluation metrics. Moving forward, a broadening of the base from which the datasets are selected is an item of particular interest. Aside from diversifying and increasing the size of the experiment, this will help mitigate the effect of the less challenging classification task. These datasets result in trivial results and were found to be ineffective for our purposes.

Finally, *ACC* has much in common with *TPR* and *FPR*. It is therefore entirely conceivable that *ACC* had an advantage in this experiment, and that repeating this experiment with alternate optimization goals, would produce new results. Thus, future experiments, which judged the training metrics based on alternative criteria, are of interest.

References

Caruana, R., & Niculescu-Mizil, A. (2004). Data mining in metric space: An empirical analysis of supervised learning performance criteria. *Proceedings of the Tenth International Conference on Knowledge Discovery and Data Mining* (pp. 69 – 78).

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*, 861–874.

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, *32*, 675–701.

Huang, J., & Ling, C. (2007). Constructing new and better evaluation measures for machine learning. *Proceedings of the Twenty-First International Joint Conference on Artificial Intelligence* (pp. 859–864).

Huang, J., Ling, C., Zhang, H., & Matwin, S. (2008). Proper model selection with significance test. *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases* (pp. 536–547).

Kononenko, I., & Bratko, I. (1991). Information-based evaluation criterion for classifier's performance. *Machine Learning*, *6*, 67–80.

Mossman, D. (1999). Three-way ROCs. *Med Decis Making*, *19*, 78–89.

Rosset, S. (2004). Model selection via the AUC. *Proceedings of the Twenty-First International Conference on Machine Learning*.