Proper Model Selection with Significance Test

Jin Huang[1], Charles X. Ling[2], Harry Zhang[3], and Stan Matwin[1]

[1] School of Information Tech. and Eng., University of Ottawa, Canada
jhuang33@site.uottawa.ca, stan@site.uottawa.ca
[2] Department of Computer Science, The University of Western Ontario, Canada
cling@csd.uwo.ca
[3] Faculty of Computer Science, University of New Brunswick, Canada
hzhang@unb.ca

Abstract. Model selection is an important and ubiquitous task in machine learning. To select models with the best future classification performance measured by a *goal metric*, an *evaluation metric* is often used to select the best classification model among the competing ones. A common practice is to use the same goal and evaluation metric. However, in several recent studies, it is claimed that using an evaluation metric (such as AUC) other than the goal metric (such as accuracy) results in better selection of the correct models. In this paper, we point out a flaw in the experimental design of those studies, and propose an improved method to test the claim. Our extensive experiments show convincingly that only the goal metric itself can most reliably select the correct classification models.

1 Introduction

Model selection is an important task in machine learning and data mining. In classification tasks model selection attempts to select the model with the best future classification performance from (possibly a large number of) competing models. For example, when we build artificial neural networks for face recognition, we may vary the number of hidden nodes and other parameters to build many classification models, and select the best one. Other examples of model selection include choosing the optimal parameter setting for the Support Vector Machines, determining the most suitable amount of pruning in building decision trees, and so on. Clearly, model selection is ubiquitous in machine learning and machine learning applications.

A leading empirical approach for model selection in classification is to use the holdout set, and proceeds as follows: to select the model with the best future classification performance measured by a *goal metric*, competing models are evaluated by an *evaluation metric*, possibly the same as the goal metric, on the holdout set.[1] A natural preference is to use the same metric for both the goal and evaluation. For example, if the goal is to obtain a classification model with

[1] Other empirical approaches are reviewed in Section 2. The goal metric is also called performance metric, and the evaluation metric is called selection metric [1].

W. Daelemans et al. (Eds.): ECML PKDD 2008, Part I, LNAI 5211, pp. 536–547, 2008.

the highest accuracy on the test sets, accuracy is used to select the most accurate model on the holdout sets. The intuition is that if your child wants to achieve the highest SAT score (the goal metric), you child should train to obtain high SAT scores (the evaluation metric) on (previous) SAT tests (the holdout sets), rather than train to obtain high scores on the GRE test (a different metric).

Rosset [2] recently conducted an empirical research on model selection for binary classification with two specific metrics: accuracy and AUC (Area Under the ROC Curve; see Appendix). He compared the suitability of AUC and accuracy as the model evaluation metrics when the goal metric is accuracy. He showed that AUC can better select the correct models than accuracy even when the goal metric is accuracy. The results are quite surprising and puzzling, and no convincing explanations were given. Nevertheless, several other papers [1,3,4] have since confirmed his finding. For example, Huang and Ling [3] studied model selection with many popular machine learning metrics and claimed that often an evaluation metric different from the goal metric can better select the correct models. Skalak and Caruana [1] used absolute loss to compare model selection abilities of various metrics, and drew similar conclusions. It now seems to be a well-regarded conclusion in the machine learning community that a different metric can do a better job in model selection.

In this paper we point out a potential flaw in the experimental design of these model selection studies: the variance of these metrics when applied to randomly sampled datasets was not taken into consideration. Suppose we have two competing models, X and Y, to be selected by an evaluation metric h. If we say X is better than Y, it should really mean that $E(h(X))$ (E is the expected value or mean of a random variable) is "reliably better" than $E(h(Y))$ on the sampled datasets, rather than simply $E(h(X)) > E(h(Y))$, as in the previous studies.[2] The same is true for the goal metric g. This is because metrics have variances when applied to randomly sampled datasets, and thus, $E(h(X)) > E(h(Y))$ by a minute amount does not really indicate that X is better than Y. Thus, a significance test must be employed in comparison so that the conclusion on which model is indeed better is reliable.

We propose an improved method for proper model selection by incorporating statistical significance tests in the comparison, and carefully re-implementing Rosset's experiments with more datasets and algorithms. Not surprisingly, "AUC can better select models measured by accuracy" is no longer true. We then include more metrics and more model selection approaches, and show convincingly that, *in all cases*, the goal metric itself is the best evaluation metric for model evaluation. We hope that with the proper model selection method proposed here, we can settle this controversial issue once for all.

2 Review of Previous Works

Model selection has been extensively studied by researchers in the machine learning and statistics communities. Here we review several relevant approaches.

[2] We normalize all metrics in this paper so that the larger the value, the better the model.

Rosset [2] performed extensive experiments to study the suitability of AUC and accuracy to select highly accurate models. He assumed that the training set is very small (10% of the original dataset), and thus no data can be used as the holdout sets in model selection. Therefore, he used "the test sets approach" in his study of model selection.

Rosset's experiments are conducted in the following way: First, the original dataset is split randomly into the training set (10%) and the test set (90%). Second, two competing models X and Y (they can be two decision trees with different pruning levels) are built on the training set. Third, X and Y's performance on the test set is estimated. This is done by splitting the test set into 100 equal-sized, stratified[3] subsets, and applying X and Y on the 100 subsets by the goal metric g (i.e., accuracy). The average (mean) g value on the 100 test subsets is denoted as $E(g(X))$ and $E(g(Y))$ respectively. If $E(g(X)) > E(g(Y))$, then model X is regarded as better than model Y on the test set, in terms of the goal metric g. If $E(g(X)) = E(g(Y))$, then they are regarded as the same. Note that no significance test is performed for the comparison, and thus, such conclusion may not be reliable.

In the next step, models X and Y are evaluated by the evaluation metric h (AUC or accuracy) on each of the 100 test subsets. For each subset, if $h(X)$ and $h(Y)$ are consistent with $E(g(X))$ and $E(g(Y))$, then h selects the model correctly. That is, if $h(X) > h(Y)$,[4] and $E(g(X)) > E(g(Y))$, then h selects the right model; otherwise h selects the wrong one. The number of times (among 100) when h selects the correct model is noted, and this represents how well h can select the correct model. The larger the number, the better the evaluation metric in model selection. Using this method, Rosset showed that AUC can better select the correct models than accuracy when the goal metric is accuracy.

Huang and Ling [3] explored the performance of model selection for classification tasks for eight popular machine learning metrics of accuracy, AUC, lift, break-even point, F-metric, average precision, RMS (Root Mean Square) error, and MXE (Mean Cross Entropy) [5] by following Rosset's method (without using the significance test). They showed that generally the metrics of RMS and MXE are the best evaluation metrics, followed by AUC, average precision, and F-metric, no matter what goal metrics are. That is, better model selection can be achieved with a different evaluation metric from the goal metric.

Skalak and Caruana [1] studied the robustness of the evaluation metric when the goal metric is unknown. They used the absolute loss for measuring the model selection error but again no variance is calculated for the loss. They showed that when the models are well calibrated for the probability outputs, and the available holdout data is limited, MXE is the most robust one, while other metrics, such as F-metric, lift and accuracy, performed poorly.

[3] Stratified subsets refer to partitions of a dataset into equal-sized subsets with the same class distribution.

[4] Note that the significance test is not performed here because each subset is just one dataset so there is no variance. But again, if $h(X) > h(Y)$ by a minute amount, it is really not reliable to conclude that X is better than Y.

[5] See Appendix for more details on some of these metrics.

In addition to "the test sets approach" for model selection [2], another (more popular) empirical approach is to use separate holdout sets to estimate model's future performance. We will study this approach extensively later in the paper. Another empirical model selection approach is called complexity-penalization, which assigns a complexity value to each model and chooses the best model that minimizes a predefined trade-off function combining the model complexity and empirical error. Many variants of this approach were proposed, including structural risk minimization [5,6], the minimum description length principle [7], and regularization [8]. Kearns and Mansour [9] theoretically and empirically compared the above two approaches and demonstrated that in some cases the holdout set approach has an advantage of small generalization errors over the complexity-penalization method. In addition, the holdout set approach is very robust and hard to beat compared to other approaches [10,11].

3 Selecting Models Properly with Significance Test

Rosset studied "the test sets approach" of model selection on accuracy and AUC on the UCI dataset "adult" [12] with naive Bayes and k-nearest neighbour (KNN) learning algorithms. In the following subsections, we first replicate Rosset's experiments (without the significance test) but we include more UCI datasets and learning models. We use five UCI datasets: "adult" (included in Rosset's experiments), "letter", "kr-vs-kp", "page blocks", and "pen digits", in our experiment. In the original datasets, only "adult" and "kr-vs-kp" are binary; the rest are multiclass. All multiclass datasets are converted into binary datasets by assigning some classes to the positive class and the rest to the negative class. The properties of the datasets are shown in Table 1. The last column is the ratio of the positive class in the original and converted binary datasets.

We choose three popular learning algorithms in our experiments: decision tree, naive Bayes and k-nearest neighbour (the last two are included in Rosset's experiments). For each learning algorithm we build two competing models with different parameter settings. For the decision tree, we build two trees with and without pruning. For naive Bayes, we use different numbers of attributes in the datasets to train two classifiers. More specifically, we use the first 8 and 10 attributes for the "adult" dataset, first 25 and 35 attributes for "kr-vs-kp", first 10 and 11 attributes for "letter", first 5 and 8 attributes for "page blocks", and

Table 1. UCI datasets used in our experiments

Dataset	Size	Attribute #	Class #	Positive Class Ratio
Adult	30162	14	2	24.8%
Kr-vs-kp	28060	36	2	47.8%
Letter	20000	16	26	38.2%
Page blocks	5473	10	5	10.2%
Pen digits	10992	16	10	30%

Table 2. Ratio of correct model selection with accuracy and AUC using Rosset's method

Dataset	**Decision tree**		**KNN**		**Naive Bayes**	
	accuracy	AUC	accuracy	AUC	accuracy	AUC
Adult	0.74	0.68	0.65	0.93	0.53	0.67
Kr-vs-kp	0.77	0.74	0.88	0.73	0.78	0.75
letter	0.54	0.57	1	1	0.57	0.41
Page blocks	0.72	0.71	0.96	0.51	0.59	0.5
Pen digits	0.65	0.63	0.99	0.88	0.54	0.7

first 8 and 10 attributes for "pen digits". For the k-nearest neighbour, we build two models with $k = 5$ and $k = 50$.

3.1 Model Selection Using Test Sets

We first replicate Rosset's experiments (but with more datasets and learning algorithms) to explore the suitability of accuracy and AUC in model selection when the goal metric is accuracy. (See Rosset's experimental method in Section 2). We report our experimental results in Table 2. Our results confirm Rosset's finding: the ratios measuring correct model selection are higher when AUC is used as the evaluation metric than when accuracy is used, for the adult dataset with KNN and naive Bayes. However, it is surprising that in almost all other cases, accuracy can better select models than AUC. More specifically, in a total of 15 cases (3 learning algorithms and 5 datasets), accuracy is better than AUC in 10 cases, and the same in one case. As Rosset used limited datasets and learning algorithms, his conclusion that AUC can better select models than accuracy when the goal metric is accuracy seems to be unreliable.

Nevertheless, Rosset's experimental design (as well as those in several other studies of model selection [1,3,4]) may lead to unreliable conclusions. That is, it is unreliable to conclude which model is better when a simple comparison is used on two metrics or two means of metrics. To improve the experimental design, we use the significance test (both the t-test and the sign test are studied) in comparison. More specifically, following the notation in Section 2, when deciding which model has a better performance on the 100 test subsets, instead of simply comparing $E(g(X))$ and $E(g(Y))$, a paired t-test with 95% confidence level is performed to see if one is larger, or if they are not statistically significant. To decide which model is better using the evaluation metric h, Rosset simply compared $h(X)$ and $h(Y)$ on each of the 100 test subsets. For one test subset, the t-test cannot be performed. We modify Rosset's approach by re-partitioning randomly the test set into 100 test subsets, and compare the average evaluation scores, $E(h(x))$ and $E(h(Y))$, with the same paired t-test to see if one model is better, or if they are indifferent. If this outcome is the same as the one with the goal metric, then h selects the right model. Notice that we use the same number of subsets (100 here) and the same t-test (both with 95% confidence level) on both evaluation

Table 3. Ratio of correct model selection with t-test using the test sets

Dataset	**Decision tree**		**KNN**		**Naive Bayes**	
	accuracy	AUC	accuracy	AUC	accuracy	AUC
Adult	0.99	0.7	0.97	0.89	0.97	0.08
Kr-vs-kp	0.92	0.37	1	0.83	0.89	0.48
letter	0.97	0.42	1	1	0.93	0.18
Page blocks	0.98	0.62	1	0	0.98	0.19
Pen digits	0.93	0.53	1	0.94	0.93	0.59

and goal metrics. Thus, the "sensitivity" on the variance of the metrics is the same. See more discussions in Section 4.

This process is repeated 100 times and we obtain the number of cases when AUC and accuracy select the right model respectively. The results are shown in Table 3. From the table, we can see that with *all* datasets and *all* learning algorithms, accuracy always does better, and in most cases much better, than AUC (except for one case they tie). Accuracy achieves a very high ratio of correctness ($\geq$ 0.89) while AUCs scores are much lower.[6] For all 15 cases (3 learning algorithms and 5 datasets), AUC tends to choose the wrong model more often (ratios of correctness < 0.5) in 7 cases. These experimental results contradict the claim that AUC performs better than accuracy in model selection using the test sets, after the t-test is employed in comparison.

So far we use the t-test as the significance test in comparing means of metrics to draw reliable conclusion about model selection. Here we show that another popular significance test, the sign test [13] [7], is equally effective in proper model selection. Table 4 lists the outcome of correct model selection using the sign test, instead of the t-test as in Table 3. The results are similar to Table 3: in all cases, accuracy can better select more accurate models than AUC. We believe that the choice of the significance test does not matter, as long as the same significance test is applied on both the evaluation and goal metrics obtained from the sampled datasets. In the rest of the paper, only results with the t-test is reported.

3.2 Model Selection Using Holdout Sets

In this section we perform experiments using the holdout set approach of model selection with the proper significance test. In this approach, the original dataset is split randomly into three non-overlap sets: the training set (10%), the holdout

[6] Note that due to sampling variations between training, holdout, and test sets, occasionally accuracy may select the wrong model with the highest accuracy. Also, due to the consistency between accuracy and AUC (see Section 4), AUC may sometimes also choose the most accurate model. Accuracy is simply more likely than AUC in selecting the most accurate models.

[7] The sign test is a non-parametric test used to compare the distribution median with a given pair of data. This test could be used as an alternative for one-sample Student t-test. Unlike the t-test, the sign test can work with non-normal distributions.

Table 4. Ratio of correct model selection with sign test using the test sets

Dataset	**Decision tree**		**KNN**		**Naive Bayes**	
	accuracy	AUC	accuracy	AUC	accuracy	AUC
Adult	0.96	0.64	0.94	0.77	0.92	0.19
Kr-vs-kp	0.93	0.42	1	0.78	0.93	0.52
letter	0.94	0.33	1	1	0.90	0.24
Page blocks	0.95	0.67	1	0	0.97	0.21
Pen digits	0.88	0.54	1	0.92	0.94	0.52

set (45%) and the test set (45%). The holdout set and the test set are further partitioned into 50 subsets. Then two competing models are built on the training set, and they are applied to the 50 test subsets by the goal metric g. Their averages are compared with the paired t-test. The same is applied to the 50 holdout subsets by the evaluation metric h. If the two outcomes are the same, then h selects the correct model. The process is repeated 100 times and the number of cases of correct model selection is noted.

The results are shown in Table 5. We can see again that *in all cases* accuracy better selects the correct models than AUC (except for one case when the two tie). These results show that accuracy is again much better in selecting the correct model than AUC if the goal metric is accuracy using the holdout set approach.

We note that the numbers in Table 5 (using the holdout sets) are generally smaller than the corresponding ones in Table 3 (using the test sets). We believe that this is because in the holdout set approach, there is no overlap between the holdout set and the test set. Thus, it is less likely to select the right models determined by the test set using the holdout set.

3.3 Model Selection with More Metrics

In this section we explore the model selection ability of other popular metrics used in machine learning. For example, Root Mean Squared error (RMS) is widely used in regression to reflect the average deviation of the predicted values from the true ones. F-measure combines precision and recall, often used in information retrieval. Here we choose five commonly used metrics as both evaluation

Table 5. Ratio of correct model selection with t-test using the holdout sets

Dataset	**Decision tree**		**KNN**		**Naive Bayes**	
	accuracy	AUC	accuracy	AUC	accuracy	AUC
Adult	0.97	0.62	0.96	0.81	0.76	0.12
Kr-vs-kp	0.83	0.44	1	0.86	0.73	0.56
letter	0.74	0.47	1	1	0.65	0.19
Page blocks	0.88	0.61	1	0	0.78	0.13
Pen digits	0.88	0.56	1	0.9	0.55	0.54

Table 6. Average ratio of correct model selection using test sets

Goal ⇓	**Evaluation ⇒**	**accuracy**	**AUC**	**F**	**RMS**	**MXE**	**Average**
accuracy		0.96*	0.46	0.62	0.57	0.47	0.57
AUC		0.42	0.97*	0.32	0.77	0.74	0.7
F		0.6	0.31	0.97*	0.43	0.35	0.39
RMS		0.63	0.75	0.44	0.94*	0.77	0.78
MXE		0.49	0.74	0.34	0.78	0.98*	0.8
Average		0.6	0.68	0.37	0.78	0.81	0.96*

and goal metrics. The five metrics are: accuracy, AUC, F-measure, Root Mean Squared error (RMS), and Mean Cross Entropy (MXE). Definitions of these metrics can be found in Appendix.

Caruana and Niculescu-Mizil [14] suggested that when the goal metric is unknown during model construction and model selection, one could use the average of several different metrics as the goal metric. Here we use the average of the five single metrics mentioned above, and use it as both the goal and evaluation metrics. Thus, there is a total of six metrics in this experiment.

We perform the same model selection experiments with the t-test as in Sections 3.1 and 3.2. Each of the six metrics is used for the goal metric as well as for the evaluation metric. The results of the test set approach are shown in Table 6, with the goal metrics in the row and evaluation metrics in the column. Each number in the table is the average of 15 scores for that pair of metrics over five datasets and three learning algorithms. For example, 0.96 in Table 6 for accuracy to select accuracy is the average of 15 numbers in Table 3 for accuracy to select accuracy over five datasets and three learning algorithms. We put a * next to the largest number in each row to indicate that not only the average (of the 15 scores) is the largest, but each individual score is also larger or the same compared to another evaluation metric (as in the case for accuracy and AUC in Table 3). We can see that numbers in the diagonal line are the largest in each row, and each largest number has a * next to it. This suggests that in general when the evaluation and goal metrics are the same, correct models can always be more reliably selected. This is also true for the average of the five single metrics: when the goal metric is the average of the five metrics, using the same metric itself (the average) as the evaluation metric is best for model selection, compared to any other single metric. The results using the holdout sets are similar, as seen in Table 7. These results generalize the conclusion we obtain in the previous sections regarding accuracy and AUC. It shows convincingly that we should always use the same metric for evaluation and goal in model selection.

Although we can use the average of several metrics as a robust goal metric when the goal metric is unknown, sometimes we must choose a single metric for model construction, optimization, and selection. Tables 6 and 7 can also tell us which single metric is best if the goal metric is the average. From the bottom row of the two tables, we can see that, for both the test set approach and the holdout set approach, MXE has the largest ratio for correct model selection, followed by RMS, AUC, accuracy, and last, F. This means that MXE is the

Table 7. Average ratio of correct model selection using holdout sets

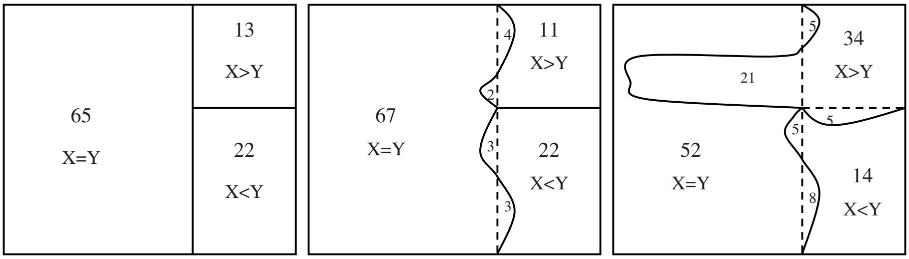
Goal ⇓ Evaluation ⇒	**accuracy**	**AUC**	**F**	**RMS**	**MXE**	**Average**
accuracy	0.85*	0.52	0.66	0.61	0.47	0.63
AUC	0.51	0.87*	0.4	0.8	0.77	0.72
F	0.67	0.39	0.91*	0.52	0.54	0.48
RMS	0.58	0.76	0.49	0.89*	0.81	0.8
MXE	0.43	0.76	0.52	0.8	0.9*	0.83
Average	0.59	0.72	0.5	0.8	0.82	0.89*

most robust metric for model evaluation if the goal metric is the average. This confirms with the conclusion of [1]. Only in this situation should we use a metric (such as MXE) different from the goal metric (such as the average) in model selection.

4 Discussion

In this section we investigate why contradicting conclusions on accuracy and AUC can be drawn with and without the t-test. We provide a detailed analysis for selecting the more accurate model by accuracy or AUC from two competing decision trees on the "pen digits" dataset using the holdout sets (as discussed in Section 3.2). We denote the two competing decision tree models as X and Y. Among 100 repeated runs (cases) of model selection with the t-test, there are 13 cases when X is better than Y, 65 cases when X equals to Y, and 22 cases when X is worse than Y, all evaluated by accuracy with the t-test on the test sets. This is regarded as the "ground truth". We depict this distribution roughly in Figure 1(a).

We then count how each of the three outcomes ($X > Y$, $X = Y$, and $X < Y$) is classified by the evaluation metrics of accuracy and AUC respectively with the t-test. The results are depicted in Figure 1 (b) and (c) respectively. From



(a) Goal metric accuracy (b) Evaluation metric accuracy (c) Evaluation metric AUC

Fig. 1. The distribution of cases that goal metric accuracy, evaluation metric accuracy and AUC evaluate models X and Y

Figure 1 (b), we can see clearly that when accuracy is used to select models, there is a small number of confusing cases between $X > Y$ and $X = Y$ (6 cases), and between $X < Y$ and $X = Y$ (6 cases). This is again due to sampling variations in the training, holdout, and test datasets. There is no confusing case between $X > Y$ and $X < Y$. However, when AUC is used as the evaluation metric, we can see from Figure 1 (c) that there is a huge number of confusing cases between $X > Y$ and $X = Y$ (26 cases), and the number of confusing cases between $X < Y$ and $X = Y$, and between $X > Y$ and $X < Y$ are also larger (13 cases and 5 cases respectively).

The explanation is that AUC and accuracy *are* different metrics. As shown in [15], although AUC and accuracy are largely consistent, there are cases when accuracy and AUC contradict each other. Furthermore, AUC is more "sensitive" (or discriminating) than accuracy [15]. That is, AUC often treats two objects with the same accuracy as different. This implies that AUC is likely to be "too sensitive" in comparing two objects when their accuracy values are statistically indifferent. We can see this from Figure 1 (c): there is a large number of $X = Y$ cases (21 cases) being identified as $X > Y$.

One might argue that overly sensitive metric may not be a problem if we only care about correctly identifying *different* models. That is, if two models are statistically different, we must identify the better one; but if they are not, it will not hurt if we say one of them is better. Under this assumption, if we calculate the correct "recall" of the $X > Y$ and $X < Y$ cases (a total of $13 + 22 = 35$ cases), we can see from Figure 1 (b) that the recall with accuracy is $(13 - 4) + (22 - 3) = 28$. On the other hand, from Figure 1 (c), the recall with AUC is $(13-5)+(22-8-5) = 17$. Thus, under the assumption that we only care about correctly identifying the statistically different models, accuracy is still better than AUC for this dataset. If one assumes that correctly identifying if a model is better or indifferent statistically is equally important in model selection, our results in previous sections show that in all cases, the goal metric should be used to select the right models, and such model selection outcomes (with the significance test) are reliable.

From this analysis, we can conclude that different metrics may contradict each other, and may have different sensitivity. Without the significance test in model selection, it is not reliable to select better models, and such a study may lead to the wrong conclusion that an evaluation metric different from the goal can better select models. When the paired t-test or sign test is used in comparing models' performance, we conclude convincingly that we should always use the goal metric to evaluate and select models.

5 Conclusion

In this paper we investigate model selection with different metrics. We first point out a flaw in the experimental design in several previous studies which led to an erroneous conclusion that a different evaluation metric can better select models. The problem lies in the lack of significance test when comparing competing models. This may result in statistically indifferent models being regarded as different, and vice versa. With the proper use of the significance test (such as the t-test and the sign test) in model selection, we show convincingly that in all cases (with six metrics, three learning algorithms, five UCI datasets, and using the test sets approach or the holdout sets approach), the same goal metric is the best evaluation metric for model selection.

Occasionally the goal metric and evaluation metric cannot be the same. For example, the goal metric may be unknown during model selection. We will study this problem further in our future work.

References

1. Skalak, D.B., Niculescu-Mizil, A., Caruana, R.: Classifier loss under metric uncertainty. In: Kok, J.N., Koronacki, J., Lopez de Mantaras, R., Matwin, S., Mladenič, D., Skowron, A. (eds.) ECML 2007. LNCS (LNAI), vol. 4701, pp. 310–322. Springer, Heidelberg (2007)
2. Rosset, S.: Model selection via the AUC. In: Proceedings of the 21st International Conference on Machine Learning (2004)
3. Huang, J., Ling, C.: Evaluating model selection abilities of performance measures. In: Proceedings of the Workshop on Evaluation Methods for Machine Learning at the 21st National Conference on Artificial Intelligence (AAAI 2006) (2006)
4. Wu, S., Flach, P., Ferri, C.: An improved model selection heuristic for AUC. In: Kok, J.N., Koronacki, J., Lopez de Mantaras, R., Matwin, S., Mladenič, D., Skowron, A. (eds.) ECML 2007. LNCS (LNAI), vol. 4701, pp. 478–489. Springer, Heidelberg (2007)
5. Vapnik, V.: Estimation of Dependences Based on Empirical Data. Springer, New York (1982)
6. Vapnik, V.: Statistical Learning Theory. Wiley, New York (1998)
7. Rissanen, J.: Stochastic complexity and modeling. Annals of Statistics 10(3), 1080–1100 (1986)
8. Moody, J.E.: The effective number of parameters: An analysis of generalization and regularization in nonlinear learning systems. In: Advances in Neural Information Processing Systems — 4 (NIPS 1992) (1992)
9. Kearns, M., Mansour, Y., Ng, A.Y., Ron, D.: An experimental and theoretical comparison of model selection methods. Machine Learning 27, 7–50 (1997)
10. Efron, B.: Computers and the theory of statistics: Thinking the unthinkable. SIAM Review 21, 460–480 (1979)
11. Weiss, S., Kulikowski, C.: Computer Systems that Learn: classification and prediction methods from statistics, neural networks, machine learning, and expert systems. Morgan Kaufmann, San Mateo (1991)
12. Blake, C., Merz, C.: UCI repository of machine learning databases. University of California, Irvine, Dept. of Information and Computer Sciences (1998), `http://www.ics.uci.edu/~mlearn/MLRepository.html`
13. Weisstein, E.W.: Fisher sign test. MathWorld–A Wolfram Web Resource
14. Caruana, R., Niculescu-Mizil, A.: Data mining in metric space: An empirical analysis of supervised learning performance criteria. In: Proceedings of the 10th ACM SIGKDD conference (2004)

15. Ling, C.X., Huang, J., Zhang, H.: AUC: a statistically consistent and more discriminating measure than accuracy. In: Proceedings of 18th International Conference on Artificial Intelligence (IJCAI-2003), pp. 519–524 (2003)
16. Hand, D.J., Till, R.J.: A simple generalisation of the area under the ROC curve for multiple class classification problems. Machine Learning 45, 171–186 (2001)

Appendix: Metrics Used in This Paper

Accuracy: Accuracy is the most commonly used performance metric in Machine Learning. For a classification task, accuracy is the percentage of the correctly classified examples in all examples.

F-metric: F-metric combines precision and recall as a single metric. It is defined as the harmonic mean of the precision and recall.

$$F = \frac{2 * precision * recall}{precision + recall}$$

AUC: The Area Under the ROC Curve, or simply AUC, is a single-number metric widely used in evaluating classification algorithms. AUC reflects the overall ranking performance of a classifier. For a binary ranked list, Hand and Till [16] present the following simple formula to calculating AUC

$$AUC = \frac{S_0 - n_0(n_0+1)/2}{n_0 n_1}$$

where S_0 is the sum of the ranked positions of all positive examples. n_0 and n_1 are the numbers of positive and negative examples.

RMS: Widely used in regression, RMS (Root Mean Square error) reflects the average deviation of all predicted values from the true values. For K instances, suppose that the true probability value and the predicted probability value for an instance I_i are $Tar(I_i)$ and $Pred(I_i)$,

$$RMS = \sqrt{\frac{1}{K}\sum_{i=1}^{K}[Tar(I_i) - Pred(I_i)]^2}$$

MXE: MXE (Mean Cross Entropy) is used to measure in average how close all predicted probabilities are to the true probabilities. It can be shown that minimizing the cross entropy gives rise the maximum likelihood hypothesis. When using the same notations as in RMS, MXE is defined as

$$MXE = -\frac{1}{K}\sum_{i=1}^{K}\{Tar(I_i) * log[Pred(I_i)] + (1 - Tar(I_i)) * log[1 - Pred(I_i)]\}$$