# Classifier Evaluation: A Need for Better Education and Restructuring

**Nathalie Japkowicz**
School of Information Technology and Engineering
University of Ottawa, Ottawa, Ontario K1N 6N5
**nat@site.uottawa.ca**

## Abstract

Classifier evaluation is often performed following a de-facto strategy that involves the use of accuracy, 10-fold cross-validation and a paired t-test. Various aspects of this strategy have previously been criticized, but, in most cases, to little avail: the strategy is still very popular. This paper questions why this is the case and suggests two explanations. On the one hand, the criticisms may have been ignored because they were issued in isolation of one another. On the other hand, they may have been ignored because they are inconvenient. The paper addresses the first explanation by unifying the various facets of this criticism within a single document. It addresses the second by issuing a couple of recommendations to the machine learning community.

## 1. Introduction

The field of classification is at a crossroad. On the one hand, it has matured to the point where it has developed a plethora of impressive and useful classification methods, each with different strengths and advantages. On the other hand, it is now overflowed by hundreds of studies trying to improve the basic methods but only marginally succeeding in doing so [1]. While the machine learning community keeps itself busy developing these new improvements, the applied world, or consumers of our research, remains sceptical about its worth. This is caused, in great part, by the fact that our often ritualized approach to classifier evaluation lacks the depth necessary to help us focus on worthwhile research improvements or convince potential users of its worth.

Indeed, classifier evaluation, while branded by most machine learning scientists as essential is, in general, poorly understood and performed automatically by blindly following a set of rules accepted by the community at large with no concerns about the meaning of these rules or the fact that they do not apply in every case. As a result of this poor understanding, a large portion of researchers has come to rely on this procedure and accept its results as near representative of the truth. This is, in fact, wrong.

Over the past ten to fifteen years, several important papers have rung the alarm bell ([2], [3], [4], [5]). While their arguments were noted, they have, for the most part been ignored. (See, for example, the survey in [5]). There may be several reasons why this has been the case. The first one, and the one mostly addressed in this paper, could be the fact that each of these authors pointed to different aspects of our evaluation framework that leaves to be desired in an isolated fashion. To address this problem, this paper seeks to gather all the issues that have been considered by the above authors and others ([6], [7], [8]) and show the different challenges that plague the framework.

The second reason, potentially, explaining why so little interest has generally been given to the arguments criticizing our evaluation framework may be of a more social nature. By this I mean that these criticisms are inconvenient and their conclusions not universally recognized and enforced by the reviewers of conference and journal articles. First, understanding and implementing new evaluation schemes is not easy and can be quite time-consuming. Second, a researcher may conclude that since no one else does it, why should he or she 'waste' his or her time with this issue. Similarly, reviewers may feel that since they are not implementing these schemes themselves, they should not hold other researchers to strict standards. Mixed with this could also be the fear that the results obtained by a new evaluation scheme may not be encouraging. Because the community as a whole has not adopted the new standards, a researcher adhering to them may be unfairly penalized as compared to one who uses less stringent evaluation methods.

Thus, the purpose of this paper is both to reiterate, within the same document the different issues regarding classifier evaluation that have previously been pointed out, but more or less ignored, and to suggest ways for the community to break away from the vicious circle in which it is caught.

I begin with a discussion of the problems emanating from the current status quo and of their proposed solutions. I, then, make a couple of recommendations regarding the steps that could be taken to ensure that classifier evaluation be done more rigorously.

## 2. Issues with Machine Learning Evaluation

For the past 20 years, with [9] suggesting the need for a greater emphasis on performance evaluation, the machine learning community has recognized the importance of proper evaluation. Over the years, a de-facto procedure has been used by the majority of researchers involved in experimental work. This procedure consists of selecting an evaluation metric (often accuracy), selecting a large enough number of domains (often chosen from the UCI Repository for machine Learning), selecting a convincing number of previously designed strong learning algorithms to be compared to one another or pitted against a new proposed method, and running (stratified or not) 10-fold cross-validation experiments on each domain, possibly, repeating these experiments several times on different shufflings of the data. Once these experiments are completed, it is customary to apply paired t-tests to all pairs of results or to all pairs of results that include the new algorithm of interest and to average the results obtained by each classifier on each domain or to record the number of wins, ties and losses experienced by each algorithm with respect to the others. (See [10], for example, for a more detailed discussion of the procedure). We now discuss, in turn, three categories of problems associated with this approach.

### 2.1 Problems with Evaluation Metrics

In the realm of all the issues related to classifier evaluation, those concerning evaluation metrics have by far been given the most attention, both in terms of discussions and following. The metric most commonly used by machine learning researchers and practitioners is accuracy [5]. Yet, accuracy suffers from a serious shortcoming: it does not take misclassification costs into consideration. This is a serious issue in practical research since there is almost always an unequal misclassification cost associated with each class.

This problem was recognized early. In [11], Kononenko and Bratko proposed an information-based approach that takes this issue into consideration, along with the questions of dealing with classifiers that issue different kind of answers (categorical, multiple, no answer or probabilistic) and comparisons on different domains. Their method did not receive large scale acceptance, possibly because it relies on knowledge of the cost-matrix and prior class probabilities, which cannot be estimated that accurately.

More successful has been the effort initiated in [4] which introduced ROC Analysis to the Machine Learning community. ROC Analysis allows an evaluator not to commit to a particular class prior distribution nor to a particular cost matrix. Instead, it displays the classifier's performance over all the possible priors and costs. ROC Analysis is graphical in nature and has not had much success in and of itself. Its associated metric, however, the Area under the Curve (AUC), has now become relatively popular especially in cases of class imbalances.

Other evaluation metrics are also, often, used in specific domains. For example, the area of text categorization often uses metrics such as Precision, Recall, and the F-Measure. In medical application, it is not uncommon to encounter results expressed in terms of sensitivity and specificity, as well as in terms of positive and negative predictive values.

More recently, [6] engaged in a comparison of classifiers on a number of domains that uses eight performance metrics divided into three categories: the *threshold metrics* (Accuracy, F-Measure and Lift), the *ordering/rank metrics* (Area under the curve, average precision and precision/recall break-even point) and the *probability metrics* (root-mean square and cross-entropy). In [12], they compared the various evaluation metrics using correlation analysis and found that root mean square is the metric that is best correlated with all the others. It can, thus, be seen as a good compromise if the particular criteria of interest to the evaluator are not clearly laid out.

Another issue relating to evaluation metrics is that of aggregation of the results obtained by different classifiers on different domains. Sometimes, the results are averaged for each classifier over all the domains. This is a mistake since the same value may take different meanings depending on the domain. Recognizing this problem, researchers sometimes use a

win/tie/loss approach, counting the number of times each classifier won over all the others, tied with the best or lost against one or more. This approach, however, ignores any kind of information pertaining to how close classifiers were to winning or tieing. The best alternative would be to refrain from aggregating the results, but since that may not always be practical, [8] attempts to provide a visualization technique for dealing with the problem.

### 2.2 Problems with Sampling approaches and Statistical Tests

The statistical test most often used in machine learning experiments is the cross-validated paired t-test. This test is usually applied without much concern about the assumptions upon which it depends (normal distribution of the data to which it is applied or sufficient data in the testing set to ensure that the assumption of a normal distribution is acceptable). Perhaps even more serious, is the fact that many researchers using it are often unaware of the true significance of this test. In particular, they do not always consider all the uncertainty revolving around statistical tests. [2] and [5] provide some alternatives to the t-test, which we now summarize.

In [2], Dietterich considers five different statistical tests. He compares these tests based on two quantities: the *Type I Error* of the test—the probability of incorrectly detecting a difference when no such difference between two classifiers exists. The *Power* of the test—the ability to detect algorithm differences when such differences do exist. His experiments suggest that the most often used statistical test, the paired difference t-test based on 10-fold cross-validation has high power as compared to the other tests, but unacceptable Type I error. On the other hand, he concludes that both *McNemar's test* and *5x2CV* present good compromises with respect to Type I error and Power.

In [5], Demšar discusses several parametric and non-parametric tests for both the comparisons of two algorithms and that of several algorithms. For the case of two algorithms, he suggests the use of the non-parametric *Wilcoxon test* which, although less powerful than the t-test when the t-test's assumptions are verified, can be more powerful when these assumptions are violated. In the case of more than two algorithms, he similarly recommends a non-parametric alternative to ANOVA, namely, the *Friedman test*.

Another related issue is that of sampling. Most researchers in the field today apply 10-fold cross-validation, which makes intuitive sense, but causes accrued uncertainty in the ensuing commonly used t-test, because the learned classifiers are not independent of each other. When cross-validation is further repeated, the independence assumption between the test sets is then violated in addition to the one concerning the classifiers. As an alternative to cross-validation, another branch of statistics that, so far, has practically eluded the machine learning community, is the newer field of re-sampling statistics. The two areas of research in re-sampling statistics are Bootstrapping and Randomization. *Bootstrapping* has attracted a bit of interest in the field (e.g., [13]) but is not, by any means, widely used. *Randomization* has been practically unnoticed. It is compared to other statistical tests in [14], but has not been used, otherwise. Re-sampling tests appear to be strong alternatives to parametric or non-parametric tests. We believe that the machine learning community should engage in more experimentation with them to establish the kind of situations in which they can be considered good alternatives.

### 2.3 Problems with our Evaluation Framework

The evaluation framework used by the machine learning community often consists of running large numbers of experiments on community shared domains such as the data sets from the UCI Repository for machine learning. There are many advantages to working in such a setting. In particular, new algorithms can easily be tested in real-world settings; problems arising in such settings can, thus, be promptly identified and focused on; and comparisons between new and old algorithms is easy since researchers share the same data sets. Unfortunately, coupled with these advantages, are a number of disadvantages that were pointed out in [3] and described below.

The first disadvantage is the *Multiplicity Effect* which concerns the execution of large numbers of experiments. In such cases, more stringent requirements need to be used to establish statistical significance than when only a small number of experiments are considered. The next disadvantage, the *Community Experiments problem* corresponds to the fact that if many researchers run the same experiments, it is possible that, by chance, some of them will obtain statistically significant results that will get published and gain undue significance. The *Repeated Tuning problem* states that in order to be valid, all tuning should be done before the test set is known, a seldomly

applied practice. Finally, the *problem of generalizing results* recognizes that it is not necessarily correct to generalize from the UCI Repository to any other data sets, given that these data sets only represent a small portion of the data sets encountered in the real-world. [3] proposes a solution to some of these problems that includes Bonferroni's adjustment and a strict testing strategy.

## 3. Practical Recommendations

We believe that the current de-facto evaluation method used in machine Learning could be improved upon through two processes: *Better Education* and *Better division between exploratory research and evaluation*.

In better educating students interested in machine learning, I advocate sensitization to the uncertainties associated with the evaluation procedure, and familiarization with the kind of tools mentioned in this paper. The education process could involve the inclusion of more material on evaluation in introductory courses in machine learning, or, as we are currently experimenting with, the creation of an advanced course in Machine Learning devoted entirely to the topic of classifier evaluation.

In seeking a better division between exploratory research and evaluation, I was inspired by the field of drug design where the researchers involved in drug design are, typically, not involved in the drug testing process. The tests are typically performed independently, once the drug design process is completed. In machine learning, we have the advantage that our experiments are simple and of low cost. With this advantage, however, comes the disadvantage of believing that we can engage in formal testing by ourselves. This, I believe, is incorrect. Any such testing will be necessarily biased. Based on these observations, my recommendation is to follow the kind of division that is materially necessary in the drug design field and divide our experimental process in three groups. More specifically, I envision a distinction between the *exploratory researchers* who would come up with new ideas and algorithms, and would not need to test their ideas very stringently, but instead would be judged on their innovations. The *evaluators* who would pick previously published algorithms from the literature and test them formally and independently of the researchers that designed them. The *evaluation designers* who would come up with new approaches to evaluation, such as those of [2], [3] [4] or [5].

## References


[1] D. Hand, 2006. Classifier Technology and the Illusion of Progress.*Statistical Sciences*.Vol 21:1, 1-15.

[2] T. Dietterich, 1998: Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation* 10, pp. 1895-1923.

[3] S. Salzberg, "On Comparing Classifiers: Pitfalls to Avoid and a Recommended Approach", *Data Mining and Knowledge Discovery*, Vol. 1, pp.317-327, 1997.

[4] Provost, F., Fawcett, T. and Kohavi, R. 1998. "The Case Against Accuracy Estimation for Comparing Induction Algorithms", *ICML'1998.*

[5] Demšar, J. 2006. "Statistical Comparisons of Classifiers over Multiple Data Sets", *Journal of machine Learning Research*, 7, pp. 1-30.

[6] Caruana, R. and Niculescu-Mizil, A., "An Empirical Comparison of Supervised Learning Algorithms", *ICML'2006*.

[7] Elkan and Zadrozny, Obtaining Calibrated Probability Estimates from Decision Trees and Naive Bayesian Classifiers. In *ICML'2001,* pp. 609-616.

[8] Japkowicz, N., Sanghi, P. and Tischer, P., Classifier Utility Visualization by Distance-Preserving Projection of High Dimensional Performance Data.**,** *ISAIM-08.*

[9] Dennis F. Kibler, Pat Langley: Machine Learning as an Experimental Science. *EWSL 1988*: 81-92.

[10] Witten, I. & Frank, E., *Data Mining: Practical Machine Learning Tools and Techniques (Second Edition)*, Morgan Kaufmann, 2005.

[11] Kononenko, I. and Bratko, I., 1991, "Information-Based Evaluation Criterion for Classifier's Performance, *Machine Learning* 6:67-80.

[12] Caruana, R. and Niculescu-Mizil, A., "Data Mining in Metric Space: An Empirical Analysis of Supervised LearningPerformanceCriteria", KDD'2004.

[13] Kohavi, R., 1995, "A Study of Cross-Validation and bootstrap for Accuracy Estimation and Model Selection", *IJCAI'95*.

[14] Jensen, D. and Cohen, P.R., "Multiple Comparisons in induction Algorithms", *Machine Learning Journal*, 2000.