2010 IEEE International Conference on Data Mining

Consequences of Variability in Classifier Performance Estimates

Troy Raeder, T. Ryan Hoens, Nitesh V. Chawla
Dept. of Computer Science and Engineering, University of Notre Dame
{traeder, thoens, nchawla}@cse.nd.edu

***Abstract*—The prevailing approach to evaluating classifiers in the machine learning community involves comparing the performance of several algorithms over a series of usually unrelated data sets. However, beyond this there are many dimensions along which methodologies vary wildly. We show that, depending on the stability and similarity of the algorithms being compared, these sometimes-arbitrary methodological choices can have a significant impact on the conclusions of any study, including the results of statistical tests. In particular, we show that performance metrics and data sets used, the type of cross-validation employed, and the number of iterations of cross-validation run have a significant, and often predictable, effect. Based on these results, we offer a series of recommendations for achieving consistent, reproducible results in classifier performance comparisons.**

***Keywords*-evaluation; reproducibility; classification**

I. Introduction

Reproducibility is imperative in modern science. Reproducibility not only includes the ability to recreate scientific experiments, but also the underlying data (where possible). In fact, many recent data mining and machine learning conferences explicitly mention repeatability as a criterion for submission. The obvious question becomes: *what is reproducibility for machine learning and data mining?* The answer is (perhaps) non-trivial. Drummond [1] defines reproducibility broadly as the ability to reproduce in independent experimentation "the idea that [a] result empirically justifies."

In the realm of supervised classification, this "idea" is often either the relative superiority of one classifier over another, or the generalized performance of a classifier. However, evaluation methodology in data mining and machine learning is far from standard. Research has produced a dizzying array of performance metrics [2], [3], each estimating classification performance differently. Since different metrics rank classifiers differently, the selection of an inappropriate metric can lead to incorrect conclusions about a classifier's suitability for a given classification task [4].

Additionally, the manner in which training and testing examples are chosen from the set of available data is something of an open question. The most popular approach is some form of *cross-validation*, in which the data set is partitioned into a series of disjoint *folds* and each fold is used to evaluate a classifier built on the rest of the data. Evaluations on individual folds are usually averaged to produce a final score.

Significant research has explored the merits of various cross-validation techniques. For example, Kohavi [5] shows that ten-fold cross-validation has a lower variance than leave-one-out cross-validation or bootstrap estimation. In separate studies, Dietterich and Alpaydin [6], [7] advocate running five separate iterations of two-fold cross-validation in order to reduce the correlation between the training sets, but there is no "standard" approach in the literature. One further aspect of evaluation that is seldom considered is that these strategies are sensitive to the *order* of instances within a data set. Presented with a different ordering of the same data, cross-validation or hold-out will choose different data for training and different data for testing, which undoubtedly yields a different evaluation of classifier performance.

This inconsistency in evaluation methodology is well-known in the machine learning community, but what is not well-studied is its impact on the reproducibility of machine learning experiments. Hanczar et al. [8] show that the variance in ROC curves computed over small data sets can significantly impact scientific conclusions. Bouckaert and Frank [9] study the consistency of statistical tests on individual data sets and recommend a corrected t-test [10] across ten iterations of ten-fold cross-validation as the least sensitive to the order of the data. Many studies (e.g., [11], [12]) offer general guidelines for evaluation but do not touch on repeatability.

Contributions: We rigorously evaluate the process of evaluation, the variability therein, the impact on statistical significance of results, and its impact on the reproducibility and the generalization of scientific conclusions. We study common methodological choices in order to understand their effects on classifier performance evaluation and classifier performance comparison. We demonstrate that under the prevailing evaluation methodology (relatively few iterations of cross-validation), the results of classifier comparisons are overly sensitive to the ordering of the data. In fact, given two decision trees, we can produce two sets of *completely opposing* statistical significance results on the same group of data sets.

In Section III, we examine the behavior of classifier performance estimates across different cross-validation schemes. A larger number of cross-validation folds increases the number of instances used for training, which makes individual training sets more "representative" of the data set as a whole. However, using more folds also increases

1550-4786/10 $26.00 © 2010 IEEE
DOI 10.1109/ICDM.2010.110

the amount of overlap between the test sets, and therefore affects the independence assumption of the results.

Section IV studies the impact of the number of *iterations* of cross-validation. We find that averaging few iterations produces highly unstable, and potentially misleading results. This problem is especially noticeable if the classifiers being compared are similar, as this instability can even affect the results of statistical tests even over many data sets. Averaging over many iterations produces a stable estimate, but one that may be biased if the distribution of individual estimates is highly skewed. We discuss a method for overcoming these issues in Section V. This method enables a robust comparison of classifiers in a reproducible manner.

Section VI compares the estimates produced by the various cross-validation schemes against performance of the classifiers on "true" test sets on a series of data sets which are either temporal in nature (such that a valid concept of "test" exists) or for which a separate test set was provided by an external entity. Finally, Section VII concludes with general observations and recommendations for generating reasonable and reproducible results.

II. Data and Methodology

Table I gives the relevant statistics about the data sets used in this paper[1]. One of the main goals when choosing the data sets to consider was ensuring not only different sizes, but also different domains and class distributions. This enables us to make broader observations and conclusions than would be possible with a more restricted pool of data sets.

We consider a number of different classifiers and evaluate their performance over n (for $n \geq 1$) iterations of k-fold cross-validation (denoted nxk-fold cross-validation) on the mentioned data sets. The cross-validation strategies we considered include: nx10-fold, nx5-fold, and nx2-fold. While there are other validation strategies, including, but not limited to, leave-one-out and bootstrap, we focus on the more commonly used cross-validation strategies in this paper. These strategies also allow us to evaluate the effects of different training and testing sizes, potentially resulting in pessimistic or optimistic performance estimates. In addition to cross-validation strategies, we also employ separate testing sets when applicable.

We note that over the course of many iterations (as $n \to \infty$) the average performance estimate for a given classifier may stabilize. We call this the **steady-state** performance estimate. Conversely, estimates produced by a single iteration of 10-fold cross-validation (1x10-fold), 2 iterations of 5-fold (2x5-fold), and 5 iterations of 2-fold (5x2-fold), we will call **canonical** estimates. The collection of estimates over each of the iterations produces an empirical **distribution** of performance estimates whose properties we later study.

Table I
Data sets used in this study.

	Data set	Inst.	Attr.	Class Dist.	Test Set
Experimental	breast-w	699	9	0.345/0.655	
	bupa	345	6	0.420/0.570	
	credit-a	690	15	0.445/0.555	
	crx	690	15	0.445/0.555	
	heart-c	303	13	0.459/0.541	
	heart-h	294	13	0.361/0.639	
	horse-colic	368	22	0.370/0.630	
	ion	351	34	0.359/0.641	
	krkp	3,196	36	0.478/0.522	
	pima	768	8	0.349/0.651	
	promoters	106	57	0.500/0.500	
	ringnorm	300	20	0.457/0.543	
	sonar	208	60	0.466/0.534	
	threenorm	300	20	0.500/0.500	
	tic-tac-toe	958	9	0.347/0.653	
	twonorm	300	20	0.493/0.506	
	vote	435	16	0.386/0.614	
Case Study	NCAAF	4,288	13	0.626/0.374	✓
	compustat	10,358	20	0.037/0.963	✓
	KDDPE	3,038	116	0.120/0.880	✓
	ozone-1h	2,170	72	0.032/0.968	✓
	ozone-8h	2,168	72	0.069/0.931	✓
	text	8,000	2,000	0.070/0.929	✓
	wrds	49,600	41	0.320/0.680	✓

In order to study the inherent variability in classifier performance estimates, we ran 500 iterations of 2-fold, 5-fold, and 10-fold cross-validation in WEKA [13]. For each iteration, the data set ordering was determined by passing a specific random number seed (0 through 499). On each of these iterations, we evaluated six classifiers: Naïve Bayes (NB) (with kernel density estimation), Multilayer Perceptron (MLP) (with 3 hidden layers), a 5-nearest-neighbor (5-NN) classifier, C4.5 [14] and Hellinger Distance decision trees (HDDT) [15], and an SMO support vector machine. Unless otherwise noted, parameters not specified remained as their default in WEKA. These particular algorithms were chosen to provide a wide range of classifiers for this study. We include two different decision tree classifiers given their different levels of skew-(in)sensitivity for the class distributions [15].

Finally, we note that while there are a wide variety of performance metrics for us to choose from, in this paper we use AUROC (Area Under the Receiver Operating Characteristic). This choice was influenced by its popularity as a general-purpose metric, and the fact that it is less skew sensitive and makes fewer assumptions about misclassification cost than other metrics (e.g., accuracy) [16]. These properties make it suitable to many domains which is central to the results of the paper. Note that the choice of AUROC does not alter the general observations of the paper.

III. Performance estimates over varying numbers of folds

When performing cross-validation, one is faced with an inherent trade-off. If the data are divided into fewer folds,

[1] For a description of the datasets, an extended version is available on our website at www.nd.edu/~dial/papers/ICDM10.pdf.

Table II
RELATIVE POSITIONS (AVERAGE RANKS) OF AUROC ESTIMATES FROM 2-FOLD, 5-FOLD, AND 10-FOLD CROSS-VALIDATION FOR DIFFERENT CLASSIFIERS. BOLD INDICATES A STATISTICALLY SIGNIFICANT ($\alpha = 0.05$) DIFFERENCE FROM 2-FOLD (CANONICAL) OR 10-FOLD (STEADY-STATE) ESTIMATES. CRITICAL DISTANCE FOR THE BONFERRONI-DUNN TEST IS 0.709.

	C4.5	HDDT	MLP	5-NN	NB	SMO
Canonical						
2-fold	2.50	2.30	2.50	2.85	2.65	2.60
5-fold	**1.60**	1.85	1.95	**1.73**	1.95	**1.65**
10-fold	1.90	1.85	**1.55**	**1.43**	**1.40**	**1.75**
Steady-State						
2-fold	**3.00**	**3.00**	**2.70**	**2.95**	**2.90**	**2.85**
5-fold	1.82	**1.90**	1.88	**1.90**	**1.95**	1.70
10-fold	1.18	1.10	1.43	1.05	1.15	1.45

each classifier is trained on a very small number of instances, meaning that one would expect poorer-performing classifiers and higher estimation variance. Dividing the data into a very large number of folds greatly increases the overlap between the training sets and tends to *understate* the estimation variance. Kohavi [5] studied the impact of cross-validation strategy on six UCI data sets and found that 2-fold and 5-fold cross-validation tended to generate "pessimistic" performance estimates. However, that study compares very few data sets and only one evaluation methodology: the average of 50 runs of a particular cross-validation scheme.

We now explore the properties of different cross-validation estimates in both the steady-state and canonical cases. To formally investigate the relationships between performance estimates with different numbers of folds, we perform the Friedman test [11] with the Bonferroni-Dunn post-hoc test. The results are summarized in Table II.

Table II evaluates the extent to which a particular cross-validation approach tends to produce *optimistic* or *pessimistic* estimates of classifier performance. Higher ranks correspond to optimistic estimates and lower ranks to pessimistic estimates. Table II provides an interesting expansion and clarification to the existing literature. We can corroborate that regardless of the classifier used or the number of iterations run, 2-fold cross-validation provides relatively conservative estimates. However, beyond this, the number of iterations plays an important role.

Among canonical estimates, 10-fold and 5-fold cross-validation are statistically indistinguishable at $\alpha = 0.05$, but 2-fold is significantly more pessimistic than 5-fold in all cases, and more pessimistic than 10-fold in half. Among steady-state estimates, 2-fold and 5-fold cross-validation are indistinguishable, but 10-fold cross-validation is substantially more optimistic. This contradicts an earlier result by Kohavi [5], which found 5-fold cross-validation to be a pessimistic estimator. We find a significant difference for 5-fold only when many iterations are averaged together. One final important observation is that none of the statements made above is universal. There are data set/classifier combinations for which 2-fold cross-validation provides the most optimistic estimate even in the steady state.

IV. VARIATIONS IN PERFORMANCE ESTIMATES

The previous section was fairly restricted in that it considered only very-short-run canonical estimates and 500-run steady-state estimates of classifier performance. The general behavior of performance estimates over an increasing series of sampling iterations is also of interest. In this section we concentrate on how each of the methods (canonical and long term) performs over a wide range of tests.

A. Reproducibility of results

One problem with performance estimate variability is ensuring the reproducibility of results. Diettrich [6] advocates 5x2-fold cross-validation since "Exploratory studies showed that using fewer or more than five replications increased the risk of type I error", but the consequences of longer-run estimates are relatively unexplored. In Figure 1 we present the performance of a number of different classifiers under repeated 10, 5, and 2-fold cross-validations. In addition to varying the cross-validation scheme chosen, we also varied the classifiers and data sets. In this way we demonstrate that the results found are not specific to any combination of factors, but instead general trends common to any set of parameters chosen when evaluating classifier performance.

1) In each case, performance estimates stabilize after a non-trivial period whose length differs for each classifier and data set.
2) Estimates based on few iterations can be unstable.
3) Conclusions drawn from the steady-state estimate, in terms of absolute or relative performance of classifiers, may differ from conclusions based on canonical estimates.

In Figure 1(a) we provide results over 500 cross-validation iterations for HDDT and C4.5 in order to show that once an estimate stabilizes it generally remains stable. In Figures 1(b) and 1(c), we plot only 50 and 100 iterations respectively in order to isolate specific regions of the graph. Figure 1(d) further restricts the number of iterations observed to the first 30 iterations to demonstrate the (potential) for high variability under a small number of iterations. On the `vote` data set, performance estimates for all three classifiers shown (C4.5, MLP and 5-NN) oscillate considerably before settling into their steady-state rank-ordering at about fifteen iterations. Importantly, all three classifiers hold the "lead" at some point within those first fifteen iterations. This seems to imply that any performance comparison based on so few iterations is tenuous at best.

In Figure 1(c), the situation is even more extreme. While the MLP separates from the pack early, C4.5 and 5-NN continue to oscillate back and forth for almost 50 iterations before finally settling into a steady-state ordering. The question that naturally arises, then, is: *how should these classifiers be evaluated?*

If reproducibility is ultimately the goal of an evaluation, it is worth defining precisely what is meant by reproducible results. Drummond [1] draws an interesting distinction between reproducibility and what he calls "replicability." Drummond defines replicability as the ability to "reproduce exactly" the results of a prior experiment, meaning that every last condition of the two experiments is equivalent. Reproducibility, on the other hand, is the ability of different researchers, under different conditions, to demonstrate "the idea that [a] result empirically justifies."

This seems like a reasonable definition of reproducibility and a laudable goal. While realizing this goal via a community effort to rerun published experiments to verify correctness seems infeasible, it is unwise to publish results supported only by a single cross-validation strategy and data ordering. The practice of comparing algorithms over several data sets from diverse domains is an important step in this regard, but Table III suggests that this alone is not sufficient.

An approach adopted in prior works (e.g., [17], [18], [19]) is to use a paired t-test across the folds of a single n-by-k-fold cross-validation run. If classifier A is significantly better across the (usually 10) folds, then the difference is deemed *significant* and counted; otherwise, it is ignored. Demšar [11] argues that such approaches decrease the reliability of subsequent statistical tests by "drawing an arbitrary line of $P < 0.05$ between what counts and what doesn't."

While Demšar raises a valid concern, we would argue that results such as the one in Figure 1(c) show that some indicator of the significance of individual performance results is necessary to prevent the (accidental or intentional) reporting of unusual cases. This is especially important when comparing two classifiers which are very similar, e.g., one is an adaptation of the other. To emphasize the severity of this concern, Table III illustrates the extent to which instability can affect performance results. It shows 5-fold AUROC estimates for C4.5 and HDDT, as well as the results of the Wilcoxon signed rank test of significance. In the first iteration, C4.5 performs significantly better than HDDT at $\alpha = 0.05$. In the second, HDDT ranks higher than C4.5 and the difference is again significant at $\alpha = 0.05$. The only difference between the two results is the random seed that was used to select the cross-validation folds. A consistent random seed was used across all data sets within an iteration.

One potentially attractive solution is to draw repeated cross-validation samples until the variance of the resulting performance estimate is sufficiently low to provide confidence in the reported result. This is difficult in practice, however, because there is an inherent dependency among the performance estimates. It has been shown in fact [20] that it is impossible to construct an unbiased estimate for the variance of cross-validation results.

Another alternative is to iterate until the performance estimates stop crossing. The challenge with this is that this point is, of course, impractical to determine with certainty as it would require vastly more computation than can be reasonably expected.

Table III
CLASSIFIER PERFORMANCE RESULTS OVER TWO SEPARATE ITERATIONS OF 10-FOLD CROSS-VALIDATION. A CONSISTENT RANDOM NUMBER SEED WAS USED ACROSS ALL DATA SETS WITHIN AN ITERATION.

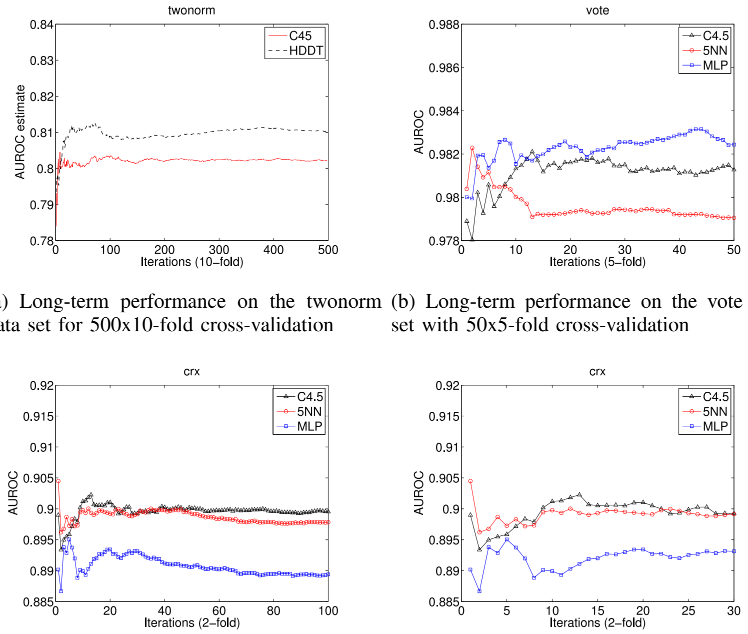
	Iteration 216		Iteration 459	
	C45	HDDT	C45	HDDT
breast-w	**0.9784**	0.9753	0.9768	**0.9820**
bupa	**0.6936**	0.6913	0.6521	**0.6531**
credit-a	**0.8996**	0.8967	**0.9044**	0.8967
crx	**0.8993**	0.8877	**0.9021**	0.8898
heart-c	**0.8431**	0.8181	0.8161	**0.8333**
heart-h	**0.8756**	0.8290	0.8376	**0.8404**
horse-colic	0.8646	**0.8848**	0.8742	**0.8928**
ion	**0.9353**	0.9301	0.9247	**0.9371**
krkp	0.9992	**0.9993**	0.9988	**0.9991**
pima	**0.7781**	0.7717	0.7661	**0.7696**
promoters	**0.8654**	0.8514	0.8676	**0.8774**
ringnorm	**0.8699**	0.8533	0.8669	**0.8727**
sonar	**0.8053**	0.7929	0.8076	**0.8127**
threenorm	**0.7964**	0.7575	**0.7419**	0.7311
tic-tac-toe	**0.9354**	0.9254	**0.9342**	0.9273
twonorm	**0.8051**	0.8023	0.7722	**0.7962**
vote	**0.9843**	0.9824	0.9828	**0.9835**
vote1	**0.9451**	0.9343	**0.9497**	0.9426
avg. rank	**1.11**	1.89	1.72	**1.28**
$\alpha = 0.10$	✓			✓
$\alpha = 0.05$	✓			✓

The remainder of the paper seeks to answer the following questions. First, to what extent do long-run performance estimates produce *reproducible* results as defined above? Second, how can we determine, in an adaptive fashion, how much iteration is sufficient? Finally, to what extent does repeated iteration produce an authoritative answer in classifier performance comparisons.

As we have seen, by relying on canonical estimates without some assurance of significance, even well-meaning researchers can arrive at vastly different conclusions. While this is true also in cases where the classifiers vary more in their inductive biases, we do not present the results here due to space constraints. This implies that in order to ensure the reproducibility of results, researchers must take care that the validation method chosen was robust against the issues we presented.

B. Bias of classifier performance estimates

In Section IV-A we investigated the performance estimates of various cross-validation schemes. One factor to consider when estimating the performance is the existence of a bias in the classifier on the data sets. As we shall see, this can have a profound impact on the estimated performance of a classifier. This is due to the fact that the standard approach to evaluating classifiers over several iterations of cross-validation (which we have used in this work and is used in all other work of which we are aware) is to compare, in some fashion, the mean performance estimates across the several



(a) Long-term performance on the twonorm data set for 500x10-fold cross-validation

(b) Long-term performance on the vote data set with 50x5-fold cross-validation

(c) Long-term performance on the crx data set for 100x2-fold cross-validation

(d) An enhanced view of the instability of short-term performance on the crx data set for 30x2-fold cross-validation

Figure 1. Typical long-run performance estimates.

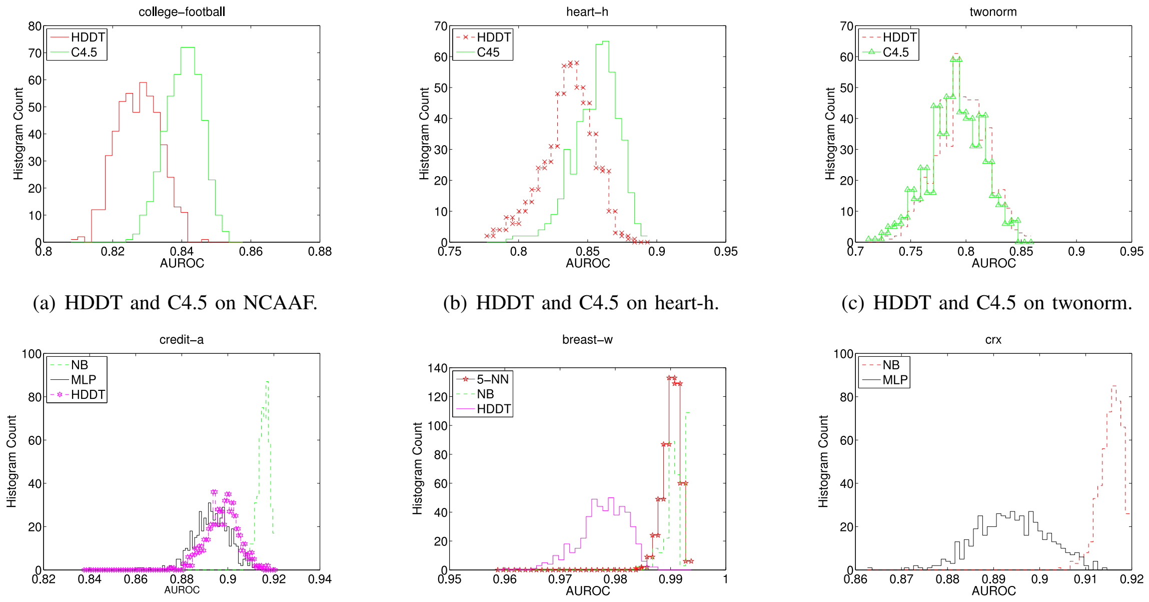
iterations. In this section we explore the consequences of this decision and the situations in which it may lead to unexpected results.

When averaging the performance of a classifier over multiple runs, the latent assumption is that the performance estimates are normally distributed. When this assumption is satisfied, the average performance can be used since it (along with the standard deviation) captures the overall performance of the classifiers. While this assumption is prevalent in empirical studies, it has not been appropriately validated in the literature. We now show that in some instances the normality assumption is (obviously) violated, and its effects on the performance estimates which can be achieved.

If class distributions are highly separable, such that many of the iterations separate the classes perfectly, then the long-run estimate will be lower than the canonical estimate with very high probability. The average of hundreds of runs is not perfect, but it is very likely that one or two randomly-chosen iterations are. This idea generalizes to situations in which the classes are never perfectly separable. For a given problem, inductive bias, and performance metric there will be some maximum achievable performance. If an algorithm is especially effective (or ineffective), it would approach this performance threshold most of the time and the probability of a random iteration performing better (worse, respectively) than the long-run average is high.

Thus, the relationship between steady-state and canonical performance estimates is ultimately a property of the *distribution* of performance estimates for a given classifier, data set, and metric. If the distribution is highly left-skewed, such that the mean is much less than the median, then long-run averages will produce lower estimates. If the distribution is right-skewed, steady-state estimates will be higher, and under a symmetric distribution they should be roughly equivalent. Since one can imagine arbitrarily skewed estimate distributions (and indeed the closer a classifier's performance is to optimal the more skewed we can expect the distribution to be), it may be preferable to compare the medians rather than the means of performance estimates that are based on a large number of cross-validation iterations.

To better explain this phenomenon, Figure 2 shows C4.5 and HDDT performance estimate distributions for four representative data sets. That is, we plot the probability of obtaining a specific AUROC value in each of these classifiers over multiple runs. In three of the four cases the distributions are left-skewed, indicating that the majority of the estimates are greater than the mean. It is interesting to note that on the `breast-w` data set, which produces a relatively symmetric distribution for all classifiers, decision trees perform poorly compared to 5-NN, and NB. In Figure 2(f) we see that NB has an obvious left skew while MLP is normally distributed, demonstrating that the normality assumption is not only violated by decision trees, but is in fact a global phenomenon. As before, the left-skewed distribution performs better than the normally distributed performance estimates in the long-term rankings. This implies that left-skewness in the performance estimate distribution may indicate that a classifier is well-suited for a problem, while symmetry may



(a) HDDT and C4.5 on NCAAF.

(b) HDDT and C4.5 on heart-h.

(c) HDDT and C4.5 on twonorm.

(d) NB, HDDT, and MLP on credit-a.

(e) 5-NN, NB, and HDDT on breast-w.

(f) NB, and MLP on crx.

Figure 2. Performance estimate distributions under AUROC (500x2-fold cross-validation).

imply that a more appropriate classifier exists.

V. Evaluating Classifiers Reproducibly

Figure 1[2] suggests that classifier performance estimates reach some steady-state under repeated iterations of cross-validation. This might be expected, as cross-validation is a "nearly unbiased" estimator of prediction error [21], meaning that it will converge to a value very close to the actual prediction error. Still, this does not necessarily mean that repeated iterations of cross-validation produce a *reproducible* result for arbitrary performance metrics. Recall that by reproducible, we mean that different researchers following the same procedure reach the same general conclusion.

To assess the reproducibility of repeated k-fold cross-validation, we calculated a series of random n-by-k-fold cross-validation estimates and measure the similarity both in terms of the absolute performance estimate and in the relative ordering of classifiers. If the procedure is reproducible, the final estimates should be "close" and the ordering of the classifiers should be preserved with high probability.

Furthermore, if iterations of cross-validation tend to be indistinguishable, it is natural to ask the question: How many iterations are necessary before the results are reproducible? The answer will surely depend on the classifier and data set under consideration since it is well-known that certain algorithms produce more stable classifiers than others [22]. Running hundreds of iterations for a classifier whose performance estimate stabilizes after ten runs is a tremendous waste of effort. It is desirable, then, to develop a criterion for stopping the iteration that allows us to avoid wasted computation. In this section, we evaluate a series of such termination criteria and show that a repeated-iteration framework tends to provide reproducible results. Based on our experiments, *we recommend that practitioners apply the Spearman rank correlation in order to determine convergence,* as it provides good repeatability with relatively little wasted computation.

Termination Criteria: Before proposing candidate termination criteria, it is useful to outline some desirable properties of such criteria. A good criterion should be generalizable so that it can be applied to any learning problem and classifier. While any criterion will naturally depend on the performance estimates seen up to that point, we would like our criteria to be independent of the magnitude of the performance estimates themselves. This is due to the fact that, due to differences in domains, we do not know a priori whether a difference of, say, 0.001 is significant or insignificant. We would prefer for this to be determined by the criterion itself. With these considerations in mind, we propose the following criteria:

Fixed number of iterations This is the simplest possible criterion: simply run a fixed number of iterations of cross-validation regardless of the data set and classifier.

Spearman rank correlation Most classification algorithms produce, for each instance, an estimate of the probability that the instance belongs to a particular class. Different iterations of cross-validation will produce different probability estimates for each instance. If a running average of the probability estimates for each instance is maintained, the *rank correlation* between the averages at successive itera-

[2] A version of this paper with larger tables and figures is available at www.nd.edu/~dial/papers/ICDM10.pdf.

tions provides a measure of convergence for the classifier. Rank correlation measures the extent to which an arbitrary monotonic function describes the relationship between two variables (i.e., the average probability estimates at two successive iterations) and is preferable to the more standard Pearson correlation because it does not assume a linear relationship between the variables. When the correlation exceeds a threshold t, we terminate the computation.

Kolmogorov-Smirnov D statistic The Kolmogorov-Smirnov D statistic measures the maximum distance between two cumulative distribution functions (CDFs) and is used as the basis of the KS test of distributional divergence. For our purposes, it is designed to measure the convergence of our samples to the "true" estimate distribution. To apply the test, we split the cross-validation iterations into two separate distributions, with the even iterations forming one distribution and the odd iterations forming the other. When the D statistic converges below some threshold t, meaning that the even and odd distributions are "similar", we terminate the computation.

We call a termination criteria "adaptive" if the number of iterations of cross-validation chosen varies based on some well-defined function.

Determining the most appropriate criterion: In evaluating termination criteria, there are two important factors to consider. The first is the consistency of the results produced, as this is the essence of reproducibility. The second is the amount of computation required to achieve the results. If method B requires twice as long to achieve essentially the same result as method A, then method A is probably superior. We quantify reproducibility in the following manner. Using the results of our 500 cross-validation runs, we apply each termination criterion 50 times to 50 random orderings of those results. We record the final performance estimate as well as the number of iterations required. We evaluate stability in the final rank-orderings of classifiers through pairwise comparisons of classifiers (c_i, c_j).

We say that a result is reproducible if the rank-ordering of the classifiers is preserved with high probability. Given n performance estimates e_{ik} for classifier c_i and n estimates e_{jk} for c_j, these estimates represent a completely reproducible comparison if the distribution of the e_{ik} is completely separate from the distribution of the e_{jk}, meaning that the rank-ordering of the two classifiers is perfectly consistent. The more the distributions overlap, the less consistent (and less reproducible) the result.

In order to quantify the separation in estimate distributions, we present the following simple overlap metric:

$$R'(i,j) = \frac{1}{n}\sum_{a=1}^{n} I(e_{ia}, e_{ja}). \qquad (1)$$

$$R = max(2 * R'(i,j) - 1, 2 * R'(j,i) - 1). \qquad (2)$$

where $I(x, y)$ is an indicator function that takes the value 1 if $x > y$, $\frac{1}{2}$ if $x = y$, and 0 if $x < y$. It is clear that $R'(i, j)$ takes on a value of 0.5 if the classifiers tie on each iteration or if one classifier "wins" exactly half the time. It may take on a value of 0 or 1 if one classifier is always superior, depending on which classifier wins. Because the superior classifier is irrelevant for our purposes and because we would like the metric to scale between zero and one, we apply the transformation in Equation 2 to obtain the final reproducibility metric.

A series of experimental results are presented in Table IV. We provide results both for a pair of classifiers that is relatively unstable (C4.5 and HDDT) and a pair of classifiers that is relatively stable (NB and SMO). All of the termination criteria have parameters which govern how long they will run. For the rank-based criterion, we iterate until the correlation is at least $1 - 10^{-4} = 0.9999$. We found that values even as high as $1 - 10^{-3}$ tend to iterate too few times, and values in the neighborhood of $1-10^{-5}$ can cause unstable classifiers such as decision trees or k-nearest neighbors to iterate forever. For fixed-iteration termination, we experiment with values of 10 and 50, and for Kolmogorov-Smirnov termination, we experiment with distributional separations of 0.1 and 0.2.

As we expect, the benefits of adaptive iteration are greatest with unstable classifiers and with two-fold cross-validation (shown in Table IV(a)), where the variance of the performance estimate distribution is greatest. Under these circumstances, using fixed iterations provides poor performance across the board for small numbers of iterations. Note, however, that the 10 iterations shown are more than those suggested by the canonical estimates. This means that papers which use 5x2 cross-validation may be difficult to reproduce if the classifiers compared are highly similar.

When the number of fixed iterations is increased to 50, reproducibility becomes similar to that of the adaptive methods in all cases. When the classifiers are relatively stable, however, this improved performance comes at an unnecessary computational cost (Table IV(c)). While the adaptive methods, KS and Rank often go far beyond 50 iterations for the unstable classifiers C4.5 and HDDT, for NB and SMO stabilization often occurs at much fewer than 50 iterations. This suggests that in such instances the canonical estimates should more accurately represent the performance of the classifiers. Due to the little overhead required to ensure reproducibility, however, we argue that these methods should nevertheless be incorporated to safeguard against false assumptions and further aid reproducibility.

For space reasons, we only included the 10-fold runs for the unstable classifiers (Table IV(b)). We note that each of the observations made about the 2-fold runs for the instable classifiers still hold. That is, the Fixed-10 iterations exhibit the least reproducibility, while Fixed-50 more closely follows the adaptive methods. Since each of the performance estimates is more stable than those found in the 2-fold runs, however, the effects are not as drastic.

Among the adaptive methods themselves, it appears that `Rank`-based termination offers superior performance to `KS`. At $t = 0.2$, `KS` and `Rank` often have similar iteration counts, and `KS` almost universally provides inferior reproducibility. At the stricter criterion of $t = 0.1$, `KS` iterates significantly more but incurs only modest improvement by doing so. Due to this minor improvement coupled with the vast disparity in the number runs required, we recommend using the `Rank` method for determining the number of iterations that each dataset should be run under for each classifier.

VI. Case Study: true test sets

All of the work presented thus far has focused solely on *relative* comparisons of cross-validation methodologies. In the absence of ground truth, we have been unable to draw any absolute conclusions. In order to address this issue, we have acquired 7 data sets that have some notion of a "true" test set. These data sets were either provided with a test set from the original data source (such as the KDD cup data sets `KDDPE`) or are intrinsically ordered. The `NCAAF`, `ozone`, and `compustat` data sets, for example, are collected over time, such that one year of data can be used for testing and all prior years for training.

There is no guarantee these data sets satisfy the stationary distribution assumption, such that it is reasonable to expect cross-validation performance to be predictive of test set performance. However, we feel that this situation is worth studying simply because *it mimics exactly the manner in which classifiers are validated in the real world.* As such, it is instructive to consider whether any over-arching conclusions can be drawn in this context, especially regarding the difference between steady-state and canonical estimates.

The results of our experiments appear in Tables V and VI, where "steady-state" estimates in this case were generated with 500 iterations of cross-validation. Note that regardless of the number of folds used, cross-validation tends to underestimate performances. This makes intuitive sense, as cross-validation classifiers are trained on less data than the final classifier, and is corroborated by Kohavi [5]. There are notable exceptions to this, however. On `compustat` for example, the performance of Naïve Bayes is substantially underestimated by all three methods, yet becomes the best-performing classifier on the true test set. This suggests that perhaps `compustat`, a temporal financial data set, undergoes a significant distribution shift between the training and test period. The fact that the best-performing classifier ranks third on the training data underscores the need for *robust* analysis techniques for real-world applications.

The differences between canonical and steady-state estimates are also enlightening. In the (randomly chosen) canonical case, the most accurate (i.e., lowest-MSE) estimates for each classifier are scattered among 5-fold and 10-fold runs. However, in the long run 10-fold cross-validation consistently produces the best estimates. We speculate that this is due to the fact that 10-fold cross-validation trains on the greatest amount of data and therefore is able to build the most comprehensive classifier. In the canonical case, random inconsistencies in the estimates allow 5-fold to come out ahead by chance. Steady-state estimates smooth out these inconsistencies and best estimator becomes clear.

VII. Discussion

One of the fundamental underpinnings of the scientific method is obtaining results which are reproducible. In the absence of reproducible results, a theory must be rejected as not having enough support. In this paper we have studied the current state of the art in comparing machine learning algorithms. We have found that while the current most popular methods for validating performance are 2x5, 5x2, and 1x10-fold cross-validation, these "canonical" methods can lead to results, which a) might not be reproducible and b) might lead to inconsistent conclusions.

To combat these shortcomings, we recommend determining the steady-state performance of the classifiers tested, and only then applying the standard statistical significance tests. In order to guide this process and aid in reproducibility, we present a series of criteria for determining an appropriate definition of "steady-state" classifier performance. We show that a simple criterion based on the *rank correlation* between instances' probability estimates over many iterations reasonably balances computation and reproducibility. Additionally, It produces substantially more repeatable results than using a small fixed number of iterations.

While studying the reproducibility of results, an important concern that arose was the "skewness" of the performance estimate distribution. We hypothesize that left-skewed distributions indicate that the classifier is well suited to the problem, while right-skewed distributions indicate a poor fit. We therefore recommend that classifiers which exhibit a left-skew should be preferred for that domain.

When developing classifiers, being able to accurately estimate a candidate classifier's performance on future data is an important concern. In Section VI we found that long-term performance estimates produced by 10-fold cross-validation were most consistent with the classifiers' measured performance on "true" test sets. With this in mind, we recommend using 10-fold cross-validation for data sets similar to those marked "experimental" in this paper, i.e., "balanced" data sets (note: a balanced data set is one for which the number of positive and negative examples are commensurate).

We also noted that while cross-validation results are often good predictors of future performance, there are notable exceptions when they fail to appropriately rank the classifiers. This is an important observation, as it shows that while a classifier can perform (perhaps significantly) better on a cross-validation set, this is no guarantee that this will be

Table IV
Reproducibility and stopping time (number of cross-validation iterations) for several different termination criteria. Results are averaged across 50 random applications of the criteria. Columns are: Rank: Rank correlation at $t = 0.9999$, Fixed-t: Fixed number of iterations, KS-t: Kolmogorov-Smirnov criterion with threshold t.

(a) Two-fold cross-validation: Comparing C45 and HDDT.

	Reproducibility					Avg. Iterations					
dataset	Fixed-10	Fixed-50	Rank	KS-0.1	KS-0.2	HTree-Rank	J48-Rank	HTree-KS-0.1	J48-KS-0.1	HTree-KS-0.2	J48-KS-0.2
breast-w	0.52	1.00	0.88	1.00	0.72	30.34	28.22	179.12	187.24	42.64	46.08
bupa	1.00	1.00	1.00	1.00	1.00	139.88	133.42	192.72	200.48	53.40	49.00
NCAAF	1.00	1.00	1.00	1.00	1.00	61.08	56.28	221.96	172.28	60.44	42.40
credit-a	0.88	1.00	1.00	1.00	0.96	43.48	47.02	191.92	202.84	40.28	52.88
crx	0.84	1.00	1.00	1.00	1.00	43.92	45.42	196.60	165.28	48.52	49.56
heart-c	0.24	0.68	0.92	0.80	0.40	71.50	69.90	203.08	168.28	45.68	46.08
heart-h	1.00	1.00	1.00	1.00	1.00	59.74	47.60	188.04	205.48	46.44	58.28
horse-colic	0.84	1.00	1.00	1.00	0.96	66.46	64.90	248.64	176.88	41.32	50.72
ion	0.44	0.52	0.60	0.72	0.36	49.82	49.72	194.96	178.60	48.76	43.24
krkp	0.08	0.28	0.92	0.36	0.20	31.40	26.54	170.80	183.80	48.76	41.80
pima	0.40	0.96	0.84	1.00	0.68	74.84	61.22	253.36	184.84	47.64	43.00
promoters	0.52	0.96	1.00	1.00	0.76	130.70	116.58	171.56	193.48	48.44	49.92
ringnorm	0.84	1.00	1.00	1.00	0.96	88.94	78.92	199.68	190.28	51.32	48.60
sonar	0.84	1.00	1.00	1.00	1.00	113.88	118.44	212.96	206.12	53.36	40.16
threenorm	0.20	0.48	0.56	0.60	0.44	152.58	148.24	213.60	179.20	44.20	36.64
tic-tac-toe	0.88	1.00	0.92	1.00	0.84	72.38	73.02	210.20	233.40	40.48	51.16
twonorm	0.44	0.88	1.00	0.96	0.76	110.08	111.76	186.52	205.16	44.92	51.92
vote	0.68	1.00	0.80	1.00	0.96	26.96	35.56	185.64	190.28	45.88	41.44
vote1	0.88	1.00	1.00	1.00	0.96	32.58	33.36	203.48	190.72	49.68	47.40

(b) Ten-fold cross-validation: Comparing C45 and HDDT.

	Reproducibility					Avg. Iterations					
dataset	Fixed-10	Fixed-50	Rank	KS-0.1	KS-0.2	HTree-Rank	J48-Rank	HTree-KS-0.1	J48-KS-0.1	HTree-KS-0.2	J48-KS-0.2
breast-w	0.84	1.00	1.00	1.00	1.00	143.62	232.20	222.48	205.24	54.96	50.96
bupa	0.48	0.96	1.00	1.00	0.84	92.36	85.10	219.28	200.68	47.64	47.28
college-football	1.00	1.00	1.00	1.00	1.00	51.80	45.32	225.68	185.88	47.08	49.16
credit-a	0.80	1.00	1.00	1.00	0.96	30.48	40.72	193.88	200.52	47.36	59.48
crx	0.68	1.00	1.00	1.00	0.96	30.70	41.10	175.32	194.80	44.96	53.76
heart-c	0.92	1.00	1.00	1.00	1.00	51.26	51.18	187.16	210.24	39.40	42.68
heart-h	1.00	1.00	1.00	1.00	1.00	47.88	60.68	193.00	215.28	40.64	47.76
horse-colic	1.00	1.00	1.00	1.00	1.00	39.02	39.94	204.48	219.88	54.76	40.64
ion	0.80	1.00	1.00	1.00	0.96	62.70	83.74	175.88	202.56	42.92	39.52
krkp	1.00	1.00	1.00	1.00	1.00	86.26	86.30	190.20	173.36	37.72	37.40
pima	1.00	1.00	1.00	1.00	1.00	58.28	48.52	201.56	198.08	40.92	52.68
promoters	0.06	0.28	0.84	0.68	0.24	75.88	69.76	224.16	200.68	45.44	43.84
ringnorm	1.00	1.00	1.00	1.00	1.00	69.90	57.26	185.76	219.00	52.12	49.96
sonar	1.00	1.00	1.00	1.00	1.00	62.98	80.28	221.44	255.20	58.32	46.72
threenorm	0.76	1.00	1.00	1.00	0.92	114.26	94.52	205.52	205.20	45.04	48.72
tic-tac-toe	1.00	1.00	1.00	1.00	1.00	63.78	64.82	223.68	237.04	50.68	55.96
twonorm	0.72	0.96	1.00	1.00	0.96	87.38	88.36	254.96	195.92	54.52	56.32
vote	0.64	0.88	1.00	1.00	0.80	74.78	122.76	245.36	240.76	52.20	48.76
vote1	1.00	1.00	1.00	1.00	1.00	76.32	46.02	205.68	177.00	48.12	42.16

(c) Two-fold cross-validation: Comparing NB and SMO.

	Reproducibility					Avg. Iterations					
dataset	Fixed-10	Fixed-50	Rank	KS-0.1	KS-0.2	NB-Rank	SMO-Rank	NB-KS-0.1	SMO-KS-0.1	NB-KS-0.2	SMO-KS-0.2
breast-w	1.00	1.00	1.00	1.00	1.00	13.20	6.00	240.92	420.92	50.24	69.44
bupa	1.00	1.00	1.00	1.00	1.00	64.92	3.08	209.56	1000.00	51.32	714.12
NCAAF	1.00	1.00	1.00	1.00	1.00	9.12	24.80	206.56	156.00	47.88	46.48
credit-a	1.00	1.00	1.00	1.00	1.00	12.64	9.20	223.52	202.24	54.44	44.68
crx	1.00	1.00	1.00	1.00	1.00	12.94	8.82	222.88	253.80	44.36	49.56
heart-c	1.00	1.00	1.00	1.00	1.00	14.90	18.30	221.08	252.04	53.36	44.68
heart-h	1.00	1.00	1.00	1.00	1.00	17.10	19.36	191.88	228.00	49.52	56.48
horse-colic	1.00	1.00	1.00	1.00	1.00	16.00	27.42	189.44	194.80	39.00	47.84
ion	1.00	1.00	1.00	1.00	1.00	13.54	12.74	242.16	227.48	48.40	49.56
krkp	1.00	1.00	0.96	1.00	1.00	7.52	12.02	191.68	164.44	42.76	45.00
pima	1.00	1.00	1.00	1.00	1.00	18.40	12.50	204.76	189.00	45.32	50.76
promoters	1.00	1.00	1.00	1.00	1.00	36.10	27.94	223.36	731.64	54.56	101.60
ringnorm	1.00	1.00	1.00	1.00	1.00	14.38	42.58	300.68	201.56	49.16	53.92
sonar	1.00	1.00	1.00	1.00	1.00	29.20	40.62	196.28	218.00	44.36	50.24
threenorm	1.00	1.00	1.00	1.00	1.00	31.80	17.46	205.40	458.96	47.20	63.16
tic-tac-toe	1.00	1.00	1.00	1.00	1.00	35.24	2.00	193.80	1000.00	45.00	1000.00
twonorm	1.00	1.00	1.00	1.00	1.00	17.02	7.30	191.16	246.56	43.36	51.84
vote	1.00	1.00	1.00	1.00	1.00	10.44	10.26	197.36	210.52	46.08	49.28
vote1	1.00	1.00	1.00	1.00	1.00	9.80	13.36	181.32	191.24	45.92	51.60

Table V
Canonical cross-validation estimates of AUROC from true-test data sets and actual performances.

	2			5			10			actual		
	HDDT	C4.5	NB	HDDT	C4.5	NB	HDDT	C4.5	NB	HDDT	C4.5	NB
KDDPE	0.771	0.769	0.760	0.794	0.784	0.785	0.776	0.789	0.787	0.813	0.763	0.750
college-f	0.829	0.837	0.908	0.837	0.850	0.909	0.846	0.854	0.909	0.842	0.850	0.914
compustat	0.816	0.777	0.641	0.841	0.802	0.705	0.849	0.800	0.706	0.832	0.807	0.862
ozone-1h	0.821	0.798	0.861	0.841	0.828	0.862	0.831	0.852	0.859	0.934	0.950	0.963
ozone-8h	0.821	0.808	0.849	0.812	0.805	0.847	0.827	0.800	0.847	0.879	0.890	0.931
text	0.945	0.941	0.888	0.957	0.949	0.901	0.955	0.955	0.904	0.958	0.945	0.908
wrds	0.966	0.969	0.878	0.968	0.970	0.867	0.968	0.971	0.863	0.993	1.000	0.991
MSE	0.040	0.045	0.079	**0.031**	**0.038**	**0.073**	0.035	**0.038**	0.074			

Table VI
STEADY-STATE CROSS-VALIDATION ESTIMATES OF AUROC FROM TRUE-TEST DATA SETS AND ACTUAL PERFORMANCES.

	2			5			10			actual		
	HDDT	C4.5	NB	HDDT	C4.5	NB	HDDT	C4.5	NB	HDDT	C4.5	NB
KDDPE	0.771	0.764	0.763	0.785	0.781	0.782	0.791	0.785	0.788	0.813	0.763	0.750
college-f	0.827	0.841	0.908	0.838	0.849	0.909	0.840	0.849	0.909	0.842	0.850	0.914
compustat	0.814	0.776	0.663	0.839	0.807	0.689	0.847	0.816	0.712	0.832	0.807	0.862
ozone-1h	0.806	0.783	0.859	0.829	0.826	0.859	0.840	0.834	0.860	0.934	0.950	0.963
ozone-8h	0.811	0.791	0.848	0.817	0.800	0.848	0.816	0.805	0.848	0.879	0.890	0.931
text	0.947	0.943	0.889	0.955	0.949	0.901	0.957	0.949	0.904	0.958	0.945	0.908
wrds	0.966	0.969	0.878	0.968	0.971	0.867	0.969	0.971	0.863	0.993	1.000	0.991
MSE	0.044	0.048	0.077	0.033	0.038	0.075	**0.032**	**0.038**	**0.073**			

realized in "real world" scenarios. The impacts of this will be studied more heavily in future work.

This paper focused mainly on AUROC, but the main conclusions of this work hold with other metrics. We will pursue this line of work in greater depth in the future. We will also include additional multi-class and imbalanced data sets in order to paint a more comprehensive picture of how data sets, evaluation metrics, and validation strategies effect how results are drawn in the machine learning community. With this in mind, the purpose of this paper is independent of the metric used. Instead we demonstrate some of the shortcomings of current machine learning evaluation techniques, and how to mitigate them. More importantly, however, we seek to open discussion in the community on how to ensure reproducibility of results, fairness of comparisons and generalizability of experimental results.

VIII. ACKNOWLEDGMENTS

Work was supported in part by the NSF Grant ECCS-0926170 and the Notebaert Premier Fellowship.

REFERENCES

[1] C. Drummond, "Replicability is not reproducibility: Nor is it good science," in *Proceedings of the Evaluation Methods for Machine Learning Workshop at the 26th ICML*, 2009.

[2] R. Caruana and A. Niculescu-Mizil, "Data mining in metric space: an empirical analysis of supervised learning performance criteria," in *Proceedings of KDD*. ACM, 2004, p. 78.

[3] C. Ferri, J. Hernández-Orallo, and R. Modriou, "An empirical comparison of performance measures for classification," *Pattern Recognition Letters*, vol. 30, pp. 27–38, 2009.

[4] D. J. Hand, "Good practice in retail credit scorecard assessment," *The Journal of the Operational Research Society*, vol. 56, no. 9, pp. 1109–1117, 2005.

[5] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proceedings of IJCAI*, vol. 14, 1995, pp. 1137–1145.

[6] T. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural computation*, vol. 10, no. 7, pp. 1895–1923, 1998.

[7] E. Alpaydin, "Combined 5 x 2 cv F test for comparing supervised classification learning algorithms," *Neural Computation*, vol. 11, no. 8, pp. 1885–1892, 1999.

[8] B. Hanczar, J. Hua, C. Sima, J. Weinstein, M. Bittner, and E. Dougherty, "Small-sample precision of ROC-related estimates," *Bioinformatics*, vol. 26, no. 6, p. 822, 2010.

[9] R. Bouckaert and E. Frank, "Evaluating the replicability of significance tests for comparing learning algorithms," in *Proceedings of PAKDD*. Springer, 2004, pp. 3–12.

[10] C. Nadeau and Y. Bengio, "Inference for the generalization error," *Machine Learning*, vol. 52, no. 3, pp. 239–281, 2003.

[11] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *J. Machine Learning Research*, vol. 7, p. 30, 2006.

[12] S. Salzberg, "On comparing classifiers: Pitfalls to avoid and a recommended approach," *Data Mining and Knowledge Discovery*, vol. 1, no. 3, pp. 317–328, 1997.

[13] G. Holmes, A. Donkin, and I. Witten, "Weka: A machine learning workbench," in *Second Australia and New Zealand Conference on Intelligent Information Systems*, 1994, pp. 357–361.

[14] J. Quinlan, *C4. 5: programs for machine learning*. Morgan Kaufmann, 1993.

[15] D. Cieslak and N. Chawla, "Learning decision trees for unbalanced data," in *Proc of ECML*, 2008.

[16] F. Provost, T. Fawcett, and R. Kohavi, "The case against accuracy estimation for comparing induction algorithms," in *Proc of ICML*, vol. 445, 1998.

[17] N. Lachiche and P. Flach, "Improving accuracy and cost of two-class and multi-class probabilistic classifiers using ROC curves," in *Proceedings of ICML*, vol. 20, 2003, p. 416.

[18] P. Melville and R. Mooney, "Constructing diverse classifier ensembles using artificial training examples," in *Proceedings of the IJCAI*, 2003, pp. 505–510.

[19] G. Tsoumakas, I. Katakis, and I. Vlahavas, "Effective voting of heterogeneous classifiers," *Machine Learning: ECML 2004*, pp. 465–476, 2004.

[20] Y. Bengio and Y. Grandvalet, "No unbiased estimator of the variance of k-fold cross-validation," *The Journal of Machine Learning Research*, vol. 5, p. 1105, 2004.

[21] B. Efron and G. Gong, "A leisurely look at the bootstrap, the jackknife, and cross-validation," *American Statistician*, vol. 37, no. 1, pp. 36–48, 1983.

[22] L. Breiman, "Bagging predictors," *Machine learning*, vol. 24, no. 2, pp. 123–140, 1996.