Machine Learning, 42, 269–286, 2001

The Effect of Instance-Space Partition on Significance

JEFFREY P. BRADFORD jeffrey.bradford@computer.org
CARLA E. BRODLEY brodley@ecn.purdue.edu
School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907, USA

Editor: Douglas Fisher

Abstract. This paper demonstrates experimentally that concluding which induction algorithm is more accurate based on the results from one partition of the instances into the cross-validation folds may lead to statistically erroneous conclusions. Comparing two decision tree induction and one naive-bayes induction algorithms, we find situations in which one algorithm is judged more accurate at the $p = 0.05$ level with one partition of the training instances but the other algorithm is judged more accurate at the $p = 0.05$ level with an alternate partition. We recommend a new significance procedure that involves performing cross-validation using multiple instance-space partitions. Significance is determined by applying the paired Student t-test separately to the results from each cross-validation partition, averaging their values, and converting this averaged value into a significance value.

Keywords: classification, comparative studies, statistical tests of significance, cross validation

1. Introduction

When developing machine learning algorithms, typically the estimated classification accuracy of the novel algorithm is compared to the estimated classification accuracy of existing algorithms. To this end, practitioners compare the accuracy of classifiers induced by two or more learning algorithms on one or more datasets. Since each dataset typically is a sample from a larger population, algorithms that induce equally accurate classifiers when trained over the entire population might yield different results when trained on the available sample. Thus, when appropriate, accuracy values should be quoted with an associated confidence interval, and statistical tests should be used to determine whether the difference in accuracy between classifiers induced by two algorithms is statistically significant.

Many procedures and statistical tests have been proposed and used for machine learning tasks to determine the significance of an accuracy difference between classifiers. Popular procedures include hold-out, cross-validation, and leave-one-out; popular statistical tests include McNemar's test, difference of two proportions, the binomial test, the Student t-test, and comparing accuracy directly. The characteristics of these procedures and tests have been investigated by other researchers (Salzberg, 1997; Dietterich, 1998).

This paper extends previous research by examining the *variation* in the significance values across different cross-validation runs. In particular, we show that the previously recommended statistical procedures (Salzberg, 1997; Dietterich, 1998) do not sufficiently account for the effects of differing instance-space partitions (across different cross-validation runs),

leading researchers to draw erroneous conclusions about which algorithm is "better". For example, the paired Student t-test can be used to test whether the mean of the differences in the accuracy values is statistically significantly different than zero. This paper shows that the significance value generated from a single run of ten-fold cross validation or from five runs of two-fold cross-validation is greatly affected by the partition of the instances into the folds; based on these results, conclusions should be not drawn based on one significance value from one partition (or set of partitions in the case of five runs of two-fold cross-validation) of the instances. Instead, a new procedure is proposed: performing multiple cross-validation runs, each using a different partition (or set of partitions) of the instance space, and applying the paired t-test separately to each partition. These multiple t-values (i.e., the value generated directly from the t-test) are combined by calculating the arithmetic mean across the multiple partitions. This resulting mean t-value is converted into a p-value by comparing it to a t-distribution. The p-value estimates rejecting the null hypothesis is an error.

The remainder of the paper is organized as follows. Section 2 reviews existing procedures and statistical tests for determining whether two algorithms induce classifiers that yield statistically significantly different accuracy estimates, and known problems with these procedures and tests. Section 3 empirically illustrates a previously unexplored problem with existing procedures, that of the variance of the significance values due to instance-space partition. Section 4 proposes a new method for comparing the accuracy of algorithms that addresses this problem. Section 5 provides conclusions and ideas for future work.

2. Review of existing procedures and tests

This section reviews existing procedures and statistical tests for determining which of two induction algorithms is statistically significantly more accurate. The induction algorithms themselves do not perform classification; that function is performed by the classifier induced by the induction algorithm. Thus, this section focuses on procedures and tests for comparing the accuracy of two *classifiers*; the algorithm that induces more-accurate classifiers is considered to be the more accurate algorithm. While it has been shown that no algorithm always outperforms other algorithms on all datasets (Wolpert, 1994), it is assumed that these algorithms are compared over a set of datasets that are similar to the dataset that will be used in practice, so higher accuracy on the testing datasets will likely lead to higher accuracy in practice.

Determining which of two classifiers is the more accurate typically involves two steps: first, a procedure is applied to measure the accuracy of the two classifiers on a given dataset; second, a statistical test is applied to determine whether the accuracy difference is caused by the known sources of variation, or whether the difference in accuracy is caused by an unknown source, which is assumed to be a difference in the classifiers. Dietterich (1998) lists four sources of variation commonly found in machine learning experiments: sampling from a larger population, the split of the data into the training/testing sets, class noise, and the internal randomness of the learning algorithm (e.g., the setting of the initial weights in neural nets, or the choice of test during decision tree induction when the estimated utility of two tests is equal).

The simplest method for estimating which classifier is more accurate is to perform a single hold-out run on each dataset, and then apply a statistical test to determine which classifier is significantly more accurate, if either. In this paper, "significant" always means "statistically significant", not "large". In hold-out, the available instances are split into two disjoint sets, a training set and a testing set. The training set is used to induce the classifier, and the testing set is used to estimate the classification accuracy of the classifier. Because the testing instances are not used for induction, if one makes the assumption that the instances are representative of the underlying population and sufficient instances are available for training to properly form the generalization, then this procedure usually provides an unbiased estimate of the accuracy of the classifier over the population. There are several tests that may be used to determine the significance of the difference in accuracy, including McNemar's test (Casella & Berger, 1990, pp. 399–400), which is an application of the sign test (Stell, 1997, pp. 568–569), and the binomial test (Snedecor & Cocharan, 1989, p. 120), also called a test for the difference of two proportions.

One shortcoming of hold-out is that it does not account for the effect of the partition of the instances into the training and testing sets. Accordingly, resampling, that is, repeating hold-out for multiple random partitions of the instances into the training and testing sets, is commonly used. The paired Student t-test may then be applied to determine significance, by applying the following equation (Casella & Berger, 1990, p. 395).

$$\frac{\bar{Z}}{\sqrt{\frac{1}{n(n-1)}\sum_{i=1}^{n}(Z_i - \bar{Z})^2}}$$

Thus, to perform resampling and apply the paired t-test, hold-out is performed for n random partitions of the dataset into the training and testing sets. For each partition, induction is performed using each algorithm on the training set, and the accuracy of the induced classifiers is estimated using the testing set. This procedure yields two accuracy values per instance-space partition, X_i and Y_i, the accuracy of the classifier induced by the first algorithm on partition i and the accuracy of the classifier induced by the second algorithm on partition i, respectively. By letting $Z_i = X_i - Y_i$, that is, the difference in accuracy between the two classifiers for partition i, and letting $\bar{Z}$ be the arithmetic mean over Z_i, the equation above may be applied to calculate the t-value. This t-value is compared to a t-distribution with $n - 1$ degrees of freedom to determine the significance (the p-value, that is, the probability of being wrong if the null hypothesis is rejected).

Unfortunately, problems have been found with the application of the paired t-test to the accuracy estimates generated by resampling. Dietterich (1998) finds applying the paired Student t-test to resampling yields an unacceptably high probability of making a type-I error (i.e., incorrectly concluding a difference in accuracy exists), and recommends that the test never be applied to accuracy estimates generated by resampling. This problem is attributed to the violation of an assumption of the t-test that the Z_i's are independent. This assumption is violated in the case of resampling because each accuracy value is generated using the same sample of instances, just partitioned differently into the training and testing sets.

Another common procedure for comparing the accuracy of classifiers is cross-validation. In n-fold cross-validation, the instances are partitioned into n disjoint folds (sets). For

each algorithm, n classifiers are induced. To induce a classifier, $n-1$ of the folds are used, leaving one remaining fold for accuracy estimation. This procedure yields n accuracy values. Cross-validation provides an unbiased estimate of the accuracy of a classifier built over $(n-1)/n\%$ of the sample. If one makes the assumption that the instances are representative of the underlying population and that sufficient instances are available for training to form an accurate generalization, then this procedure usually provides an unbiased estimate of the accuracy of the classifier over the population. To determine whether any difference in the estimated accuracies of the two classifiers is significant, the paired t-test can be applied. Each fold yields two accuracy values, X_i and Y_i, resulting in n pairs of accuracy values. By letting Z_i be the difference in accuracy between the two classifiers for fold i, that is, $X_i - Y_i$, the equation above may be applied. Dietterich (1998) finds the application of the paired t-test to accuracy estimates generated via cross-validation has a slightly elevated probability of type-I error, but a decreased probability of type-II error (i.e., decreased probability of incorrectly accepting the null hypothesis). This is again attributed to the lack of independence, due to the fact that $(n-2)/(n-1)\%$ of the instances overlap between each pair of training folds. However, in contrast to resampling, each instance is used only once for accuracy estimation.

Rather than applying the paired t-test to one run of n-fold cross-validation data, Dietterich (1998) introduces a method named 5×2CV. To perform a test for significance using 5×2CV, two-fold cross validation is run for five partitions of the instances, yielding ten pairs of accuracy values, as does a single run of ten-fold cross validation. Significance can be determined by applying the following equation.

$$\frac{Z_1}{\sqrt{\frac{2}{10}\sum_{i=1}^{10}(Z_i - \bar{Z})^2}}$$

The primary difference between the equation for 5×2CV versus that for n-fold cross-validation is the numerator: in the former, it is the difference in accuracy only for the first fold of the first partition of the instances; in the latter, the numerator is the mean of the difference in accuracy between the two algorithms over the n folds. Dietterich shows that the 5×2CV test has a distribution approximating a t-distribution with five degrees of freedom.

3. Empirical evaluation of significance tests

This section shows that the procedures and tests discussed in the previous section do not sufficiently account for the variation in accuracy due to differing instance-space partitions. Two pairs of algorithms are used, one pair of similar algorithms and one pair of dissimilar algorithms. Two procedures are used, one run of ten-fold cross-validation (1×10CV) and five runs of two-fold cross-validation (5×2CV). The paired t-test is used to measure significance.

The first algorithm uses the top-down decision tree induction algorithm in $\mathcal{MLC}$++ (Kohavi et al., 1996) but uses a different pruning algorithm that we developed. This pruning algorithm is based on the cost function proposed by Mansour (1997); the difference is that

while this proposed cost function requires a user-settable pruning factor, our implementation automatically determines the pruning factor as was done in the cost-complexity pruning algorithm described in CART (Breiman et al., 1984). This algorithm will be referred to as `TDDT-1`. The second algorithm induces the same decision tree as the first algorithm, but bases its pruning decision on the pessimistic correction factor used in C4.5 (Quinlan, 1993). This algorithm will be referred to as the `TDDT-2` algorithm. We chose two similar algorithms to increase the difficulty of finding a significant difference.

The third algorithm used is the discretized `naive-bayes` classifier (Kohavi, 1996) provided in $\mathcal{MLC}$++; it is a standard naive-bayes classifier that first discretizes continuous attributes. The `naive-bayes` algorithm has a different representational and preference bias from the decision tree algorithms, increasing the likelihood of there being a difference between the classifiers induced by the two algorithms.

Due to the violation of statistical independence assumptions of performing an all-pairs comparison, only the `TDDT-1` algorithm will be compared to the `TDDT-2` algorithm and to the `naive-bayes` algorithm; the `TDDT-2` algorithm will not be compared directly to the `naive-bayes` algorithm. The accuracy of the three algorithms on the twenty-three datasets listed in Table 1 were determined; these datasets are available in the UCI repository (Merz & Murphy, 1999).

To measure the effect of instance-space partition on the significance values, 55 partitions of the instances were generated; these partitions were generated by random non-stratified sampling. A different random sample generated a different partition. [Using stratified sampling did not consistently increase or decrease the mean or variation of the significance values, and therefore results for stratified sampling are not presented.] For each partition, one run of ten-fold cross-validation and five runs of two-fold cross validation were performed using each of the three algorithms. The paired t-test was applied separately to each of the ten pairs of accuracy values per partition, yielding 55 significance values. The paired t-test is applied to test the hypothesis that the accuracy of both algorithms over the population is the same (the null hypothesis) against the alternative hypothesis that the accuracy is different. Thus, a two-tailed test is applied, so significance at the $p = 0.05$ level requires a threshold of $\alpha = 0.025$. The significance values were plotted as follows (see Figure 1). For each dataset the 55 significance values were sorted, assigning each significance value a rank between 1 and 55. Each significance value was then plotted, with the rank along the x-axis and the significance value along the y-axis. To reduce confusion, the significance values for each dataset are plotted separately. Each dataset is represented by an abbreviation along the x-axis; the name of the associated dataset can be found in Table 1. Along the y-axis (i.e., the significance value), values above the $p = 0.5$ line mean that `TDDT-1` achieved higher accuracy and values below $p = 0.5$ line mean that the comparison algorithm achieved higher accuracy; a value of $p = 0.5$ means that both algorithms achieved the same accuracy.

The results for `TDDT-1` versus `TDDT-2` using the 1×10CV test for significance (i.e., applying the paired t-test to one run of ten-fold cross-validation) are shown in Figure 1. The immediate result to note is the wide range of significance values. This range of values has an important practical implication: the (random) choice of partition can change the conclusion as to whether a difference in accuracy exists or even which algorithm is better.

Table 1. Column 1 shows dataset abbreviation used in figures 1–4. Column 2 reports the names of the datasets used, and Column 3 reports the total number of instances available for training and testing. Columns 4–9 show the accuracy for each algorithm on each procedure averaged over the 55 instance-space partitions.

	Dataset		TDDT-1		TDDT-2		naive-bayes	
Abbr	Name	Size	1×10	5×2	1×10	5×2	1×10	5×2
bc	breast-cancer	286	71.6	70.5	73.4	72.1	72.8	72.1
br	breast	699	94.5	94.0	95.1	94.4	97.2	97.3
cl	cleve	303	75.6	73.5	75.7	74.3	83.3	82.8
cr	crx	690	85.3	84.9	84.8	83.8	86.0	85.7
di	diabetes	768	74.7	73.3	72.7	71.4	75.5	74.1
ge	german-org	1000	72.4	71.5	72.0	71.0	73.2	71.7
gl	glass	214	65.8	63.1	67.7	64.6	71.1	60.2
g2	glass2	163	79.1	73.6	79.4	74.8	81.8	74.3
he	heart	270	76.9	74.4	75.5	75.1	82.9	81.9
hc	horse-colic	368	83.7	82.1	84.2	82.2	80.3	79.9
hy	hypothyroid	3163	99.2	99.1	99.2	99.1	98.4	98.4
io	ionosphere	351	89.3	87.7	88.7	87.6	89.5	89.3
ln	labor-neg	57	79.5	75.0	82.2	78.3	88.5	85.0
m1	monk1-full	432	94.7	86.1	92.7	86.6	75.0	73.9
m2	monk2	432	66.5	65.9	67.1	66.1	66.1	62.3
m3	monk3-full	432	100.0	99.7	100.0	99.3	97.2	97.2
mu	mushroom	8124	100.0	100.0	100.0	100.0	95.5	94.8
so	soybean-large	683	92.6	88.7	93.4	89.3	89.8	86.7
tt	tic-tac-toe	958	84.6	80.5	84.4	80.9	69.7	70.7
ve	vehicle	846	71.4	67.7	73.1	69.8	60.6	57.9
vo	vote	435	95.2	95.1	95.0	95.0	90.1	90.0
wa	waveform-21	5000	77.0	76.0	76.3	75.6	80.9	80.6
zo	zoo	101	90.4	86.1	91.9	89.5	93.3	91.4

We divide the datasets into four groups, depending on the impact of the partition on the conclusion drawn. The first group contains those datasets for which different partitions change which algorithm is considered better; that is, for these datasets a difference is found at the $p = 0.05$ level, but changing the instance-space partition may change *which* algorithm is judged more accurate. Two datasets (heart and tic-tac-toe) are in group 1. The second group contains datasets for which different partitions change whether a difference exists; that is, for these datasets changing the instance partition may change from concluding that a difference exists at the $p = 0.05$ level (i.e., reject the null hypothesis) to concluding that no significant difference in accuracy exists (i.e., do not reject the null hypothesis). Thirteen datasets (breast-cancer, breast, crx, diabetes, german, glass, glass2, horse-colic, ionosphere, monk1-full, soybean-large, vehicle, and waveform-21) are in group 2. The third group contains datasets for which different partitions do not affect the conclusion drawn,

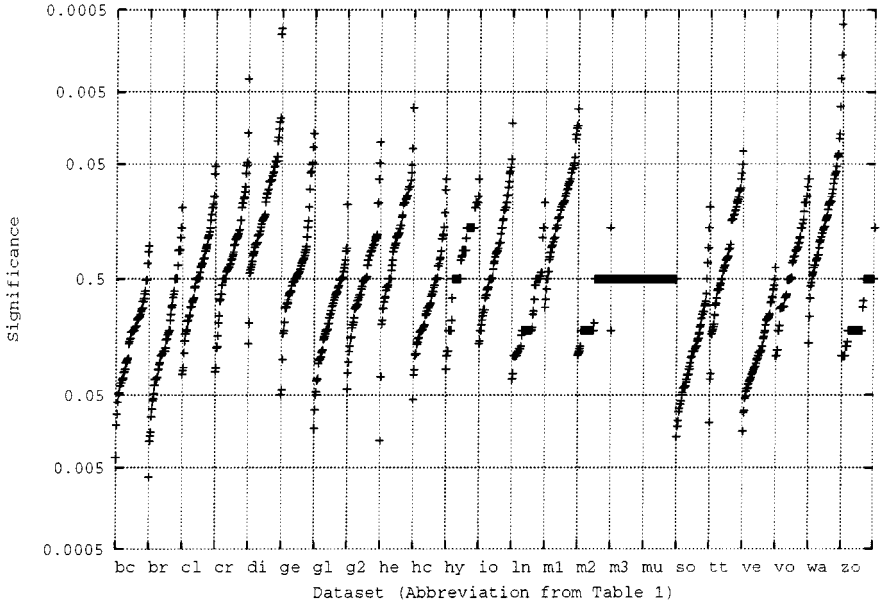

Figure 1. Significance values of the accuracy difference between the two decision tree algorithms, TDDT-1 and TDDT-2, using the 1×10CV test for significance. Values above the $p = 0.5$ line mean TDDT-1 yielded higher accuracy, values below the $p = 0.5$ line mean TDDT-2 yielded higher accuracy.

although there still exists a range of significance values; that is, for these datasets changing the instance-space partition may change the significance value, but the difference is never significant at the $p = 0.05$ level. Six datasets (cleve, hypothyroid, labor-neg, monk2, vote, and zoo) are in group 3. The fourth, and final, group contains datasets for which the instance partition did not tend to affect the accuracy. Two datasets (monk3-full and mushroom) are in group 4.

These results show that, for datasets for which different partitions yield different accuracy estimates,

> *one should not claim significance over the target population based on only one partition of the instances.*

The results for comparing the TDDT-1 algorithm to the naive-bayes algorithm are shown in Figure 2. Again, the instance-space partition impacts the significance values. For fourteen datasets (breast, cleve, crx, diabetes, german, glass, glass2, heart, horse-colic, hypothyroid, labor-neg, monk2, soybean-large, and zoo), changing the instance-space partition may change the conclusion from one algorithm is more accurate with $p = 0.05$ (i.e., reject the null hypothesis) to no significant difference in accuracy exists (i.e., not reject the null hypothesis). These graphs show that the initial random partition of the instances into the folds can substantially affect the significance value reported, and that reporting the significance from only one partition may lead to erroneous conclusions.

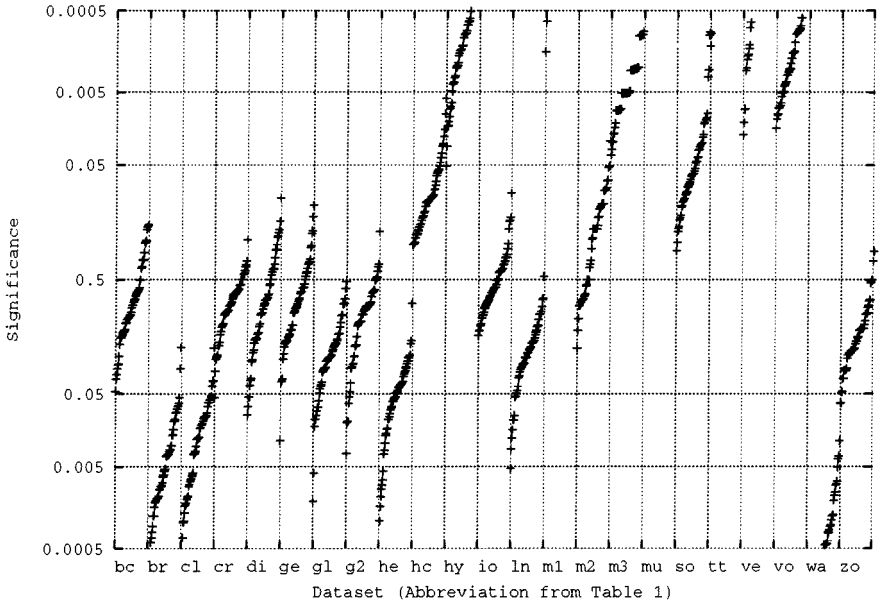

Figure 2. Significance values of the accuracy difference between the `TDDT-1` algorithm and the `naive-bayes` algorithm, using the 1×10CV test for significance. Values above the $p = 0.5$ line mean the `TDDT-1` algorithm yielded higher accuracy, values below the $p = 0.5$ line mean the naive bayes algorithm yielded higher accuracy. Missing values for monk1-full, mushroom, and tic-tac-toe are `TDDT-1` more accurate with significance $p < 0.0005$.

The experiments reported in Figure 1 and Figure 2 were repeated, except that 5×2CV was used instead of 1×10CV; that is, five runs of two-fold cross-validation were performed, and the paired t-test was applied to determine significance. The significance values are shown in Figure 3 for the `TDDT-1` algorithm versus the `TDDT-2` algorithm, and in Figure 4 for the `TDDT-1` algorithm versus the `naive-bayes` algorithm.

Using the 5×2CV test generally resulted in the significance value indicating the accuracy difference is less significant (i.e., more likely to accept the null hypothesis), in agreement with Dietterich (1998); it additionally resulted in a slightly smaller variation in the significance values. The smaller variation was generally advantageous, in that changing the partition has less impact on the conclusion drawn. Consider tic-tac-toe. When using 1×10CV, changing the partition may result in changing which algorithm judged is more accurate, as the significance values ranged from the `TDDT-2` algorithm more accurate at the $p = 0.02$ level to the `TDDT-1` algorithm more accurate at the $p = 0.03$ level. In contrast, when using 5×2CV none of the differences in accuracy were significant at the $p = 0.05$ level. However, for most datasets the range of significance values were similar to those from 1×10CV. Moreover, the difference in significance values from one partition to another was still large enough that drawing conclusions based on only one run of 5×2CV may lead to erroneous claims. For example, for the horse-colic and soybean-large datasets, changing the instance-space partition may change which algorithm is judged more accurate.

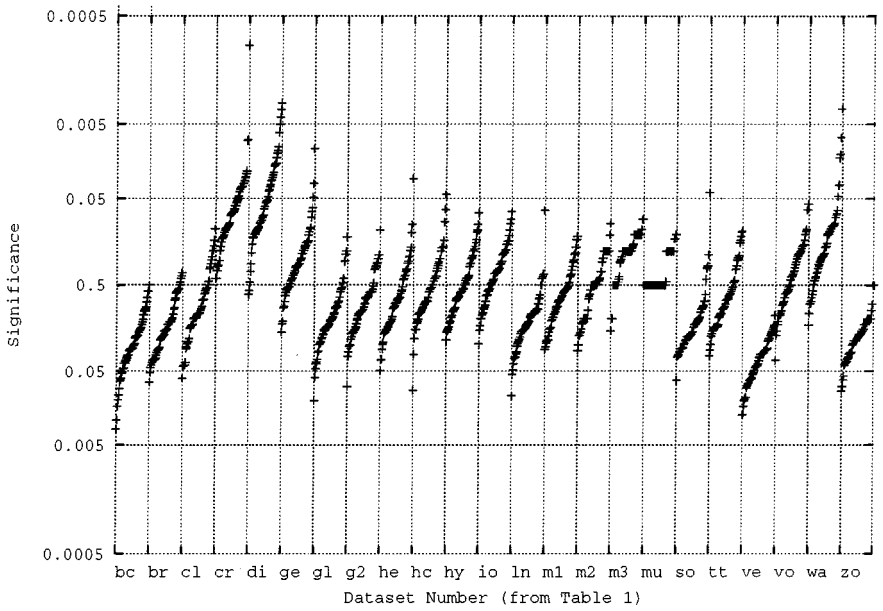

Figure 3. Significance values of the accuracy difference between the two decision tree algorithms, TDDT-1 and TDDT-2, using the 5×2CV test for significance. Values above the $p = 0.5$ line mean TDDT-1 yielded higher accuracy, values below the $p = 0.5$ line mean TDDT-2 yielded higher accuracy.

One important difference between 1×10CV and 5×2CV is that out of 69 possibilities (23 datasets times 3 algorithms), only twice, breast and tic-tac-toe for the naive-bayes algorithm, were the accuracy values from the two-fold cross-validation in 5×2CV higher than the accuracy values from the ten-fold cross-validation in 1×10CV. This is to be expected, as 5×2CV uses fewer instances to induce the classifier (50% versus 90% of the available sample). While for the majority of datasets this decrease in accuracy did not greatly affect conclusions regarding which algorithm is more accurate, in some cases it did. For example, consider the TDDT-1 algorithm versus the TDDT-2 algorithm on monk1-full. Applying 1×10CV (Figure 1) results in TDDT-1 generally inducing more accurate classifiers, with significance values ranging from TDDT-1 more accurate with $p = 0.005$ to TDDT-2 more accurate with $p = 0.25$. However, applying 5×2CV (Figure 3), the results ranged from the TDDT-2 algorithm more accurate at the $p = 0.10$ level to the TDDT-1 algorithm more accurate at the $p = 0.20$ level. The reason for this change was that in switching from 1×10CV to 5×2CV, the average accuracy for the TDDT-2 algorithm dropped 6.1 percentage points from 92.7% to 86.6%; in contrast, the average accuracy for the TDDT-1 algorithm dropped a much larger 8.6 percentage points from 94.7% to 86.1%. Running 1×10CV on monk3-full and mushroom results in all folds yielding 100% accuracy on both decision tree algorithms, so there is no difference between the two algorithms (Figure 1); in contrast, when performing 5×2CV some folds yield lower accuracy values, as low as 95%, thus causing a difference in the significance values (see Figure 3), although the difference is never significant at the $p = 0.05$ level.

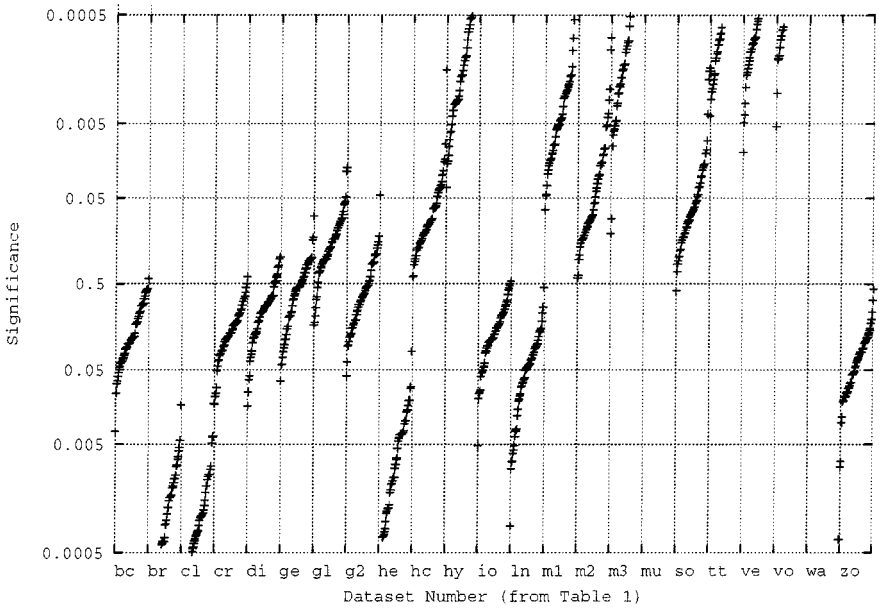

Figure 4. Significance values of the accuracy difference between TDDT-1 and the naive bayes algorithm, using the 5×2CV test for significance. Values above the $p = 0.5$ line mean TDDT-1 yielded higher accuracy, values below the $p = 0.5$ line mean the naive bayes algorithm yielded higher accuracy. Missing values for mushroom are TDDT-1 more accurate with significance $p < 0.0005$ and for tic-tac-toe are the naive-bayes algorithm more accurate with significance $p < 0.0005$.

To be fair, the 5×2CV test was not intended to estimate the accuracy of an algorithm, just which of two algorithms is more accurate. However, ideally we want to know which algorithm is more accurate over the population. Determining which algorithm is more accurate on 90% of the available instances is closer to the ideal than determining which algorithm is more accurate on 50% of the available instances. Thus, if using 50% of the available instances yields accuracy values comparable to using 90%, 5×2CV seems preferable to 1×10CV, as it results in decreased variation in the significance values. When this is not the case, 1×10CV seems preferable, because it results in accuracy values closer to what one would expect over the population.

4. Proposed significance procedure

When comparing algorithms, presenting significance values in the graph form as was done in Section 3 is unsatisfactory; some summary statistic is desired. That is, a procedure is desired to convert the multiple values from multiple instance-space partitions into a single significance value. In this section we propose a procedure for generating such a summary statistic: average the t-values and convert this averaged t-value to a p-value. Empirical and theoretical evidence for the validity of the proposed approach is provided. Next, two other candidate summary statistics are considered: (i) converting the significance values

to a χ^2 distribution and (ii) applying the Wilcoxon signed rank test to the significance values. These two procedures are mentioned only for completeness; our discussion points to problems with these two procedures due to the lack of independence of the significance values, and thus we do not recommend their use in this situation. This section concludes by comparing the procedure proposed in this section to procedures that are used to account for the familywise-error-rate. Note that the framework to account for multiple comparisons (Utgoff et al., 1997) is not appropriate to this situation, as each accuracy value is being compared only once.

To generate the summary statistic, 1×10CV or 5×2CV is performed repeatedly using different instance space partitions, and for each partition the t-value is calculated using the existing methodology summerized in Section 2. The arithmetic mean of these t-values (before converting them into significance values) is calculated, and is then converted into a significance value by comparing it to a Student-t random variable with the same number of degrees of freedom as would have been used had only a single partition been used. Table 2 shows the significance values calculated using this method for each dataset from Figures 1–4. The number of partitions that should be used depends on two factors: how close the significance values are to the chosen threshold required to claim significance (e.g., $p = 0.05$) and the variation of the significance values caused by the instance-space partition. Clearly, the closer the mean t-value is to the chosen threshold and the greater the variation in the significance value, the more partitions that should be used. A precise number can be calculated by assuming the t-values are normally distributed, and using the standard one-tailed test to see whether sufficient partitions have been used to provide the confidence that one is accounting for the effects of instance space partition. That is, for n measured significance values P_i (each using a different partition, and before converting to a p-value), a confidence of $p = \alpha_1$ to judge whether a significant difference exists between the algorithms, and a confidence of $p = \alpha_2$ for using sufficient partitions, the test is

$$\frac{\bar{P} - t_{\alpha_1, n-1}}{\sqrt{\frac{1}{n(n-1)} \sum_{i=1}^{n} (P_i - \bar{P})^2}} > t_{\alpha_2, n-1}$$

where $t_{\alpha,n}$ is the value for the Student t-distribution with n degrees of freedom and a confidence of $p = \alpha$. (Freund & Wilson, 1997, pp. 158–160).

Empirically, we found that fitting a normal curve to the significance values resulted in a very poor fit; we assume this is caused by the fact that the distribution of significance values (the p-values) is not symmetric when the variation is caused by differing instance-space partitions. However, we found empirically that fitting a normal curve to the *t-values* resulted in a very good fit. Figure 5 shows the data from Figure 1 (`TDDT-1` versus the `TDDT-2` algorithm using 1×10CV) with the addition of a normal curve for each dataset. Only four representative datasets are shown (crx, heart, labor-neg, and mushroom); other datasets look similar.

Theoretically, we believe the reason a normal curve fits the t-values well is because a t-distribution with many degrees of freedom converges to a normal distribution. The ten degrees of freedom used in this case appears to be sufficient. While it may be the case that a t-distribution would fit the data better, because only the mean is of interest, assuming either a normal distribution or a t-distribution results in the same averaged value. Note that the

Table 2. Column 1 reports the names of the datasets. Columns 2–5 show the significance of the t-values averaged over the 55 partitions. Positive values indicate the TDDT-1 algorithm is more accurate; negative values indicate the other algorithm is more accurate. *Emphasised* values are significant at the $p = 0.05$ level. [Recall that as this is a two-tailed test, significance at the $p = 0.05$ level requires a threshold of 0.025.]

	TDDT-1 versus TDDT-2		TDDT-1 versus Naive-Bayes	
Dataset name	10×1CV	5×2CV	10×1CV	5×2CV
breast-cancer	−0.13	−0.050	−0.32	−0.077
breast	−0.14	−0.12	*−0.0018*	*−0.00001*
cleve	0.47	−0.25	*−0.0035*	*−0.00002*
crx	0.31	0.029	−0.30	−0.089
diabetes	0.086	*0.018*	−0.30	−0.20
german-org	0.36	0.26	−0.31	−0.36
glass	−0.23	−0.18	−0.071	0.15
glass2	−0.50	−0.28	−0.19	−0.37
heart	0.25	−0.34	*−0.016*	*−0.00015*
horse-colic	−0.36	−0.45	0.065	0.038
hypothyroid	0.37	−0.49	*0.00048*	*0.00003*
ionosphere	0.32	0.40	−0.47	−0.078
labor-neg	−0.24	−0.15	−0.075	*−0.010*
monk1-full	0.10	−0.36	*0.00000*	*0.00017*
monk2	−0.31	−0.46	0.24	*0.011*
monk3-full	−0.49	0.22	*0.0021*	*0.00000*
mushroom	0.50	0.37	*0.00000*	*0.00000*
soybean-large	−0.10	−0.14	0.026	*0.018*
tic-tac-toe	0.35	−0.35	*0.00000*	*0.00000*
vehicle	−0.10	*−0.023*	*0.00004*	*0.00000*
vote	0.41	0.41	*0.00074*	*0.00000*
waveform-21	0.12	0.13	*−0.00010*	*−0.00000*
zoo	−0.24	−0.072	−0.14	*−0.019*

central limit theorem (Casella & Berger, 1990, pp. 216–217) does not apply in this case, as the central limit theorem applies only to the *means* of random variables; it does not apply to the *significance* values.

Using the proposed method, we are now in a better position to show the possibility of erroneous conclusions when using only one partition of the instances, and to reemphasize the importance of considering multiple partitions. Table 3 shows the conclusions (i.e., first algorithm is more accurate, no difference in accuracy, second algorithm is more accurate, for $p = 0.05$) for each dataset in Figures 1–4. For most datasets the averaged significance value is far from the chosen threshold of $p = 0.05$; thus, for these datasets most or all of the partitions yield the same conclusion as to whether a significant difference exists at the $p = 0.05$ level. Examples include monk2, monk3, and mushroom when comparing the two

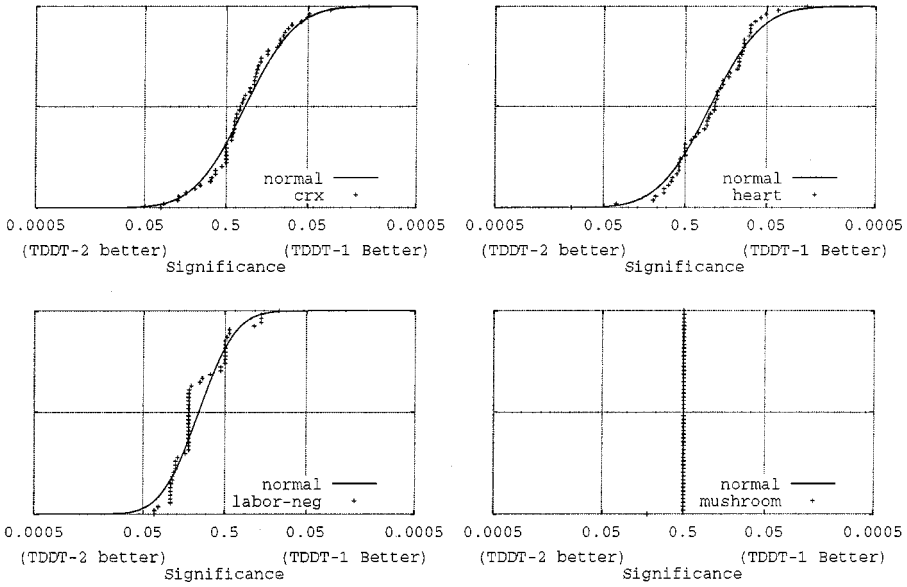

Figure 5. Significance values from figure 1 with the addition of a normal curve.

decision tree algorithms, and tic-tac-toe, vehicle, and vote when comparing the first decision tree algorithm to the `naive-bayes` algorithm. For these datasets, running for only a couple of partitions would most likely suffice. In contrast, when the averaged significance value is close to the chosen threshold, the instance-space partition greatly affects the conclusion as to whether a significant difference exists. Examples include vehicle for `TDDT-1` versus `TDDT-2` using the 5×2CV test and heart for `TDDT-1` versus `naive-bayes` for the 1×10CV test.

Another metric for determining the practical impact of the instance-space partition is to determine how frequently using one partition yields a different conclusion than does using multiple partitions. This was determined by, for each of the 5060 tests (23 datasets times 4 comparisons times 55 partitions), comparing the conclusion as to whether a significance difference exists at the $p = 0.05$ level to the conclusion when the significance value is generated by averaging over the 55 instance-space partitions. Table 4 shows the number of correct and incorrect conclusions due solely to the instance-space partition, assuming the averaged values in Table 2 are the correct significance values. Over the 5060 tests, a total of 496 incorrect conclusions would have been made, for an overall error rate of 9.8% due solely to the instance-space partition.

Two other potential procedures for generating a summary statistic from multiple significance values are to assume a χ^2 random variable or to use the Wilcoxon signed rank test. In the next three paragraphs, we describe each in turn, and then discuss why they are unsuitable in this situation. Assume one performs n identical independent experiments over the same population, and for each experiment calculates P_i, that is, the significance value from the ith experiment. Then $-2\sum_{i=1}^{n} \ln P_i$ has a χ^2_{2n} distribution (Stell, 1997,

Table 3. Column 1 reports the names of the datasets. Each group of three numbers shows the number of partitions (out of 55) that yield a given conclusion, either the first algorithm is more accurate for $p = 0.05$, no significant difference exists at the $p = 0.05$ level, or the second algorithm is more accurate for $p = 0.05$. The conclusion based on the averaged value (Table 2) is **bolded**. For example, for breast-cancer dataset comparing the two TDDT algorithms using the 10×1CV test for significance, for 4 partitions the `TDDT-1` algorithm was more accurate at the $p = 0.05$ level, for 51 partitions no significant difference in accuracy existed at the $p = 0.05$ level, and for no partitions the `TDDT-2` algorithm was more accurate at the $p = 0.05$ level.

	`TDDT-1` versus `TDDT-2`						`TDDT-1` versus `Naive-Bayes`					
Dataset name	10×1CV			5×2CV			10×1CV			5×2CV		
breast-cancer	4	**51**	0	16	**39**	0	0	**55**	0	12	**43**	0
breast	8	**47**	0	5	**50**	0	**53**	2	0	**55**	0	0
cleve	0	**55**	0	6	**49**	0	**51**	4	0	**55**	0	0
crx	0	**52**	3	0	**31**	24	1	**54**	0	8	**47**	0
diabetes	0	**44**	11	0	26	**29**	4	**51**	0	5	**50**	0
german-org	0	**51**	4	0	**51**	4	1	**54**	0	3	**52**	0
glass	2	**53**	0	5	**50**	0	10	**45**	0	0	**49**	6
glass2	0	**53**	2	1	**54**	0	6	**49**	0	2	**52**	1
heart	1	**52**	2	1	**53**	1	**28**	27	0	**54**	1	0
horse-colic	1	**54**	0	1	**53**	1	0	**45**	10	0	**35**	20
hypothyroid	0	**55**	0	0	**55**	0	0	1	**54**	0	0	**55**
ionosphere	0	**53**	2	0	**55**	0	0	**55**	0	12	**43**	0
labor-neg	0	**55**	0	4	**51**	0	10	**45**	0	**34**	21	0
monk1-full	0	**48**	7	0	**55**	0	0	0	**55**	0	1	**54**
monk2	0	**55**	0	0	**55**	0	0	**54**	1	0	27	**28**
monk3-full	0	**55**	0	0	**55**	0	0	0	**55**	0	2	**53**
mushroom	0	**55**	0	0	**55**	0	0	0	**55**	0	0	**55**
soybean-large	9	**46**	0	1	**53**	1	0	**31**	24	0	**29**	26
tic-tac-toe	1	**53**	1	0	**55**	0	0	0	**55**	0	0	**55**
vehicle	6	**49**	0	27	**28**	0	0	0	**55**	0	0	**55**
vote	0	**55**	0	0	**54**	1	0	0	**55**	0	0	**55**
waveform-21	0	**46**	9	0	**48**	7	**55**	0	0	**55**	0	0
zoo	0	**55**	0	10	**45**	0	2	**53**	0	**28**	27	0

pp. 483–484), and can be tested for significance against a χ^2_{2n} random variable. Say one performs three identical independent experiments (each using an independent sample from the same population) that yield significance values $p = \{0.15, 0.10, 0.10\}$. None of these experiments are significant at the $p = 0.05$ level on their own, but all three *hint* at rejecting the null hypothesis. They are combined as follows. $-2\sum_{i=1}^{3} \ln P_i$ is calculated to yield $-2(-1.897 + -2.303 + -2.303) = 13.006$. This value is tested for significance as a χ^2_6 distribution, yielding a significance value of $p = 0.05$. A theoretical drawback of this test is that the value of $-2\sum_{i=1}^{n} \ln P_i$ depends on which algorithm is taken as the reference

Table 4. Number of correct and incorrect conclusions over all 5060 comparisons from Table 3.

		Correct		Incorrect	
Comparison:	TDDT-1 versus	Accept null hypothesis	Reject null hypothesis	Type-I	Type-II
TDDT-2	1×10CV	1192	0	73	0
	5×2CV	1066	56	89	54
Naive-Bayes	1×10CV	591	571	69	34
	5×2CV	371	717	69	108

algorithm, which is an undesirable property, and one that the paired t-test does not share. Although algorithm A more accurate with $p = \{0.15, 0.10, 0.10\}$ is equivalent statistically to algorithm B more accurate with $p = \{0.85, 0.90, 0.90\}$, when applying this test, in the first case $-2\sum_{i=1}^{3} \ln P_i$ yields $p = 0.05$, while in the second case $-2\sum_{i=1}^{3} \ln P_i$ yields $-2(-0.163 + -0.105 + -0.105) = 0.747 \Rightarrow p = 0.99 \neq 1 - 0.05$.

The second possible test is the Wilcoxon signed rank test (Freund & Wilson, 1997, pp. 596–598) (Stell, 1997, pp. 569–570). The intuition behind applying this test is that if algorithm A outperforms algorithm B on all (or almost all) of the instance-space partitions, even if the differences are not significant at the $p = 0.05$ level, then algorithm A is better. The Wilcoxon signed rank test formalizes this intuition as follows. Say one performs three identical independent experiments that yield significance values $p = \{-0.20, 0.30, 0.05\}$. [That is, algorithm A is more accurate once with significance $p = 0.20$, and algorithm B is more accurate twice with significance $p = \{0.30, 0.05\}$.] The values are sorted without regard to sign, yielding $\{0.30, -0.20, 0.05\}$. Each value is given its rank, yielding $\{1, -2, 3\}$. The positive ranks and negative ranks are summed separately, and the smaller is kept. In this case, the sum of the positive ranks is 4, while the sum of the negative rank is 2, so the rank sum is 2. To determine significance, this value is compared via a table (for small n) or using the normal distribution (for large n). For this example, the difference is not significant.

Both of these tests rely on the assumption that the significance values to be combined are independent. If one is trying to estimate the significance over the population (the typical goal), the significance values in Section 3 fail to meet this assumption, as each experiment uses the same set of instances, just partitioned differently. This lack of independence causes a fatal problem when applying these two tests *in this situation*: by increasing the number of partitions used, the significance of any accuracy difference between two algorithms, no matter how small, can be made arbitrarily significant when using the χ^2 test or the Wilcoxon signed rank test. This should not occur, as using multiple partitions accounts for only one of the four sources of variation listed in the beginning of Section 2. Consider the case in which the instance-space partition does not affect the significance value, and algorithm A is slightly more accurate than algorithm B with significance of, say, $p = 0.25$. Note that we did not find such a dataset in our experiments, although some datasets (e.g., monk1 in Figure 1) were close. With 55 partitions the χ^2 method yields $-2\sum_{i=1}^{55} \ln 0.25 = 152.4924$. Comparing this value to a χ^2_{110} distribution yields a significance value of $p = 0.005$. The Wilcoxon signed rank test suffers from this problem to an even greater extent. Only six arrangements are required to conclude a significant difference exists at the $p = 0.05$ level,

and only eight arrangements are required to conclude a significant difference exists at the $p = 0.01$ level. It is for this reason that we cannot recommend the use of either test. In contrast, the proposed procedure does not suffer from this problem. If algorithm A is more accurate than algorithm B with significance $p = 0.25$ for all instance-space partitions, then, independent of the number of partitions used, the averaged significance value will be $p = 0.25$. Thus, it is the averaging process of choice.

Lastly, note that this procedure does not addresses the multiplicity effect or the familywise error rate (Feelders & Verkooijen, 1996; Salzberg, 1997), If one performs n *independent* trials, the probability of finding significance at the $p = \alpha$ level by chance is $1-(1-\alpha)^n$. The expected number, assuming independent events, is $n\alpha$. For even moderate values of n and α, one has a high probability of erroneously concluding that a significant difference exists (i.e., making a type-I error). For example, Section 3 shows results for $n = 23$ and $\alpha = 0.05$, leading to a 69% chance of making a type-I error, and an expected number of 1.2. Salzberg (1997) recommends the Bonferroni adjustment to correct for the multiplicity effect. In the Bonferroni adjustment, α is set small enough so that the probability of making a type-I error over n datasets is acceptably low. The reason that our proposed procedure does not address the multiplicity effect is that we are investigating the variation in the significance values due to instance-space partition. Thus, even if no variation in the significance values exists, a significance value of $p = 0.05$ still implies a 5% probability of incorrectly rejecting the null hypothesis when the null hypothesis is true, and a correction should still be applied.

5. Conclusions and future work

The major conclusion of this work is to empirically illustrate a problem with a popular method of determining which of two algorithms yields more accurate classifiers. In particular, the partition of the instances into the folds greatly affects the significance value when applying the paired Student t-test to accuracy results from one run of ten-fold cross-validation (1×10CV) and from five runs of two-fold cross-validation (5×2CV). Based on the results presented in this paper, we recommend that significance results not be reported for only one partition of the instances. Instead, we recommend performing multiple cross-validation runs, determining the t-value of the accuracy difference of each run, averaging the t-values, converting the averaged t-value value into a significance value, and reporting this significance value.

An important focus for future work is the development of a better significance test. Significance tests are used to account for the sources of variation inherent in any experimental procedure, such as sampling from a larger population, class noise, the internal randomness of the learning algorithm, and the split of the data into the training/testing sets (Dietterich, 1998). The procedure presented in this paper removes this last source of variation. One has the intuitive feeling that the significance value should be better than the average we present. More work is required to carefully define what the significance of the averaged t-values measures, and how it should be adjusted to account for using multiple partitions of the instances.

Another focus is to determine why different partitions result in different significance values, and either correct the source or modify the algorithm to deal with them better.

Finding outliers or errors (Brodley & Friedl, 1999) may be examples of correcting the source; option decision trees (Kohavi & Kunz, 1997) and bagging (Breiman, 1996) may be examples of algorithmic changes.

Lastly, an interesting relationship to examine is the one between the results we present and aggregate techniques that use the results from many random partitions of the instances, such as bagging (Breiman, 1996), boosting (Freund & Schapire, 1996), and ECOC (Dietterich & Bakiri, 1995). Bagging improves algorithms that have high variation. Our purpose in using multiple partitions is to account for the effect of the instance-space partition on the accuracy, allowing better determination of the accuracy difference between two algorithms.

Acknowledgments

We thank our editor Doug Fisher, the anonymous referees, and Ron Kohavi for their many suggestions, which significantly improved this paper. This work was made possible in part by an equipment grant from SGI. Jeffrey Bradford's research was supported by an Intel Foundation Doctoral Fellowship. Carla Brodley's research was supported by the NSF under Grant Number IIS-9733573.

References

Breiman, L. (1996). Bagging predictors. *Machine Learning, 24*(2), 123–140.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and Regression Trees*. Belmont, CA: Wadsworth International Group.

Brodley, C. E. & Friedl, M. A. (1999). Identifying mislabeled training data. *Journal of Artificial Intelligence Research, 11*, 131–167.

Casella, G. & Berger, R. L. (1990). *Statistical Inference*. Belmont, CA: Duxbury Press.

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation, 10*(7), 1895–1924.

Dietterich, T. G. & Bakiri, G. (1995). Solving multiclass learning problems via error-correcting output codes. *Journal of Artificial Intelligence Research, 2*, 263–286.

Feelders, A. & Verkooijen, W. (1996). On the statistical comparison of inductive learning methods. In *Fifth International Workshop on Artificial Intelligence and Statistics*. Ft. Lauderdale, FL. Proceedings available as Fisher, D. H. & Lenz, H.-J., Ed. (1996), *Learning from Data: Artificial Intelligence and Statistics V*. New York, NY: Springer.

Freund, R. J. & Wilson, W. J. (1997). *Statistical Methods, revised edition*. San Diego, CA: Academic Press.

Freund, Y. & Schapire, R. (1996). Experiments with a new boosting algorithm. In *Proceedings of the Thirteenth International Conference on Machine Learning* (pp. 146–156). San Mateo, CA.

Kohavi, R. (1996). Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* (pp. 202–207). Portland, OR.

Kohavi, R. & Kunz, C. (1997). Option decision trees with majority votes. In *Proceedings of the Fourteenth International Conference on Machine Learning* (pp. 161–169). Nashville, TN.

Kohavi, R., Sommerfield, D., & Dougherty, J. (1996). Data mining using $\mathcal{MLC}$++: A machine learning library in C++. In *Tools with Artificial Intelligence* (pp. 234–245). `http://www.sgi.com/Technology/mlc`.

Mansour, Y. (1997). Pessimistic decision tree pruning based on tree size. In *Proceedings of the Fourteenth International Conference on Machine Learning* (pp. 195–201). Nashville, TN.

Merz, C. J. & Murphy, P. M. (1997). UCI repository of machine learning databases. `http://www.ics.uci.edu/~mlearn/MLRepository.html`.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. San Mateo, California: Morgan Kaufmann.

Salzberg, S. (1997). On comparing classifiers: Pitfalls to avoid and a recommended approach. *Data Mining and Knowledge Discovery, 1*(3), 317–328.

Snedecor, G. W. & Cochran, W. G. (1989). *Statistical Methods, eighth edition*. Ames, Iowa: Iowa State University Press.

Steel, R. G. D. (1997). *Principles and Procedures of Statistics: A Biometrical Approach*. New York, NY: McGraw-Hill.

Utgoff, P., Berkman, N. C., & Clouse, J. A. (1997). Decision tree induction based on efficient tree restructuring. *Machine Learning, 29*(1), 5–44.

Wolpert, D. H. (1994). Off-training set error and a priori distinctions between learning algorithms. Technical Report SFI TR 94-12-123, The Santa Fe Institute.

Received
Accepted
Final manuscript