Journal of Machine Learning Research 7 (2006) 1–30 Submitted 8/04; Revised 4/05; Published 1/06

Statistical Comparisons of Classifiers over Multiple Data Sets

Janez Demšar JANEZ.DEMSAR@FRI.UNI-LJ.SI
Faculty of Computer and Information Science
Tržaška 25
Ljubljana, Slovenia

Editor: Dale Schuurmans

Abstract

While methods for comparing two learning algorithms on a single data set have been scrutinized for quite some time already, the issue of statistical tests for comparisons of more algorithms on multiple data sets, which is even more essential to typical machine learning studies, has been all but ignored. This article reviews the current practice and then theoretically and empirically examines several suitable tests. Based on that, we recommend a set of simple, yet safe and robust non-parametric tests for statistical comparisons of classifiers: the Wilcoxon signed ranks test for comparison of two classifiers and the Friedman test with the corresponding post-hoc tests for comparison of more classifiers over multiple data sets. Results of the latter can also be neatly presented with the newly introduced CD (critical difference) diagrams.

Keywords: comparative studies, statistical methods, Wilcoxon signed ranks test, Friedman test, multiple comparisons tests

1. Introduction

Over the last years, the machine learning community has become increasingly aware of the need for statistical validation of the published results. This can be attributed to the maturity of the area, the increasing number of real-world applications and the availability of open machine learning frameworks that make it easy to develop new algorithms or modify the existing, and compare them among themselves.

In a typical machine learning paper, a new machine learning algorithm, a part of it or some new pre- or postprocessing step has been proposed, and the implicit hypothesis is made that such an enhancement yields an improved performance over the existing algorithm(s). Alternatively, various solutions to a problem are proposed and the goal is to tell the successful from the failed. A number of test data sets is selected for testing, the algorithms are run and the quality of the resulting models is evaluated using an appropriate measure, most commonly classification accuracy. The remaining step, and the topic of this paper, is to statistically verify the hypothesis of improved performance.

The following section explores the related theoretical work and existing practice. Various researchers have addressed the problem of comparing two classifiers on a single data set and proposed several solutions. Their message has been taken by the community, and the overly confident paired t-tests over cross validation folds are giving place to the McNemar test and 5×2 cross validation. On the other side, comparing multiple classifiers over multiple data sets—a situation which is even more common, especially when general performance and not the performance on certain specific

problem is tested—is still theoretically unexplored and left to various *ad hoc* procedures that either lack statistical ground or use statistical methods in inappropriate ways. To see what is used in the actual practice, we have studied the recent (1999-2003) proceedings of the International Conference on Machine Learning. We observed that many otherwise excellent and innovative machine learning papers end by drawing conclusions from a matrix of, for instance, McNemar's tests comparing all pairs of classifiers, as if the tests for multiple comparisons, such as ANOVA and Friedman test are yet to be invented.

The core of the paper is the study of the statistical tests that could be (or already are) used for comparing two or more classifiers on multiple data sets. Formally, assume that we have tested k learning algorithms on N data sets. Let c_i^j be the performance score of the j-th algorithm on the i-th data set. The task is to decide whether, based on the values c_i^j, the algorithms are statistically significantly different and, in the case of more than two algorithms, which are the particular algorithms that differ in performance. We will not record the variance of these scores, $\sigma_{c_i^j}$, but will only assume that the measured results are "reliable"; to that end, we require that enough experiments were done on each data set and, preferably, that all the algorithms were evaluated using the same random samples. We make no other assumptions about the sampling scheme.

In Section 3 we shall observe the theoretical assumptions behind each test in the light of our problem. Although some of the tests are quite common in machine learning literature, many researchers seem ignorant about what the tests actually measure and which circumstances they are suitable for. We will also show how to present the results of multiple comparisons with neat space-friendly graphs. In Section 4 we shall provide some empirical insights into the properties of the tests.

2. Previous Work

Statistical evaluation of experimental results has been considered an essential part of validation of new machine learning methods for quite some time. The tests used have however long been rather naive and unverified. While the procedures for comparison of a pair of classifiers on a single problem have been proposed almost a decade ago, comparative studies with more classifiers and/or more data sets still employ partial and unsatisfactory solutions.

2.1 Related Theoretical Work

One of the most cited papers from this area is the one by Dietterich (1998). After describing the taxonomy of statistical questions in machine learning, he focuses on the question of deciding which of the two algorithms under study will produce more accurate classifiers when tested on a given data set. He examines five statistical tests and concludes the analysis by recommending the newly crafted 5×2cv t-test that overcomes the problem of underestimated variance and the consequently elevated Type I error of the more traditional paired t-test over folds of the usual k-fold cross validation. For the cases where running the algorithm for multiple times is not appropriate, Dietterich finds McNemar's test on misclassification matrix as powerful as the 5×2cv t-test. He warns against t-tests after repetitive random sampling and also discourages using t-tests after cross-validation. The 5×2cv t-test has been improved by Alpaydın (1999) who constructed a more robust 5×2cv F test with a lower type I error and higher power.

Bouckaert (2003) argues that theoretical degrees of freedom are incorrect due to dependencies between the experiments and that empirically found values should be used instead, while Nadeau and Bengio (2000) propose the corrected resampled t-test that adjusts the variance based on the overlaps between subsets of examples. Bouckaert and Frank (Bouckaert and Frank, 2004; Bouckaert, 2004) also investigated the replicability of machine learning experiments, found the 5×2cv t-test dissatisfactory and opted for the corrected resampled t-test. For a more general work on the problem of estimating the variance of k-fold cross validation, see the work of Bengio and Grandvalet (2004).

None of the above studies deal with evaluating the performance of multiple classifiers and neither studies the applicability of the statistics when classifiers are tested over multiple data sets. For the former case, Salzberg (1997) mentions ANOVA as one of the possible solutions, but afterwards describes the binomial test with the Bonferroni correction for multiple comparisons. As Salzberg himself notes, binomial testing lacks the power of the better non-parametric tests and the Bonferroni correction is overly radical. Vázquez et al. (2001) and Pizarro et al. (2002), for instance, use ANOVA and Friedman's test for comparison of multiple models (in particular, neural networks) on a single data set.

Finally, for comparison of classifiers over multiple data sets, Hull (1994) was, to the best of our knowledge, the first who used non-parametric tests for comparing classifiers in information retrieval and assessment of relevance of documents (see also Schütze et al., 1995). Brazdil and Soares (2000) used average ranks to compare classification algorithms. Pursuing a different goal of choosing the optimal algorithm, they do not statistically test the significance of differences between them.

2.2 Testing in Practice: Analysis of ICML Papers

We analyzed the papers from the proceedings of five recent International Conferences on Machine Learning (1999-2003). We have focused on the papers that compare at least two classifiers by measuring their classification accuracy, mean squared error, AUC (Beck and Schultz, 1986), precision/recall or some other model performance score.

The sampling methods and measures used for evaluating the performance of classifiers are not directly relevant for this study. It is astounding though that classification accuracy is usually still the only measure used, despite the voices from the medical (Beck and Schultz, 1986; Bellazzi and Zupan, 1998) and the machine learning community (Provost et al., 1998; Langley, 2000) urging that other measures, such as AUC, should be used as well. The only real competition to classification accuracy are the measures that are used in the area of document retrieval. This is also the only field where the abundance of data permits the use of separate testing data sets instead of using cross validation or random sampling.

Of greater interest to our paper are the methods for analysis of differences between the algorithms. The studied papers published the results of two or more classifiers over multiple data sets, usually in a tabular form. We did not record how many of them include (informal) statements about the overall performance of the classifiers. However, from one quarter and up to a half of the papers include some statistical procedure either for determining the optimal method or for comparing the performances among themselves.

The most straightforward way to compare classifiers is to compute the average over all data sets; such averaging appears naive and is seldom used. Pairwise t-tests are about the only method used for assessing statistical significance of differences. They fall into three categories: only two methods

	1999	2000	2001	2002	2003
Total number of papers	54	152	80	87	118
Relevant papers for our study	19	45	25	31	54
Sampling method [%]					
cross validation, leave-one-out	22	49	44	42	56
random resampling	11	29	44	32	54
separate subset	5	11	0	13	9
Score function [%]					
classification accuracy	74	67	84	84	70
classification accuracy - *exclusively*	68	60	80	58	67
recall, precision...	21	18	16	25	19
ROC, AUC	0	4	4	13	9
deviations, confidence intervals	32	42	48	42	19
Overall comparison of classifiers [%]	53	44	44	26	45
averages over the data sets	0	4	6	0	10
t-test to compare two algorithms	16	11	4	6	7
pairwise t-test one vs. others	5	11	16	3	7
pairwise t-test each vs. each	16	13	4	6	4
counts of wins/ties/losses	5	4	0	6	9
counts of *significant* wins/ties/losses	16	4	8	16	6

Table 1: An overview of the papers accepted to International Conference on Machine Learning in years 1999—2003. The reported percentages (the third line and below) apply to the number of papers relevant for our study.

are compared, one method (a new method or the base method) is compared to the others, or all methods are compared to each other. Despite the repetitive warnings against multiple hypotheses testing, the Bonferroni correction is used only in a few ICML papers annually. A common non-parametric approach is to count the number of times an algorithm performs better, worse or equally to the others; counting is sometimes pairwise, resulting in a matrix of wins/ties/losses count, and the alternative is to count the number of data sets on which the algorithm outperformed all the others. Some authors prefer to count only the differences that were statistically significant; for verifying this, they use various techniques for comparison of two algorithms that were reviewed above.

This figures need to be taken with some caution. Some papers do not explicitly describe the sampling and testing methods used. Besides, it can often be hard to decide whether a specific sampling procedure, test or measure of quality is equivalent to the general one or not.

3. Statistics and Tests for Comparison of Classifiers

The overview shows that there is no established procedure for comparing classifiers over multiple data sets. Various researchers adopt different statistical and common-sense techniques to decide whether the differences between the algorithms are real or random. In this section we shall examine

several known and less known statistical tests, and study their suitability for our purpose from the point of what they actually measure and of their safety regarding the assumptions they make about the data.

As a starting point, two or more learning algorithms have been run on a suitable set of data sets and were evaluated using classification accuracy, AUC or some other measure (see Tables 2 and 6 for an example). We do not record the variance of these results over multiple samples, and therefore assume nothing about the sampling scheme. The only requirement is that the compiled results provide reliable estimates of the algorithms' performance on each data set. In the usual experimental setups, these numbers come from cross-validation or from repeated stratified random splits onto training and testing data sets.

There is a fundamental difference between the tests used to assess the difference between two classifiers on a single data set and the differences over multiple data sets. When testing on a single data set, we usually compute the mean performance and its variance over repetitive training and testing on random samples of examples. Since these samples are usually related, a lot of care is needed in designing the statistical procedures and tests that avoid problems with biased estimations of variance.

In our task, multiple resampling from each data set is used only to assess the performance score and not its variance. The sources of the variance are the differences in performance over (independent) data sets and not on (usually dependent) samples, so the elevated Type 1 error is not an issue. Since multiple resampling does not bias the score estimation, various types of cross-validation or leave-one-out procedures can be used without any risk.

Furthermore, the problem of correct statistical tests for comparing classifiers on a single data set is not related to the comparison on multiple data sets in the sense that we would first have to solve the former problem in order to tackle the latter. Since running the algorithms on multiple data sets naturally gives a sample of independent measurements, such comparisons are even simpler than comparisons on a single data set.

We should also stress that the "sample size" in the following section will refer to the number of data sets used, not to the number of training/testing samples drawn from each individual set or to the number of instances in each set. The sample size can therefore be as small as five and is usually well below 30.

3.1 Comparisons of Two Classifiers

In the discussion of the tests for comparisons of two classifiers over multiple data sets we will make two points. We shall warn against the widely used t-test as usually conceptually inappropriate and statistically unsafe. Since we will finally recommend the Wilcoxon (1945) signed-ranks test, it will be presented with more details. Another, even more rarely used test is the sign test which is weaker than the Wilcoxon test but also has its distinct merits. The other message will be that the described statistics measure differences between the classifiers from different aspects, so the selection of the test should be based not only on statistical appropriateness but also on what we intend to measure.

3.1.1 AVERAGING OVER DATA SETS

Some authors of machine learning papers compute the average classification accuracies of classifiers across the tested data sets. In words of Webb (2000), "it is debatable whether error rates in different domains are commensurable, and hence whether averaging error rates across domains is very mean-

ingful". If the results on different data sets are not comparable, their averages are meaningless. A different case are studies in which the algorithms are compared on a set of related problems, such as medical databases for a certain disease from different institutions or various text mining problems with similar properties.

Averages are also susceptible to outliers. They allow classifier's excellent performance on one data set to compensate for the overall bad performance, or the opposite, a total failure on one domain can prevail over the fair results on most others. There may be situations in which such behaviour is desired, while in general we probably prefer classifiers that behave well on as many problems as possible, which makes averaging over data sets inappropriate.

Given that not many papers report such averages, we can assume that the community generally finds them meaningless. Consequently, averages are also not used (nor useful) for statistical inference with the z- or t-test.

3.1.2 PAIRED T-TEST

A common way to test whether the difference between two classifiers' results over various data sets is non-random is to compute a paired t-test, which checks whether the average difference in their performance over the data sets is significantly different from zero.

Let c_i^1 and c_i^2 be performance scores of two classifiers on the i-th out of N data sets and let d_i be the difference $c_i^2 - c_i^1$. The t statistics is computed as $\overline{d}/\sigma_{\overline{d}}$ and is distributed according to the Student distribution with $N-1$ degrees of freedom.

In our context, the t-test suffers from three weaknesses. The first is commensurability: the t-test only makes sense when the differences over the data sets are commensurate. In this view, using the paired t-test for comparing a pair of classifiers makes as little sense as computing the averages over data sets. The average difference $\overline{d}$ equals the difference between the averaged scores of the two classifiers, $\overline{d} = \overline{c^2} - \overline{c^1}$. The only distinction between this form of the t-test and comparing the two averages (as those discussed above) directly using the t-test for unrelated samples is in the denominator: the paired t-test decreases the standard error $\sigma_{\overline{d}}$ by the variance between the data sets (or, put another way, by the covariance between the classifiers).

Webb (2000) approaches the problem of commensurability by computing the geometric means of relative ratios, $(\prod_i c_i^1/c_i^2)^{1/N}$. Since this equals to $e^{1/N\sum_i(\ln c_i^1 - \ln c_i^2)}$, this statistic is essentially the same as the ordinary averages, except that it compares logarithms of scores. The utility of this transformation is thus rather questionable. Quinlan (1996) computes arithmetic means of relative ratios; due to skewed distributions, these cannot be used in the t-test without further manipulation. A simpler way of compensating for different complexity of the problems is to divide the difference by the average score, $d_i = \frac{c_i^1 - c_i^2}{(c_i^1 + c_i^2)/2}$.

The second problem with the t-test is that unless the sample size is large enough ($\sim$ 30 data sets), the paired t-test requires that the differences between the two random variables compared are distributed normally. The nature of our problems does not give any provisions for normality and the number of data sets is usually much less than 30. Ironically, the Kolmogorov-Smirnov and similar tests for testing the normality of distributions have little power on small samples, that is, they are unlikely to detect abnormalities and warn against using the t-test. Therefore, for using the t-test we need normal distributions because we have small samples, but the small samples also prohibit us from checking the distribution shape.

	C4.5	C4.5+m	difference	rank
adult (sample)	0.763	0.768	+0.005	3.5
breast cancer	0.599	0.591	−0.008	7
breast cancer wisconsin	0.954	0.971	+0.017	9
cmc	0.628	0.661	+0.033	12
ionosphere	0.882	0.888	+0.006	5
iris	0.936	0.931	−0.005	3.5
liver disorders	0.661	0.668	+0.007	6
lung cancer	0.583	0.583	0.000	1.5
lymphography	0.775	0.838	+0.063	14
mushroom	1.000	1.000	0.000	1.5
primary tumor	0.940	0.962	+0.022	11
rheum	0.619	0.666	+0.047	13
voting	0.972	0.981	+0.009	8
wine	0.957	0.978	+0.021	10

Table 2: Comparison of AUC for C4.5 with m = 0 and C4.5 with m tuned for the optimal AUC. The columns on the right-hand illustrate the computation and would normally not be published in an actual paper.

The third problem is that the t-test is, just as averaging over data sets, affected by outliers which skew the test statistics and decrease the test's power by increasing the estimated standard error.

3.1.3 WILCOXON SIGNED-RANKS TEST

The Wilcoxon signed-ranks test (Wilcoxon, 1945) is a non-parametric alternative to the paired t-test, which ranks the differences in performances of two classifiers for each data set, ignoring the signs, and compares the ranks for the positive and the negative differences.

Let d_i again be the difference between the performance scores of the two classifiers on i-th out of N data sets. The differences are ranked according to their absolute values; average ranks are assigned in case of ties. Let R^+ be the sum of ranks for the data sets on which the second algorithm outperformed the first, and R^- the sum of ranks for the opposite. Ranks of $d_i = 0$ are split evenly among the sums; if there is an odd number of them, one is ignored:

$$R^+ = \sum_{d_i>0} \text{rank}(d_i) + \frac{1}{2}\sum_{d_i=0} \text{rank}(d_i) \qquad R^- = \sum_{d_i<0} \text{rank}(d_i) + \frac{1}{2}\sum_{d_i=0} \text{rank}(d_i).$$

Let T be the smaller of the sums, $T = \min(R^+, R^-)$. Most books on general statistics include a table of exact critical values for T for N up to 25 (or sometimes more). For a larger number of data sets, the statistics

$$z = \frac{T - \frac{1}{4}N(N+1)}{\sqrt{\frac{1}{24}N(N+1)(2N+1)}}$$

is distributed approximately normally. With $\alpha = 0.05$, the null-hypothesis can be rejected if z is smaller than -1.96.

Let us illustrate the procedure on an example. Table 2 shows the comparison of AUC for C4.5 with m (the minimal number of examples in a leaf) set to zero and C4.5 with m tuned for the optimal AUC. For the latter, AUC has been computed with 5-fold internal cross validation on training examples for $m \in \{0,1,2,3,5,10,15,20,50\}$. The experiments were performed on 14 data sets from the UCI repository with binary class attribute. We used the original Quinlan's C4.5 code, equipped with an interface that integrates it into machine learning system Orange (Demšar and Zupan, 2004), which provided us with the cross validation procedures, classes for tuning arguments, and the scoring functions. We are trying to reject the null-hypothesis that both algorithms perform equally well.

There are two data sets on which the classifiers performed equally (lung-cancer and mushroom); if there was an odd number of them, we would ignore one. The ranks are assigned from the lowest to the highest absolute difference, and the equal differences (0.000, ± 0.005) are assigned average ranks.

The sum of ranks for the positive differences is $R^+ = 3.5+9+12+5+6+14+11+13+8+10+1.5 = 93$ and the sum of ranks for the negative differences equals $R^- = 7+3.5+1.5 = 12$. According to the table of exact critical values for the Wilcoxon's test, for a confidence level of $\alpha = 0.05$ and $N = 14$ data sets, the difference between the classifiers is significant if the smaller of the sums is equal or less than 21. We therefore reject the null-hypothesis.

The Wilcoxon signed ranks test is more sensible than the t-test. It assumes commensurability of differences, but only qualitatively: greater differences still count more, which is probably desired, but the absolute magnitudes are ignored. From the statistical point of view, the test is safer since it does not assume normal distributions. Also, the outliers (exceptionally good/bad performances on a few data sets) have less effect on the Wilcoxon than on the t-test.

The Wilcoxon test assumes continuous differences d_i, therefore they should not be rounded to, say, one or two decimals since this would decrease the power of the test due to a high number of ties.

When the assumptions of the paired t-test are met, the Wilcoxon signed-ranks test is less powerful than the paired t-test. On the other hand, when the assumptions are violated, the Wilcoxon test can be even more powerful than the t-test.

3.1.4 COUNTS OF WINS, LOSSES AND TIES: SIGN TEST

A popular way to compare the overall performances of classifiers is to count the number of data sets on which an algorithm is the overall winner. When multiple algorithms are compared, pairwise comparisons are sometimes organized in a matrix.

Some authors also use these counts in inferential statistics, with a form of binomial test that is known as the sign test (Sheskin, 2000; Salzberg, 1997). If the two algorithms compared are, as assumed under the null-hypothesis, equivalent, each should win on approximately $N/2$ out of N data sets. The number of wins is distributed according to the binomial distribution; the critical number of wins can be found in Table 3. For a greater number of data sets, the number of wins is under the null-hypothesis distributed according to $N(N/2, \sqrt{N}/2)$, which allows for the use of z-test: if the number of wins is at least $N/2 + 1.96\sqrt{N}/2$ (or, for a quick rule of a thumb, $N/2 + \sqrt{N}$), the algorithm is significantly better with $p < 0.05$. Since tied matches support the null-hypothesis we should not discount them but split them evenly between the two classifiers; if there is an odd number of them, we again ignore one.

#data sets	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
$w_{0.05}$	5	6	7	7	8	9	9	10	10	11	12	12	13	13	14	15	15	16	17	18	18
$w_{0.10}$	5	6	6	7	7	8	9	9	10	10	11	12	12	13	13	14	14	15	16	16	17

Table 3: Critical values for the two-tailed sign test at $\alpha = 0.05$ and $\alpha = 0.10$. A classifier is significantly better than another if it performs better on at least w_α data sets.

In example from Table 2, C4.5+m was better on 11 out of 14 data sets (counting also one of the two data sets on which the two classifiers were tied). According to Table 3 this difference is significant with $p < 0.05$.

This test does not assume any commensurability of scores or differences nor does it assume normal distributions and is thus applicable to any data (as long as the observations, *i.e.* the data sets, are independent). On the other hand, it is much weaker than the Wilcoxon signed-ranks test. According to Table 3, the sign test will not reject the null-hypothesis unless one algorithm almost always outperforms the other.

Some authors prefer to count only the significant wins and losses, where the significance is determined using a statistical test on each data set, for instance Dietterich's 5×2cv. The reasoning behind this practice is that "some wins and losses are random and these should not count". This would be a valid argument if statistical tests could distinguish between the random and non-random differences. However, statistical tests only measure the improbability of the obtained experimental result if the null hypothesis was correct, which is not even the (im)probability of the null-hypothesis.

For the sake of argument, suppose that we compared two algorithms on one thousand different data sets. In each and every case, algorithm A was better than algorithm B, but the difference was never significant. It is true that for each single case the difference between the two algorithms can be attributed to a random chance, but how likely is it that one algorithm was just lucky in all 1000 out of 1000 independent experiments?

Contrary to the popular belief, counting only significant wins and losses therefore does not make the tests more but rather less reliable, since it draws an arbitrary threshold of $p < 0.05$ between what counts and what does not.

3.2 Comparisons of Multiple Classifiers

None of the above tests was designed for reasoning about the means of multiple random variables. Many authors of machine learning papers nevertheless use them for that purpose. A common example of such questionable procedure would be comparing seven algorithms by conducting all 21 paired t-tests and reporting results like "algorithm A was found significantly better than B and C, and algorithms A and E were significantly better than D, while there were no significant differences between other pairs". When so many tests are made, a certain proportion of the null hypotheses is rejected due to random chance, so listing them makes little sense.

The issue of multiple hypothesis testing is a well-known statistical problem. The usual goal is to control the *family-wise error*, the probability of making at least one Type 1 error in any of the comparisons. In machine learning literature, Salzberg (1997) mentions a general solution for the

problem of multiple testing, the Bonferroni correction, and notes that it is usually very conservative and weak since it supposes the independence of the hypotheses.

Statistics offers more powerful specialized procedures for testing the significance of differences between multiple means. In our situation, the most interesting two are the well-known ANOVA and its non-parametric counterpart, the Friedman test. The latter, and especially its corresponding Nemenyi post-hoc test are less known and the literature on them is less abundant; for this reason, we present them in more detail.

3.2.1 ANOVA

The common statistical method for testing the differences between more than two related sample means is the *repeated-measures ANOVA* (or *within-subjects ANOVA*) (Fisher, 1959). The "related samples" are again the performances of the classifiers measured across the same data sets, preferably using the same splits onto training and testing sets. The null-hypothesis being tested is that all classifiers perform the same and the observed differences are merely random.

ANOVA divides the total variability into the variability between the classifiers, variability between the data sets and the residual (error) variability. If the between-classifiers variability is significantly larger than the error variability, we can reject the null-hypothesis and conclude that there *are* some differences between the classifiers. In this case, we can proceed with a post-hoc test to find out which classifiers actually differ. Of many such tests for ANOVA, the two most suitable for our situation are the Tukey test (Tukey, 1949) for comparing all classifiers with each other and the Dunnett test (Dunnett, 1980) for comparisons of all classifiers with the control (for instance, comparing the base classifier and some proposed improvements, or comparing the newly proposed classifier with several existing methods). Both procedures compute the standard error of the difference between two classifiers by dividing the residual variance by the number of data sets. To make pairwise comparisons between the classifiers, the corresponding differences in performances are divided by the standard error and compared with the critical value. The two procedures are thus similar to a t-test, except that the critical values tabulated by Tukey and Dunnett are higher to ensure that there is at most 5 % chance that one of the pairwise differences will be erroneously found significant.

Unfortunately, ANOVA is based on assumptions which are most probably violated when analyzing the performance of machine learning algorithms. First, ANOVA assumes that the samples are drawn from normal distributions. In general, there is no guarantee for normality of classification accuracy distributions across a set of problems. Admittedly, even if distributions are abnormal this is a minor problem and many statisticians would not object to using ANOVA unless the distributions were, for instance, clearly bi-modal (Hamilton, 1990). The second and more important assumption of the repeated-measures ANOVA is sphericity (a property similar to the homogeneity of variance in the usual ANOVA, which requires that the random variables have equal variance). Due to the nature of the learning algorithms and data sets this cannot be taken for granted. Violations of these assumptions have an even greater effect on the post-hoc tests. ANOVA therefore does not seem to be a suitable omnibus test for the typical machine learning studies.

We will not describe ANOVA and its post-hoc tests in more details due to our reservations about the parametric tests and, especially, since these tests are well known and described in statistical literature (Zar, 1998; Sheskin, 2000).

		ANOVA		
		$p < 0.01$	$0.01 \leq p \leq 0.05$	$0.05 < p$
Friedman	$p < 0.01$	16	1	0
test	$0.01 \leq p \leq 0.05$	4	1	4
	$0.05 < p$	0	2	28

Table 4: Friedman's comparison of his test and the repeated-measures ANOVA on 56 independent problems (Friedman, 1940).

3.2.2 FRIEDMAN TEST

The Friedman test (Friedman, 1937, 1940) is a non-parametric equivalent of the repeated-measures ANOVA. It ranks the algorithms for each data set separately, the best performing algorithm getting the rank of 1, the second best rank 2..., as shown in Table 6. In case of ties (like in iris, lung cancer, mushroom and primary tumor), average ranks are assigned.

Let r_i^j be the rank of the j-th of k algorithms on the i-th of N data sets. The Friedman test compares the average ranks of algorithms, $R_j = \frac{1}{N}\sum_i r_i^j$. Under the null-hypothesis, which states that all the algorithms are equivalent and so their ranks R_j should be equal, the Friedman statistic

$$\chi_F^2 = \frac{12N}{k(k+1)}\left[\sum_j R_j^2 - \frac{k(k+1)^2}{4}\right]$$

is distributed according to χ_F^2 with $k-1$ degrees of freedom, when N and k are big enough (as a rule of a thumb, $N > 10$ and $k > 5$). For a smaller number of algorithms and data sets, exact critical values have been computed (Zar, 1998; Sheskin, 2000).

Iman and Davenport (1980) showed that Friedman's χ_F^2 is undesirably conservative and derived a better statistic

$$F_F = \frac{(N-1)\chi_F^2}{N(k-1)-\chi_F^2}$$

which is distributed according to the F-distribution with $k-1$ and $(k-1)(N-1)$ degrees of freedom. The table of critical values can be found in any statistical book.

As for the two-classifier comparisons, the (non-parametric) Friedman test has theoretically less power than (parametric) ANOVA when the ANOVA's assumptions are met, but this does not need to be the case when they are not. Friedman (1940) experimentally compared ANOVA and his test on 56 independent problems and showed that the two methods mostly agree (Table 4). When one method finds significance at $p < 0.01$, the other shows significance of at least $p < 0.05$. Only in 2 cases did ANOVA find significant what was insignificant for Friedman, while the opposite happened in 4 cases.

If the null-hypothesis is rejected, we can proceed with a post-hoc test. The Nemenyi test (Nemenyi, 1963) is similar to the Tukey test for ANOVA and is used when all classifiers are compared to each other. The performance of two classifiers is significantly different if the corresponding average ranks differ by at least the critical difference

$$CD = q_\alpha\sqrt{\frac{k(k+1)}{6N}}$$

#classifiers	2	3	4	5	6	7	8	9	10
$q_{0.05}$	1.960	2.343	2.569	2.728	2.850	2.949	3.031	3.102	3.164
$q_{0.10}$	1.645	2.052	2.291	2.459	2.589	2.693	2.780	2.855	2.920

(a) Critical values for the two-tailed Nemenyi test

#classifiers	2	3	4	5	6	7	8	9	10
$q_{0.05}$	1.960	2.241	2.394	2.498	2.576	2.638	2.690	2.724	2.773
$q_{0.10}$	1.645	1.960	2.128	2.241	2.326	2.394	2.450	2.498	2.539

(b) Critical values for the two-tailed Bonferroni-Dunn test; the number of classifiers include the control classifier.

Table 5: Critical values for post-hoc tests after the Friedman test

where critical values q_α are based on the Studentized range statistic divided by $\sqrt{2}$ (Table 5(a)).

When all classifiers are compared with a control classifier, we can instead of the Nemenyi test use one of the general procedures for controlling the family-wise error in multiple hypothesis testing, such as the Bonferroni correction or similar procedures. Although these methods are generally conservative and can have little power, they are in this specific case more powerful than the Nemenyi test, since the latter adjusts the critical value for making $k(k-1)/2$ comparisons while when comparing with a control we only make $k-1$ comparisons.

The test statistics for comparing the i-th and j-th classifier using these methods is

$$z = (R_i - R_j) \Big/ \sqrt{\frac{k(k+1)}{6N}} \,.$$

The z value is used to find the corresponding probability from the table of normal distribution, which is then compared with an appropriate α. The tests differ in the way they adjust the value of α to compensate for multiple comparisons.

The Bonferroni-Dunn test (Dunn, 1961) controls the family-wise error rate by dividing α by the number of comparisons made ($k-1$, in our case). The alternative way to compute the same test is to calculate the CD using the same equation as for the Nemenyi test, but using the critical values for $\alpha/(k-1)$ (for convenience, they are given in Table 5(b)). The comparison between the tables for Nemenyi's and Dunn's test shows that the power of the post-hoc test is much greater when all classifiers are compared only to a control classifier and not between themselves. We thus should not make pairwise comparisons when we in fact only test whether a newly proposed method is better than the existing ones.

For a contrast from the single-step Bonferroni-Dunn procedure, step-up and step-down procedures sequentially test the hypotheses ordered by their significance. We will denote the ordered p values by p_1, p_2, ..., so that $p_1 \leq p_2 \leq \ldots \leq p_{k-1}$. The simplest such methods are due to Holm (1979) and Hochberg (1988). They both compare each p_i with $\alpha/(k-i)$, but differ in the order

of the tests.[1] Holm's step-down procedure starts with the most significant p value. If p_1 is below $\alpha/(k-1)$, the corresponding hypothesis is rejected and we are allowed to compare p_2 with $\alpha/(k-2)$. If the second hypothesis is rejected, the test proceeds with the third, and so on. As soon as a certain null hypothesis cannot be rejected, all the remaining hypotheses are retained as well. Hochberg's step-up procedure works in the opposite direction, comparing the largest p value with α, the next largest with $\alpha/2$ and so forth until it encounters a hypothesis it *can* reject. All hypotheses with smaller p values are then rejected as well.

Hommel's procedure (Hommel, 1988) is more complicated to compute and understand. First, we need to find the largest j for which $p_{n-j+k} > k\alpha/j$ for all $k = 1..j$. If no such j exists, we can reject all hypotheses, otherwise we reject all for which $p_i \leq \alpha/j$.

Holm's procedure is more powerful than the Bonferroni-Dunn's and makes no additional assumptions about the hypotheses tested. The only advantage of the Bonferroni-Dunn test seems to be that it is easier to describe and visualize because it uses the same CD for all comparisons. In turn, Hochberg's and Hommel's methods reject more hypotheses than Holm's, yet they may under some circumstances exceed the prescribed family-wise error since they are based on the Simes conjecture which is still being investigated. It has been reported (Holland, 1991) that the differences between the enhanced methods are in practice rather small, therefore the more complex Hommel method offers no great advantage over the simple Holm method.

Although we here use these procedures only as post-hoc tests for the Friedman test, they can be used generally for controlling the family-wise error when multiple hypotheses of possibly various types are tested. There exist other similar methods, as well as some methods that instead of controlling the family-wise error control the number of falsely rejected null-hypotheses (false discovery rate, FDR). The latter are less suitable for the evaluation of machine learning algorithms since they require the researcher to decide for the acceptable false discovery rate. A more complete formal description and discussion of all these procedures was written, for instance, by Shaffer (1995).

Sometimes the Friedman test reports a significant difference but the post-hoc test fails to detect it. This is due to the lower power of the latter. No other conclusions than that some algorithms do differ can be drawn in this case. In our experiments this has, however, occurred only in a few cases out of one thousand.

The procedure is illustrated by the data from Table 6, which compares four algorithms: C4.5 with `m` fixed to 0 and `cf` (confidence interval) to 0.25, C4.5 with `m` fitted in 5-fold internal cross validation, C4.5 with `cf` fitted the same way and, finally, C4.5 in which we fitted both arguments, trying all combinations of their values. Parameter `m` was set to 0, 1, 2, 3, 5, 10, 15, 20, 50 and `cf` to 0, 0.1, 0.25 and 0.5.

Average ranks by themselves provide a fair comparison of the algorithms. On average, C4.5+m and C4.5+m+cf ranked the second (with ranks 2.000 and 1.964, respectively), and C4.5 and C4.5+cf the third (3.143 and 2.893). The Friedman test checks whether the measured average ranks are significantly different from the mean rank $R_j = 2.5$ expected under the null-hypothesis:

$$\chi_F^2 = \frac{12 \cdot 14}{4 \cdot 5}\left[(3.143^2 + 2.000^2 + 2.893^2 + 1.964^2) - \frac{4 \cdot 5^2}{4}\right] = 9.28$$

$$F_F = \frac{13 \cdot 9.28}{14 \cdot 3 - 9.28} = 3.69.$$

1. In the usual definitions of these procedures k would denote the number of hypotheses, while in our case the number of hypotheses is $k-1$, hence the differences in the formulae.

	C4.5	C4.5+m	C4.5+cf	C4.5+m+cf
adult (sample)	0.763 (4)	0.768 (3)	0.771 (2)	0.798 (1)
breast cancer	0.599 (1)	0.591 (2)	0.590 (3)	0.569 (4)
breast cancer wisconsin	0.954 (4)	0.971 (1)	0.968 (2)	0.967 (3)
cmc	0.628 (4)	0.661 (1)	0.654 (3)	0.657 (2)
ionosphere	0.882 (4)	0.888 (2)	0.886 (3)	0.898 (1)
iris	0.936 (1)	0.931 (2.5)	0.916 (4)	0.931 (2.5)
liver disorders	0.661 (3)	0.668 (2)	0.609 (4)	0.685 (1)
lung cancer	0.583 (2.5)	0.583 (2.5)	0.563 (4)	0.625 (1)
lymphography	0.775 (4)	0.838 (3)	0.866 (2)	0.875 (1)
mushroom	1.000 (2.5)	1.000 (2.5)	1.000 (2.5)	1.000 (2.5)
primary tumor	0.940 (4)	0.962 (2.5)	0.965 (1)	0.962 (2.5)
rheum	0.619 (3)	0.666 (2)	0.614 (4)	0.669 (1)
voting	0.972 (4)	0.981 (1)	0.975 (2)	0.975 (3)
wine	0.957 (3)	0.978 (1)	0.946 (4)	0.970 (2)
average rank	3.143	2.000	2.893	1.964

Table 6: Comparison of AUC between C4.5 with `m` = 0 and C4.5 with parameters `m` and/or `cf` tuned for the optimal AUC. The ranks in the parentheses are used in computation of the Friedman test and would usually not be published in an actual paper.

With four algorithms and 14 data sets, F_F is distributed according to the F distribution with $4-1=3$ and $(4-1)\times(14-1)=39$ degrees of freedom. The critical value of $F(3,39)$ for $\alpha=0.05$ is 2.85, so we reject the null-hypothesis.

Further analysis depends upon what we intended to study. If no classifier is singled out, we use the Nemenyi test for pairwise comparisons. The critical value (Table 5(a)) is 2.569 and the corresponding CD is $2.569\sqrt{\frac{4\cdot5}{6\cdot14}}=1.25$. Since even the difference between the best and the worst performing algorithm is already smaller than that, we can conclude that the post-hoc test is not powerful enough to detect any significant differences between the algorithms.

At p=0.10, CD is $2.291\sqrt{\frac{4\cdot5}{6\cdot14}}=1.12$. We can identify two groups of algorithms: the performance of pure C4.5 is significantly worse than that of C4.5+m and C4.5+m+cf. We cannot tell which group C4.5+cf belongs to. Concluding that it belongs to both would be a statistical nonsense since a subject cannot come from two different populations. The correct statistical statement would be that *the experimental data is not sufficient to reach any conclusion regarding C4.5+cf*.

The other possible hypothesis made before collecting the data could be that it is possible to improve on C4.5's performance by tuning its parameters. The easiest way to verify this is to compute the CD with the Bonferroni-Dunn test. In Table 5(b) we find that the critical value $q_{0.05}$ for 4 classifiers is 2.394, so CD is $2.394\sqrt{\frac{4\cdot5}{6\cdot14}}=1.16$. C4.5+m+cf performs significantly better than C4.5 ($3.143-1.964=1.179>1.16$) and C4.5+cf does not ($3.143-2.893=0.250<1.16$), while C4.5+m is just below the critical difference, but close to it ($3.143-2.000=1.143\approx1.16$). We can conclude that the experiments showed that fitting `m` seems to help, while we did not detect any significant improvement by fitting `cf`.

For the other tests we have to compute and order the corresponding statistics and p values. The standard error is $SE=\sqrt{\frac{4\cdot5}{6\cdot14}}=0.488$.

i	classifier	$z = (R_0 - R_i)/SE$	p	α/i
1	C4.5+m+cf	$(3.143-1.964)/0.488 = 2.416$	0.016	0.017
2	C4.5+m	$(3.143-2.000)/0.488 = 2.342$	0.019	0.025
3	C4.5+cf	$(3.143-2.893)/0.488 = 0.512$	0.607	0.050

The Holm procedure rejects the first and then the second hypothesis since the corresponding p values are smaller than the adjusted α's. The third hypothesis cannot be rejected; if there were any more, we would have to retain them, too.

The Hochberg procedure starts from the bottom. Unable to reject the last hypothesis, it check the second last, rejects it and among with it all the hypotheses with smaller p values (the top-most one).

Finally, the Hommel procedure finds that $j = 3$ does not satisfy the condition at $k = 2$. The maximal value of j is 2, and the first two hypotheses can be rejected since their p values are below $\alpha/2$.

All step-down and step-up procedure found C4.5+cf+m and C4.5+m significantly different from C4.5, while the Bonferroni-Dunn test found C4.5 and C4.5+m too similar.

3.2.3 CONSIDERING MULTIPLE REPETITIONS OF EXPERIMENTS

In our examples we have used AUCs measured and averaged over repetitions of training/testing episodes. For instance, each cell in Table 6 represents an average over five-fold cross validation. Could we also consider the variance, or even the results of individual folds?

There are variations of the ANOVA and the Friedman test which can consider *multiple observations per cell* provided that the observations are independent (Zar, 1998). This is not the case here, since training data in multiple random samples overlaps. We are not aware of any statistical test that could take this into account.

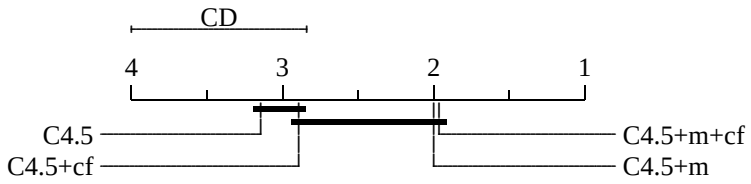
3.2.4 GRAPHICAL PRESENTATION OF RESULTS

When multiple classifiers are compared, the results of the post-hoc tests can be visually represented with a simple diagram. Figure 1 shows the results of the analysis of the data from Table 6. The top line in the diagram is the axis on which we plot the average ranks of methods. The axis is turned so that the lowest (best) ranks are to the right since we perceive the methods on the right side as better.

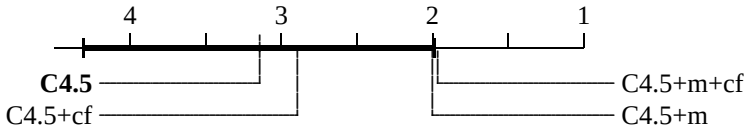
When comparing all the algorithms against each other, we connect the groups of algorithms that are not significantly different (Figure 1(a)). We also show the critical difference above the graph.

If the methods are compared to the control using the Bonferroni-Dunn test we can mark the interval of one CD to the left and right of the average rank of the control algorithm (Figure 1(b)). Any algorithm with the rank outside this area is significantly different from the control. Similar graphs for the other post-hoc tests would need to plot a different adjusted critical interval for each classifier and specify the procedure used for testing and the corresponding order of comparisons, which could easily become confusing.

For another example, Figure 2 graphically represents the comparison of feature scoring measures for the problem of keyword prediction on five domains formed from the Yahoo hierarchy studied by Mladenić and Grobelnik (1999). The analysis reveals that Information gain performs significantly worse than Weight of evidence, Cross entropy Txt and Odds ratio, which seem to have



(a) Comparison of all classifiers against each other with the Nemenyi test. Groups of classifiers that are not significantly different (at $p = 0.10$) are connected.



(b) Comparison of one classifier against the others with the Bonferroni-Dunn test. All classifiers with ranks outside the marked interval are significantly different ($p < 0.05$) from the control.

Figure 1: Visualization of post-hoc tests for data from Table 6.

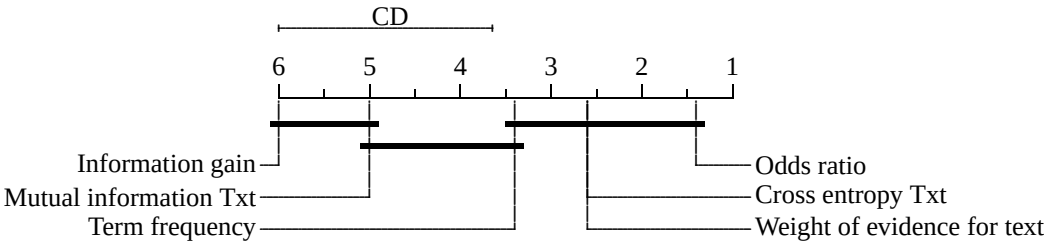


Figure 2: Comparison of recalls for various feature selection measures; analysis of the results from the paper by Mladenić and Grobelnik (1999).

equivalent performances. The data is not sufficient to conclude whether Mutual information Txt performs the same as Information gain or Term Frequency, and similarly, whether Term Frequency is equivalent to Mutual information Txt or to the better three methods.

4. Empirical Comparison of Tests

We experimentally observed two properties of the described tests: their replicability and the likelihood of rejecting the null-hypothesis. Performing the experiments to answer questions like "which statistical test is most likely to give the correct result" or "which test has the lowest Type 1/Type 2 error rate" would be a pointless exercise since the proposed inferential tests suppose different kinds

of commensurability and thus compare the classifiers from different aspects. The "correct answer", rejection or non-rejection of the null-hypothesis, is thus not well determined and is, in a sense, related to the choice of the test.

4.1 Experimental Setup

We examined the behaviour of the studied tests through the experiments in which we repeatedly compared the learning algorithms on sets of ten randomly drawn data sets and recorded the p values returned by the tests.

4.1.1 DATA SETS AND LEARNING ALGORITHMS

We based our experiments on several common learning algorithms and their variations: C4.5, C4.5 with m and C4.5 with cf fitted for optimal accuracy, another tree learning algorithm implemented in Orange (with features similar to the original C4.5), naive Bayesian learner that models continuous probabilities using LOESS (Cleveland, 1979), naive Bayesian learner with continuous attributes discretized using Fayyad-Irani's discretization (Fayyad and Irani, 1993) and kNN (k=10, neighbour weights adjusted with the Gaussian kernel).

We have compiled a sample of forty real-world data sets,[2] from the UCI machine learning repository (Blake and Merz, 1998); we have used the data sets with discrete classes and avoided artificial data sets like Monk problems. Since no classifier is optimal for all possible data sets, we have simulated experiments in which a researcher wants to show particular advantages of a particular algorithm and thus selects a corresponding compendium of data sets. We did this by measuring the classification accuracies of the classifiers on all data sets in advance by using ten-fold cross validation. When comparing two classifiers, samples of ten data sets were randomly selected so that the probability for the data set i being chosen was proportional to $1/(1+e^{-kd_i})$, where d_i is the (positive or negative) difference in the classification accuracies on that data set and k is the bias through which we can regulate the differences between the classifiers.[3] Whereas at $k = 0$ the data set selection is random with the uniform distribution, with higher values of k we are more likely to select the sets that favour a particular learning method. Note that choosing the data sets with knowing their success (as estimated in advance) is only a simulation, while the researcher would select the data sets according to other criteria. Using the described procedure in practical evaluations of algorithms would be considered cheating.

We decided to avoid "artificial" classifiers and data sets constructed specifically for testing the statistical tests, such as those used, for instance, by Dietterich (1998). In such experimental procedures some assumptions need to be made about the real-world data sets and the learning algorithms, and the artificial data and algorithms are constructed in a way that mimics the supposed real-world situation in a controllable manner. In our case, we would construct two or more classifiers with a prescribed probability of failure over a set of (possible imaginary) data sets so that we could,

2. The data sets used are: adult, balance-scale, bands, breast cancer (haberman), breast cancer (lju), breast cancer (wisc), car evaluation, contraceptive method choice, credit screening, dermatology, ecoli, glass identification, hayes-roth, hepatitis, housing, imports-85, ionosphere, iris, liver disorders, lung cancer, lymphography, mushrooms, pima indians diabetes, post-operative, primary tumor, promoters, rheumatism, servo, shuttle landing, soybean, spambase, spect, spectf, teaching assistant evaluation, tic tac toe, titanic, voting, waveform, wine recognition, yeast.
3. The function used is the logistic function. It was chosen for its convenient shape; we do not claim that such relation actually occurs in practice when selecting the data sets for experiments.

knowing the correct hypothesis, observe the Type 1 and 2 error rates of the proposed statistical tests.

Unfortunately, we do not know what should be our assumptions about the real world. To what extent are the classification accuracies (or other measures of success) incommensurable? How (ab)normal is their distribution? How homogenous is the variance? Moreover, if we do make certain assumptions, the statistical theory is already able to tell the results of the experiments that we are setting up. Since the statistical tests which we use are theoretically well understood, we do not need to test the tests but the compliance of the real-world data to their assumptions. In other words, we know, from the theory, that the t-test on a small sample (that is, on a small number of data sets) requires the normal distribution, so by constructing an artificial environment that will yield non-normal distributions we can make the t-test fail. The real question however is whether the *real world* distributions are normal enough for the t-test to work.

Cannot we test the assumptions directly? As already mentioned in the description of the t-test, the tests like the Kolmogorov-Smirnov test of normality are unreliable on small samples where they are very unlikely to detect abnormalities. And even if we did have suitable tests at our disposal, they would only compute the degree of (ab)normality of the distribution, non-homogeneity of variance *etc*, and not the sample's suitability for t-test.

Our decision to use real-world learning algorithms and data sets in unmodified form prevents us from artificially setting the differences between them by making them intentionally misclassify a certain proportion of examples. This is however compensated by our method of selecting the data sets: we can regulate the differences between the learning algorithms by affecting the data set selection through regulating the bias k. In this way, we perform the experiments on real-world data sets and algorithms, and yet observe the performance of the statistics at various degrees of differences between the classifiers.

4.1.2 MEASURES OF POWER AND REPLICABILITY

Formally, the power of a statistical test is defined as the probability that the test will (correctly) reject the false null-hypothesis. Since our criterion of what is actually false is related to the selection of the test (which should be based on the kind of differences between the classifiers we want to measure), we can only observe the probability of the rejection of the null-hypothesis, which is nevertheless related to the power.

We do this in two ways. First, we set the significance level at 5% and observe in how many experiments out of one thousand does a particular test reject the null-hypothesis. The shortcoming of this is that it observes only the behaviour of statistics at around $p = 0.05$ (which is probably what we are interested in), yet it can miss a bigger picture. We therefore also observed the average p values as another measure of "power" of the test: the lower the values, the more likely it is for a test to reject the null-hypothesis at a set confidence level.

The two measures for assessing the power of the tests lead to two related measures of replicability. Bouckaert (2004) proposed a definition which can be used in conjuction with counting the rejections of the null-hypothesis. He defined the replicability as the probability that two experiments with the same pair of algorithms will produce the same results, that is, that both experiments accept or reject the null-hypothesis, and devised the optimal unbiased estimator of this probability,

$$R(e) = \sum_{1 \le i < j \le n} \frac{I(e_i = e_j)}{n(n-1)/2}$$

where e_i is the outcome of the i-th experiment out of n (e_i is 1 if the null-hypothesis is accepted, 0 if it is not) and I is the indicator function which is 1 if its argument is true and 0 otherwise. Bouckaert also describes a simpler way to compute $R(e)$: if the hypothesis was accepted in p and rejected in q experiments out of n, $R(e)$ equals $(p(p-1)+q(q-1))/n(n-1)$. The minimal value of R, 0.5, occurs when $p=q=n/2$, and the maximal, 1.0, when either p or q is zero.

The disadvantage of this measure is that a statistical test will show a low replicability when the difference between the classifiers is marginally significant. When comparing two tests of different power, the one with results closer to the chosen α will usually be deemed as less reliable.

When the power is estimated by the average of p values, the replicability is naturally defined through their variance. The variance of p is between 0 and 0.25; the latter occurs when one half of p's equals zero and the other half equals one.[4] To allow for comparisons with Bouckaert's $R(e)$, we define the replicability with respect to the variance of p as

$$R(p) = 1 - 2 \cdot \mathrm{var}(p) = 1 - 2\frac{\sum_i (p_i - \overline{p})^2}{n-1}.$$

A problem with this measure of replicability when used in our experimental procedure is that when the bias k increases, the variability of the data set selection decreases and so does the variance of p. The size of the effect depends on the number of data sets. Judged by the results of the experiments, our collection of forty data sets is large enough to keep the variability practically unaffected for the used values of k (see the left graph in Figure 4.c; if the variability of selections decreased, the variance of p could not remain constant).

The described definitions of replicability are related. Since $I(e_i = e_j)$ equals $1-(e_i-e_j)^2$, we can reformulate $R(e)$ as

$$R(e) = \sum_{1 \le i < j \le n} \frac{1-(e_i-e_j)^2}{n(n-1)/2} = 1 - \sum_i \sum_j \frac{(e_i-e_j)^2}{n(n-1)} = 1 - \sum_i \sum_j \frac{((e_i-\overline{e})-(e_j-\overline{e}))^2}{n(n-1)}.$$

From here, it is easy to verify that

$$R(e) = 1 - 2\frac{\sum_i (e_i - \overline{e})^2}{n-1}.$$

The fact that Bouckaert's formula is the optimal unbiased estimator for $R(e)$ is related to $\sum_i (e_i - \overline{e})^2/(n-1)$ being the optimal unbiased estimator of the population variance.

4.2 Comparisons of Two Classifiers

We have tested four statistics for comparisons of two classifiers: the paired t-test on absolute and on relative differences, the Wilcoxon test and the sign test. The experiments were run on 1000 random selections of ten data sets, as described above.

The graphs on the left hand side of Figure 3 show the average p values returned by the tests as a function of the bias k when comparing C4.5-cf, naive Bayesian classifier and kNN (note that the scale is turned upside down so the curve rises when the power of the test increases). The graphs on the right hand side show the number of experiments in which the hypothesis was rejected at

4. Since we estimate the population variance from the sample variance, the estimated variance will be higher by $0.25/(n-1)$. With any decent number of experiments, the difference is however negligible.

$\alpha = 5\%$. To demonstrate the relation between power (as we measure it) and Bouckaert's measure of replicability we have added the right axis that shows $R(e)$ corresponding to the number of rejected hypothesis.

Note that at $k = 0$ the number of experiments in which the null hypothesis is rejected is not 50%. Lower settings of k do not imply that both algorithms compared should perform approximately equally, but only that we do not (artificially) bias the data sets selection to favour one of them. Therefore, at $k = 0$ the tests reflect the number of rejections of the null-hypothesis on a completely random selection of data sets from our collection.

Both variations of the t-test give similar results, with the test on relative differences being slightly, yet consistently weaker. The Wilcoxon signed-ranks test gives much lower p values and is more likely to reject the null-hypothesis than t-tests in almost all cases. The sign test is, as known from the theory, much weaker than the other tests.

The two measures of replicability give quite different results. Judged by $R(p)$ (graphs on the left hand side of Figure 4), the Wilcoxon test exhibits the smallest variation of p values. For a contrast, Bouckaert's $R(e)$ (right hand side of Figure 4) shows the Wilcoxon test as the least reliable. However, the shape of the curves on these graphs and the right axes in Figure 3 clearly show that the test is less reliable (according to $R(e)$) when the p values are closer to 0.05, so the Wilcoxon test seems unreliable due to its higher power keeping it closer to p=0.05 than the other tests.

Table 7 shows comparisons of all seven classifiers with k set to 15. The numbers below the diagonal show the average p values and the related replicability $R(p)$, and the numbers above the diagonal represent the number of experiments in which the null-hypothesis was rejected at $\alpha = 5\%$ and the related $R(e)$. The table again shows that the Wilcoxon test almost always returns lower p values than other tests and more often rejects the null hypothesis. Measured by $R(p)$, the Wilcoxon test also has the highest replicability. $R(e)$, on the other hand, again prefers other tests with p values farther from the critical 0.05.

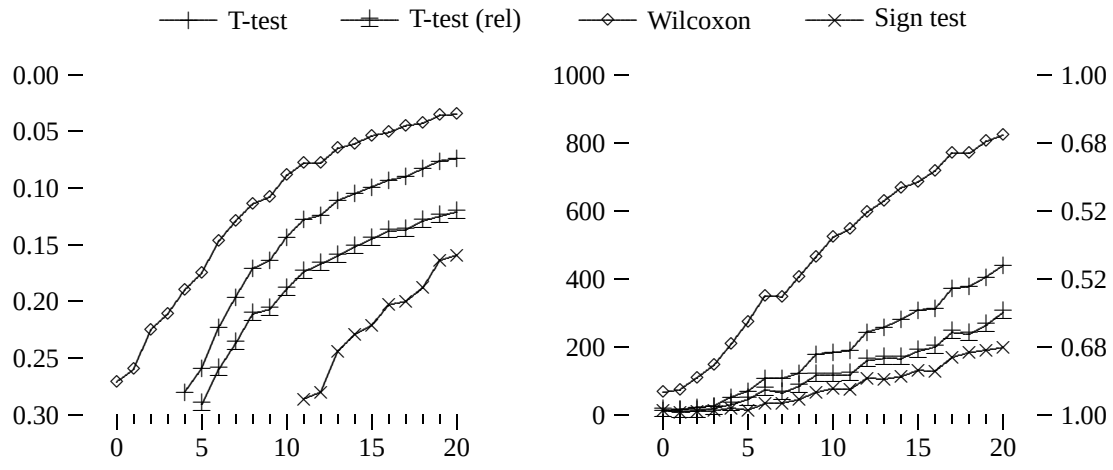
Overall, it is known that parametric tests are more likely to reject the null-hypothesis than the non-parametric unless their assumptions are violated. Our results suggest that the latter is indeed happening in machine learning studies that compare algorithms across collections of data sets. We therefore recommend using the Wilcoxon test, unless the t-test assumptions are met, either because we have many data sets or because we have reasons to believe that the measure of performance across data sets is distributed normally. The sign test, as the third alternative, is too weak to be generally useful.

Low values of $R(e)$ suggest that we should ensure the reliability of the results (especially when the differences between classifiers are marginally significant) by running the experiments on as many appropriate data sets as possible.

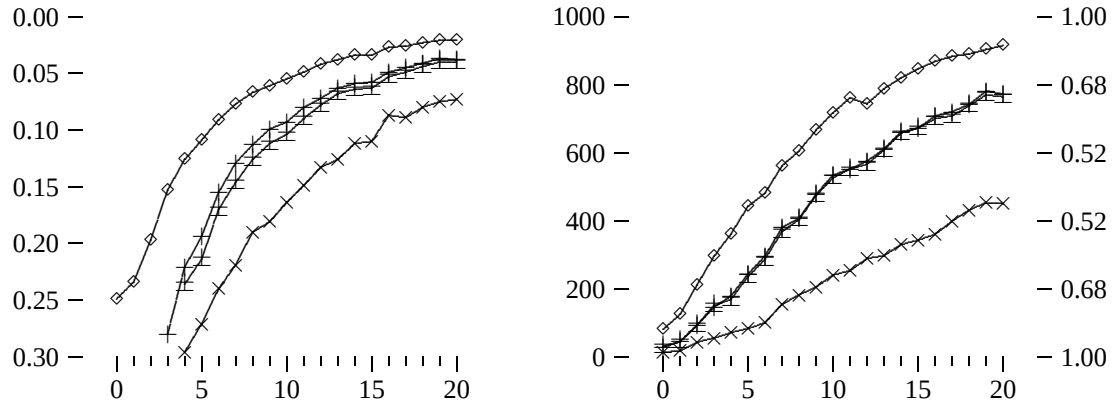
4.3 Comparisons of Multiple Classifiers

For comparison of multiple classifiers, samples of data sets were selected with the probabilities computed from the differences in the classification accuracy of C4.5 and naive Bayesian classifier with Fayyad-Irani discretization. These two classifiers were chosen for no particular reason; we have verified that the choice has no practical effect on the results.

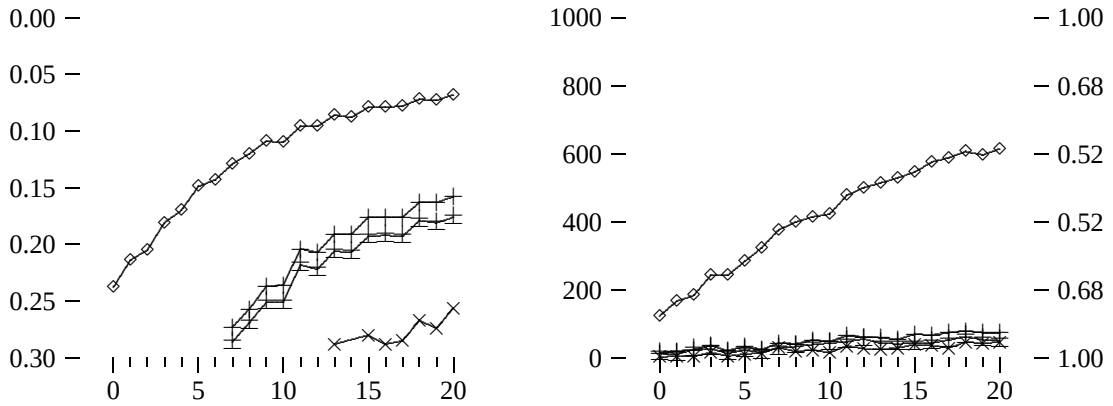
Results are shown in Figure 5. When the algorithms are more similar (at smaller values of k), the non-parametric Friedman test again appears stronger than the parametric, ANOVA. At greater



(a) C45-cf vs. naive Bayes

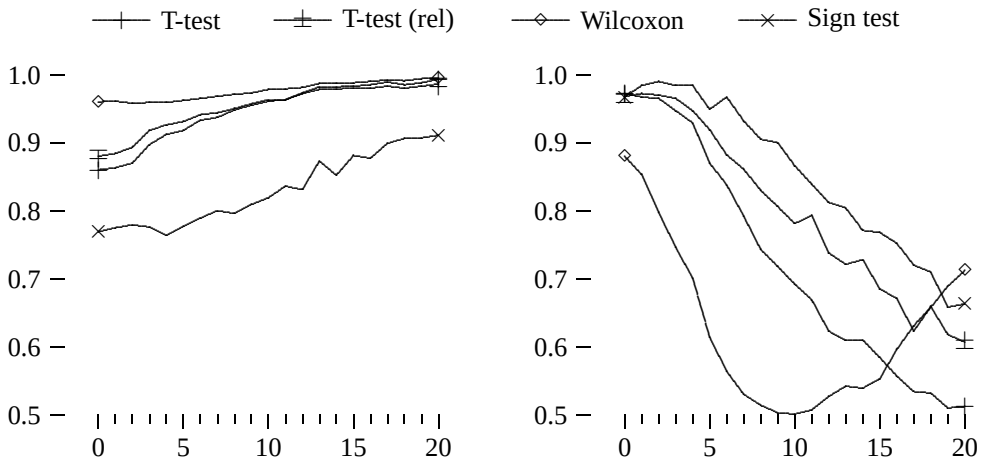


(b) C45-cf vs. kNN

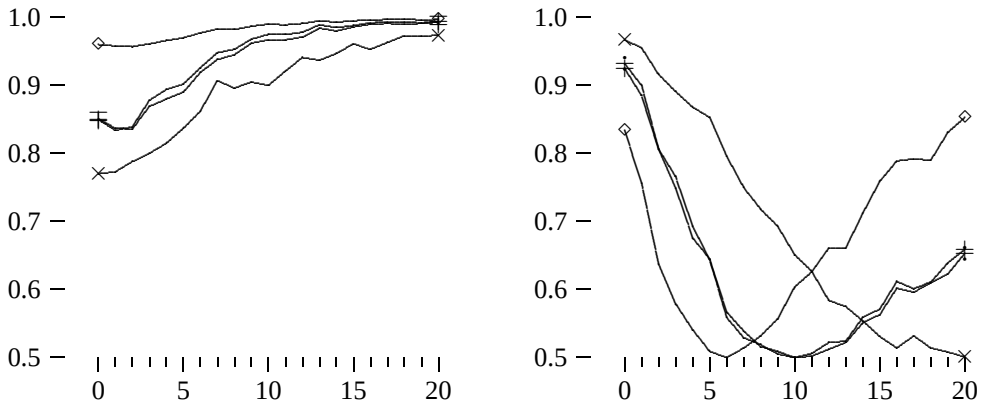


(c) Naive Bayes vs. kNN

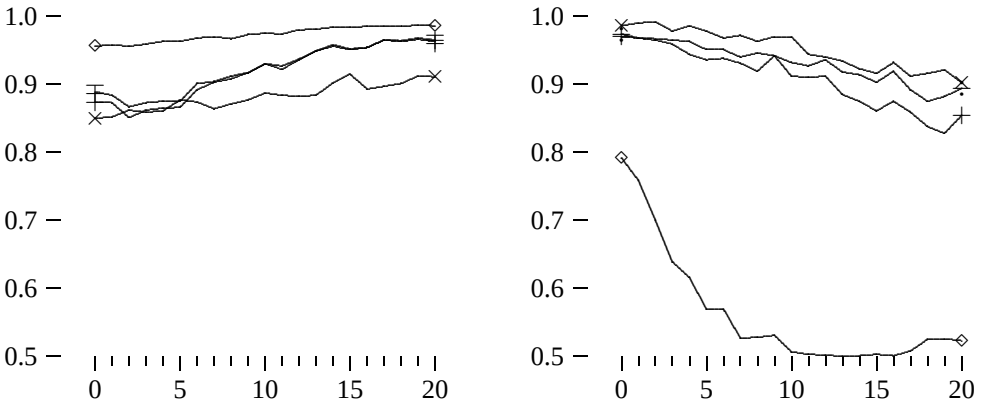
Figure 3: Power of statistical tests for comparison of two classifiers. Left: p values as a function of bias (k). Right: number of times the hypothesis was rejected (left axis) and the Bouckaert's R (right axis).



(a) C45-cf vs. naive Bayes



(b) C45-cf vs. kNN



(c) Naive Bayes vs. kNN

Figure 4: Replicability of tests for comparison of two classifiers: variance-based $R(p)$ (left) and Bouckaert's $R(e)$ (right).

	c45	c45-m	c45-cf	tree	bayes	disc-bayes	knn
c45		154/.74	709/.59	818/.70	178/.71	0/1.00	151/.74
c45-m	.16/.96		307/.57	909/.83	300/.58	0/1.00	376/.53
c45-cf	.05/.99	.10/.98		758/.63	335/.55	0/1.00	167/.72
tree	.04/.98	.02/1.00	.05/.98		679/.56	162/.73	592/.52
bayes	.15/.96	.12/.97	.11/.97	.05/.99		0/1.00	2/1.00
disc-bayes	.41/.92	.20/.95	.28/.92	.18/.94	.20/.97		981/.96
knn	.16/.96	.10/.98	.14/.97	.06/.99	.35/.94	.01/1.00	

(a) Paired t-test

	c45	c45-m	c45-cf	tree	bayes	disc-bayes	knn
c45		75/.86	592/.52	809/.69	181/.70	0/1.00	184/.70
c45-m	.17/.96		238/.64	848/.74	314/.57	0/1.00	438/.51
c45-cf	.06/.99	.11/.98		729/.60	361/.54	0/1.00	216/.66
tree	.04/.99	.03/1.00	.06/.98		662/.55	79/.85	584/.51
bayes	.16/.95	.12/.97	.11/.97	.05/.99		0/1.00	1/1.00
disc-bayes	.36/.94	.20/.96	.27/.94	.19/.95	.24/.98		970/.94
knn	.14/.96	.09/.98	.13/.97	.06/.99	.35/.95	.01/1.00	

(b) Paired t-test on relative differences

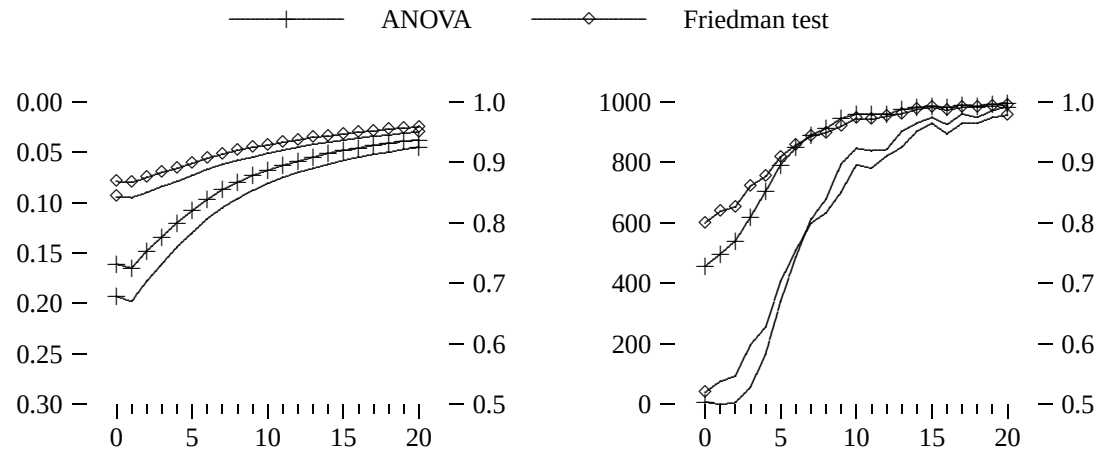
	c45	c45-m	c45-cf	tree	bayes	disc-bayes	knn
c45		521/.50	884/.79	897/.82	662/.55	81/.85	618/.53
c45-m	.08/.98		774/.65	983/.97	710/.59	351/.54	750/.62
c45-cf	.03/1.00	.04/.99		854/.75	804/.68	172/.71	720/.60
tree	.02/1.00	.01/1.00	.03/1.00		915/.84	521/.50	920/.85
bayes	.06/.99	.05/.99	.04/.99	.02/1.00		94/.83	102/.82
disc-bayes	.22/.96	.11/.98	.16/.97	.08/.98	.18/.97		999/1.00
knn	.07/.98	.04/.99	.05/.99	.02/1.00	.22/.96	.00/1.00	

(c) Wilcoxon signed-ranks test

	c45	c45-m	c45-cf	tree	bayes	disc-bayes	knn
c45		157/.74	323/.56	653/.55	171/.72	48/.91	110/.80
c45-m	.21/.90		205/.67	863/.76	299/.58	156/.74	256/.62
c45-cf	.10/.98	.16/.93		513/.50	423/.51	95/.83	229/.65
tree	.05/.99	.02/1.00	.09/.97		460/.50	210/.67	486/.50
bayes	.19/.89	.13/.94	.08/.97	.08/.97		0/1.00	1/1.00
disc-bayes	.29/.89	.18/.93	.25/.89	.18/.93	.52/.78		850/.74
knn	.25/.85	.14/.93	.15/.93	.07/.97	.45/.86	.01/1.00	

(d) Sign test

Table 7: Tests for comparisons of two classifiers: average p-values and $R(p)$ (below diagonal), and the number of null-hypothesis rejections and $R(e)$ (above diagonal).



(a) Average p values (left axis) and $R(p)$ (no symbols on lines, right axis)

(b) Number of experiments in which the null-hypothesis was rejected (left axis) and the corresponding $R(e)$ (no symbols on lines, right axis)

Figure 5: Comparison of ANOVA and Friedman test

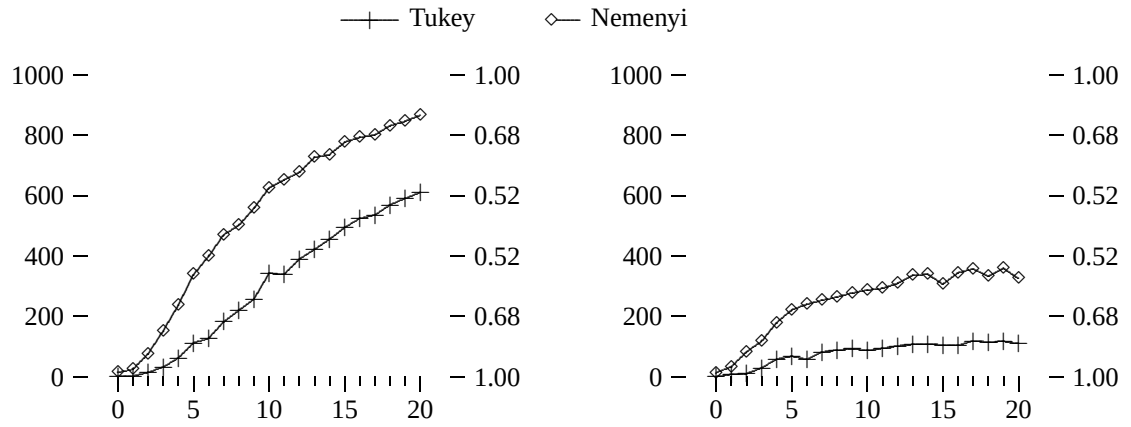
differences between the algorithms (k at around 10, in our experimental setup), ANOVA catches up and the two tests give similar results.

Replicability of the Friedman test is higher than that of ANOVA when measured by $R(p)$ and, due to the similar power of the tests, comparable when measured by $R(e)$. Altogether, replicability seems somewhat smaller than the replicability of the tests for comparisons of two classifiers. Therefore, as common sense would suggest, when comparing multiple classifiers, it is even more important to conduct the tests on as many data sets as possible.

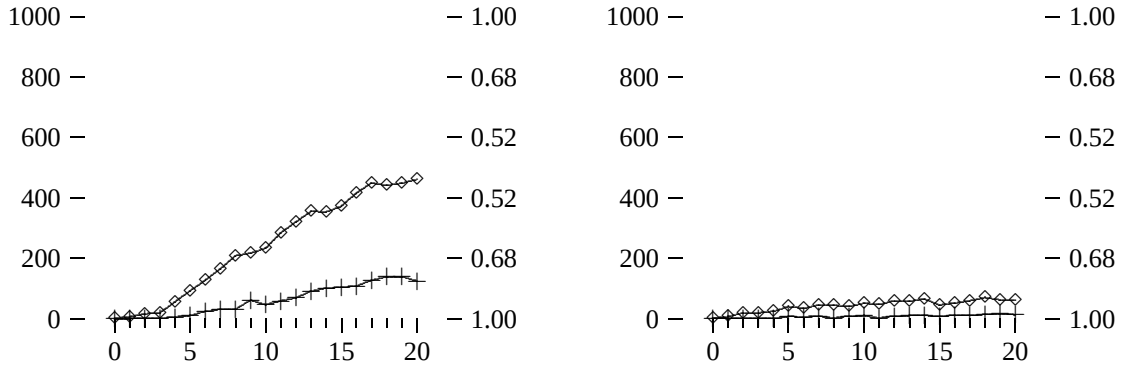
Figure 6 shows the comparison between the parametric Tukey and the non-parametric Nemenyi test. We counted the number of times they rejected the equivalence of C4.5-cf and naive Bayesian classifier and the equivalence of C4.5-cf and kNN (the comparison between the naive Bayesian classifier and kNN, which was included in previous tests, was less interesting since the null hypothesis was very seldom rejected). The two graphs on the left represent experiments in which the selection was based on the differences between the two algorithms compared on the graph, while for the right two graphs we used differences between the C4.5-cf and the average of the other six classifiers tested. In all cases, we have compared all seven algorithms, but presented only the number of rejections for the pair on the graph. The non-parametric test again more often rejects the null-hypothesis than the parametric one.

We do not show the p values and the corresponding replicabilities since they cannot always be computed or compared in all procedures due to different orders of testing.

Figure 7 compares post hoc tests for comparisons with a control classifier, using the same two ways of selecting data sets as in Figure 6. When the differences are large, the power of all tests is comparable, while when they are smaller the number of rejections for the parametric test seems to lag behind (we have observed this same pattern on other combinations of algorithms). The order of the non-parametric tests is as expected from the theory, although it is interesting to note that the Holm and Hochberg tests give practically equal results.



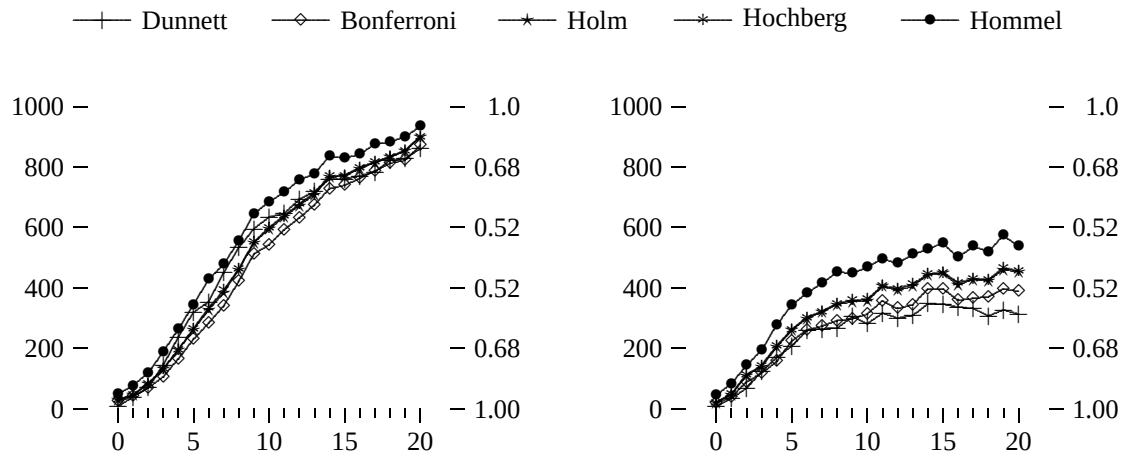
(a) C45-cf vs. naive Bayes



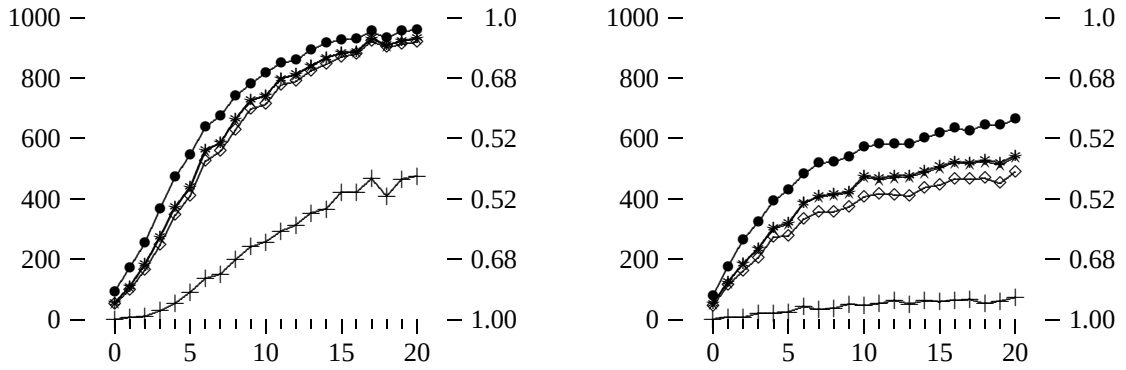
(b) C45-cf vs. kNN

Figure 6: Power of statistical tests for comparison of multiple classifiers. Bias is defined by the difference in performance of the two classifiers on the graph (left) or between the C4.5-cf and all other classifiers (right). The left scale on each graph gives the number of times the hypothesis was rejected and the right scale gives the corresponding $R(e)$.

These experiments again seem to favour the non-parametric tests over the parametric ones although not always as convincingly as in the case of comparisons of two classifiers. Due to the theoretical and practical advantages of the Friedman test (ease of computation and interpretation, the ability to present the overall performance of classifiers in form of ranks instead of the dubious averages), the Friedman test should be preferred over ANOVA. The corresponding non-parametric post-hoc tests give similar results, so it is upon the researcher to decide whether the slightly more powerful Hommel test is worth the complexity of its calculation as compared to the much simpler Holm test.



(a) C45-cf vs. naive Bayes



(b) C45-cf vs. kNN

Figure 7: Power of statistical tests for comparison of multiple classifiers with a control. Bias is defined by the difference in performance of the two classifiers on the graph (left) or between the C4.5-cf and the average of all other classifiers (right). The left scale on each graph gives the number of times the hypothesis was rejected and the right scale gives the corresponding Bouckaert's *R*.

5. Conclusion

Our analysis of the papers from the past International Conferences on Machine Learning has shown that many authors feel that the algorithms they propose should be compared over a set of problems and that the results can be used for drawing general conclusions. There is however no golden standard for making such comparisons and the tests performed often have dubious statistical foundations and lead to unwarranted and unverified conclusions.

While comparisons using a single data set are pestered by the biased variance estimations due to dependencies between the samples of examples drawn from the data set, in comparisons over multiple data set the variance comes from the differences between the data sets, which are usually

independent. Our setup is therefore free from the elevated Type 1 error that is common on the single data set testing procedures. The problems with the multiple data set tests are quite different, even in a sense complementary: the measurements from different data sets are usually incommensurate, and the normality of their distributions and the homogeneity of variance is questionable at best.

We theoretically and empirically analyzed three families of statistical tests that can be used for comparing two or more classifiers over multiple data sets: parametric tests (the paired t-test and ANOVA), non-parametric tests (the Wilcoxon and the Friedman test) and the non-parametric test that assumes no commensurability of the results (sign test). In the theoretical part, we specifically discussed the possible violations of the tests' assumptions by a typical machine learning data. Based on the well known statistical properties of the tests and our knowledge of the machine learning data, we concluded that the non-parametric tests should be preferred over the parametric ones.

We have observed the behaviour of the proposed statistics on several real-world classifiers and data sets. We varied the differences between the classifiers by biasing the selection of data sets, and measured the likelihood of rejection of the null-hypothesis and the replicability of the test. We have indeed found that the non-parametric tests are more likely to reject the null-hypothesis, which hints at the presence of outliers or violations of assumptions of the parametric tests and confirms our theoretical misgivings about them. The empirical analysis also shows that replicability of the tests might be a problem, thus the actual experiments should be conducted on as many data sets as possible.

In the empirical study we provided no analysis of Type 1/Type 2 error rates. The main reason for this is that the correct result—rejection or non-rejection of the null-hypothesis—is not well defined and depends upon the kind of difference between the algorithms we intend to measure. Besides, conducting the experiments in which we knew the true hypotheses would require artificial data sets and classifiers with the prescribed probabilities and distributions of errors. For this we would need to make some assumptions about the real-world distributions; these assumptions are, however, exactly what we were testing in the first place.

Overall, the non-parametric tests, namely the Wilcoxon and Friedman test are suitable for our problems. They are *appropriate* since they assume some, but limited commensurability. They are *safer than parametric tests* since they do not assume normal distributions or homogeneity of variance. As such, they can be applied to classification accuracies, error ratios or any other measure for evaluation of classifiers, including even model sizes and computation times. Empirical results suggest that they are also *stronger than the other tests studied*. The latter is particularly true when comparing a pair of classifiers.

We have proposed a visual representation of the post-hoc analysis when multiple classifiers are compared. CD diagrams are "space-friendly" and thus suitable when the length of the paper is an issue, yet they present the order of the algorithms, the magnitude of differences between them (in terms of ranks) and the significance of the observed differences much more clearly than it can be done in textual or in a pure numerical form.

There is an alternative opinion among statisticians that significance tests should not be performed at all since they are often misused, either due to misinterpretation or by putting too much stress on their results (Cohen, 1994; Schmidt, 1996; Harlow and Mulaik, 1997). Our stance is that statistical tests provide certain reassurance about the validity and non-randomness of the published results. For that to be true, they should be performed correctly and the resulting conclusions should be drawn cautiously. On the other hand, statistical tests should not be the deciding factor for or against publishing the work. Other merits of the proposed algorithm that are beyond the grasp of

statistical testing should also be considered and possibly even favoured over pure improvements in predictive power.

Acknowledgments

I wish to thank my colleagues from Artificial Intelligence Laboratory and Cognitive Modeling Laboratory at the Faculty of Computer and Information Science in Ljubljana, Slovenia, especially Blaž Zupan and Aleks Jakulin for their invaluable remarks and suggestions.

References

E. Alpaydın. Combined 5×2 F test for comparing supervised classification learning algorithms. *Neural Computation*, 11:1885–1892, 1999.

J. R. Beck and E. K. Schultz. The use of ROC curves in test performance evaluation. *Arch Pathol Lab Med*, 110:13–20, 1986.

R. Bellazzi and B. Zupan. Intelligent data analysis in medicine and pharmacology: a position statement. In *IDAMAP Workshop Notes at the 13th European Conference on Artificial Intelligence, ECAI-98*, Brighton, UK, 1998.

Y. Bengio and Y. Grandvalet. No unbiased estimator of the variance of k-fold cross-validation. *Journal of Machine Learning Research*, 5:1089–1105, 2004.

C. L. Blake and C. J. Merz. UCI repository of machine learning databases, 1998. URL `http://www.ics.uci.edu/~mlearn/MLRepository.html`.

R. R. Bouckaert. Choosing between two learning algorithms based on calibrated tests. In T. Fawcett and N. Mishra, editors, *Machine Learning, Proceedings of the Twentieth International Conference (ICML 2003), August 21-24, 2003, Washington, DC, USA*. AAAI Press, 2003.

R. R. Bouckaert. Estimating replicability of classifier learning experiments. In C Brodley, editor, *Machine Learning, Proceedings of the Twenty-First International Conference (ICML 2004)*. AAAI Press, 2004.

R. R. Bouckaert and E. Frank. Evaluating the replicability of significance tests for comparing learning algorithms. In D. Honghua, R. Srikant, and C. Zhang, editors, *Advances in Knowledge Discovery and Data Mining, 8th Pacific-Asia Conference, PAKDD 2004, Sydney, Australia, May 26-28, 2004, Proceedings*. Springer, 2004.

P. B. Brazdil and C. Soares. A comparison of ranking methods for classification algorithm selection. In *Proceedings of 11th European Conference on Machine Learning*. Springer Verlag, 2000.

W. Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74:329–336, 1979.

J. Cohen. The earth is round ($p < .05$). *American Psychologist*, 49:997 1003, 1994.

J. Demšar and B. Zupan. *Orange: From Experimental Machine Learning to Interactive Data Mining, A White Paper*. Faculty of Computer and Information Science, Ljubljana, Slovenia, 2004.

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

O. J. Dunn. Multiple comparisons among means. *Journal of the American Statistical Association*, 56:52–64, 1961.

C. W. Dunnett. A multiple comparison procedure for comparing several treatments with a control. *Journal of American Statistical Association*, 50:1096–1121, 1980.

U. M. Fayyad and K. B. Irani. Multi-interval discretization of continuous valued attributes for classification learning. In *Proceedings of the 13th International Joint Conference on Artificial Intelligence*, pages 1022–1029, Chambery, France, 1993. Morgan-Kaufmann.

R. A. Fisher. *Statistical methods and scientific inference (2nd edition)*. Hafner Publishing Co., New York, 1959.

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

M. Friedman. A comparison of alternative tests of significance for the problem of m rankings. *Annals of Mathematical Statistics*, 11:86–92, 1940.

L. C. Hamilton. *Modern Data Analysis: A First Course in Applied Statistics*. Wadsworth, Belmont, California, 1990.

L. L. Harlow and S. A. Mulaik, editors. *What If There Were No Significance Tests?* Lawrence Erlbaum Associates, July 1997.

Y. Hochberg. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, 75: 800–803, 1988.

B. Holland. On the application of three modified Bonferroni procedures to pairwise multiple comparisons in balanced repeated measures designs. *Computational Statistics Quarterly*, 6:219–231, 1991.

S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6:65–70, 1979.

G. Hommel. A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika*, 75:383–386, 1988.

D. A. Hull. *Information Retrieval Using Statistical Classification*. PhD thesis, Stanford University, November 1994.

R. L. Iman and J. M. Davenport. Approximations of the critical region of the Friedman statistic. *Communications in Statistics*, pages 571–595, 1980.

P. Langley. Crafting papers on machine learning. In *Proc. of Seventeenth International Conference on Machine Learning (ICML-2000)*, 2000.

D. Mladenić and M. Grobelnik. Feature selection for unbalanced class distribution and naive bayes. In I. Bratko and S. Džeroski, editors, *Machine Learning, Proceedings of the Sixteenth International Conference (ICML 1999), June 27-30, 2002, Bled, Slovenia*, pages 258–267. Morgan Kaufmann, 1999.

C. Nadeau and Y. Bengio. Inference for the generalization error. *Advances in Neural Information Processing Systems*, 12:239–281, 2000.

P. B. Nemenyi. *Distribution-free multiple comparisons*. PhD thesis, Princeton University, 1963.

J. Pizarro, E. Guerrero, and P. L. Galindo. Multiple comparison procedures applied to model selection. *Neurocomputing*, 48:155–173, 2002.

F. Provost, T. Fawcett, and R. Kohavi. The case against accuracy estimation for comparing induction algorithms. In J. Shavlik, editor, *Proceedings of the Fifteenth International Conference on Machine Learning (ICML-1998)*, pages 445–453, San Francisco, CA, 1998. Morgan Kaufmann Publishers.

J. R. Quinlan. Bagging, boosting, and c4.5. In *Proc. Thirteenth National Conference on Artificial Intelligence*, pages 725–730, Portland, OR, 1996. AAAI Press.

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

F. L. Schmidt. Statistical significance testing and cumulative knowledge in psychology. *Psychological Methods*, 1:115–129, 1996.

H. Schütze, D. A. Hull, and J. O. Pedersen. A comparison of classifiers and document representations for the routing problem. In E. A. Fox, P. Ingwersen, and R. Fidel, editors, *SIGIR'95, Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 229–237. ACM Press, 1995.

J. P. Shaffer. Multiple hypothesis testing. *Annual Review of Psychology*, 46:561–584, 1995.

D. J. Sheskin. *Handbook of parametric and nonparametric statistical procedures*. Chapman & Hall/CRC, 2000.

J. W. Tukey. Comparing individual means in the analysis of variance. *Biometrics*, 5:99–114, 1949.

E. G. Vázquez, A. Y. Escolano, and J. P. Junquera P. G. Riaño. Repeated measures multiple comparison procedures applied to model selection in neural networks. In *Proc. of the 6th Intl. Conf. On Artificial and Natural Neural Networks (IWANN 2001)*, pages 88–95, 2001.

G. I. Webb. Multiboosting: A technique for combining boosting and wagging. *Machine Learning*, 40:159–197, 2000.

F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics*, 1:80–83, 1945.

J. H. Zar. *Biostatistical Analysis (4th Edition)*. Prentice Hall, Englewood Clifs, New Jersey, 1998.