NOTE Communicated by Thomas Dietterich

Combined 5×2 cv F Test for Comparing Supervised Classification Learning Algorithms

Ethem Alpaydın
IDIAP, CP 592 CH-1920 Martigny, Switzerland
and
Department of Computer Engineering, Boğaziçi University, TR-80815 Istanbul, Turkey

Dietterich (1998) reviews five statistical tests and proposes the 5×2 cv t test for determining whether there is a significant difference between the error rates of two classifiers. In our experiments, we noticed that the 5×2 cv t test result may vary depending on factors that should not affect the test, and we propose a variant, the combined 5×2 cv F test, that combines multiple statistics to get a more robust test. Simulation results show that this combined version of the test has lower type I error and higher power than 5×2 cv proper.

1 Introduction

Given two learning algorithms and a training set, we want to test if the two algorithms construct classifiers that have the same error rate on a test example. The way we proceed is as follows: Given a labeled sample, we divide it into a training set and a test set (or many such pairs), train the two algorithms on the training set, and test them on the test set. We define a statistic computed from the errors of the two classifiers on the test set, which if our assumption that they do have the same error rate (the null hypothesis) holds, obeys a certain distribution. We then check the probability that the statistic we compute actually has a high enough probability of being drawn from that distribution. If so, we accept the hypothesis; otherwise we reject it and say that the two algorithms generate classifiers of different error rates. If we reject when no difference exists, we incur a type I error. If we accept when a difference exists, we incur a type II error. $1 - Pr\{\text{Type II error}\}$ is called the power of the test and is the probability of detecting a difference when a difference exists.

Dietterich (1998) analyzes in detail five statistical tests and concludes that two of them, McNemar's test and a new test, the 5×2 cv t test, have low type I error and reasonable power. He proposes to use McNemar's test if, due to high computational cost, the algorithms can be executed only once. For algorithms that can be executed 10 times, he proposes to use the 5×2 cv t test.

2 5×2 cv Test

In the 5×2 cv t test, proposed by Dietterich (1998), we perform five replications of twofold cross-validation. In each replication, the data set is divided into two equal-sized sets. $p_i^{(j)}$ is the difference between the error rates of the two classifiers on fold $j = 1, 2$ of replication $i = 1, \ldots, 5$. The average on replication i is $\overline{p}_i = (p_i^{(1)} + p_i^{(2)})/2$, and the estimated variance is $s_i^2 = (p_i^{(1)} - \overline{p}_i)^2 + (p_i^{(2)} - \overline{p}_i)^2$.

Under the null hypothesis, $p_i^{(j)}$ is the difference of two identically distributed proportions and, ignoring the fact that these proportions are not independent, $p_i^{(j)}$ can be treated as approximately normal distributed with zero mean and unknown variance σ^2 (Dietterich, 1998). Then $p_i^{(j)}/\sigma$ is approximately unit normal. If we assume $p_i^{(1)}$ and $p_i^{(2)}$ are independent normals (which is not strictly true because their training and test sets are not drawn independently of each other), then s_i^2/σ^2 has a chi-square distribution with one degree of freedom. If each of the s_i^2 is assumed to be independent (which is not true because they are all computed from the same set of available data), then

$$M = \frac{\sum_{i=1}^{5} s_i^2}{\sigma^2}$$

has a chi-square distribution with 5 degrees of freedom. If $Z \sim \mathcal{Z}$ and $X \sim \mathcal{X}_n^2$ and Z and X are independent, then

$$T_n = \frac{Z}{\sqrt{X/n}}$$

has a t-distribution with n degrees of freedom. Therefore, ignoring the various assumptions and approximations described above,

$$t = \frac{p_1^{(1)}}{\sqrt{M/5}} = \frac{p_1^{(1)}}{\sqrt{\sum_{i=1}^{5} s_i^2/5}} \tag{2.1}$$

is approximately t-distributed with 5 degrees of freedom (Dietterich, 1998). We reject the hypothesis that the two classifiers have the same error rate with 95 percent confidence if t is greater than 2.571.

We note that the numerator $p_1^{(1)}$ is arbitrary; actually there are 10 different values that can be placed in the numerator—$p_i^{(j)}$, $j = 1, 2$, $i = 1, \ldots, 5$—leading to 10 possible statistics

$$t_i^{(j)} = \frac{p_i^{(j)}}{\sqrt{\sum_{i=1}^{5} s_i^2/5}} \,. \tag{2.2}$$

Changing the numerator corresponds to changing the order of replications or folds and should not affect the result of the test. A first experiment

Table 1: Comparison of the 5×2 cv t Test with Its Combined Version.

	LP vs. MLP	
	5×2 cv $t_i^{(j)}$ Rejects Out of 10	Combined 5×2 cv F Rejects
GLASS	0	No
WINE	0	No
IRIS	2	No
THYROID	2	No
VOWEL	2	No
ODR	8	Yes
DIGIT	7	Yes
PEN	10	Yes

Notes: LP is a linear perceptron, and MLP is a multilayer perceptron with one hidden layer. Just changing the order of folds or replications (using a different numerator), the 5×2 cv t test sometimes give different results, whereas the combined version takes into account all 10 statistics and averages over this variability.

is done on eight data sets to measure the effect of changing the numerator where we compare a single-layer perceptron (LP) with a multilayer perceptron (MLP) with one hidden layer. ODR and DIGIT are two data sets on optical handwritten digit recognition, and PEN is a data set on pen-based handwritten digit recognition. (These three data sets are available from the author. The other data sets are from the UCI repository; Merz & Murphy, 1998).

As shown in Table 1, depending on which of the 10 numerators we use—that is, which of the 10 $t_i^{(j)}$, $j = 1, 2, i = 1, \ldots, 5$ we calculate—the test sometimes accepts and sometimes rejects the hypothesis. That is, if we change the order of folds or replications, we get different test results, a disturbing result since this order is not a function of the error rates of the algorithms and clearly should not affect the result of the test.

3 Combined 5×2 cv F test

A new test that combines the results of the 10 possible statistics promises to be more robust. If $p_i^{(j)}/\sigma \sim \mathcal{Z}$, then $\left(p_i^{(j)}\right)^2/\sigma^2 \sim \mathcal{X}_1^2$ and

$$N = \frac{\sum_{i=1}^{5}\sum_{j=1}^{2}\left(p_i^{(j)}\right)^2}{\sigma^2}$$

Table 2: Average and Standard Deviations of Error Rates on Test Folds of a Linear Perceptron and Multilayer Perceptrons with Different Number of Hidden Units.

	LP	MLP	MLP	MLP
IRIS	3.75, 2.05	3: 3.85, 2.57	10: 3.18, 1.95	20: 2.77, 1.73
WINE	2.84, 1.66	3: 2.86, 2.02	10: 2.57, 1.61	20: 2.63, 1.61
GLASS	38.66, 4.03	5: 37.52, 4.21	10: 35.81, 4.32	20: 35.04, 4.19
VOWEL	38.70, 2.48	5: 36.86, 2.86	10: 27.69, 2.60	20: 22.48, 2.37
ODR	5.31, 1.08	10: 5.14, 1.07	20: 3.16, 0.78	
THYROID	4.61, 0.38	10: 4.26, 0.34		

Note: The numbers of hidden units are given before the colon.

is chi-square with 10 degrees of freedom. If $X_1 \sim \mathcal{X}_n^2$ and $X_2 \sim \mathcal{X}_m^2$ and if X_1 and X_2 are independent, then (Ross, 1987)

$$\frac{X_1/n}{X_2/m} \sim F_{n,m}.$$

Therefore, we have

$$f = \frac{N/10}{M/5} = \frac{\sum_{i=1}^{5}\sum_{j=1}^{2}\left(p_i^{(j)}\right)^2}{2\sum_{i=1}^{5} s_i^2} \tag{3.1}$$

is approximately F distributed with 10 and 5 degrees of freedom, assuming N and M are independent (which is not true). For example, we reject the hypothesis that the algorithms have the same error rate with 0.95 confidence if the statistic f is greater than 4.74. Looking at Table 1, we see that the combined version combines the 10 statistics and is more robust; it is as if the combined version "takes a majority vote" over the 10 possible 5×2 cv t test results. Note that computing the f statistic brings no additional cost.

4 Comparing Type I and Type II Errors

On six data sets we trained a one-layer LP and MLPs with different numbers of hidden units to check for type I and type II errors. The average and standard deviation of test error rates for LP and MLP are given in Table 2. The probabilities are computed as proportions of rejects over 1000 runs.

To compare the type I error of 5×2 cv test with its combined version, we use two MLPs with equal numbers of hidden units. Thus the hypothesis is true, and any reject is a type I error. On six data sets using different numbers of hidden units, we have designed 15 experiments of 1000 runs each. In each run, we have a 5×2 cv t test result (see equation 2.1) and one combined

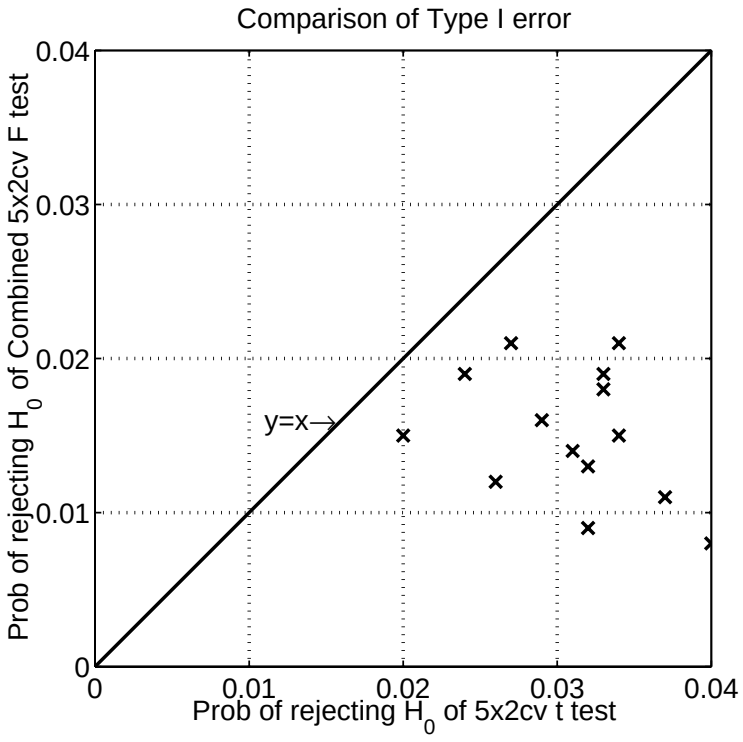

Figure 1: Comparison of type I errors of two tests. All the points are under the $y = x$ line; the combined test leads to lower type I error. All of these type I errors should be at 0.05 if the statistical tests were exactly correct instead of being approximate.

5×2 cv F result (see equation 3.1). As shown in Figure 1, the combined test has a lower probability of rejecting the hypothesis that the classifiers have the same error rate when the hypothesis is true and thus has lower type I error. The reject probabilities are given in Table 3.

To compare the type II error of the two tests, we take two classifiers that are different: an LP and an MLP with hidden units. Again on six data sets using different numbers of hidden units, we have designed 15 experiments of 1000 runs each, where in each run, we have a 5×2 cv t test result and a combined 5×2 cv F result. Reject probabilities with the 5×2 cv t test and the combined 5×2 cv F test are given in Table 3.

As shown in Figure 2, the combined test has a lower probability of rejecting the hypothesis when the two classifiers have similar error rates (lower type II error) and a larger probability of rejecting when they are different

Table 3: Probabilities of Rejecting the Null Hypothesis.

		MLP vs. MLP (Type I error)		LP vs. MLP (Type II error)	
	Hidden Units	5×2 cv	Combined 5×2 cv	5×2 cv	Combined 5×2 cv
IRIS	3	0.032	0.009	0.037	0.007
	10	0.040	0.008	0.029	0.007
	20	0.029	0.016	0.023	0.013
WINE	3	0.037	0.011	0.033	0.018
	10	0.032	0.013	0.031	0.024
	20	0.047	0.016	0.033	0.016
GLASS	5	0.034	0.021	0.025	0.021
	10	0.026	0.012	0.063	0.039
	20	0.047	0.015	0.070	0.075
VOWEL	5	0.033	0.018	0.050	0.027
	10	0.027	0.021	0.722	0.970
	20	0.034	0.015	0.962	1.000
ODR	10	0.033	0.019	0.025	0.019
	20	0.024	0.019	0.364	0.557
THYROID	10	0.031	0.014	0.041	0.031

Note: When comparing two MLPs with equal number of hidden units, any reject is a type I error, and when comparing an LP with an MLP, if their accuracies are different, any reject is lower type II error and implies higher power.

(higher power). The normalized difference in error rate between two classifiers is computed as

$$z = \frac{\overline{e_{lp}} - \overline{e_{mlp}}}{s_{mlp}}$$

where $\overline{e_{mlp}}$, s_{mlp} are the average and standard deviation of error rate of the MLP over the test folds. Note that z is an approximate measure for what we are trying to test: whether the two classifiers have different error rates.

A small difference in error rate implies that the different algorithms construct two similar classifiers with similar error rates; thus the hypothesis should not be rejected. For a large difference, the classifiers have different error rates, and the hypothesis should be rejected.

Dietterich (personal communication) has tested the 5×2 cv F test on three tasks from Dietterich (1998): worst-case, EXP6, and letter recognition. He has also found that the 5×2 cv F test has lower type I error and better power than the 5×2 cv t test.

5 Conclusions

This article has introduced the 5×2 cv F test, which averages over the variability due to replication and fold order that cause problems for the 5×2 cv t test. The simulations have shown that the combined 5×2 cv F

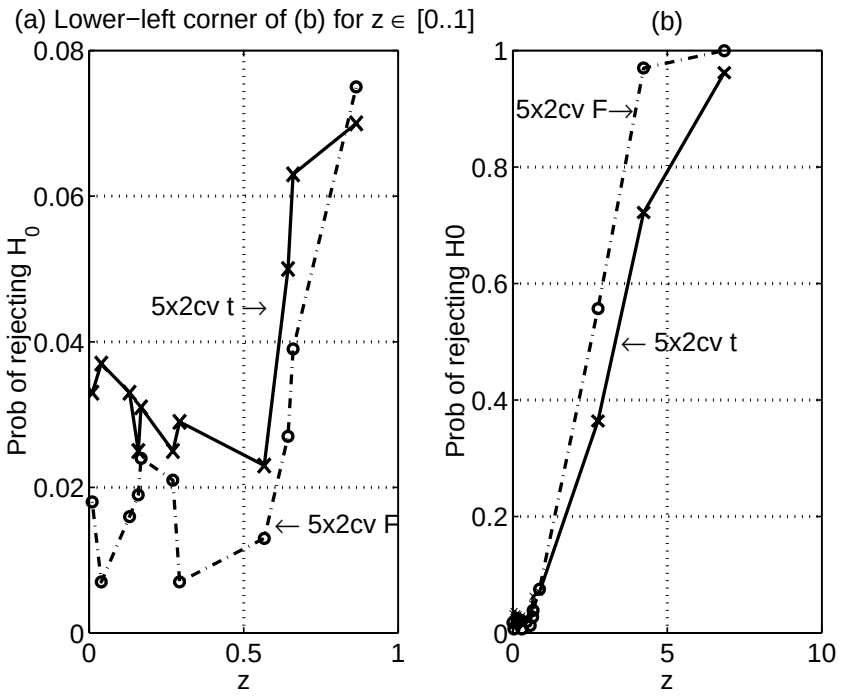

Figure 2: Comparison of type II errors of two tests. (a) zooms the lower left corner of (b) for small z, the normalized distance between the error rates of the two classifiers. The combined test has a lower probability of rejecting the hypothesis when the two classifiers have similar error rates and larger when they are different.

test has a lower risk of type I error and higher power than the 5×2 cv t test. Furthermore, the 5×2 cv F test can be computed from the same information as the 5×2 cv t test, so it adds no computational cost.

Acknowledgments

I thank Tom Dietterich for sharing the results of his comparisons of the 5×2 cv t and F tests, his careful reading of the manuscript of this article, and his comments, which greatly improved the presentation. I also thank Eddy Mayoraz, Frédéric Gobry, and Miguel Moreira for stimulating discussions on statistical tests.

References

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation, 10*, 1895–1923.

Merz, C. J., Murphy, P. M. (1998). UCI repository of machine learning databases. Available at: http://www.ics.uci.edu/∼mlearn/MLRepository.html.

Ross, S. M. (1987). *Introduction to probability and statistics for engineers and scientists*. New York: John Wiley.

Received June 17, 1998; accepted January 4, 1999.