Statistical Tests for Comparing Classifiers

Presented at Robuddies on March 18, 2010
by Sancho McCann

Outline

- Some background on comparing two classifiers (Dietterich, 1998)
- Comparing multiple classifiers on multiple data sets (Demsar, 2006)

The question

Given two learning algorithms A and B, and a small data set S, is there a difference in their classification performance when trained on data sets of the same size as S?

The null hypothesis: there is no difference in the performance of the two algorithms.

Possible errors

- Type I error: false positive, reject the null hypothesis when the null hypothesis is true
- Type II error: false negative, fail to reject the null hypothesis when it is false

A good test

- A good test procedure will not be fooled by differences that are observed by chance (low type I error)
- A good test procedure will detect true differences if they exist (high power, low type II error)

Sources of variation

1. Random variation of the test set
2. Random variation due to selection of training data
3. Internal randomness of the learning algorithm
4. Random classification error

Sources of variation

1. Random variation of the test set
2. Random variation due to selection of training data
3. Internal randomness of the learning algorithm
4. Random classification error

Dealing with 1, 4: the procedure must account for the size of the test set and consequences of changes

Dealing with 2, 3: the algorithm must be executed multiple times and measure the variation in the performance

5 tests

- McNemar's test
- Test for the difference of two proportions
- Resampled paired t-test
- k-fold cross-validated paired t-test
- 5x2 c.v. paired t-test

McNemar's test

Split the available data

Run algorithms A and B
on the training data

To get classifiers f_A and f_B

f_A f_B

Test f_A and f_B on the test data,
record results in a table:

n_{00} = # misclassified by both	n_{01} = # misclassified by A, not B
n_{10} = # misclassified by B, not A	n_{11} = # misclassified by neither

McNemar's test

n_{00} = # misclassified by both	n_{01} = # misclassified by A, not B
n_{10} = # misclassified by B, not A	n_{11} = # misclassified by neither

Under the null hypothesis, the error rates are the same. The expected counts are:

n_{00}	$(n_{01} + n_{10})/2$
$(n_{01} + n_{10})/2$	n_{11}

This statistic is distributed as chi-squared with 1 degree of freedom:

$$\chi^2 = \frac{(|n_{01} - n_{10}| - 1)^2}{n_{01} + n_{10}}$$

Resampled t-test

Randomly split the data into train and test for each trial i (30 trials)

Run algorithms A and B on the training data

To get classifiers f_A and f_B

$f_A^{(i)}$ $f_B^{(i)}$

p_A^i and p_B^i are the misclassification rates during trial i, then

assume $p^{(i)} = p_A^{(i)} - p_B^{(i)}$ are drawn independently from a normal distribution

Then run a Student's t-test by computing:

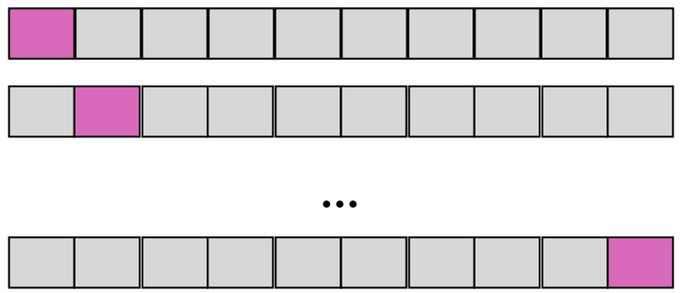
$$t = \frac{\overline{p} \cdot \sqrt{n}}{\sqrt{\frac{\sum_{i=1}^{n}(p^{(i)} - \overline{p})^2}{n-1}}}$$

k-fold c.v. paired t-test

Like the resampled t-test, but differs in how the splits are prescribed

Divide the data into k disjoint sets of equal size

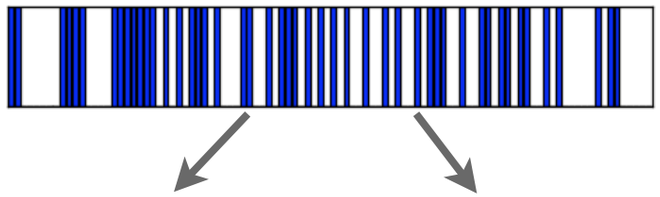
Conduct k trials, using a different set as the test set and the remainder as training



Test sets are now independent between trials, but there is still a lot of overlap between training sets.

5x2 c.v. paired t-test

Perform 5 runs of 2-fold cross validation



Split the data into two folds

fold 1	fold 2	
training	testing	Train A and B on 1st fold, test on 2nd to get $p_A^{(1)}, p_B^{(1)}$
testing	training	Train A and B on 2nd fold, test on 1st to get $p_A^{(2)}, p_B^{(2)}$

This gives two estimates of the difference and an estimated variance

$$p^{(1)} = p_A^{(1)} - p_B^{(1)}$$

$$p^{(2)} = p_A^{(2)} - p_B^{(2)}$$

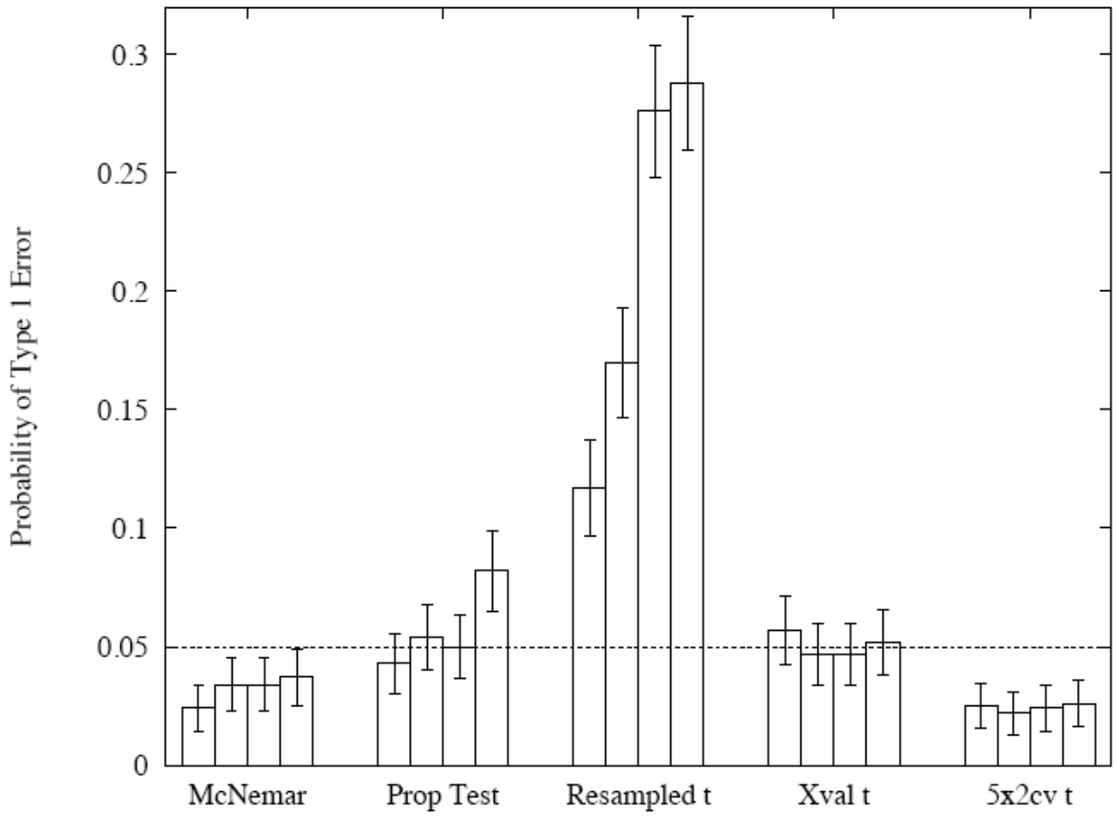
$$s^2 = (p^{(1)} - \overline{p})^2 + (p^{(2)} - \overline{p})^2$$

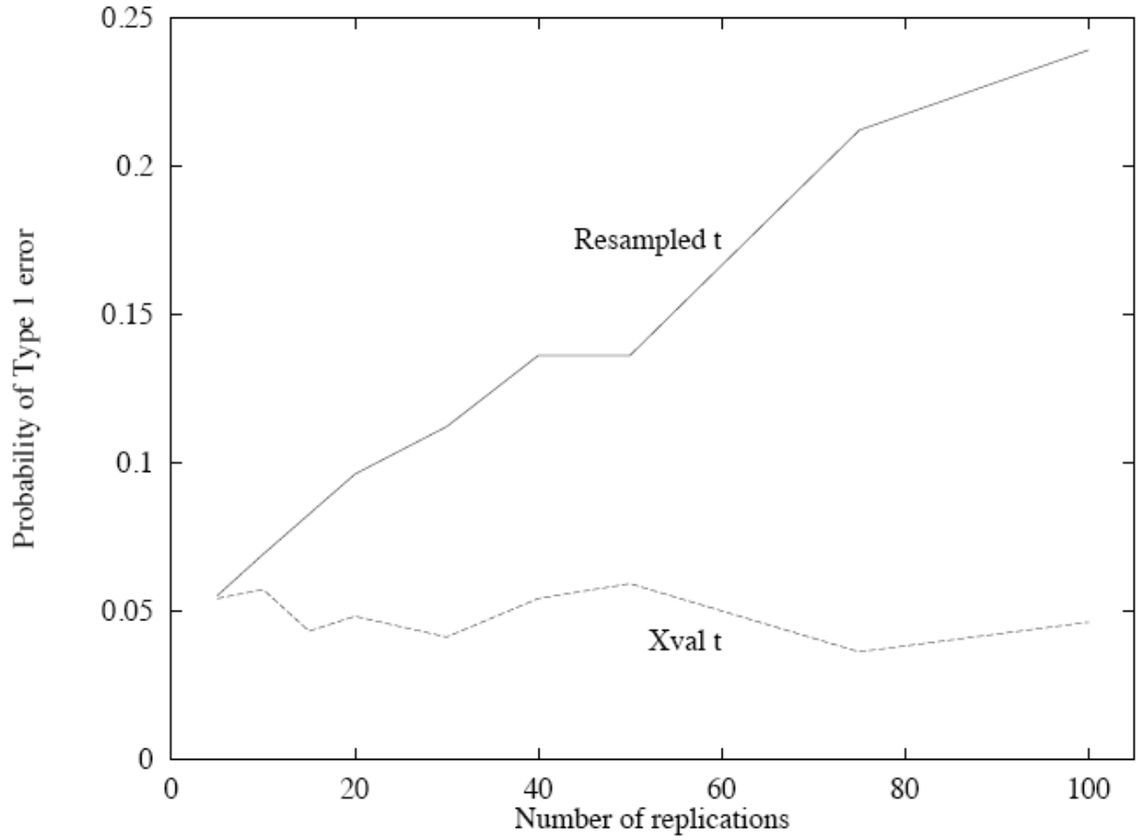
x 5

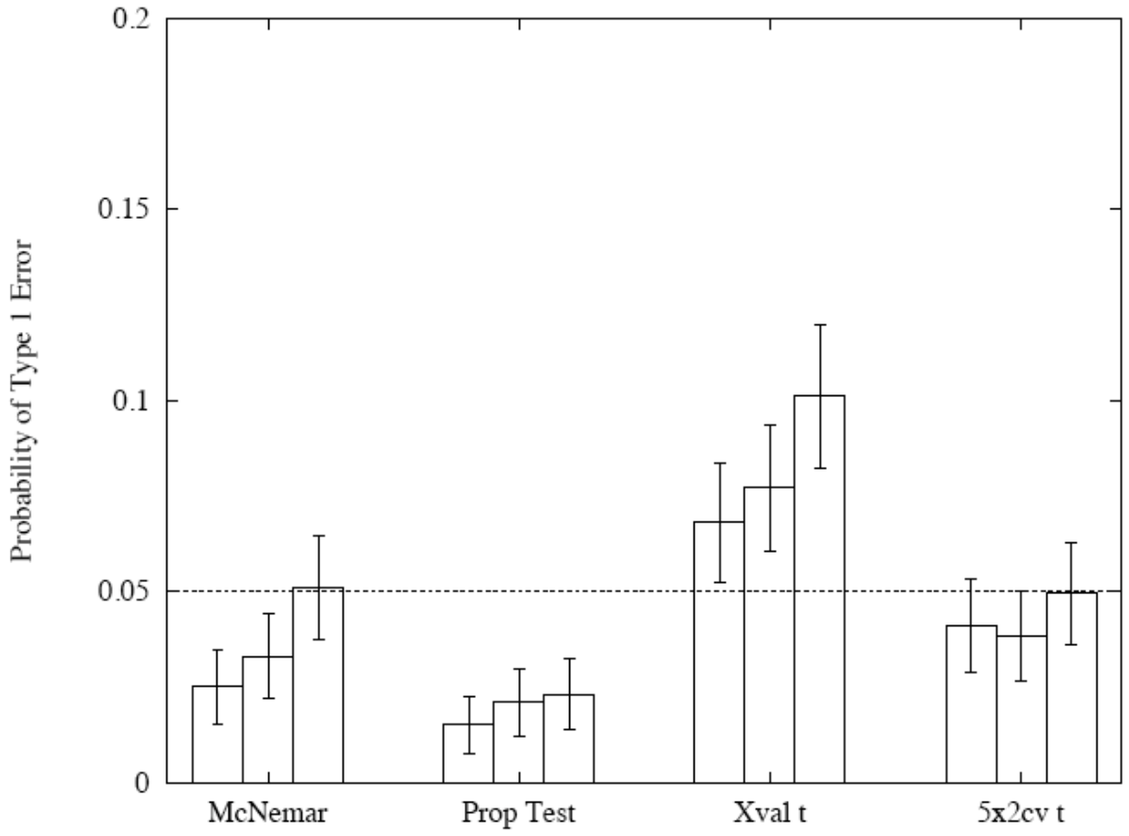
5x2 c.v. paired t-test

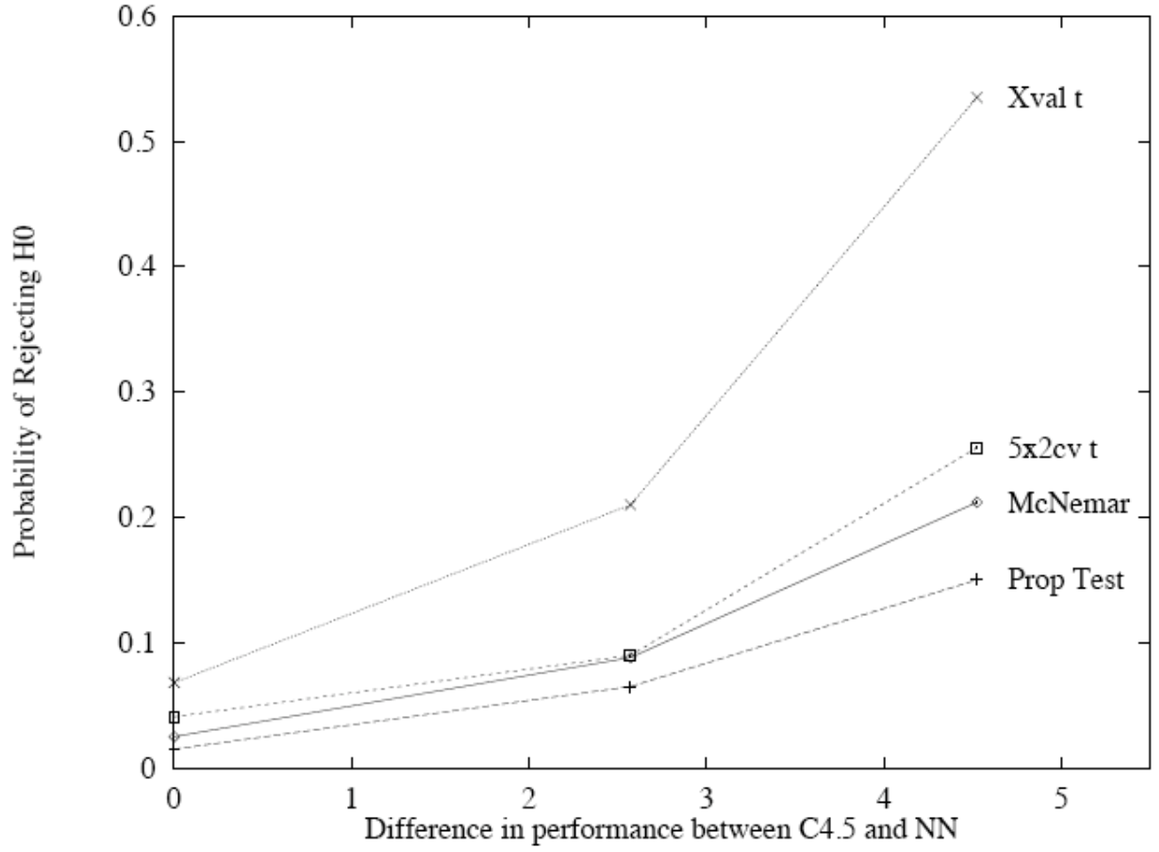
Under the null hypothesis, this test statistic has approximately a t distribution with 5 degrees of freedom

$$\tilde{t} = \frac{p_1^{(1)}}{\sqrt{\frac{1}{5}\sum_{i=1}^{5} s_i^2}}$$


Probability of Type 1 Error
0.3
0.25
0.2
0.15
0.1
0.05
0
McNemar
Prop Test
Resampled t
Xval t
5x2cv t


Probability of Type 1 error
0.25
0.2
0.15
0.1
0.05
0
0
20
40
60
80
100
Number of replications
Resampled t
Xval t


Probability of Type 1 Error
0.2
0.15
0.1
0.05
0
McNemar
Prop Test
Xval t
5x2cv t


Probability of Rejecting H0
0.6
0.5
0.4
0.3
0.2
0.1
0
0
1
2
3
4
5
Difference in performance between C4.5 and NN
Xval t
5x2cv t
McNemar
Prop Test

Recommendations

- The uncorrected resampled t-test, and the t-test over cross validation folds have elevated type I error rates
- If you can afford to run an algorithm 10 times, use the 5x2 c.v. test
- If you can only run an algorithm once, use McNemar's test

Since 1998

- 5x2 c.v. test and McNemar's test replacing t-tests
- 5x2 c.v. test criticized as not replicable enough
- Corrected versions of the resampled t-test that account for overlap have been proposed (Nadeau and Bengio)

Comparing across multiple data sets

The context

Tested k algorithms on N data sets, letting c_i^j be the score of the jth algorithm on the ith dataset

	C4.5	C4.5+m
adult (sample)	0.763	0.768
breast cancer	0.599	0.591
breast cancer wisconsin	0.954	0.971
cmc	0.628	0.661
ionosphere	0.882	0.888
iris	0.936	0.931
liver disorders	0.661	0.668
lung cancer	0.583	0.583
lymphography	0.775	0.838
mushroom	1.000	1.000
primary tumor	0.940	0.962
rheum	0.619	0.666
voting	0.972	0.981
wine	0.957	0.978

2-classifiers, multiple data sets

- Averaging over data sets
- Paired t-test
- Wilcoxon signed ranks test
- Counts sign test

Wilcoxon signed ranks test

	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

Rank differences by their absolute value

Let R+ be the sum of the ranks where algorithm 2 was best

Let R- be the sum of the ranks where algorithm 1 was best

Let T be min(R+, R-)

There's a table for less than 25 datasets, or use:

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

Counts-of-wins sign test

Count the number of data sets on which an algorithm is an overall winner.

Under the null hypothesis, each algorithm should win on N/2 of the N data sets.

The number of wins is distributed according to a binomial distribution and critical values can be looked up in a table.

Counts-of-wins sign test

Critical values

#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

This is a weaker (lower power) test
than the Wilcoxon signed ranks test.

For larger numbers of data sets, the number of wins is distributed with a normal distribution: $\mathcal{N}(N/2, \sqrt{N}/2)$

Multiple classifiers, multiple data sets

- Cautions
- Context
- ANOVA
- Friedman Test

Cautions

- The two-classifier tests are not suited to comparing multiple classifiers
- Doing all pair-wise comparisons and simply listing significant differences elevates the rate of type I error
- Need to control family-wise error rate across all hypothesis tests

Context

	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

Null hypothesis: all classifiers perform the same and the observed differences are merely random

Friedman Test

- Test the null hypothesis
- If the null-hypothesis is rejected, proceed with post-hoc tests to check for differences between individual classifiers

Friedman Test

	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

Compute the average rank for each classifier

Friedman Test

Compute the test statistic and test the null hypothesis

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

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

This is distributed with an F-distribution with k-1 and (k-1)(N-1) degrees of freedom

Post-hoc tests

- Either test each classifier against each other classifier ($\binom{k}{2}$ hypotheses), or
- Test each classifier against a baseline or control classifier (k-1 hypotheses)

Post-hoc test: each vs each

- The Nemenyi test compares all classifiers to each other to test for significant differences
- Two classifiers have significantly different performance if their average ranks differ by at least the critical difference

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

#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

Post-hoc tests: vs control

- When comparing k-1 of the classifiers against a control, or baseline classifier, other methods are more powerful
- The Bonferroni-Dunn test adjusts the target α by dividing by the number of comparisons made: (k-1)
- Easiest way to do this is compute the Nemenyi CD, but use different q_α

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

#classifiers	2	3	4	5	6	7	8	9	10
$q_{0.05}$	1.960	2.241	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

Post-hoc tests: vs control

Multi-step methods compute a p-value for each hypothesis and multiple adjustments of critical values.

The test statistic for comparing the ith and jth classifier is:

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

This has a standard normal distribution, so a p-value can be determined.

Post-hoc tests: vs control

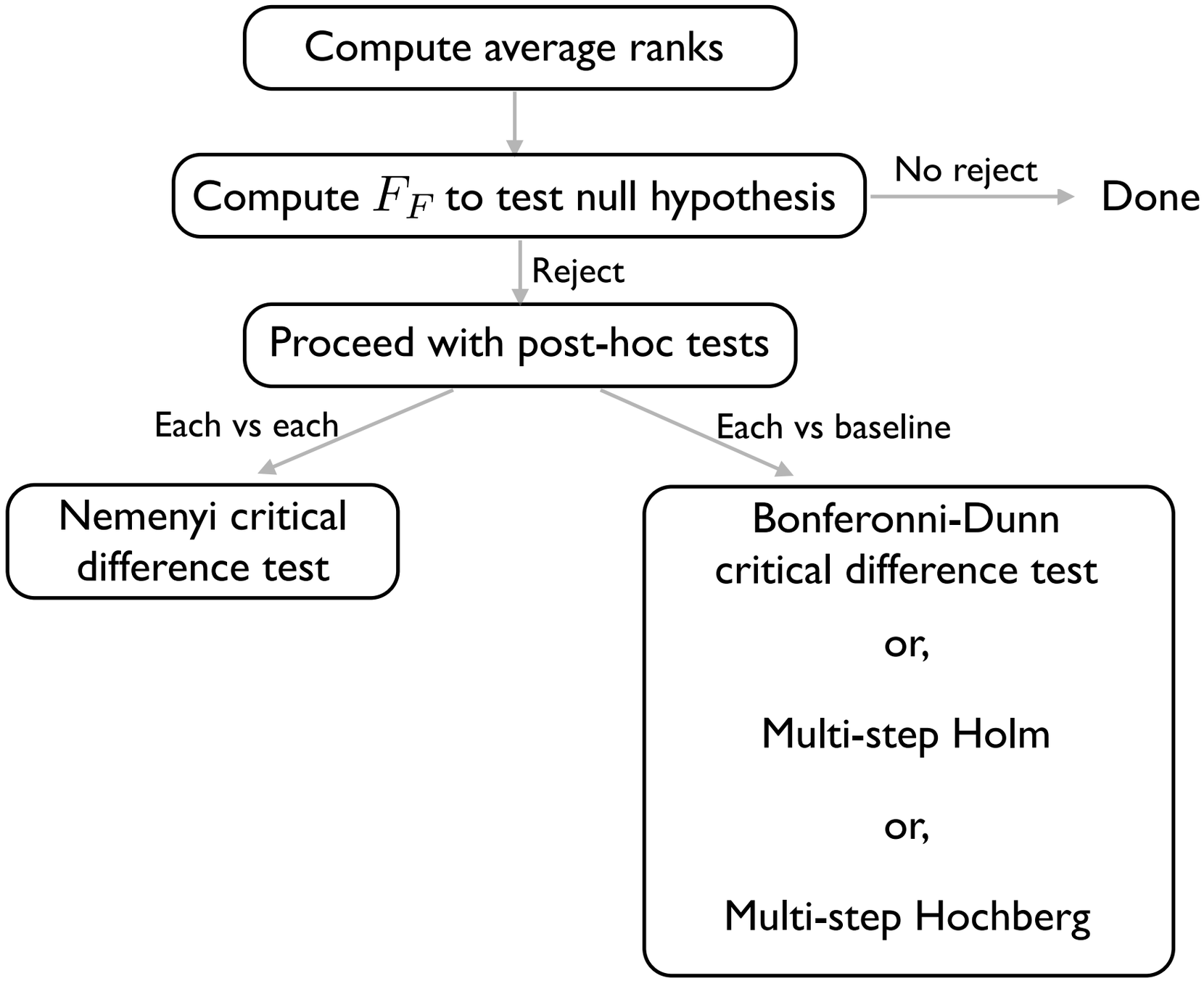
Order hypotheses by their p-values and compare against an adjusted alpha.

Holm's method steps down, rejecting until first failure

Hochberg's method steps up, finding first rejection, then rejecting all null hypotheses with smaller p-values

i	hypothesis	p	$\alpha/(k-i)$	
1	D == A	0.016	0.017	reject
2	B == A	0.019	0.025	reject
3	C == A	0.607	0.050	no reject

Friedman Summary



Example

Example

	C4.5	C4.5+m	C4.5+cf	C4.5+m+cf
average rank	3.143	2.000	2.893	1.964

$$\chi_F^2 = \frac{12N}{k(k+1)}\left[\sum_{j=1}^{k} R_j^2 - \frac{k(k+1)^2}{4}\right] = 9.28 \qquad F_F = \frac{(N-1)\chi_F^2}{N(k-1)-\chi_F^2} = 3.69$$

The critical value of F(3,39) at a = 0.05 is 2.85, so we reject the null hypothesis

#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

The Nemenyi test at a = 0.05 for 4 classifiers has a CD of

$$\mathrm{CD} = q_{0.05}\sqrt{\frac{k(k+1)}{6N}} = 2.569\sqrt{\frac{2 \cdot 5}{6 \cdot 14}} = 1.25$$

This post-hoc test isn't powerful enough to detect any significant differences

The Nemenyi test at a = 0.10 for 4 classifiers has a CD of

$$\mathrm{CD} = q_{0.10}\sqrt{\frac{k(k+1)}{6N}} = 2.291\sqrt{\frac{2 \cdot 5}{6 \cdot 14}} = 1.12$$

C4.5 is significantly worse than C4.5+m and C4.5+m+cf.
The data is not sufficient to reach any conclusion regarding C4.5+cf

	C4.5	C4.5+m	C4.5+cf	C4.5+m+cf
average rank	3.143	2.000	2.893	1.964

Test modifications against control

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

Both Holm (step down) and Hochberg (step up) methods fail to reject the last null hypothesis, but reject the others

Experimental results from paper

Recommendations

- Non-parametric tests preferred
- For 2 classifiers across multiple data sets, prefer the Wilcoxon signed ranks test
- For multiple classifiers across multiple data sets, prefer the Friedman test and associated post-hoc tests

Summary

Case	**Recommendation**
1 data set, 2 classifiers	McNemnar, 5x2 c.v., or a corrected version of the resampled t-test
Multiple data sets, 2 classifiers	Wilcoxon signed rank test
Multiple data sets, multiple classifiers	Friedman test and associated post-hoc tests

References and additional reading

Bengio, Y. and Grandvalet, Y. No Unbiased Estimator of the Variance of K-Fold Cross-Validation. *J. Mach. Learn. Res.* 5 (Dec. 2004), 1089-1105.

Demšar, J. Statistical Comparisons of Classifiers over Multiple Data Sets. *J. Mach. Learn. Res.* 7 (Dec. 2006), 1-30.

Dietterich, T. G. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Comput.* 10, 7 (Oct. 1998), 1895-1923.

Nadeau, C. and Bengio, Y. Inference for the Generalization Error. *Mach. Learn.* 52, 3 (Sep. 2003), 239-281.

Siegfried, T. Odds Are, It's Wrong: Science fails to face the shortcomings of statistics. *ScienceNews.* 177, 7 (Mar. 27, 2010), 26.