Journal Club
“Statistical Comparisons of Classifiers Over Multiple Data Sets” (Demšar, 2006)

Lukas Burk[1,2]

[1]Leibniz Institute for Prevention Research and Epidemiology – BIPS
[2]LMU Munich

2024-05-13

BioWimium

Demšar (2006)

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

Context

- The year is 2006

Context

- The year is 2006
- Machine learning is happening

Context

- The year is 2006
- Machine learning is happening
- New algorithms published every 4 seconds

Context

- The year is 2006
- Machine learning is happening
- New algorithms published every 4 seconds
- Authors compare their proposed method against SOTA
  …using whichever means necessary, appropriate, valid.

Context

- The year is 2006
- Machine learning is happening
- New algorithms published every 4 seconds
- Authors compare their proposed method against SOTA
  …using whichever means necessary, appropriate, valid.

Context

- The year is 2006
- Machine learning is happening
- New algorithms published every 4 seconds
- Authors compare their proposed method against SOTA
  ...using whichever means necessary, appropriate, valid.

Comparing things is hard(citation needed)

Motivation and Setting

- Goal: Compare k classification algorithms on N datasets

Motivation and Setting

- Goal: Compare k classification algorithms on N datasets
- Common hypothesis:
  Does new algorithm perform better than established methods?

Motivation and Setting

- Goal: Compare k classification algorithms on N datasets
- Common hypothesis:
  Does new algorithm perform better than established methods?
- Comparing 2 classifiers on 1 dataset insufficient

Motivation and Setting

- Goal: Compare k classification algorithms on N datasets
- Common hypothesis:
  Does new algorithm perform better than established methods?
- Comparing 2 classifiers on 1 dataset insufficient
- Comparing multiple classifiers on multiple datasets: More difficult

Setting

- Evaluation produces score c_i^j for j-th algorithm on the i-th dataset

Setting

- Evaluation produces score c_i^j for j-th algorithm on the i-th dataset
- Scores: Accuracy, AUC or similar measure

Setting

- Evaluation produces score c_i^j for j-th algorithm on the i-th dataset
- Scores: Accuracy, AUC or similar measure
  - No recorded variance $\Rightarrow$ no assumptions about resampling scheme

Setting

- Evaluation produces score c_i^j for j-th algorithm on the i-th dataset
- Scores: Accuracy, AUC or similar measure
  - No recorded variance $\Rightarrow$ no assumptions about resampling scheme
  - Resampling for stability of scores only

Setting

- Evaluation produces score c_i^j for j-th algorithm on the i-th dataset
- Scores: Accuracy, AUC or similar measure
  - No recorded variance $\Rightarrow$ no assumptions about resampling scheme
  - Resampling for stability of scores only
- Algorithms evaluated on same datasets

Setting

- Evaluation produces score c_i^j for j-th algorithm on the i-th dataset
- Scores: Accuracy, AUC or similar measure
  - No recorded variance $\Rightarrow$ no assumptions about resampling scheme
  - Resampling for stability of scores only
- Algorithms evaluated on same datasets
  - Sample size here “Number of datasets in benchmark”

Setting

- Evaluation produces score c_i^j for j-th algorithm on the i-th dataset
- Scores: Accuracy, AUC or similar measure
  - No recorded variance $\Rightarrow$ no assumptions about resampling scheme
  - Resampling for stability of scores only
- Algorithms evaluated on same datasets
  - Sample size here "Number of datasets in benchmark"
  - $\Rightarrow$ Datasets are independent, scores are not

Comparing 2 classifiers

- Paired t-test
  - Highest power when assumptions met
  - Assumes commensurability of scores (questionable)
  - Normality, outliers

Comparing 2 classifiers

- Paired t-test
  - Highest power when assumptions met
  - Assumes commensurability of scores (questionable)
  - Normality, outliers
- Wilcoxon signed rank test
  - Only assumes commensurability of ranks

Comparing 2 classifiers

- Paired t-test
  - Highest power when assumptions met
  - Assumes commensurability of scores (questionable)
  - Normality, outliers
- Wilcoxon signed rank test
  - Only assumes commensurability of ranks
- Sign test: Not even bothering with this one

Comparing multiple classifiers

General scheme:

1. Perform global test to detect if any two algorithms differ at all
2. If (1) is signif., perform post-hoc test to detect which algorithms differ in particular

Global tests

Repeated measures ANOVA

- Assumes normality of scores
- Assumes sphericity ($\approx$ homoskedasticity)

Friedman test

- Non-parametric analogue to rmANOVA
- Uses ranks from best (1) to worst (k), averages for ties
- Test statistic $Fr \sim F(k-1,(k-1)(N-1))$

Post-hoc tests

Choices of all pairwise or one-to-many tests
in either parametric or nonparametric flavors:

Type	All Pairwise	One-to-many
Parametric	Tukey	Dunnet
Nonparametric	Nemenyi	Bonferroni-Dunn

Nemenyi

Critical differences between two algorithms calculated as

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

- Critical values q_α based on studentized range statistic
- If difference in average ranks exceeds CD, they are signif. different

Bonferroni-Dunn

Test statistic (approx. normal) is calculated based on average ranks (R) for algorithms i and j

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

- Much greater power when comparing against baseline
- Can use any other method to control for FWER (Bonferroni, Holm, Hochberg, ...)
- Using Bonferroni-Dunn gives constant CD, easier to visualize

Example 1

Comparing 4 algorithms across 14 datasets

	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.

Example 1: Result (Nemenyi, all pairwise)

- CD = $2.569\sqrt{\frac{4\cdot 5}{6\cdot 14}} = 1.25$ (for $\alpha = 0.05$)
    - Difference between best and worst is already smaller
    - Test not powerful enough
- CD = 1.12 (for $\alpha = 0.1$):
    - Conclude that C4.5 is worse than C4.5+m and C4.5+m+cf
        - Can't make statement about C4.5+cf

Example 1: Result (BD, one-to-many)

- Hypothesis: Does tuning `m` and/or `cf` help compared to baseline C4.5?
- CD = 1.16

C4.5 vs. C4.5+m+cf $\rightarrow 3.143 - 1.964 = 1.179 > 1.16$

C4.5 vs. C4.5+cf $\rightarrow 3.143 - 2.893 = 0.250 < 1.16$

C4.5 vs. C4.5+m $\rightarrow 3.143 - 2.000 = 1.143 \approx 1.16$

Example 1: Result (BD, one-to-many)

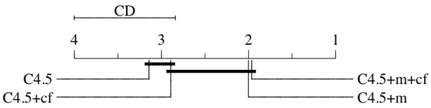
- Hypothesis: Does tuning `m` and/or `cf` help compared to baseline C4.5?
- CD = 1.16

$$
\begin{aligned}
&\text{C4.5 vs. C4.5+m+cf} && \rightarrow 3.143 - 1.964 = 1.179 > 1.16 \\
&\text{C4.5 vs. C4.5+cf} && \rightarrow 3.143 - 2.893 = 0.250 < 1.16 \\
&\text{C4.5 vs. C4.5+m} && \rightarrow 3.143 - 2.000 = 1.143 \approx 1.16
\end{aligned}
$$

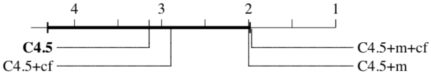
- Conclude that tuning `m` helps, `cf` probably not

Example 1: Critical Difference plots

All pairwise comparisons (top) vs. baseline comparison (bottom)



(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.

Empirical comparison of tests

- Comparing various algorithms on 10 randomly drawn out of pool of 40 real world datasets

Empirical comparison of tests

- Comparing various algorithms on 10 randomly drawn out of pool of 40 real world datasets
- No formal assessment of Type I / II error as correct test decision unclear

Empirical comparison of tests

- Comparing various algorithms on 10 randomly drawn out of pool of 40 real world datasets
- No formal assessment of Type I / II error as correct test decision unclear
- Measured performance of all algorithms on all datasets before experiment

Empirical comparison of tests

- Comparing various algorithms on 10 randomly drawn out of pool of 40 real world datasets
- No formal assessment of Type I / II error as correct test decision unclear
- Measured performance of all algorithms on all datasets before experiment
    - Introduce bias term $k \geq 0$ to adjust difference between algorithms, affects selection of datasets

Empirical comparison of tests

- Comparing various algorithms on 10 randomly drawn out of pool of 40 real world datasets
- No formal assessment of Type I / II error as correct test decision unclear
- Measured performance of all algorithms on all datasets before experiment
  - Introduce bias term $k \geq 0$ to adjust difference between algorithms, affects selection of datasets
  - $k = 0$ corresponds to random choice of datasets

Empirical comparison of tests

- Comparing various algorithms on 10 randomly drawn out of pool of 40 real world datasets
- No formal assessment of Type I / II error as correct test decision unclear
- Measured performance of all algorithms on all datasets before experiment
  - Introduce bias term $k \geq 0$ to adjust difference between algorithms, affects selection of datasets
  - $k = 0$ corresponds to random choice of datasets
  - Allows testing different hypothesis

Empirical comparison of tests

- Comparing various algorithms on 10 randomly drawn out of pool of 40 real world datasets
- No formal assessment of Type I / II error as correct test decision unclear
- Measured performance of all algorithms on all datasets before experiment
    - Introduce bias term $k \geq 0$ to adjust difference between algorithms, affects selection of datasets
    - $k = 0$ corresponds to random choice of datasets
    - Allows testing different hypothesis
- Calculate average p-values based on 1000 replicates

Measures of reliability

1. Variance of p values: $R(p) = 1 - 2 \cdot \mathrm{Var}(p)$

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

where e_i is outcome of i-th experiment out of n (1 if accepted, 0 otherwise)

Measures of reliability

1. Variance of p values: $R(p) = 1 - 2 \cdot \mathrm{Var}(p)$
2. Measure based on Bouckaert (2004):

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

where e_i is outcome of i-th experiment out of n (1 if accepted, 0 otherwise)

Measures of reliability

1. Variance of p values: $R(p) = 1 - 2 \cdot \mathrm{Var}(p)$
2. Measure based on Bouckaert (2004):

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

where e_i is outcome of i-th experiment out of n (1 if accepted, 0 otherwise)

- $R(e) = 0.5$ if # of rejected equals number of accepted

Measures of reliability

1. Variance of p values: $R(p) = 1 - 2 \cdot \mathrm{Var}(p)$
2. Measure based on Bouckaert (2004):

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

where e_i is outcome of i-th experiment out of n (1 if accepted, 0 otherwise)

- $R(e) = 0.5$ if # of rejected equals number of accepted
- $R(e) = 1$ if # of rejected or accepted is 0 respectively

Measures of reliability

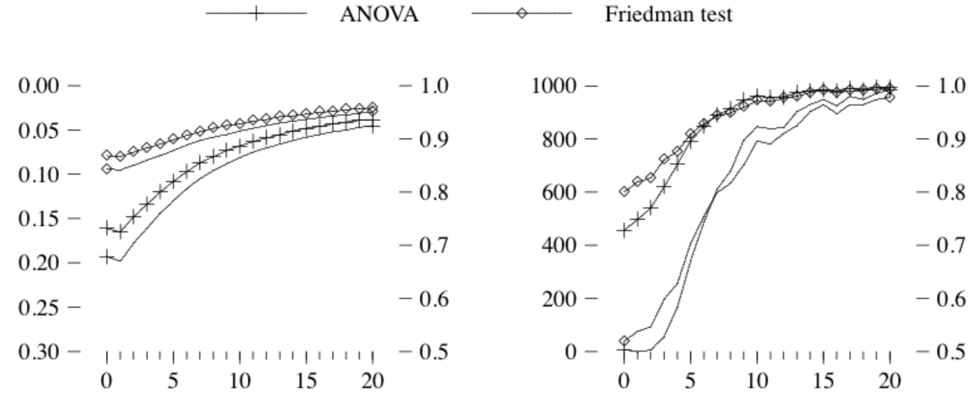
1. Variance of p values: $R(p) = 1 - 2 \cdot \mathrm{Var}(p)$
2. Measure based on Bouckaert (2004):

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

where e_i is outcome of i-th experiment out of n (1 if accepted, 0 otherwise)

- $R(e) = 0.5$ if # of rejected equals number of accepted
- $R(e) = 1$ if # of rejected or accepted is 0 respectively
- Will show low replicability if e.g. p-values fluctuate closely around 0.05

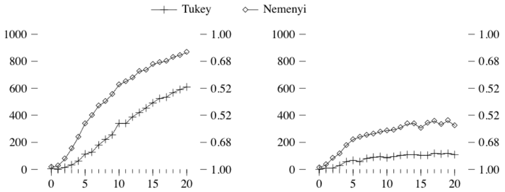
Results



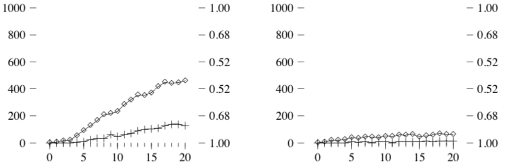
(a) Average p values (left axis) and $R(p)$ (no symbols on lines, right axis)

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

Results



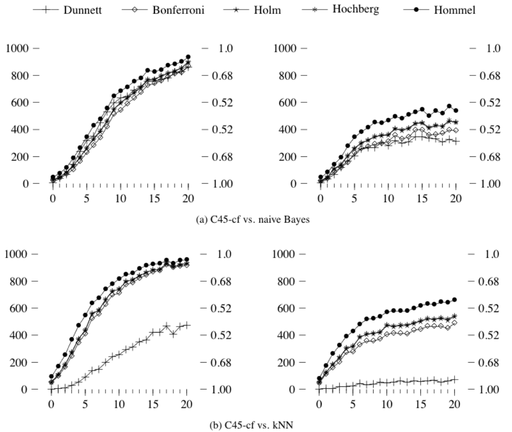
(a) C45-cf vs. naive Bayes



(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 $R(e)$.

Results



(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

Conclusion

- Nonparametric tests more likely to reject H0
- Hints at violated assumptions of parametric tests

Conclusion

- Nonparametric tests more likely to reject H0
- Hints at violated assumptions of parametric tests

Nonparametric tests:

- Appropriate as they assume limited commensurability
- Safer than parametric tests (assumptions)
- Stronger than parametric tests here, especially for pairwise tests

Thank you for your attention!

www.leibniz-bips.de/en

Contact
Lukas Burk
Leibniz Institute for Prevention Research
and Epidemiology – BIPS
Achterstraße 30
D-28359 Bremen
burk@leibniz-bips.de