The Design and Analysis of Benchmark Experiments – Part II: Analysis

Torsten Hothorn  Achim Zeileis  Friedrich Leisch  Kurt Hornik

Friedrich–Alexander–Universität Erlangen–Nürnberg
http://www.imbe.med.uni-erlangen.de/~hothorn/
Benchmark Experiments

A comparison of algorithms with respect to certain performance measures is of special interest in the following problems

- select the best out of a set of candidates,
- identify groups of algorithms with the same performance,
- test whether any useful structure is inherent in the data or
- demonstrate equivalence of two algorithms.
Illustrating Example

Stabilization of a Linear Discriminant Analysis (LDA) by using low-dimensional Principal Component (PC-\(q\)) scores (Läuter, 1992; Läuter et al., 1998; Kropf, 2000) for Glaucoma diagnosis (Hothorn et al., 2003; Mardin et al., 2003).

Laser-scanning images from 98 patients and 98 controls \((n = 196)\), \(p = 62\) numeric input variables.

Data generating process: The empirical distribution function \(\hat{Z}_n\).

Performance measure: Out-of-bootstrap misclassification error.
**Question:** Does the performance distribution $\hat{P}_{\text{LDA}}(\hat{Z}_n)$ of a LDA using the original $p$ input variables differ from the performance distribution $\hat{P}_{\text{sLDA}}(\hat{Z}_n)$ of a stabilized LDA?

**Experiment:** Draw $B$ samples $\mathcal{L}^b$ from the data generating process $\hat{Z}_n$ and compute $\hat{p}_{\text{LDA},b}$ and $\hat{p}_{\text{sLDA},b}$, the misclassification errors evaluated on the out-of-bootstrap observations.
**Inference**

\[ H_0 : \hat{P}_{LDA}(\hat{Z}_n) = \hat{P}_{sLDA}(\hat{Z}_n) \]

**Problem:** We do not know anything about the performances, except that parametric assumptions are surely not appropriate.

**Solution:** Dispose the performance distributions by conditioning on all permutations of the labels for each bootstrap sample.
Inference

\[ T = \sum_{b=1}^{B} \hat{p}_{LDA,b} - \hat{p}_{sLDA,b} = B(\bar{p}_{LDA,\cdot} - \bar{p}_{sLDA,\cdot}) \]

The conditional distribution of the test statistic \( T \) under the conditions described by \( H_0 \) can be used to construct a permutation test.

In our case, the \( P \)-value based on the asymptotic conditional distribution is \( p < 0.001 \) and therefore \( H_0 \) can be rejected.
A Regression Example

Exactly the same methodology can be applied to regression problems with univariate numeric responses. Example: Can additional randomness via Random Forests improve Bagging for the Boston Housing data?

House prices for $n = 506$ houses near Boston, $p = 13$ input variables.

Data generating process: The empirical distribution function $\hat{Z}_n$.

Performance measure: Out-of-bootstrap mean squared error.
Inference

The null-hypothesis of equal performance distributions can be rejected ($P$-value < 0.001).

The estimated difference of the mean square error of Bagging compared to Random Forests is 0.969 with confidence limits (0.633, 1.305).
Comparison of Multiple Algorithms

When multiple algorithms are under test, we are interested in both a global test and a multiple test procedure showing where the differences, if any, come from. Example: Breast Cancer data with tumor classification from \( n = 699 \) observations with \( p = 9 \) inputs.

Comparison of sLDA, Support Vector Machine, Random Forests and Bundling (Hothorn and Lausen, 2003).

Data generating process: The empirical distribution function \( \hat{Z}_n \).

Performance measure: Out-of-bootstrap misclassification error.
Inference

Again, the global hypothesis

\[ H_0 : \hat{P}_1(\hat{Z}_n) = \ldots = \hat{P}_K(\hat{Z}_n) \]

can be rejected (\( P \)-value < 0.001).

**Problem:** Which differences ‘cause’ the rejection of \( H_0 \)?

**Solution:** One can avoid complicated closed testing procedures by computing confidence intervals after mapping the \( B \)-block design into a \( K \)-sample problem via alignment (Hájek et al., 1999).
Alignment

When we look at the performance measure of algorithm $k$ in the $b$th sample drawn from the data generating process, we might want to write

$$p_{kb} = \mu + \beta_b + \gamma_k + \varepsilon_{kb}$$

where $\mu$ corresponds to the performance of the Bayes-rule, $\beta_b$ is the error induced by the $b$ sample and $\gamma_k$ is the error of the $k$th algorithm, the quantity we are primarily interested in, $\varepsilon$ indicates an error term.
Alignment (cont’d)

The aligned performance measures $p_{kb}^\star$ cover the difference of the performance of the $k$th algorithm from the average performance of all $K$ algorithms:

$$p_{kb}^\star = p_{kb} - \bar{p}_b = (\gamma_k + \varepsilon_{kb}) - \frac{1}{K} \sum_{k=1}^{K} (\gamma_k + \varepsilon_{kb})$$

For classification problems, $p_{k_1b}^\star - p_{k_2b}^\star$ is the difference of the misclassification error.
Alignment (cont’d)

The aligned random variables are not independent but exchangeable for each of the $b$ samples and are independent between samples.

Therefore, (asymptotic) permutation test procedures can be used to assess the deviations from the global null-hypothesis.

For example, asymptotic simultaneous confidence intervals for Tukey-contrasts can be used for an all-pair comparison of the $K$ algorithms under test.
Asymptotic Tukey Confidence Sets

- RF vs. Bund
- SVM vs. Bund
- sLDA vs. Bund
- SVM vs. RF
- sLDA vs. RF
- sLDA vs. SVM

95% two-sided confidence intervals
Classical Tests?

We advocate usage of permutation tests, but what about more classical tests?

Consider a paired comparison of sLDA vs. SVM for the Breast Cancer data:

- Permutation test: $T = 1.488$, $p = 0.776$
- $t$ test: $t = 0.284$, $p = 0.777$
- Wilcoxon signed rank test: $W = 18216$, $p < 0.001$
Rank Tests: A Warning

Tests like the Wilcoxon signed rank test are constructed for the null-hypothesis ‘the difference of the performance measures is symmetrically distributed around zero’. For non-symmetric distributions this leads to a complete desaster.

Look at $n = 500$ realizations of a skewed random variable

$$
\frac{X - d}{\sqrt{2d}}
$$

with expectation zero and unit variance with $X \sim \chi^2_d$. 


**Lifetime Analysis Problems**

Appropriate performance measures for censored responses are by no means obvious and still a matter of debate (Henderson, 1995; Graf et al., 1999; Molinaro et al., 2003). We use the Brier score for censored data suggested by Graf et al. (1999).

Example: Predictive performance of the Kaplan-Meier estimator, a single survival tree and Bagging of survival trees (Hothorn et al., 2004) measured for $n = 686$ women enrolled in the German Breast Cancer Study (Group 2).
Kaplan-Meier vs. Single Tree

![Graph showing comparisons between Kaplan-Meier and Single Tree results.](image-url)
Kaplan-Meier vs. Bagging

Benchmark Experiments

2004/03/11
Asymptotic Dunnett Confidence Sets

95% two-sided confidence intervals

-0.025 -0.015 -0.005 0.005

Tree vs. KaplanMeier
Bagging vs. KaplanMeier
Interpretation

Predictions derived from the estimated Kaplan-Meier curve don’t take any information covered by the input variables into account. A test for the hypothesis

\[
\text{there is no (detectable) relationship between the input variables and the response}
\]

can therefore be performed by comparing the performance of the simple Kaplan-Meier curve with the performance of the best tools available for predicting survival times.
Conclusion

When comparing the performance of $K$ algorithms it is appropriate to treat the $B$ samples from the data generating process as blocks.

Standard statistical test procedures can be used to compare arbitrary performance measures for multiple algorithms.

Some classical parametric and non-parametric procedures are only sub-optimal, we advocate procedures based on the conditional distribution of test statistics for inference.
References


