Benchmark Experiments

A Tool for Analyzing Statistical Learning Algorithms

Manuel J. A. Eugster

Institut für Statistik
Ludwig-Maximiliams-Universität München

Rigorosum, 16. März 2011

Benchmark experiments

In general, benchmarking is the **process** of comparing individual **objects** which compete in a specific **field of activity**; and the comparisons are based on number computed by **performance measures**.

Data generating process:

Given is a data generating process DGP. We draw $b = 1, \ldots, B$ independent and identically distributed learning samples:

$$\mathfrak{L}^1 = \{z_1^1, \ldots, z_n^1\} \sim DGP$$
$$\vdots$$
$$\mathfrak{L}^B = \{z_1^B, \ldots, z_n^B\} \sim DGP$$

Candidate algorithms:

There are $K > 1$ candidate algorithms a_k $(k = 1, \ldots, K)$ available; for each algorithm, $a_k(\,\cdot \mid \mathfrak{L}^b) \sim \mathcal{A}_k(DGP)$ is the fitted model based on a learning sample $\mathfrak{L}^b$.

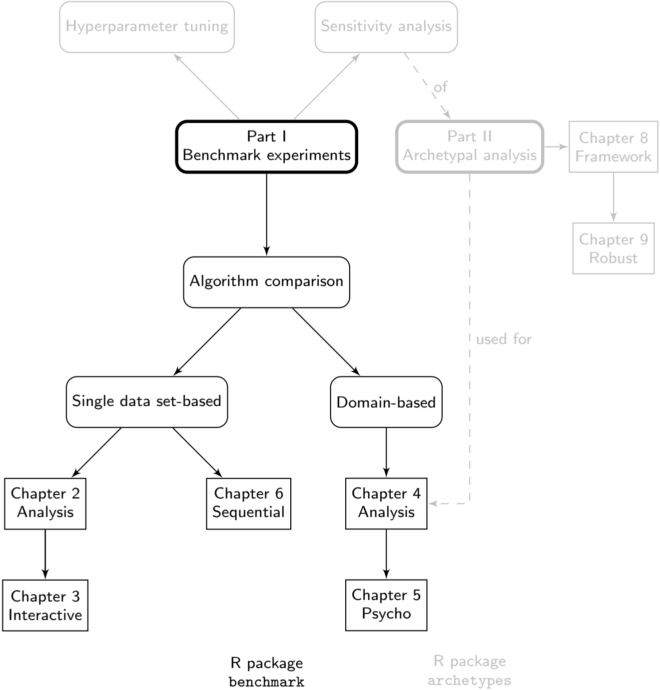
(*) Following Hothorn, Leisch, Zeileis, and Hornik (2005).

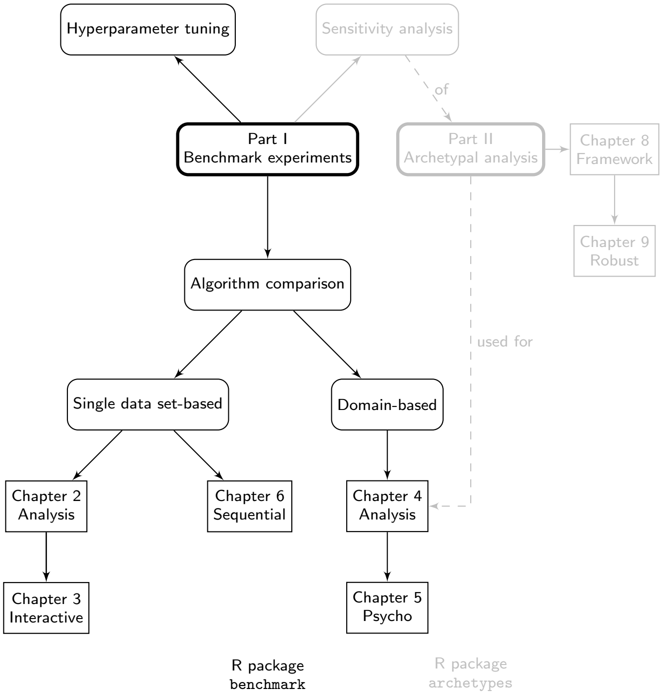
Performance measure:
The performance of the candidate algorithm a_k when provided with the learning samples $\mathfrak{L}^b$ is measured by a scalar function $p(\cdot)$:

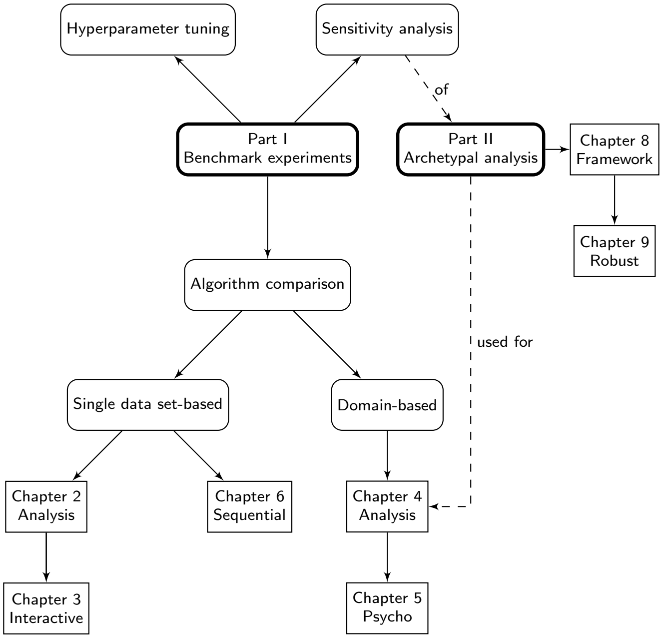
$$p_{bk} = p(a_k, \mathfrak{L}^b) \sim \mathcal{P}_k(DGP)$$

Empirical performance measure:
An estimation of the generalization performance of algorithm a_k learned on learning sample $\mathfrak{L}^b$ is based on a test sample $\mathfrak{T}^b \sim DGP$:

$$\hat{p}_{bk} = \hat{p}(a_k, \mathfrak{L}^b, \mathfrak{T}^b) \sim \hat{\mathcal{P}}_k(DGP)$$

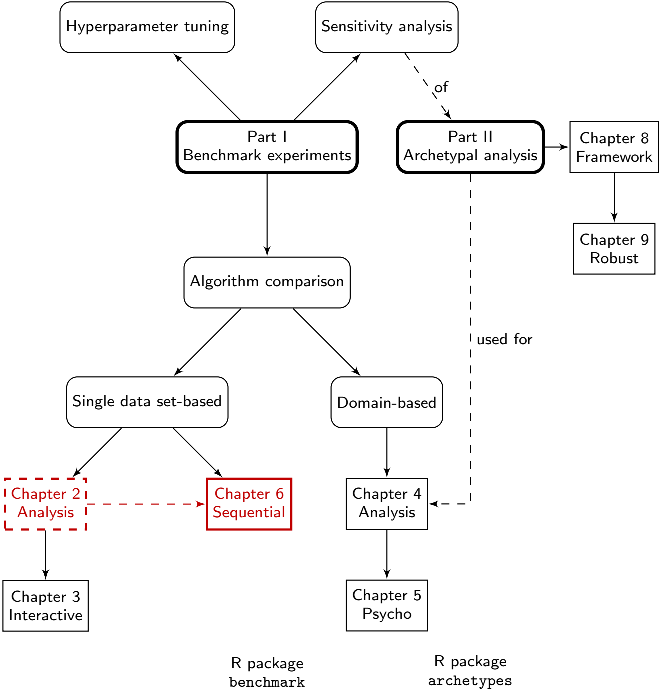

Hyperparameter tuning
Sensitivity analysis
of
Part I
Benchmark experiments
Part II
Archetypal analysis
Chapter 8
Framework
Chapter 9
Robust
Algorithm comparison
used for
Single data set-based
Domain-based
Chapter 2
Analysis
Chapter 6
Sequential
Chapter 4
Analysis
Chapter 3
Interactive
Chapter 5
Psycho
R package
benchmark
R package
archetypes


Hyperparameter tuning
Sensitivity analysis
of
Part I
Benchmark experiments
Part II
Archetypal analysis
Chapter 8
Framework
Chapter 9
Robust
Algorithm comparison
used for
Single data set-based
Domain-based
Chapter 2
Analysis
Chapter 6
Sequential
Chapter 4
Analysis
Chapter 3
Interactive
Chapter 5
Psycho
R package
benchmark
R package
archetypes

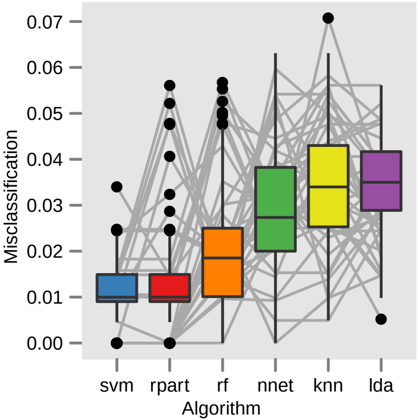


R package
`benchmark`

R package
`archetypes`


Hyperparameter tuning
Sensitivity analysis
of
Part I
Benchmark experiments
Part II
Archetypal analysis
Chapter 8
Framework
Chapter 9
Robust
Algorithm comparison
used for
Single data set-based
Domain-based
Chapter 2
Analysis
Chapter 6
Sequential
Chapter 4
Analysis
Chapter 3
Interactive
Chapter 5
Psycho
R package
benchmark
R package
archetypes

Analysis of benchmark experiments

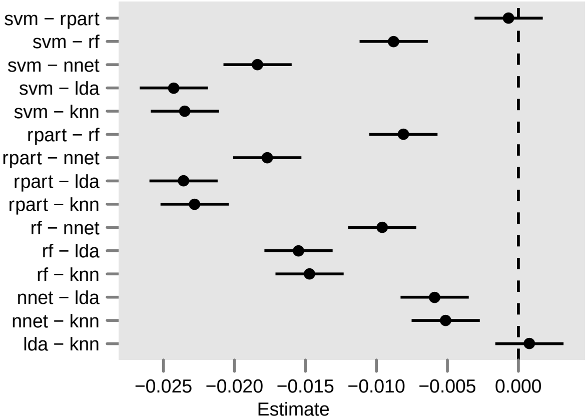


Classification problem `monks3` with $B = 250$ replications, bootstrapping as resampling scheme to generate the learning samples $\mathfrak{L}^b$, and the out-of-bag scheme for $\mathfrak{T}^b$.

Inference:
Given the K different random samples $\{\hat{p}_{1k}, \ldots, \hat{p}_{Bk}\}$ with B iid samples drawn from the distributions $\hat{\mathcal{P}}_k(DGP)$ the null hypothesis of interest for most problems is:

$$H_0 \; : \; \hat{\mathcal{P}}_1 = \cdots = \hat{\mathcal{P}}_K$$

Test procedure:
Use an appropriate non-parametric (e.g., Friedman test based) or parametric (e.g., linear mixed-effects model based) test procedure T to find significant pairwise differences.



Pairwise test decisions based on the 95% simultaneous confidence intervals computed for a linear mixed-effects model of the misclassification error using Tukey contrasts.

Preference relation:

$a_k \prec a_{k'}$ – algorithm a_k is better than $a_{k'}$

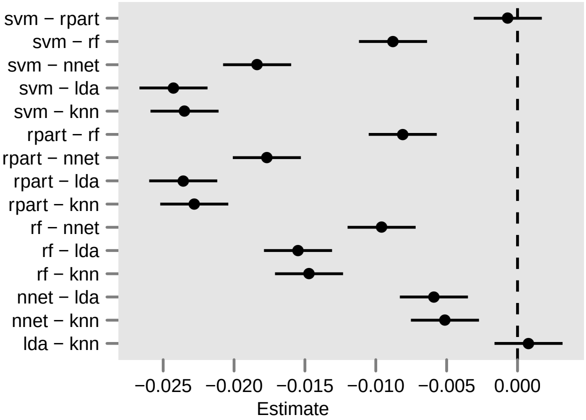
$a_k \sim a_{k'}$ – algorithm a_k is equally to $a_{k'}$

An arbitrary pairwise comparison induces a mathematical relation R which we interpret as preference relation:

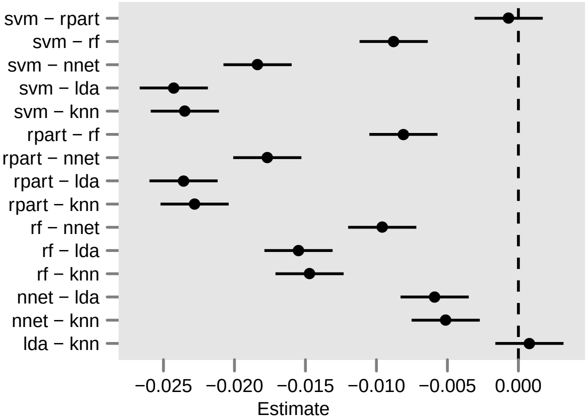
$$(a_k \; R \; a_{k'}) \Rightarrow a_k \sim a_{k'}$$

or

$$(a_k \; R \; a_{k'}) \Rightarrow a_k \prec a_{k'}$$



Relation R is "significantly better ($\alpha = 0.05$)": (`svm` R `rf`), (`svm` R `nnet`), ...



Relation R is "significantly better ($\alpha = 0.05$)": ($\texttt{svm}\ R\ \texttt{rf}$), ($\texttt{svm}\ R\ \texttt{nnet}$), ...

Preference relation (strict part):

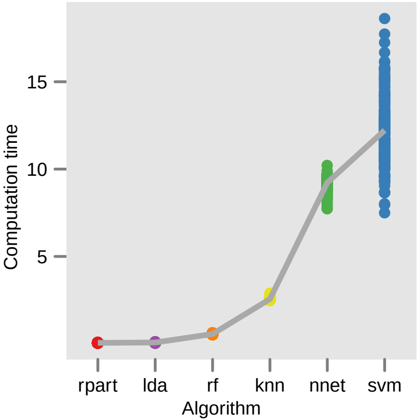
$$\texttt{svm} \sim \texttt{rpart} \prec \texttt{rf} \prec \texttt{nnet} \prec \texttt{knn} \sim \texttt{lda}$$

Preference combination:
Aggregate an ensemble of preference relations, each based on a performance measure of interest, using consensus decision-making methods:

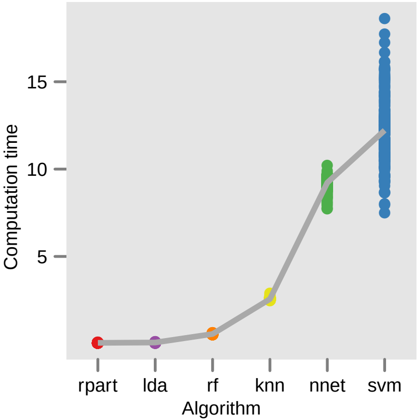
$$\{R_1, \dots, R_{J'}\} \Rightarrow_w \bar{R}$$

Aggregation methods are, for example, Borda count, Condorcet approaches, optimization methods.

(*) For details see Eugster, Hothorn, and Leisch (2010a) and, e.g., Hornik and Meyer (2007).



Relation R is "$<$ (mean$(\cdot)$, $\epsilon = 0.1$)": (`rpart` R `rf`), (`rpart` R `knn`), ...



Computation time ($w = 0.2$):	$\texttt{rpart} \sim \texttt{lda} \prec \texttt{rf} \prec \texttt{knn} \prec \texttt{nnet} \prec \texttt{svm}$
Misclassification ($w = 1$):	$\texttt{svm} \sim \texttt{rpart} \prec \texttt{rf} \prec \texttt{nnet} \prec \texttt{knn} \sim \texttt{lda}$
Consensus (linear order):	$\texttt{rpart} \prec \texttt{svm} \prec \texttt{rf} \prec \texttt{nnet} \prec \texttt{lda} \prec \texttt{knn}$

Sound benchmark experiment framework to compute a statistically correct order of the candidate algorithms, but ...

Sound benchmark experiment framework to compute a statistically correct order of the candidate algorithms, but ...

... in most benchmark experiments, it is based on a **freely chosen** number of replications B:

$$
\begin{aligned}
\mathfrak{L}^1 &= \{z_1^1, \ldots, z_n^1\} \sim DGP \\
&\vdots \\
\mathfrak{L}^B &= \{z_1^B, \ldots, z_n^B\} \sim DGP
\end{aligned}
$$

Sequential/adaptive benchmarking

For $b = 1, \ldots, B$:

1. Draw learning sample $\mathfrak{L}^b$.

2. Measure performance p_{bk} of the $k = 1, \ldots, K$ candidate algorithms.

Execute test procedure T on the K performance estimations $\{p_{1k}, \ldots, p_{Bk}\}$ and make a decision for a given α.

- Benchmark experiments are considered as **fixed-sample** experiments.
- The nature of benchmark experiments is **sequential**.

Do

1. Draw learning sample $\mathfrak{L}^b$.

2. Measure performance p_{bk} of the $k = 1, \ldots, K$ candidate algorithms.

3. Execute test procedure T on the K performance estimations $\{p_{1k}, \ldots, p_{bk}\}$.

While no decision for a given α (and $b \leq B$).

- Sequential/adaptive benchmarking enables
  (1) to monitor the benchmark experiment, and
  (2) to make a decision – to *stop* or to *go on*.

Exemplar benchmark experiments

(1) $\mathfrak{L}$ is the Pima Indians Diabetes data set; **(2)** $\mathfrak{L}^b$ by bootstrapping; **(3)** linear discriminant analysis (`lda`), support vector machine with $C = 1.00$ (`svm1`), support vector machine with $C = 1.01$ (`svm2`), random forest (`rf`); **(4)** misclassification on the out-of-bag samples; **(5)** $B = 100$.

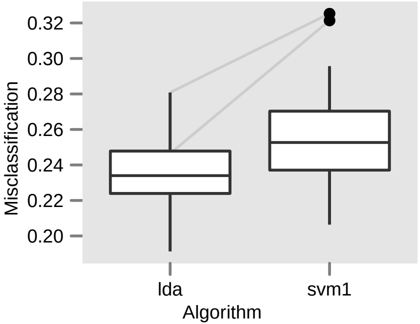
$\Rightarrow$ compare two algorithms at a time, i.e., test if algorithm a_1 is better than algorithm a_2.

(6) Wilcoxon Signed Rank test, $\alpha = 0.05$.

Monitoring

Goal: Observe and interpret the test result, mainly the *p*-value, on the accumulating performance measures.

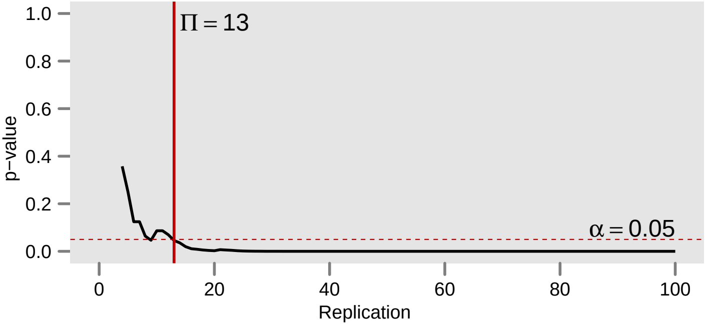
Scenario 1 – Different algorithm performances:



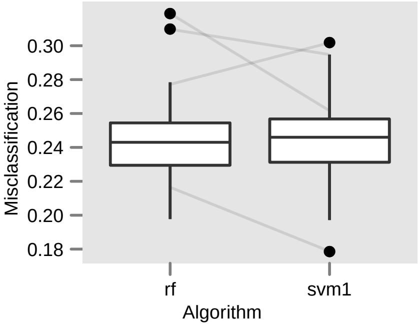
p-value $= 1.312e - 12$

$\Rightarrow$ `lda` $\prec$ `svm1`

Scenario 1 – Different algorithm performances:



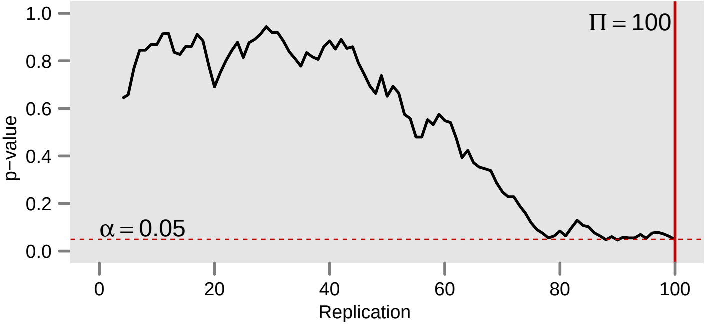
Scenario 2 – Similar algorithm performances:



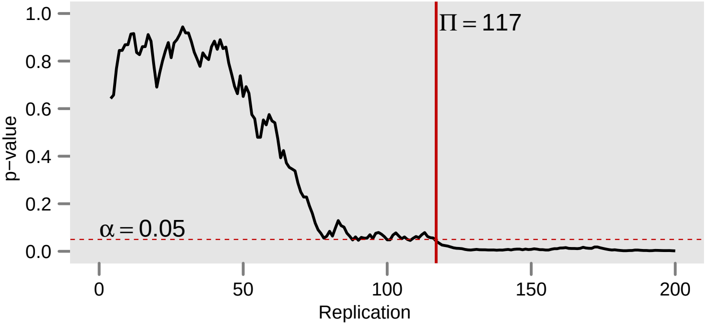
p-value $= 0.0484$

$\Rightarrow$ $\texttt{rf} \prec \texttt{svm1}$

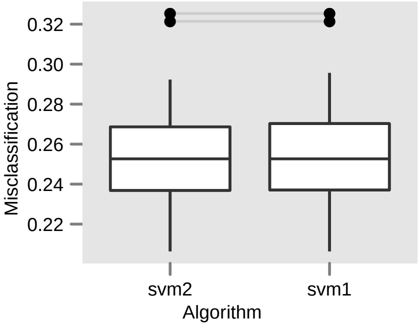
Scenario 2 – Similar algorithm performances:



Scenario 2 – Similar algorithm performances:



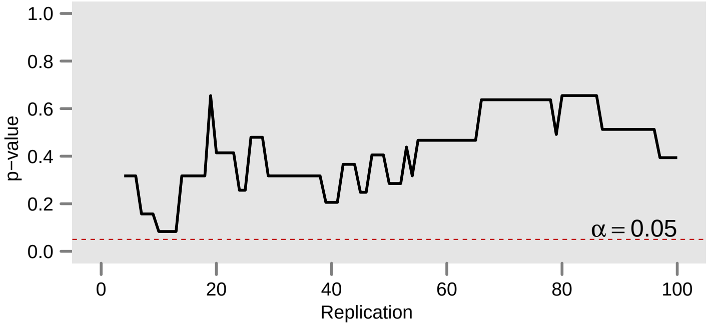
Scenario 3 – Equal algorithm performances:



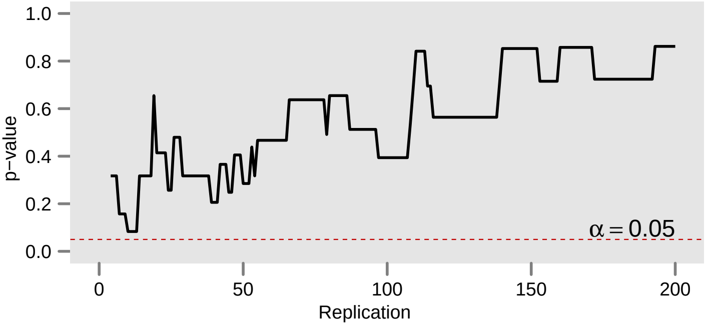
p-pvalue $= 0.9302$

$\Rightarrow$ `svm2` $\approx$ `svm1`

Scenario 3 – Equal algorithm performances:



Scenario 3 – Equal algorithm performances:



Interpretation

Point consecutively significance:

$$\Pi_{\text{Scenario 1}} = 13,\ \Pi_{\text{Scenario 2}} = 117,\ \Pi_{\text{Scenario 3}} = \infty$$

Measure of "how big the difference" is – indicator for relevance?

Decision making

Goal: Execute a benchmark experiment as long as needed – either until H_0 is rejected or H_0 is "accepted" (failed to reject).

Repeated significance testing:

Testing not once but multiple times causes the inflation of the probability for the error of the first kind, i.e., the probability of rejecting the global null hypothesis when in fact this hypothesis is true; known as **alpha inflation**.

(*) First addressed by Armitage, McPherson, and Rowe (1969).

Analyses on accumulating data

Sequential: Sample observations one by one; the test is executed after each new observation – the experiment can be stopped at any point.

Group sequential: Sample groups of observations; the test is executed after each group – the experiment can be stopped after each group.

Adaptive: Group sequential with more flexibility, e.g., to change hypothesis, group sample size, etc.

(*) Following Vandemeulebroecke (2008).

Recursive combination tests

Adaptive test procedure based on the recursive application of two-stage combination tests; and a *p*-value function to combine *p*-values from two stages.

(*) Defined by Brannath, Posch, and Bauer (2002).

Two-stage combination tests:

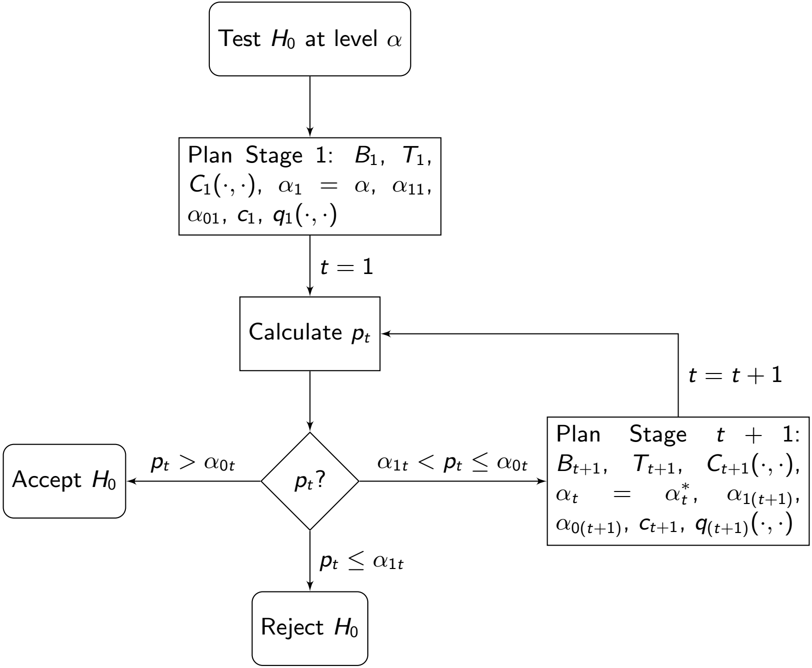
Test a one-sided null hypothesis H_0 at level α using two stages.

Stage 1: B_1 learning samples, decision boundaries α_{01} and α_{11} $(0 \leq \alpha_{11} < \alpha < \alpha_{01} \leq 1)$.

$$\text{Decision} = \begin{cases} \text{reject } H_0, & p_1 \leq \alpha_{11} \\ \text{accept } H_0, & p_1 > \alpha_{01} \\ \text{perform Stage 2}, & \alpha_{11} < p_1 \leq \alpha_{01} \end{cases}$$

Stage 2: B_2 learning samples.

$$\text{Decision} = \begin{cases} \text{reject } H_0, & C(p_1, p_2) \leq c \\ \text{undecidable}, & \text{otherwise} \end{cases}$$



Test H_0 at level α
Plan Stage 1: B_1, T_1, $C_1(\cdot,\cdot)$, $\alpha_1 = \alpha$, α_{11}, α_{01}, c_1, $q_1(\cdot,\cdot)$
$t = 1$
Calculate p_t
$t = t + 1$
$p_t?$
$p_t > \alpha_{0t}$
Accept H_0
$\alpha_{1t} < p_t \leq \alpha_{0t}$
Plan Stage $t+1$: B_{t+1}, T_{t+1}, $C_{t+1}(\cdot,\cdot)$, $\alpha_t = \alpha_t^*$, $\alpha_{1(t+1)}$, $\alpha_{0(t+1)}$, c_{t+1}, $q_{(t+1)}(\cdot,\cdot)$
$p_t \leq \alpha_{1t}$
Reject H_0

Fisher's product test:

$$C(p_{t-1}, p_t) = p_{t-1} \cdot p_t$$

Critical value:

$$c_t = \frac{\alpha_t^* - \alpha_{1t}}{\ln \alpha_{0t} - \ln \alpha_{1t}}$$

Conditional significance level:

$$\alpha_1^* = \alpha,\ \alpha_t^* = \frac{c_{t-1}}{p_{t-1}}$$

Conditional decision boundaries:

$$\alpha_{1t} < \alpha_t^* \leq \alpha_{0t}$$

Global p-value (after t stages):

$$p = q(p_1, p_2)$$

with

$$q(p_1, p_2) = \begin{cases} p_1, & p_1 \leq \alpha_{11} \text{ or } p_1 > \alpha_{01} \\ \alpha_{11} + p_1 \cdot p_2 \cdot (\ln \alpha_{01} - \ln \alpha_{11}), & p_1 \in (\alpha_{11}, \alpha_{01}] \\ & \text{ and } p_1 \cdot p_2 \leq \alpha_{11} \\ p_1 \cdot p_2 + p_1 \cdot p_2 \cdot (\ln \alpha_{01} - \ln p_1 \cdot p_2), & p_1 \in (\alpha_{11}, \alpha_{01}] \\ & \text{ and } p_1 \cdot p_2 \geq \alpha_{11} \end{cases}$$

and

$$p_2 = q(p_2, \ldots, q(p_{t-1}, p_t))$$

Recursive combination tests in benchmark experiments:

There are differences between clinical trials – where sequential analysis is usually applied – and benchmark experiments:

Recursive combination tests in benchmark experiments:

There are differences between clinical trials – where sequential analysis is usually applied – and benchmark experiments:

Automatized: Benchmark experiments are computer experiments; decisions that are to be made in the interim and planning phases need to be automatized soundly.

Recursive combination tests in benchmark experiments:

There are differences between clinical trials – where sequential analysis is usually applied – and benchmark experiments:

Automatized: Benchmark experiments are computer experiments; decisions that are to be made in the interim and planning phases need to be automatized soundly.

Multiobjective: Usually more than two candidate algorithms are compared to a set of performance measures; perform replications until a decision is made for each combination.

Recursive combination tests in benchmark experiments:

There are differences between clinical trials – where sequential analysis is usually applied – and benchmark experiments:

Automatized: Benchmark experiments are computer experiments; decisions that are to be made in the interim and planning phases need to be automatized soundly.

Multiobjective: Usually more than two candidate algorithms are compared to a set of performance measures; perform replications until a decision is made for each combination.

Approximative: Using resampling schemes lead to non-independent p-values; however, the p-values' correlation vanishes with increasing data set size, we can rely on the asymptotic.

Recursive combination tests in benchmark experiments:

There are differences between clinical trials – where sequential analysis is usually applied – and benchmark experiments:

Automatized: Benchmark experiments are computer experiments; decisions that are to be made in the interim and planning phases need to be automatized soundly.

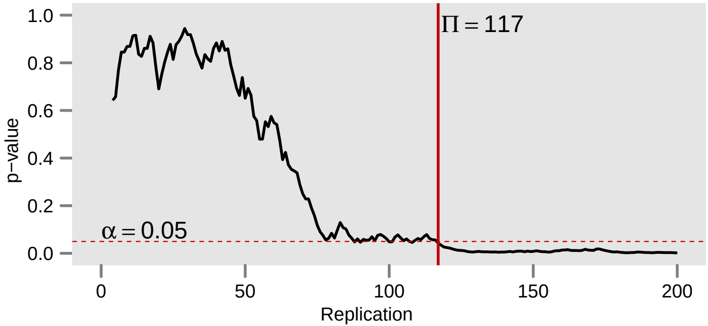
Multiobjective: Usually more than two candidate algorithms are compared to a set of performance measures; perform replications until a decision is made for each combination.

Approximative: Using resampling schemes lead to non-independent p-values; however, the p-values' correlation vanishes with increasing data set size, we can rely on the asymptotic.

Decidable: It is easy and (comparatively) cheap to make additional replications until a final decision is reached; i.e., to reject or accept H_0.

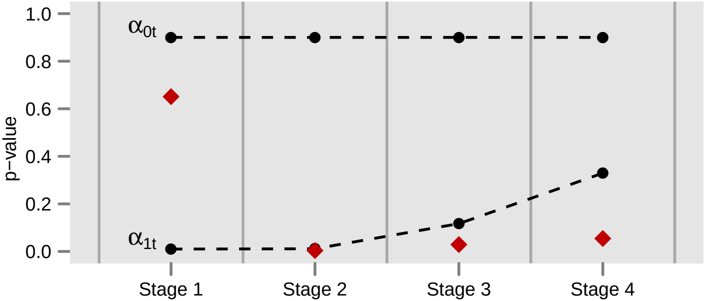
Scenario 2 – Similar algorithm performances:

Split the scenario into four stages, i.e., $B_t = \frac{B}{4} = 50$; $\alpha = \alpha_1^* = 0.05$, $\alpha_{11} = 0.01$, $\alpha_{01} = 0.9$; and the rule for the conditional decision boundaries is $\alpha_{0t} = \frac{\alpha_t^*}{1.2}$ and $\alpha_{1t} = \alpha_{1(t-1)}$.



Scenario 2 – Similar algorithm performances:

Split the scenario into four stages, i.e., $B_t = \frac{B}{4} = 50$; $\alpha = \alpha_1^* = 0.05$, $\alpha_{11} = 0.01$, $\alpha_{01} = 0.9$; and the rule for the conditional decision boundaries is $\alpha_{0t} = \frac{\alpha_t^*}{1.2}$ and $\alpha_{1t} = \alpha_{1(t-1)}$.

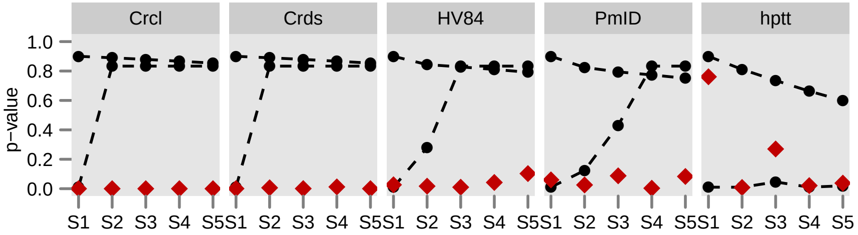


Application example

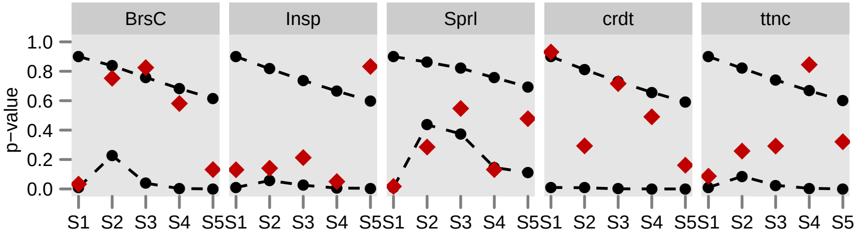
Benchmark experiments on 20 UCI data sets with $B = 250$; we now take a look at the decisions of the recursive combination tests for the two leader algorithms of each data set.

The non-parametric (one-sided) Wilcoxon Signed Rank test as T_t and Fisher's Product test as combination test C_t. Split each experiment into five stages, i.e., $t = 1, \ldots, 5$, $B_t = \frac{B}{5} = 50$, and define $\alpha = \alpha_1^* = 0.05$, $\alpha_{11} = 0.01$, $\alpha_{01} = 0.9$. The rules for the conditional decision boundaries are defined as $\alpha_{0t} = \frac{\alpha_t^*}{1.2}$ and $\alpha_{1t} = \alpha_{1(t-1)} - \frac{\|\alpha_{1(t-1)} - \alpha_t^*\|}{10}$.

Data sets with significant differences (14 data sets):



Data sets with non-significant differences (6 data sets):



Summary

Taking the sequential nature of a benchmark experiment into account, enables to monitor and to make decisions during the execution of the experiment.

Monitoring:

- Point of consecutively significance.

Decision making:

- Statistical justification for the number of replications.
- For the UCI application example – early stopping in case of significant decisions and no "well-founded" decisions otherwise.

Outlook

- Less flexibility to the benefit of more efficiency?
  Strictly sequential approaches or, e.g., the group sequential approach CRP by Müller and Schäfer (2001).
- Multiobjective framework.
  Allow more than one performance measure.
- Framework stability.
  Investigate “all” possible test decisions under rearrangements of the individual replications (permutations).

Publications – Benchmark experiments

Manuel J. A. Eugster. *benchmark: Benchmark Experiments Toolbox*, 2011. URL `http://cran.r-project.org/package=benchmark`. R package version 0.3-2.

Manuel J. A. Eugster and Friedrich Leisch. Bench plot and mixed effects models: First steps toward a comprehensive benchmark analysis toolbox. In Paula Brito, editor, *Compstat 2008—Proceedings in Computational Statistics*, pages 299–306. Physica Verlag, Heidelberg, Germany, 2008. ISBN 978-3-7908-2083-6. Preprint available from `http://epub.ub.uni-muenchen.de/3206/`.

Manuel J. A. Eugster and Friedrich Leisch. Exploratory analysis of benchmark experiments – an interactive approach. *Computational Statistics*, 2010. doi: 10.1007/s00180-010-0227-z. Accepted for publication on 2010-06-08, preprint available from `http://epub.ub.uni-muenchen.de/10604/`.

Manuel J. A. Eugster, Torsten Hothorn, and Friedrich Leisch. Exploratory and inferential analysis of benchmark experiments. Under review, preprint available from `http://epub.ub.uni-muenchen.de/4134/`, 2010a.

Manuel J. A. Eugster, Torsten Hothorn, and Friedrich Leisch. Domain-based benchmark experiments: Exploratory and inferential analysis. Under review, preprint available from `http://epub.ub.uni-muenchen.de/4134/`, 2010b.

Manuel J. A. Eugster, Friedrich Leisch, and Carolin Strobl. (Psycho-)analysis of benchmark experiments – a formal framework for investigating the relationship between data sets and learning algorithms. Under review, preprint available from `http://epub.ub.uni-muenchen.de/11425/`, 2010c.

Publications – Archetypal analysis

Manuel J. A. Eugster. *archetypes: Archetypal Analysis*, 2010. URL `http://cran.r-project.org/package=archetypes`. R package version 2.0-2.

Manuel J. A. Eugster and Friedrich Leisch. From Spider-man to Hero – archetypal analysis in R. *Journal of Statistical Software*, 30(8):1–23, 2009. URL `http://www.jstatsoft.org/v30/i08`.

Manuel J. A. Eugster and Friedrich Leisch. Weighted and robust archetypal analysis. *Computational Statistics and Data Analysis*, 55(3):1215–1225, 2011. doi: 10.1016/j.csda.2010.10.017. Preprint available from `http://epub.ub.uni-muenchen.de/11498/`.

References

P. Armitage, C. K. McPherson, and B. C. Rowe. Repeated significane test on accumulating data. *Journal of the Royal Statistical Society*, 132(2), 1969.

Werner Brannath, Martin Posch, and Peter Bauer. Recursive combination tests. *Journal of the American Statistical Association*, 97(457):236–244, 2002. doi: 10.1198/016214502753479374.

Kurt Hornik and David Meyer. Deriving consensus rankings from benchmarking experiments. In R. Decker and H.-J. Lenz, editors, *Advances in Data Analysis (Proceedings of the 30th Annual Conference of the Gesellschaft für Klassifikation e.V., Freie Universität Berlin, March 8–10, 2006*, Studies in Classification, Data Analysis, and Knowledge Organization, pages 163–170. Springer-Verlag, 2007. doi: 10.1007/978-3-540-70981-7_19.

Torsten Hothorn, Friedrich Leisch, Achim Zeileis, and Kurt Hornik. The design and analysis of benchmark experiments. *Journal of Computational and Graphical Statistics*, 14(3):675–699, 2005. doi: 10.1198/106186005X59630.

Hans-Helge Müller and Helmut Schäfer. Adaptive group sequential designs for clinical trials: Combining the advantages of adaptive and of classical group sequential approaches. *Biometrics*, 57(3):886–891, 2001. doi: 10.1111/j.0006-341X.2001.00886.x.

Marc Vandemeulebroecke. Group sequential and adaptive designs – a review of basic concepts and points of discussion. *Biometrical Journal*, 50(3), 2008. doi: 10.1002/bimj.200710436.