(Psycho-)Analysis of Benchmark Experiments

A Formal Framework for Investigating the Relationship between Data Sets and Learning Algorithms

Manuel J. A. Eugster

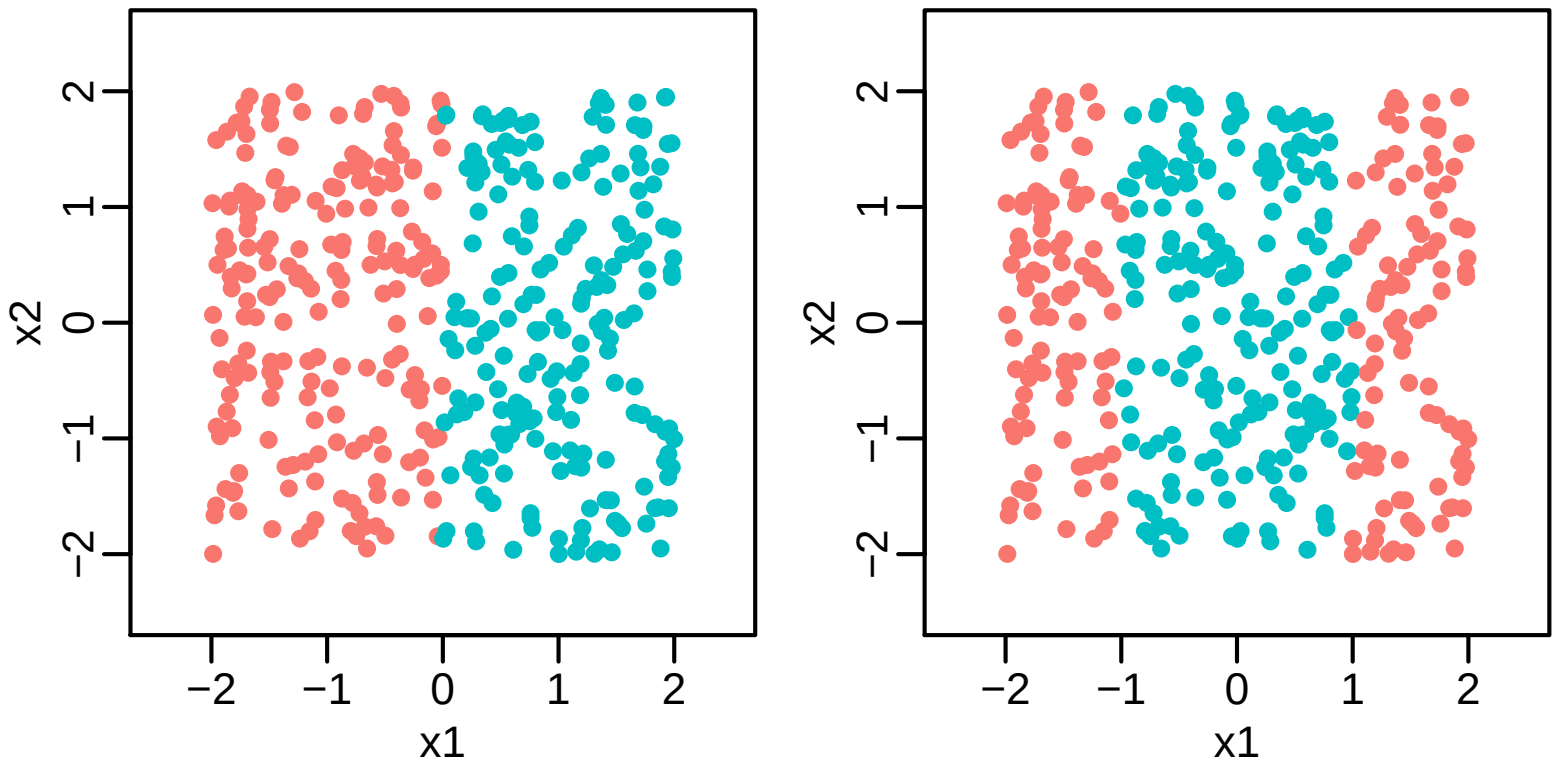
Joint work with Carolin Strobl and Friedrich Leisch

HIIT seminar, October 22, 2012

Are there systematic differences in the performances of learning algorithms related to specific dataset characteristics?

Important when illustrating

- the properties of existing methods, or
- the superiority of newly developed algorithms.



Support vector machines (`svm`), linear discriminant analysis (`lda`) and quadratic discriminant analysis (`qda`)

The presented framework combines the advantages of three well-established approaches:

1. *Benchmark experiments* from statistical and machine learning to evaluate the performance of the algorithms;
2. *Statistical and information-theoretic measures* from meta learning to describe the data sets; and
3. recursive partitioning of *Bradley-Terry models* from psychology to capture the differences in the performance of the algorithms on data sets with similar characteristics.

1. Benchmark experiments

Data generating process:

Given is a data set $\mathfrak{L} = \{z_1, \ldots, z_N\}$. We draw $b = 1, \ldots, B$ learning samples of size n using a resampling scheme (e.g., bootstrapping or subsampling):

$$\mathfrak{L}^b = \{z_1^b, \ldots, z_n^b\}$$

Candidate algorithms:

There are $K > 1$ candidate algorithms a_k, $k = 1, \ldots, K$, available for the solution of the underlying problem. For each algorithm, $a_k(\cdot \mid \mathfrak{L}^b)$ is the fitted model based on the sample $\mathfrak{L}^b$:

$$a_k(\cdot \mid \mathfrak{L}^b) \sim \mathcal{A}_k(\mathfrak{L})$$

(*) Following Hothorn et al. (2005) and Eugster et al. (2012)

Performance measure:

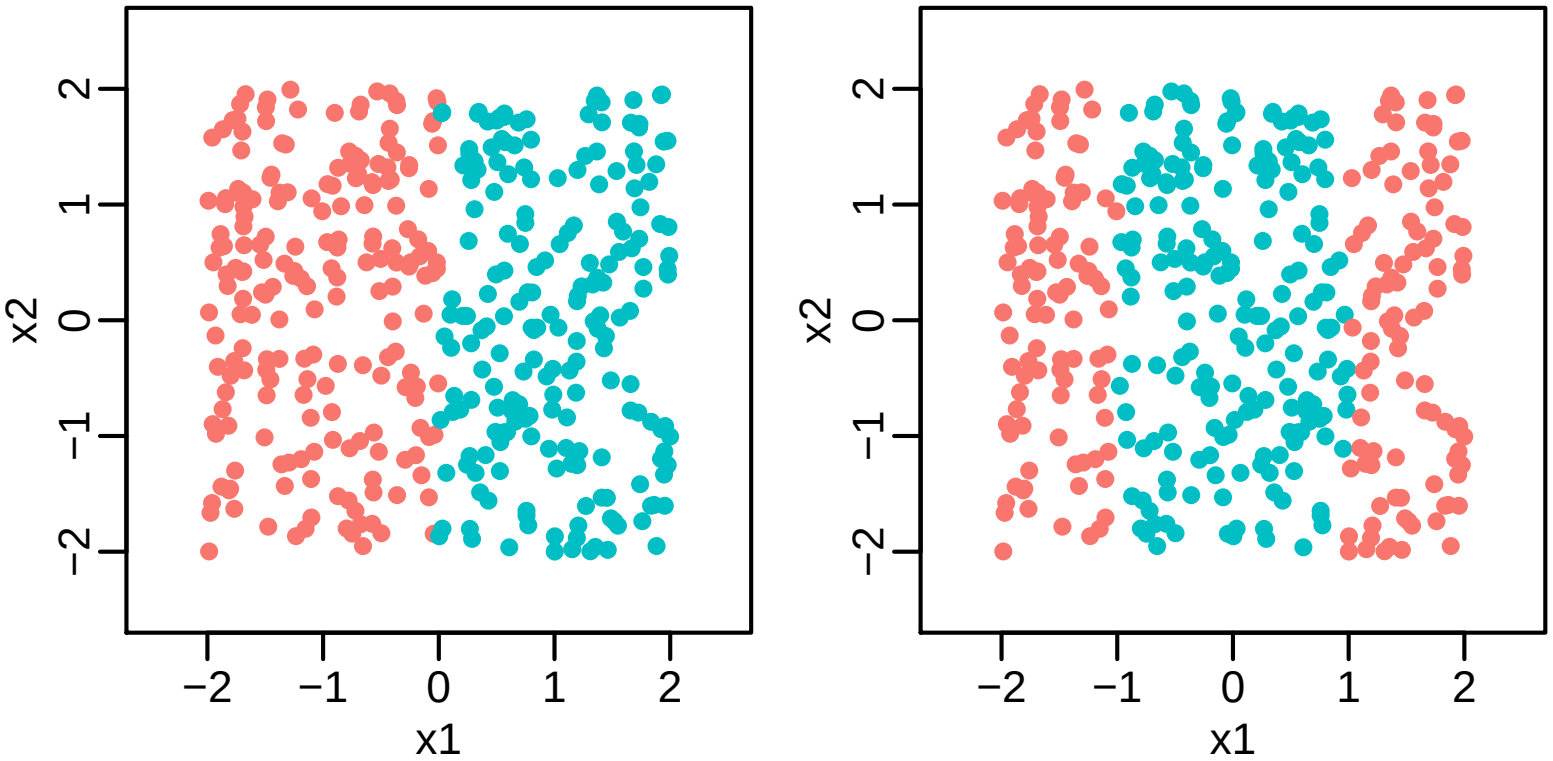
The performance of the candidate algorithm a_k when provided with the learning sample $\mathfrak{L}^b$ is measured by a scalar function p:

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

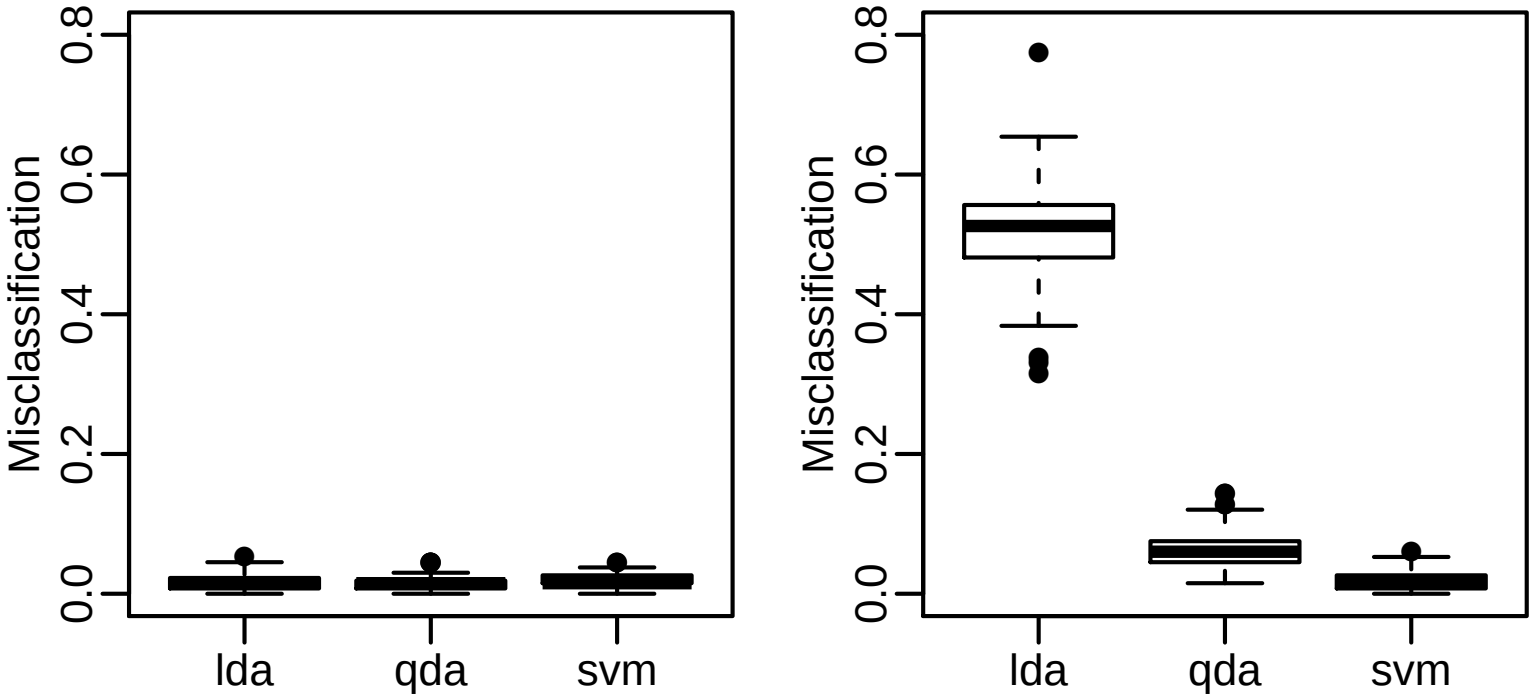
Empirical performance measure:

An estimation of the generalization performance of the candidate algorithm a_k learned on learning sample $\mathcal{L}^b$ is based on a test sample $\mathfrak{T}^b$ (often $\mathcal{L} \setminus \mathcal{L}^b$):

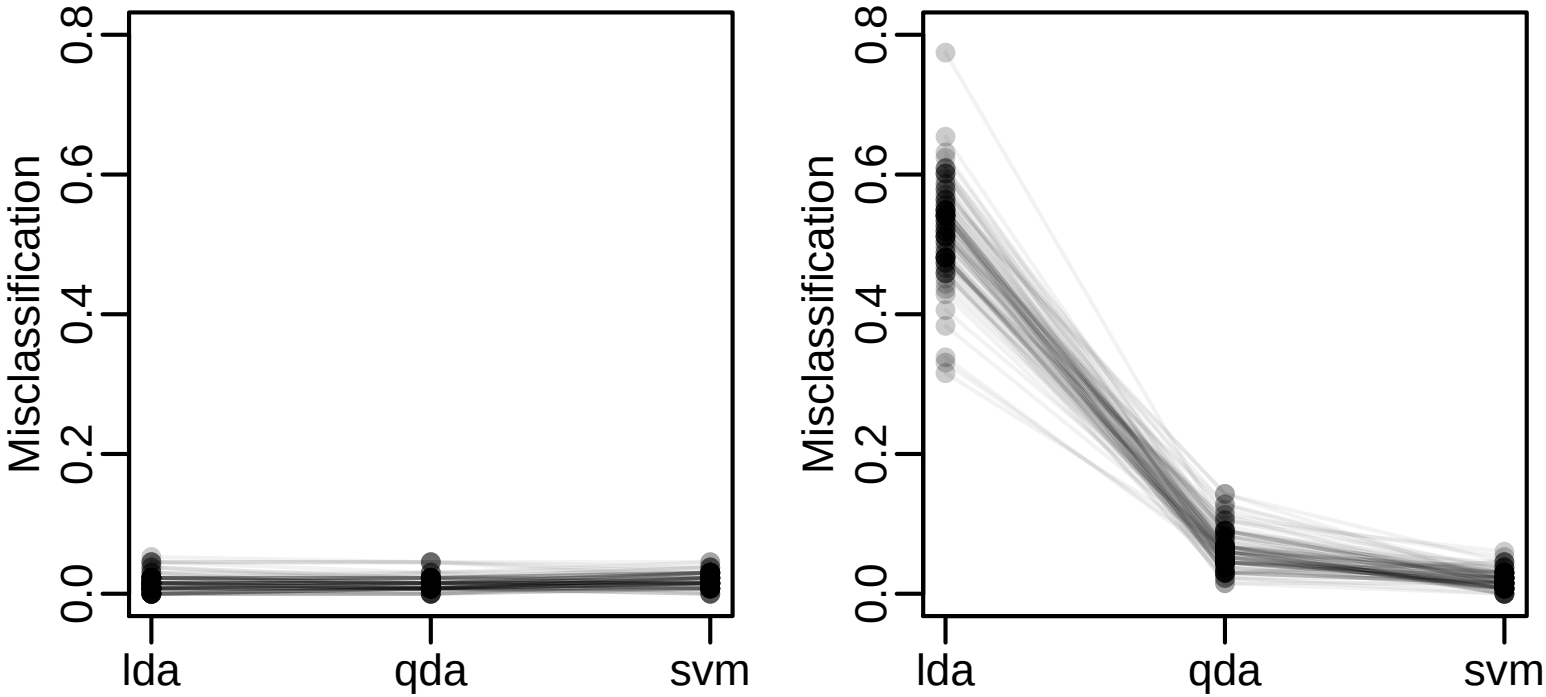
$$\hat{p}_{bk} = \hat{p}_{\mathfrak{T}^b}(a_k, \mathfrak{L}^b) \sim \hat{\mathcal{P}}_k(\mathfrak{L})$$



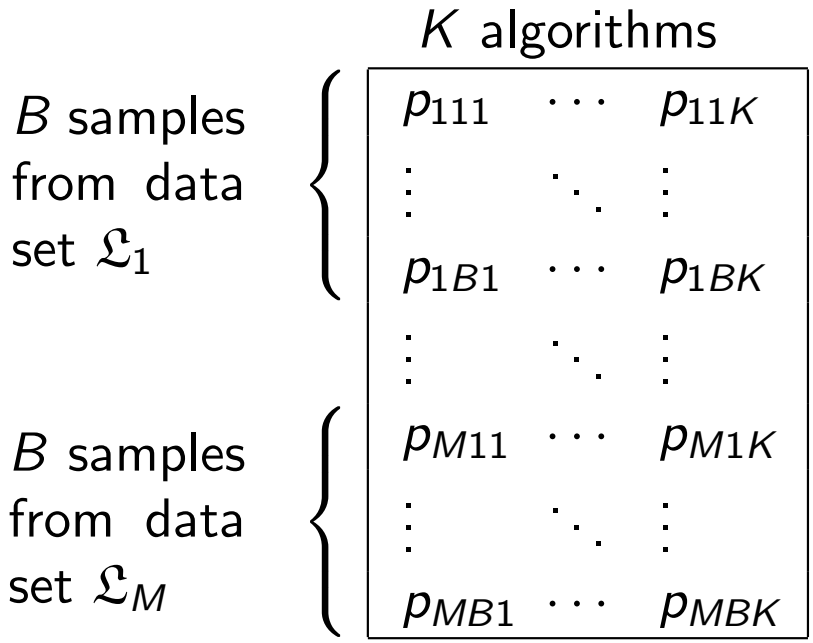
Missclassification of the candidate algorithms on the two illustrative datasets based on $B = 100$ with 2/3-subsampling as resampling scheme and p being the misclassification error.



Missclassification of the candidate algorithms on the two illustrative datasets based on $B = 100$ with 2/3-subsampling as resampling scheme and p being the misclassification error.



Missclassification of the candidate algorithms on the two illustrative datasets based on $B = 100$ with 2/3-subsampling as resampling scheme and p being the misclassification error.


K algorithms
B samples from data set $\mathfrak{L}_1$
B samples from data set $\mathfrak{L}_M$
p_{111} $\cdots$ p_{11K}
p_{1B1} $\cdots$ p_{1BK}
p_{M11} $\cdots$ p_{M1K}
p_{MB1} $\cdots$ p_{MBK}

3a. Preferene scaling

Pairwise comparisons:

Each comparison of the performance of two algorithms has three possible outcomes:

1. the first algorithm wins
2. the second algorithm wins, or
3. both algorithms perform equally (i.e., a tie).

Performance relations:

The relation $R(p_{mbk}, p_{mbk'})$ describes one of the outcomes (1), (2) or (3) of the comparison of algorithms k and k' on sample b drawn from data set m.

(*) Critchlow and Fligner (1991)

$$
\begin{array}{l}
 \\
B \text{ samples} \\
\text{from data} \\
\text{set } \mathfrak{L}_1 \\
 \\
 \\
B \text{ samples} \\
\text{from data} \\
\text{set } \mathfrak{L}_M
\end{array}
\left\{
\begin{array}{c}
\overbrace{\phantom{R(p_{111}, p_{112}) \cdots R(p_{11K-1}, p_{11K})}}^{K \cdot (K-1)/2 \text{ comparisons}} \\
R(p_{111}, p_{112}) \cdots R(p_{11K-1}, p_{11K}) \\
\vdots \quad \ddots \quad \vdots \\
R(p_{1B1}, p_{1B2}) \cdots R(p_{1BK-1}, p_{1BK}) \\
\vdots \quad \ddots \quad \vdots \\
R(p_{M11}, p_{M12}) \cdots R(p_{M1K-1}, p_{M1K}) \\
\vdots \quad \ddots \quad \vdots \\
R(p_{MB1}, p_{MB2}) \cdots R(p_{MBK-1}, p_{MBK})
\end{array}
\right.
$$

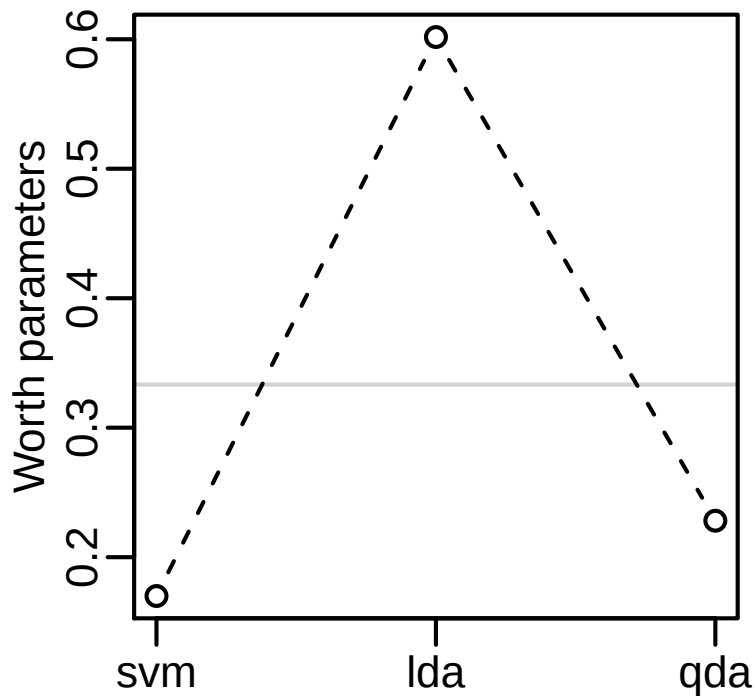
Bradley-Terry model:

According to the Bradley-Terry model the three possible outcomes have the probabilities:

$$
\begin{aligned}
P\left(R(p_{mbk}, p_{mbk'}) = \text{"}a_k \text{ wins"}\right) &= \frac{\pi_k}{\pi_k + \pi_{k'} + \nu\sqrt{\pi_k \pi_{k'}}}, \\
P\left(R(p_{mbk}, p_{mbk'}) = \text{"}a_{k'} \textit{wins"}\right) &= \frac{\pi_{k'}}{\pi_k + \pi_{k'} + \nu\sqrt{\pi_k \pi_{k'}}}, \\
P\left(R(p_{mbk}, p_{mbk'}) = \text{"tie"}\right) &= \frac{\nu\sqrt{\pi_k \pi_{j'}}}{\pi_k + \pi_{k'} + \nu\sqrt{\pi_k \pi_{k'}}},
\end{aligned}
$$

where the $\pi_k \geq 0$, $k = 1, \ldots, K$, are the parameters indicating the strength of each algorithm, and $\nu \geq 0$ is a discrimination constant governing the probability of ties.

The parameters are estimated via maximum likelihood; see, e.g., Strobl et al. (2011).



Worth parameters—describing the location of the objects on the preference scale—of a global BT model for the two illustrative datasets.

2. Statistical and information-theoretic measures

Dataset characterization:
Given some user-specified characteristics, data set characterization can be formalized as follows.

1. *map* each data set into its individual characterization space ($\mathfrak{L}$ the space of all data sets, $\mathfrak{L} \in \mathfrak{L}$):

$$map\colon \mathfrak{L} \to \mathbb{R}^* \quad \text{with} \quad \mathfrak{L} \mapsto x^*$$

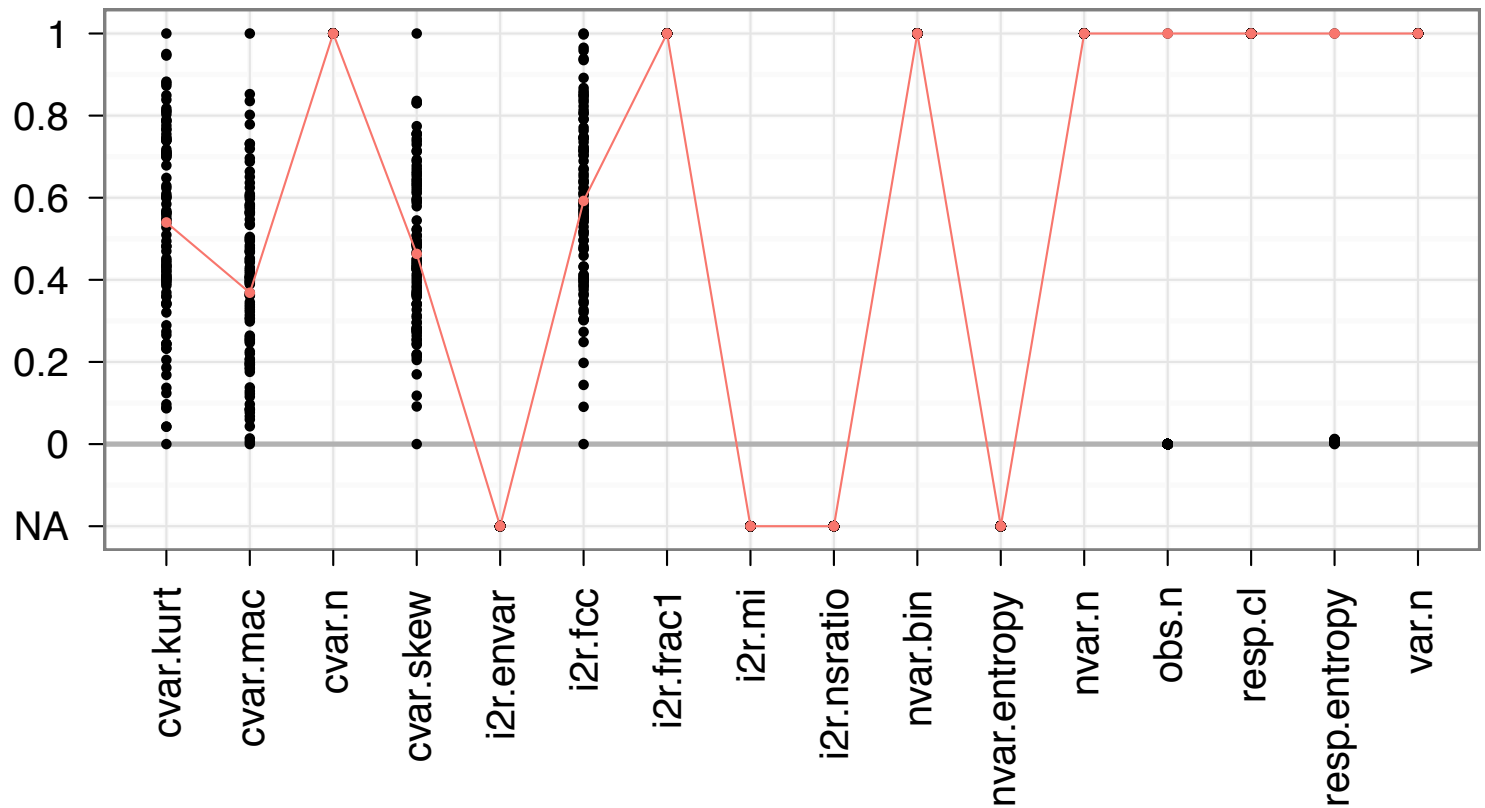
2. *reduce* the individual characterization spaces into one common characterization space where all data sets are comparable, i.e. a metric can be defined:

$$red\colon \mathbb{R}^* \to \mathbb{R}^d \quad \text{with} \quad x^* \mapsto x^d$$

3. A specific data set characterization then consists of a set of characteristics $\{(map_1, red_1), \ldots, (map_J, red_J)\}$, and for a given data set $\mathfrak{L}$, its characterization is the vector $c = (c_1, \ldots, c_J)$ with $c_j = red_j(map_j(\mathfrak{L}))$, $j = 1, \ldots, J$.

Characteristic	Description	ds1	ds2
obs.n	number of observations	400	400
var.n	number of variables	2	2
nvar.n	number of nominal variables	0	0
nvar.entropy	mean nominal variable entropy		
nvar.bin	number of binary variables	0	0
cvar.n	number of continuous variables	2	2
cvar.mac	mean multiple attribute correlation	0.06	0.06
resp.cl	number of response classes	2	2
resp.entropy	mean response entropy	5.93	5.93
i2r.fcc	**first canonical correlation**	**0.86**	**0.04**
i2r.frac1	variation from first linear discriminant	1.00	1.00
i2r.mi	mean mutual information		
⋮			

STATLOG (King et al., 1995) characteristics computed for the two illustrative data sets; e.g., map_j is "first canonical correlation for each continuous variable" and red_j is "mean".



Relative variation of the characterizations of the 100 drawn samples in case of `ds1`. The red line marks the characterization of the original data set; NA means that this characteristic is not available on this data set.

	$K \cdot (K-1)/2$ comparisons			J characteristics		
B samples from data set $\mathfrak{L}_1$	$R(p_{111}, p_{112})$	$\cdots$	$R(p_{11K-1}, p_{11K})$	c_{111}	$\cdots$	c_{11J}
	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
	$R(p_{1B1}, p_{1B2})$	$\cdots$	$R(p_{1BK-1}, p_{1BK})$	c_{1B1}	$\cdots$	c_{1BJ}
	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
B samples from data set $\mathfrak{L}_M$	$R(p_{M11}, p_{M12})$	$\cdots$	$R(p_{M1K-1}, p_{M1K})$	c_{M11}	$\cdots$	c_{M1J}
	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
	$R(p_{MB1}, p_{MB2})$	$\cdots$	$R(p_{MBK-1}, p_{MBK})$	c_{MB1}	$\cdots$	c_{MBJ}

3b. Recursive partitioning of BT models

Model-based recursive partitioning:

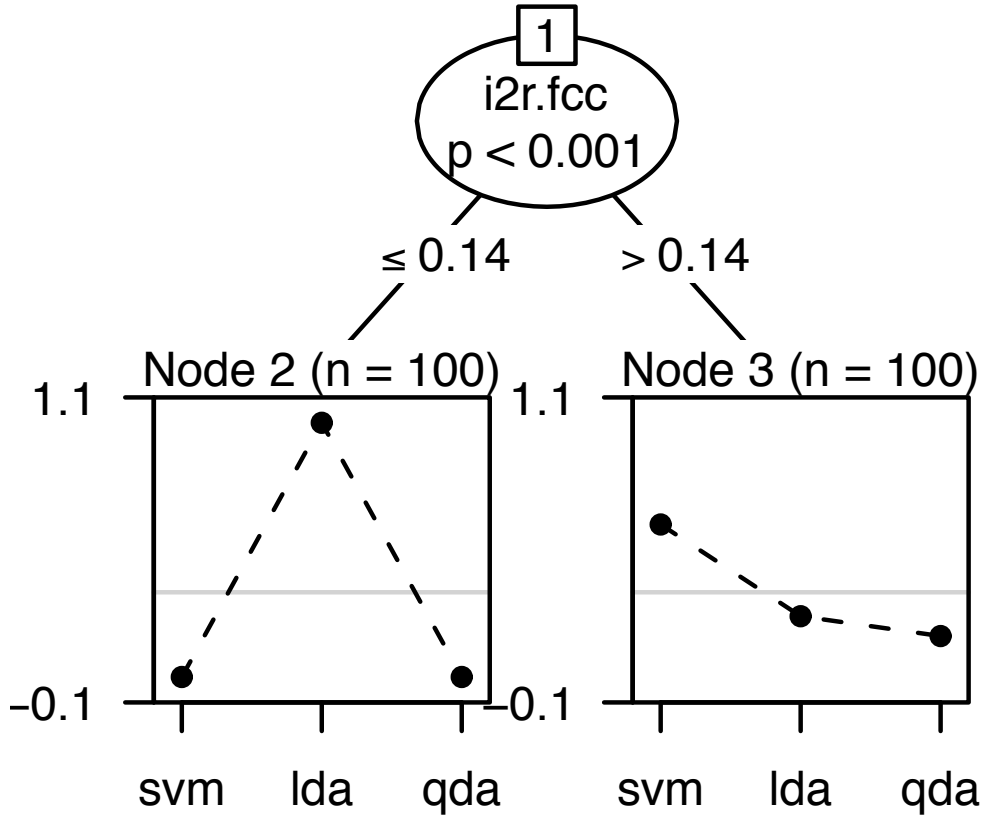
- **CART** detects groups of observations **with different values of a response variable** by means of recursively splitting the feature space.

- **Model-based recursive partitioning** detects groups of observations **which vary in the parameters of a certain model of interest** by means of recursively splitting the feature space.

(*) Breiman et al. (1984), Zeileis et al. (2008) and Strobl et al. (2011)

Algorithm:

1. Fit a Bradley-Terry model for the paired comparisons of the algorithms based on all data sets in the current node (starting with the root node including all data sets).
2. Assess the stability of the Bradley-Terry model parameters with respect to each characteristic of the data sets.
3. If there is significant instability in the model parameters, split the data sets in two nodes along the characteristic with the strongest instability, and use the cutpoint with the highest improvement of the model fit.
4. Repeat steps 1–3 recursively in the resulting nodes until there are no more significant instabilities (or the number of data sets left in a node falls below a given stopping value).

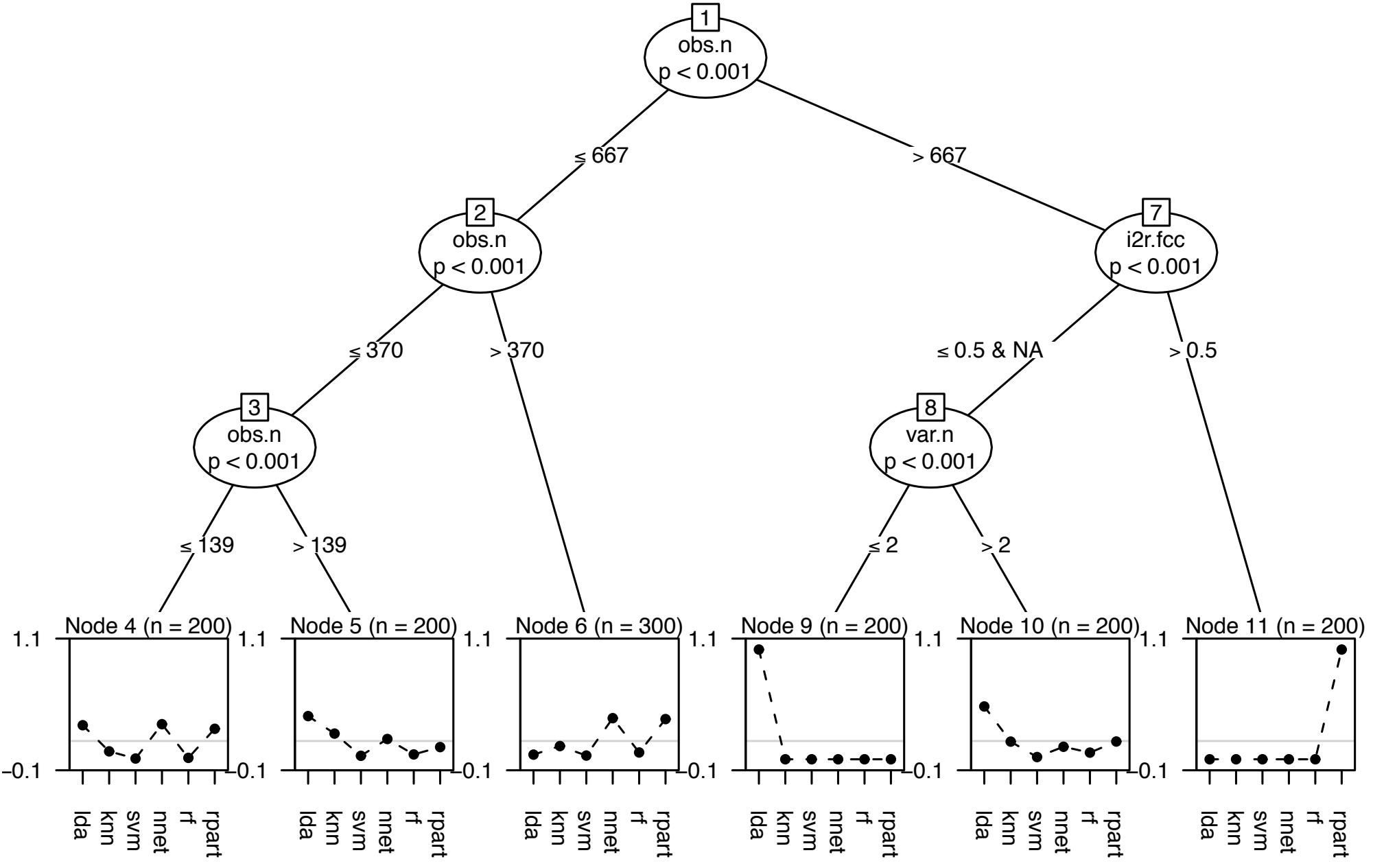


The first canonical correlation `i2r.fcc`—that indicates whether the data set is linearly separable—is correctly identified as the characteristic that induces a significant change in the performance-ranking of the algorithms.

Application example

UCI Machine Learning Repository:

- $M = 13$ **data sets**, that are all binary classification problems but cover a wide range of data set characteristics.
- $J = 16$ dataset **characteristics** based on the STATLOG project.
- $K = 6$ candidate **algorithms** (lda, knn, rpart, svm, nnet, rf).
- Misclassification as **performance measure** p; and $B = 100$ samples using 2/3-subsampling without replacement as **resampling** scheme.


1
obs.n
p < 0.001
≤ 667
> 667
2
obs.n
p < 0.001
7
i2r.fcc
p < 0.001
≤ 370
> 370
≤ 0.5 & NA
> 0.5
3
obs.n
p < 0.001
8
var.n
p < 0.001
≤ 139
> 139
≤ 2
> 2
Node 4 (n = 200)
Node 5 (n = 200)
Node 6 (n = 300)
Node 9 (n = 200)
Node 10 (n = 200)
Node 11 (n = 200)
1.1
−0.1
lda
knn
svm
nnet
rf
rpart

Challenge

The main challenge of this approach is the selection of the “right” data set characteristics.

$\overset{TODO}{\Rightarrow}$ Create a huge collection of all kind of characteristics.

The benefit of this approach is that—at least from the set of characteristics provided—the relevant ones are selected for splitting automatically.

Summary

- A statistically sound way for illustrating the properties of existing methods or the superiority of new algorithms.

- Key advantages (because of the recursive partitioning of Bradley-Terry models):

  1. Easy interpretation by means of visualization;
  2. Capturing of potentially complex interactions between different data set characteristics; and
  3. Selection of only those characteristics from a potentially large number of data set characteristics that correspond to a significant change in the performances of the algorithms.

The proposed method is implemented in the R **benchmark** package as of version 0.3-4.

Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. *Classification and Regression Trees*. Chapman and Hall, 1984. ISBN 0412048418.

Douglas E. Critchlow and Michael A. Fligner. Paired comparison, triple comparison, and ranking experiments as generalized linear models, and their implementation on GLIM. *Psychometrika*, 56(3):517–533, 1991. doi: 10.1007/BF02294488.

Manuel J. A. Eugster. **benchmark***: Benchmark Experiments Toolbox*, 2012. URL `http://CRAN.R-project.org/package=benchmark`. R package version 0.3-4.

Manuel J. A. Eugster, Torsten Hothorn, and Friedrich Leisch. Domain-based benchmark experiments: Exploratory and inferential analysis. *Austrian Journal of Statistics*, 41(1):5–26, 2012. URL `http://www.stat.tugraz.at/AJS/ausg121/121Leisch.pdf`.

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.

Ross D. King, C. Feng, and A. Sutherland. STATLOG: Comparison of classification algorithms on large real-world problems. *Applied Artifcial Intelligence*, 9:289–333, 1995. doi: 10.1080/08839519508945477.

Carolin Strobl, Florian Wickelmaier, and Achim Zeileis. Accounting for individual differences in Bradley-Terry models by means of recursive partitioning. *Journal of Educational and Behavioral Statistics*, 36(2), 2011. doi: 10.3102/1076998609359791.

Achim Zeileis, Torsten Hothorn, and Kurt Hornik. Model-based recursive partitioning. *Journal of Computational and Graphical Statistics*, 17(2):492–514, 2008. doi: 10.1198/106186008X319331.