Tree Algorithms in Data Mining: Comparison of rpart and RWeka . . . and Beyond

Achim Zeileis

http://statmath.wu.ac.at/~zeileis/

Motivation

- For publishing new tree algorithms, benchmarks against established methods are necessary.
- When developing the tools in **party**, we benchmarked against **rpart**, the open-source implementation of CART.
- Statistical journals were usually happy with that.
- Usual comment from machine learners: *You have to benchmark against C4.5, it's much better than CART!*
- Quinlan provided source code for C4.5, but not with a license that would allow usage.
- **Weka** had an open-source Java implementation, but hard to access from R.
- When we developed **RWeka**, we finally were able to set up some benchmark with CART and C4.5 within R.

Tree algorithms

- CART/RPart (**rpart**): Classification and regression trees (Breiman, Friedman, Olshen, Stone 1984). Cross-validation-based cost-complexity pruning:
  - RPart0: Best prediction error.
  - RPart1: Highest complexity parameter within 1 standard error.
- C4.5/J4.8 (**RWeka**): C4.5 (Quinlan, 1993). Determine size by confidence threshold C and minimal leaf size M:
  - J4.8: Standard heuristics $C = 0.25$, $M = 2$.
  - J4.8(cv): Cross-validation for $C = 0.01, \ldots, 0.5$, $M = 2, \ldots, 20$.
- QUEST (**LohTools**): Quick, unbiased and efficient statistical trees (Loh, Shih 1997). Popularized concept of unbiased recursive partitioning in statistics. Hand-crafted convenience interface to original binaries.
- CTree (**party**): Conditional inference trees (Hothorn, Hornik, Zeileis 2006). Unbiased recursive partitioning based on permutation tests.

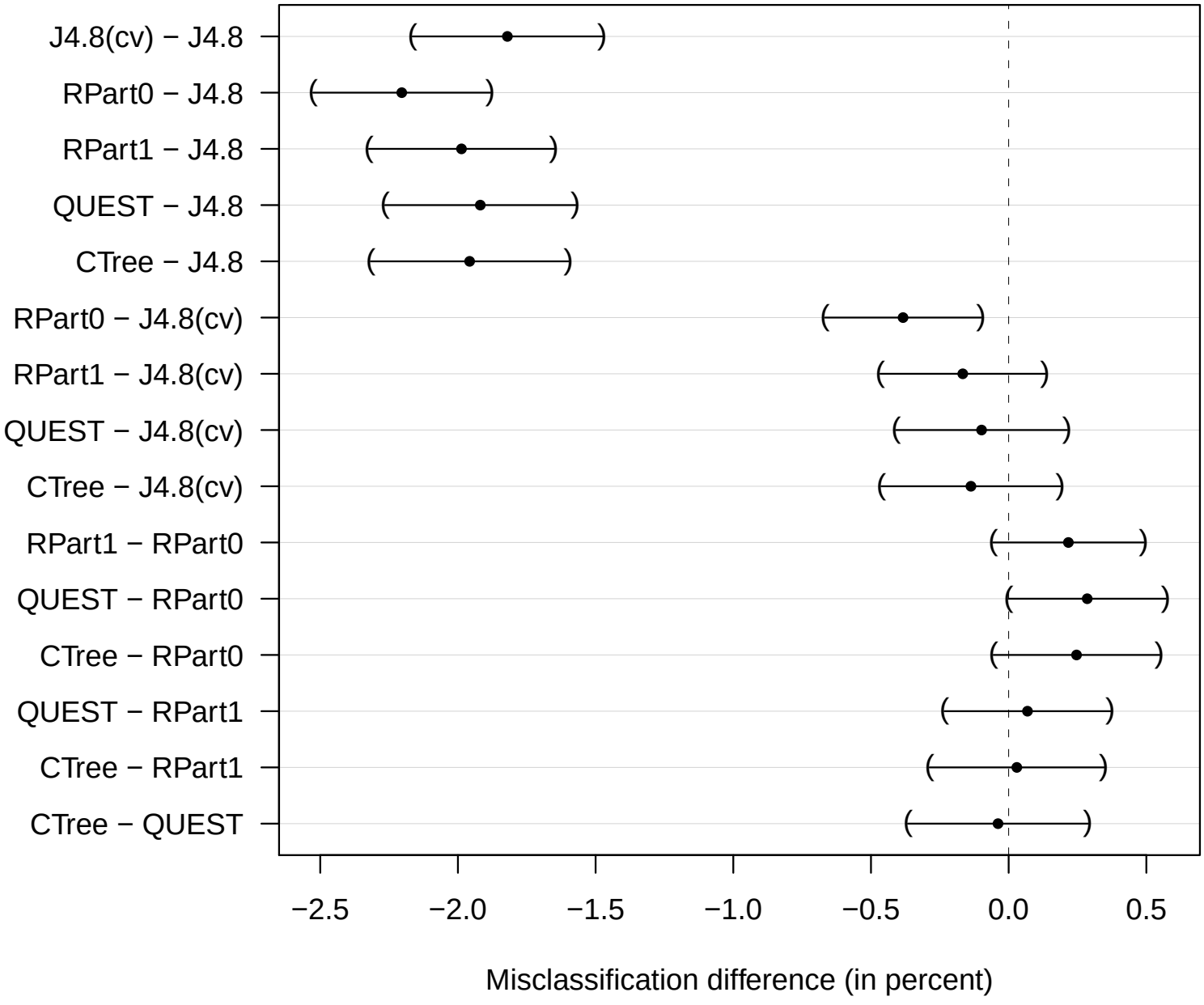
UCI data sets (mlbench)

Data set	# of obs.	# of cat. inputs	# of num. inputs
breast cancer	699	9	–
chess	3196	36	–
circle $*$	1000	–	2
credit	690	–	24
heart	303	8	5
hepatitis	155	13	6
house votes 84	435	16	–
ionosphere	351	1	32
liver	345	–	6
Pima Indians diabetes	768	–	8
promotergene	106	57	–
ringnorm $*$	1000	–	20
sonar	208	–	60
spirals $*$	1000	–	2
threenorm $*$	1000	–	20
tictactoe	958	9	–
titanic	2201	3	–
twonorm $*$	1000	–	20

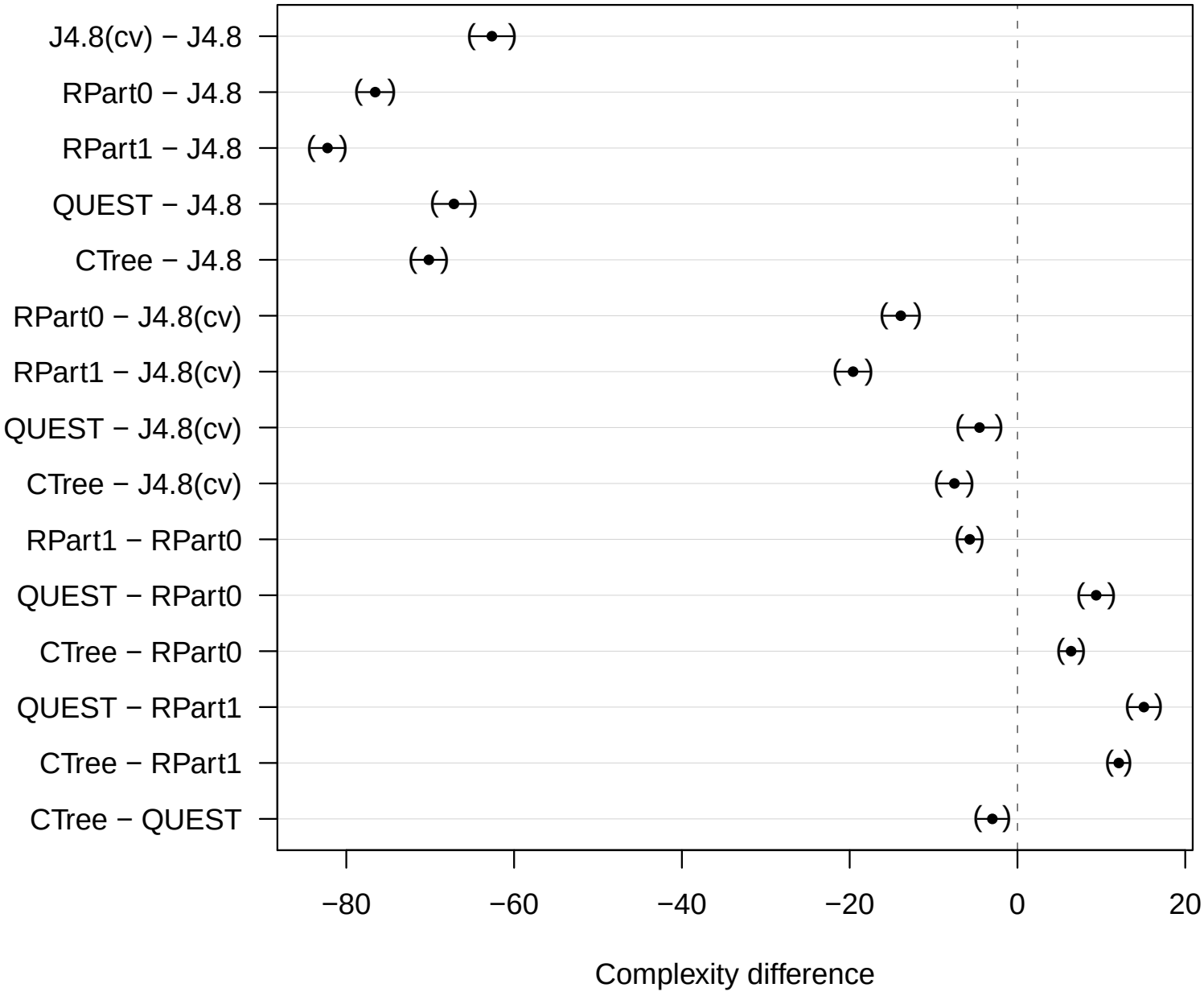
Analysis

- 6 tree algorithms.
- 18 data sets.
- 500 bootstrap samples for each combination.
- Performance measure: Out-of-bag misclassification rate.
- Complexity measure: Number of splits + number of leafs.
- Individual results: Simultaneous pairwise confidence intervals (Tukey all-pair comparisons).
- Aggregated results: Bradley-Terry model (Alternatively: median linear consensus ranking, ...).

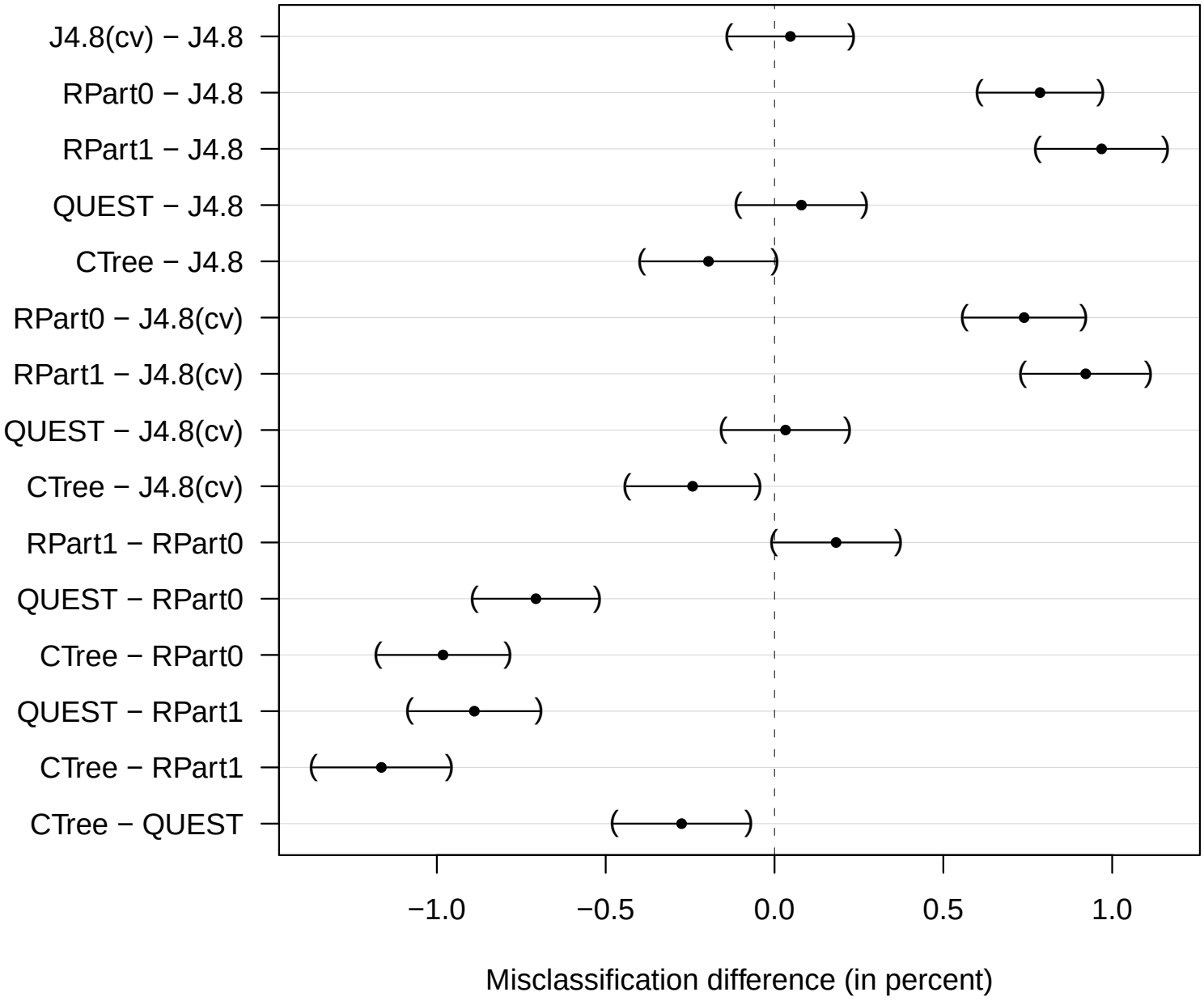
Individual results: Pima Indian diabetes



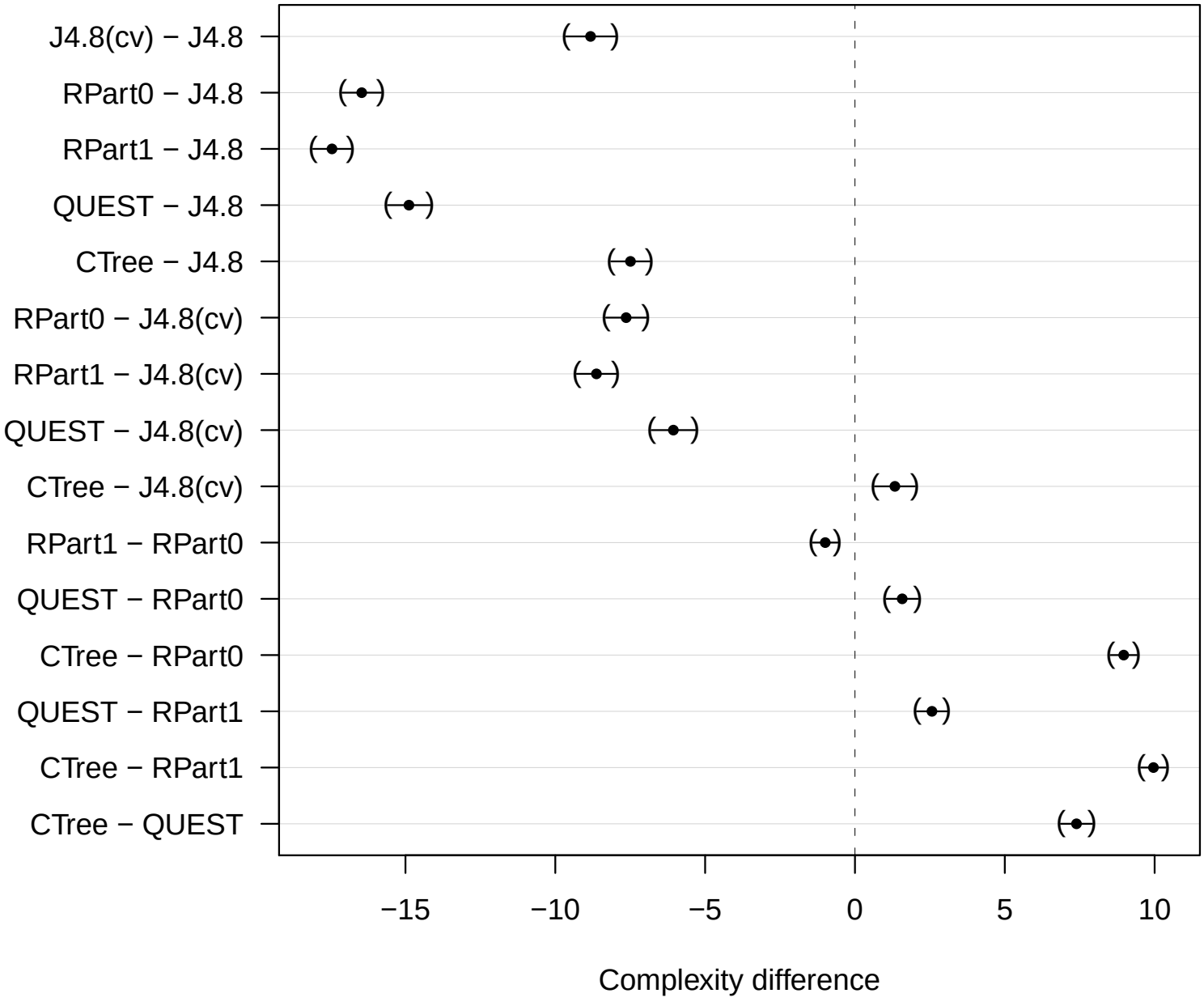
Individual results: Pima Indian diabetes



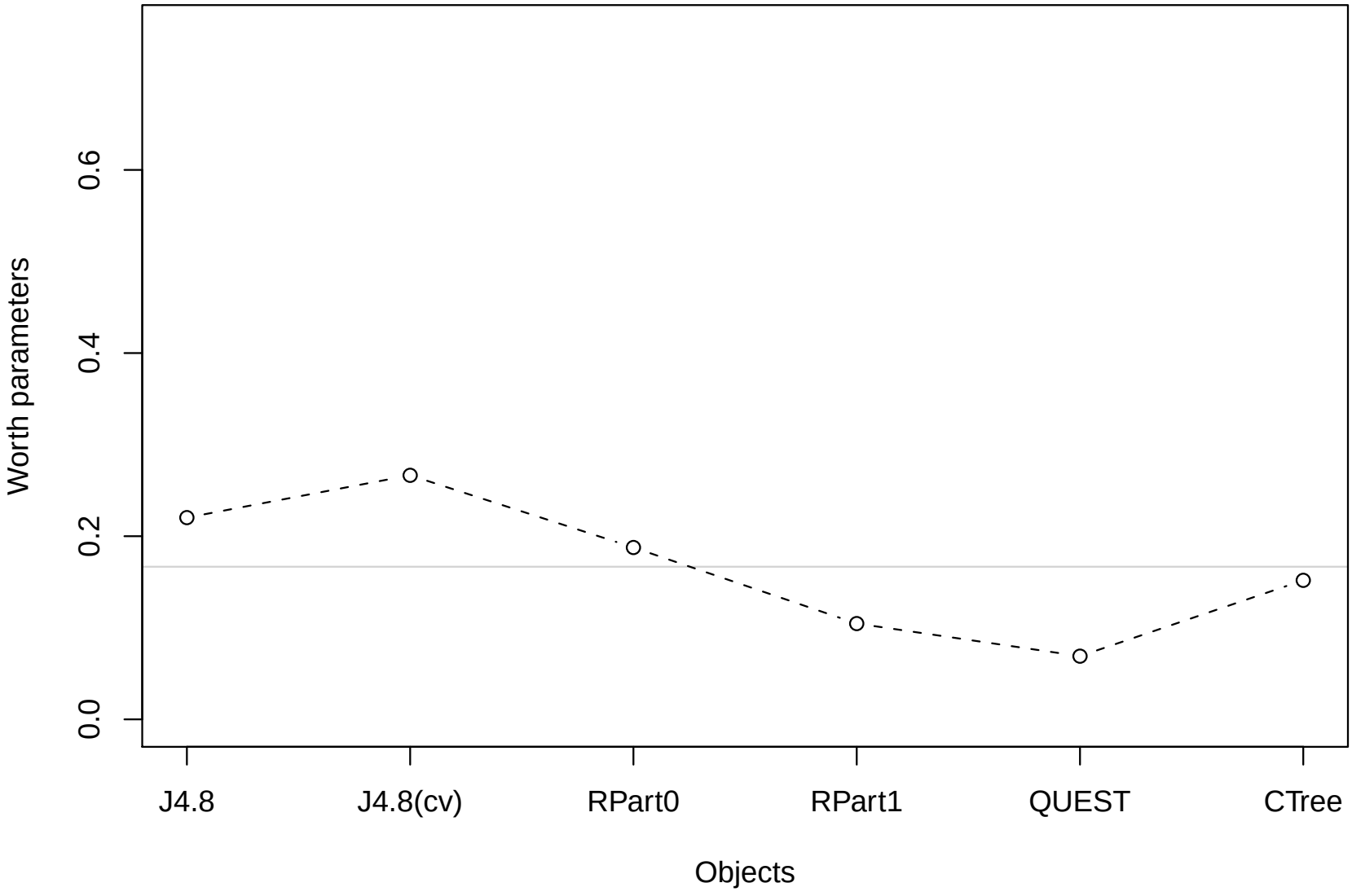
Individual results: Breast cancer



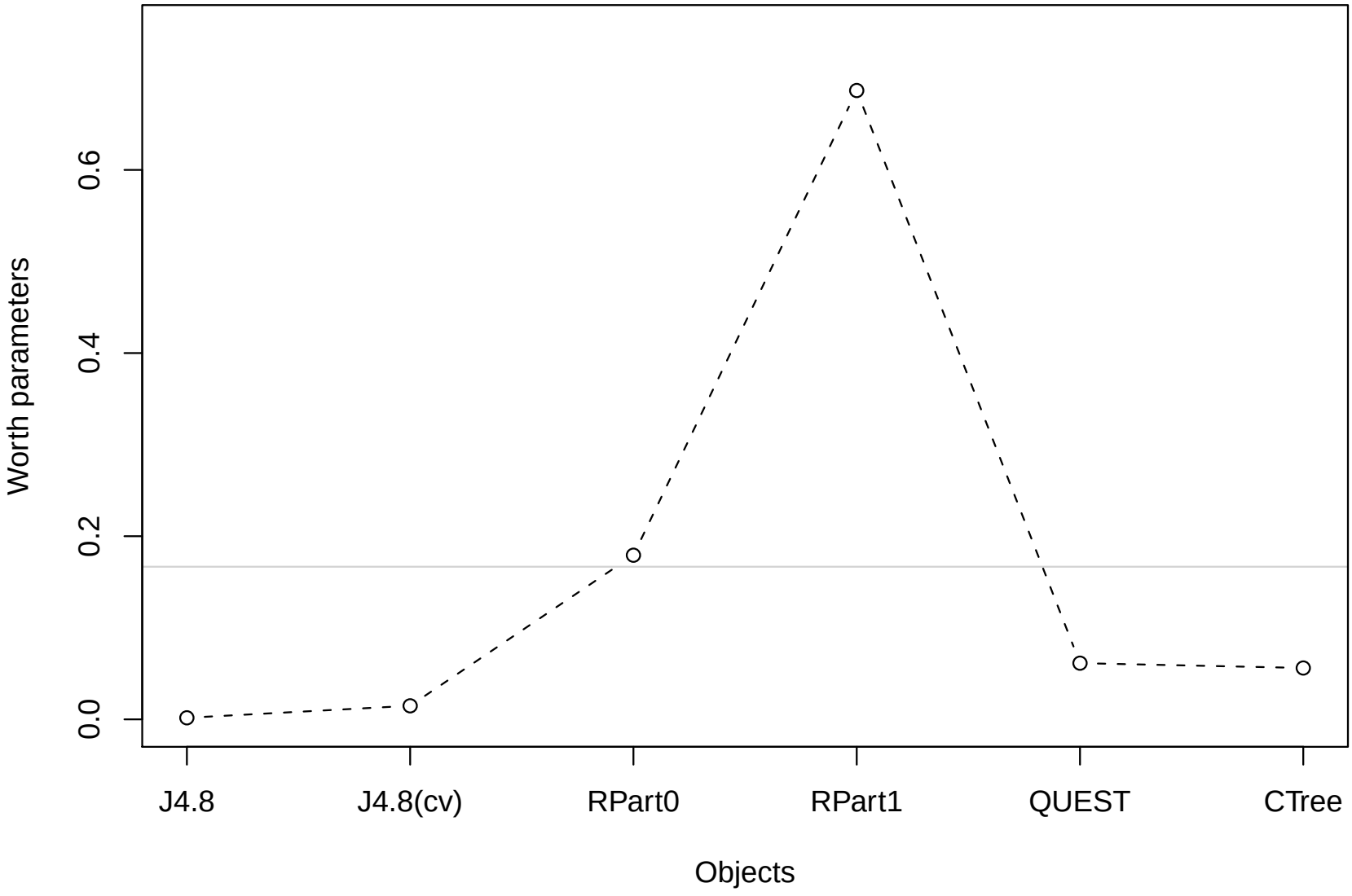
Individual results: Breast cancer



Aggregated results: Misclassification



Aggregated results: Complexity



Summary

- No clear preference between CART/RPart and C4.5/J4.8.
- Other tree algorithms perform similarly well.
- Cross-validated trees perform better than their counterparts.
- 1-standard error rule does not seem to be supported.

And now for something different:

- Before: Pairwise comparisons *of* tree algorithms.
- Now: Tree algorithm *for* pairwise comparison data.

Model-based recursive partitioning

Generic algorithm:

1. Fit parametric model for Y.
2. Assess stability of the model parameters over each splitting variable Z_j.
3. Split sample along the Z_{j^*} with strongest association: Choose breakpoint with highest improvement of the model fit.
4. Repeat steps 1–3 recursively in the subsamples until no more significant instabilities.

Application: Use Bradley-Terry models in step 1.

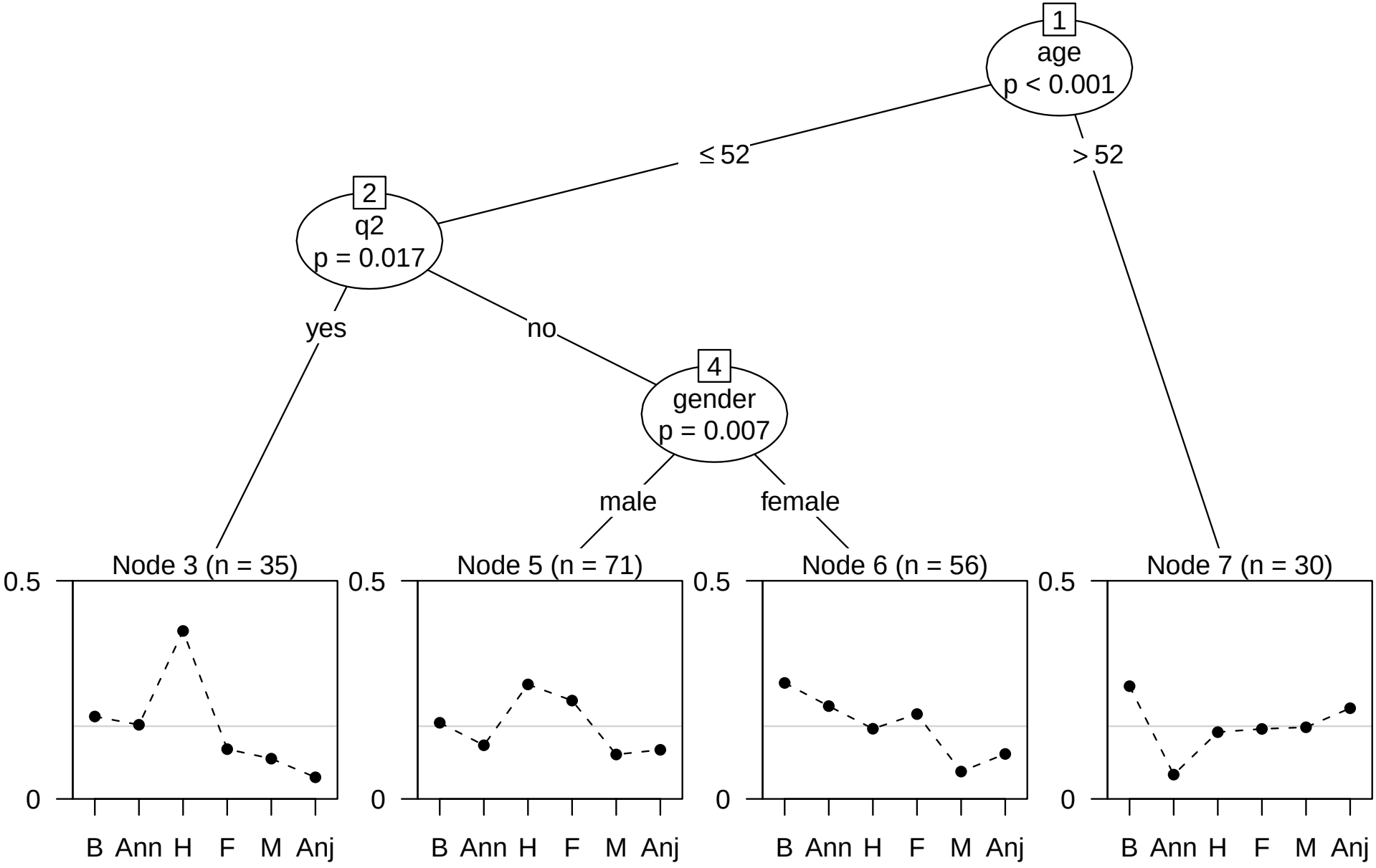
Implementation: psychotree on R-Forge.

Germany's Next Topmodel

- Study at Department of Psychology, Universität Tübingen.
- 192 subjects rated the attractiveness of candidates in 2nd season of Germany's Next Topmodel.
- 6 finalists: Barbara Meier, Anni Wendler, Hana Nitsche, Fiona Erdmann, Mandy Graff and Anja Platzer.
- Pairwise comparison (with forced choice).
- Subject covariates: Gender, age, questions about interest in the show.

Germany’s Next Topmodel

Germany's Next Topmodel



References

Hothorn T, Leisch F, Zeileis A, Hornik K (2005). "The Design and Analysis of Benchmark Experiments." *Journal of Computational and Graphical Statistics*, **14**(3), 675–699. doi:10.1198/106186005X59630

Schauerhuber M, Zeileis A, Meyer D (2008). "Benchmarking Open-Source Tree Learners in R/**RWeka**." In C Preisach, H Burkhardt, L Schmidt-Thieme, R Decker (eds.), *Data Analysis, Machine Learning and Applications (Proceedings of the 31st Annual Conference of the Gesellschaft für Klassifikation e.V., Albert-Ludwigs-Universität Freiburg, March 7–9, 2007)*. pp. 389–396.

Hornik K, Buchta C, Zeileis A (2009). "Open-Source Machine Learning: R Meets **Weka**." *Computational Statistics*, **24**(2), 225–232. doi:10.1007/s00180-008-0119-7

Strobl C, Wickelmaier F, Zeileis A (2009). "Accounting for Individual Differences in Bradley-Terry Models by Means of Recursive Partitioning." *Technical Report 54*, Department of Statistics, Ludwig-Maximilians-Universität München. URL http://epub.ub.uni-muenchen.de/10588/