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

Achim Zeileis[1,*]

1. Department of Statistics and Mathematics, Wirtschaftsuniversität Wien, Austria

* Contact author: Achim.Zeileis@R-project.org

Keywords: Benchmark, C4.5, CART, RWeka, Bradley-Terry model.

The two most popular classification tree algorithms in machine learning and statistics—C4.5 and CART—are compared in a benchmark experiment together with two other more recent constant-fit tree learners from the statistics literature (QUEST, conditional inference trees). The study assesses both misclassification error and model complexity on bootstrap replications of 18 different benchmark datasets (see Schauerhuber *et al.*, 2008, for details), employing the benchmarking framework of Hothorn *et al.* (2005). The study is carried out in the R system for statistical computing, made possible by means of the **RWeka** package (Hornik *et al.*, 2009) which interfaces R to the open-source machine learning toolbox **Weka**. Both algorithms are found to be competitive in terms of misclassification error—with the performance difference clearly varying across data sets. However, C4.5 tends to grow larger and thus more complex trees.

As Bradley-Terry models are used to aggregate the *paired comparisons of the tree algorithms* across the 18 different data sets, an outlook is provided for a *tree algorithm for paired comparison data* as recently introduced by Strobl *et al.* (2009).

References

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.

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, Hornik K (2008). "Benchmarking Open-Source Tree Learners in R/**RWeka**." In *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.

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/.