Variable Selection Bias in Classification Trees and Ensemble Methods

Carolin Strobl, Achim Zeileis,

Anne-Laure Boulesteix and Torsten Hothorn

Carolin.Strobl@stat.uni-muenchen.de

Standard classification tree algorithms, such as CART (Breiman, Friedman, Olshen, and Stone, 1984) and C4.5 (Quinlan, 1993), are known to be biased in variable selection, e.g. when potential predictor variables vary in their number of categories. The variable selection bias evident for predictor variables with different numbers of categories in binary splitting algorithms is due to a multiple testing effect: When potential predictor variables vary in their number of categories, and thus in their number of potential cutpoints, those variables that provide more potential cutpoints are more likely to be selected by chance. This effect can be demonstrated for the rpart routine, which is an implementation of the CART algorithm in R.

Ensemble methods have been introduced to increase the prediction accuracy of weak base learners such as classification trees. However, when biased classification trees are employed as base learners in ensemble methods variable selection bias is carried forward. Simulation results are presented that show variable selection bias for the gbm routine for boosting and for the randomForests routine. Both ensemble methods provide variable importance measures for variable selection purposes that are biased when potential predictor variables vary in their number of categories:

Unsurprisingly, variable importance measures that are based on the individual trees' biased impurity measures are again biased. But also variable importance measures based on the decrease of prediction accuracy after permutation (Breiman, 1998) are biased. This bias can partially be explained by the fact that variables that are preferred in the biased individual trees acquire more influential positions close to the root node and have more effect on the prediction accuracy.

Variable selection bias in individual classification trees can be eliminated by using split selection criteria that account for multiple testing and sample size effects (Strobl, Boulesteix, and Augustin, 2005; Hothorn, Hornik, and Zeileis, 2006). However, empirical experiments suggest that certain resampling schemes, for example the bootstrap used in several random-forest-like ensemble methods, may itself induce preferences towards variables with many splits, even when unbiased classification trees are used as base learners.

We give an overview over sources of variable selection bias in individual classification trees and its effects on variable importance measures in ensemble methods based on classification trees. The underlying mechanisms are illustrated by means of simulation

studies.

References

Breiman, L. (1998). Arcing classifiers. *The Annals of Statistics 26(3)*, 801–849.

Breiman, L., J. Friedman, R. Olshen, and C. Stone (1984). *Classification and Regression Trees*. New York: Chapman and Hall.

Hothorn, T., K. Hornik, and A. Zeileis (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics (to appear)*.

Quinlan, R. (1993). *C4.5: Programms for Machine Learning*. San Francisco: Morgan Kaufmann Publishers Inc.

Strobl, C., A.-L. Boulesteix, and T. Augustin (2005). Unbiased split selection for classification trees based on the Gini Index. *SFB-Discussion Paper 464, Department of Statistics, University of Munich LMU*.