Model-Based Boosting: Unbiased Variable Selection and Model Choice

Benjamin Hofner
Institut für Medizininformatik, Biometrie und Epidemiologie;
Friedrich-Alexander-Universität Erlangen-Nürnberg
Email: benjamin.hofner@imbe.med.uni-erlangen.de

Vortrag im Rahmen des Institutskolloquiums
10. Juni 2009
Institut für Statistik, LMU München

Variable selection and model choice are of major concern in many applications, especially in high-dimensional settings. Boosting (for an overview see Bühlmann and Hothorn (2007)) is a useful method for model fitting with intrinsic variable selection and model choice. However, a central problem remains: Variable selection is biased if the covariates are of very different nature. An important example is given by models that try to make use of continuous and categorical covariates at the same time. Especially if the number of categories increases, categorical covariates offer an increased flexibility and thus are preferred over continuous covariates (with linear effects). A closely related problem is model choice, where one tries to choose between different modeling alternatives for one covariate. The choice between linear or smooth effects is a classical example. The two competitors have different degrees of freedom (1 df for the linear effect and considerably more than 1 df for the smooth effect). Hence, smooth effects are preferably selected.

To make categorical covariates comparable to linear effects in the boosting framework one could use ridge penalized base-learners (i.e, modeling components) with 1 df in this case. To overcome the problem of different degrees of freedom of, e.g., linear and smooth effects Kneib et al. (2008) proposed a model choice scheme, which utilizes a decomposition of P-spline base-learners. An empirical evaluation of both approaches making use of the R package **mboost** (Hothorn et al., 2009).

References

Bühlmann, P. and Hothorn, T. (2007). Boosting algorithms: Regularization, prediction and model fitting, *Statistical Science* **22**: 477–505.

Hothorn, T., Bühlmann, P., Kneib, T., Schmid, M. and Hofner, B. (2009). *mboost: Model-Based Boosting*. R package version 1.1-1.
URL: *http://cran.R-project.org/web/packages/mboost*

Kneib, T., Hothorn, T. and Tutz, G. (2008). Variable selection and model choice in geoadditive regression models, *Biometrics* . (accepted).