Prognoseintervalle und Quantils-Boosting - Eine Simulationsstudie

Andreas Mayr, Nora Fenske, Torsten Hothorn

Andreas Mayr
Institut für Medizininformatik, Biometrie und Epidemiologie,
Friedrich-Alexander-Universität Erlangen-Nürnberg

Vortrag im Rahmen des Institutskolloquiums
30. Juni 2010, 16:15 Uhr
Seminarraum, Ludwigstraße 33 I

Prediction intervals are a useful tool to express uncertainty in the prediction of future or unobserved realizations of the response variable in a regression setting. Standard approaches typically assume an underlying distribution function and use the variance of the estimation method to compute boundaries around the expected mean.

Meinshausen (2006) suggested to use quantile regression forests to construct nonparametric prediction intervals for new observations. He adopted this generalization of random forests to estimate not only the conditional mean but the full conditional distribution of a response variable and, therefore, also conditional quantiles. Compared with classical methods, the resulting intervals have the advantage that they do not depend on distributional assumptions and are computable for high-dimensional data sets.

In this talk, we present an adaptation of gradient boosting algorithms to compute intervals based on additive quantile regression (Fenske et al., 2009), as available in the R package mboost (Hothorn et al., 2010). The boundaries of prediction intervals are modeled by applying nonparametric quantile regression with linear as well as smooth effects, fitted by componentwise boosting providing intrinsic variable selection and model choice. The main advantage of this highly flexible approach is that it allows to quantify and to interpret the influence of single covariates on the response and on the prediction accuracy.

We found that the correct interpretation of prediction intervals involves the risk of running into a severe pitfall in practice since only the conditional view based on fixed predictor variables is adequate to prove the correct coverage of the proposed intervals. Hence, we analyze simulated data sets to evaluate the accuracy of our methods and show a real-life example to emphasize their practical relevance.

Keywords: Prediction Inference, Boosting, Random Forests, Quantile Regression

References

Fenske, N., Kneib, T. and Hothorn, T. (2009). Identifying risk factors for severe childhood malnutrition by boosting additive quantile regression, *Technical Report, Department of Statistics, University of Munich* **52.**

Hothorn, T., Buehlmann, P., Kneib, T., Schmid, M. and Hofner, B. (2010). *Model-Based Boosting.* R package version 2.0-4.

Meinshausen, N. (2006). Quantile regression forests, *Journal of Machine Learning Research* **7:** 983–999.