Why and how to use random forest variable importance measures (and how you shouldn't)

Carolin Strobl
Ludwig-Maximilians-Universität
München

Achim Zeileis
Wirtschaftsuniversität
Wien

Abstract

Random forests are becoming increasingly popular in many scientific fields, especially in genetics and bioinformatics, for assessing the importance of predictor variables in high dimensional settings. Advantages of random forests in these areas are that they can cope with "small n large p" problems, complex interactions and even highly correlated predictor variables. The talk gives a short introduction to the rationale of random forests and the their variable importance measures as well as the two random forest implementations offered in the R system for statistical computing: randomForest in the package of the same name by Breiman *et al.* (2006) and cforest in the package party by Hothorn *et al.* (2008). Moreover, recent research issues are addressed:

- Solutions are presented for bias in random forest variable importance measures towards, e.g., predictor variables with many categories (Strobl, Boulesteix, Zeileis, and Hothorn 2007) and correlated predictor variables (Archer and Kimes 2008).
- Currently suggested tests for random forest variable importance measures (Breiman and Cutler 2008; Rodenburg *et al.* 2008) are critically discussed in an outlook.

Keywords: Feature selection, screening, variable importance, Gini index, CART, bagging.

References

Archer KJ, Kimes RV (2008). "Empirical characterization of random forest variable importance measures." *Computational Statistics & Data Analysis*, **52**(4), 2249–2260.

Breiman L, Cutler A (2008). "Random Forests – Classification Manual (website accessed in 1/2008)." http://www.math.usu.edu/~adele/forests/.

Breiman L, Cutler A, Liaw A, Wiener M (2006). *Breiman and Cutler's Random Forests for Classification and Regression*. R package version 4.5-16, URL http://CRAN.R-project.org/package=randomForest.

Hothorn T, Hornik K, Zeileis A (2008). "party: A Laboratory for Recursive Part(y)itioning." R package version 0.9-96, URL http://CRAN.R-project.org/package=party.

Rodenburg W, Heidema AG, Boer JM, Bovee-Oudenhoven IM, Feskens EJ, Mariman EC, Keijer J (2008). "A Framework to Identify Physiological Responses in Microarray Based Gene Expression Studies: Selection and Interpretation of Biologically Relevant Genes." *Physiological Genomics*, **33**(1), 78–90.

Strobl C, Boulesteix AL, Zeileis A, Hothorn T (2007). "Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution." *BMC Bioinformatics*, **8:25**.