Party on! – A new, conditional variable importance measure for random forests available in `party`

Carolin Strobl[1,*] **, Achim Zeileis**[2]

1. Department of Statistics, Ludwig-Maximilians-Universität München
2. Department of Statistics and Mathematics, Wirtschaftsuniversität Wien

* Contact author: carolin.strobl@stat.uni-muenchen.de

Keywords: Permutation importance, variable selection, spurious correlation.

Random forests have become very popular in many scientific fields because they can cope with "small n large p" problems involving complex interactions. Random forest variable importance measures have been suggested as screening tools, e.g., for gene expression studies. However, these variable importance measures have been shown to be biased in favor of predictor variables of certain types and towards correlated predictor variables.

While the former issue could be addressed straightforwardly in `party` by means of unbiased split selection and resampling schemes (Strobl et al., 2007), in the case of correlated predictors the original permutation importance is highly misleading, creating a new source of bias in interpretations drawn from random forests. Therefore, Strobl et al. (2008) recently suggested a solution for this problem in the form of a new, conditional permutation importance measure. Starting from version 0.9-994, this new measure is available in the `party` package.

In the talk, the rationale and application of this new measure is outlined and illustrated by means of a toy example. Moreover, some hands-on advice is given for sensibly using and interpreting random forests in R.

References

C. Strobl, A.-L. Boulesteix, A. Zeileis and T. Hothorn (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8:25.

C. Strobl, A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, 9:307.

T. Hothorn, K. Hornik, C. Strobl and A. Zeileis (2009). `party`: A Laboratory for Recursive Partytioning. `http://CARN.R-project.org/package=party`.

C. Strobl, T. Hothorn and A. Zeileis (2009). Party on! A new, conditional variable importance measure for random forests available in the `party` package. *Technical report (submitted).* `http://epub.ub.uni-muenchen.de/9387/1/techreport.pdf`.