LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN INSTITUT FÜR STATISTIK

# Deriving interpretable thresholds for variable importance in random forests by permutation

Maria Blanco, Tim Müller, Laura Schlieker, Armin Ott, Hannes Buchner

(Staburo GmbH, Munich)

10.05.2023, 16.00 (c.t.)

Department of Statistics, Ludwigstr. 33, Room 144
and online via Zoom (Link)
(Meeting-ID: 913-2473-4411; Password: StatsCol22)

In the context of clinical research and in particular precision medicine the identification of predictive or prognostic biomarkers is of utmost importance. Especially when dealing with high-dimensional data discriminating between informative and uninformative variables plays a crucial role. Machine Learning approaches, and especially Random Forests, are promising approaches in this situation as the Variable Importance (VIMP) of a Random Forest can serve as a decision guidance for the identification of potentially relevant variables.

Many different approaches for Random Forest VIMP have been proposed and evaluated (e.g., Speiser et al. 2019). One of the algorithms is the well-performing Boruta method (Kursa and Rudnicki 2010), which adds permutated - and thus uninformative - versions of each variable (so-called shadow variables) to the set of predictors. Based on that, it classifies the covariables into three possible importance categories: confirmed, tentative or rejected.

We propose a variation of the Boruta method and evaluate the relevance of the variables based on different criteria. Our method is independent of the simulations runs and compares the VIMP of each covariate directly with its permuted version. In addition, the uninformative versions are generated by permutating the rows of the dataset, which preserves the relationship between the original variables. For evaluating the importance of the features, we use different criteria, e.g., descriptive statistics of the shadow VIMPs and controlling for the FWER or for the FDR, given a significance level.

We examine our method with simulations similar to Degenhardt et al. 2019 and compare its results to the Boruta algorithm. Furthermore, we apply it on public available datasets of varying sizes to illustrate its real-life behaviour.

In our approach, the user is guided by several criteria summarized in one visual presentation and has therefore the flexibility to be more conservative by picking all important covariates or more permissive by taking only the most important ones, depending on the nature of their problem.

**References:**

[1] Speiser, J. L., Miller, M. E., Tooze, J., & Ip, E. (2019). A Comparison of Random Forest Variable Selection Methods for Classification Prediction Modeling. Expert systems with applications, 134, 93–101.

[2] Kursa, M. B., & Rudnicki, W. R. (2010). Feature Selection with the Boruta Package. Journal of Statistical Software, 36(11), 1–13.

[3] Degenhardt, F., Seifert, S., & Szymczak, S. (2019). Evaluation of variable selection methods for random forests and omics data sets. Briefings in bioinformatics, 20(2), 492–503.

**About Staburo and the authors:**
Staburo GmbH is a data science company, specialized in statistical consulting, programming and bioinformatics for healthcare projects. Our core competencies include Clinical Statistics, Translational Medicine & Biomarkers, Phase I & Pharmacokinetics/-Dynamics, Data Transparency & Disclosure Services, Health Technology Assessment and Bioinformatics. Our customers are international pharmaceutical companies (7 of the top 20), CROs, biotech companies and medical device manufacturers.
Hannes Buchner studied Statistics and received his PhD at LMU, he has co-founded Staburo and currently works as Managing Director and Head of Biostatistics and Data Science.
Laura Schlieker, Director Biostatistics with Focus Area Biomarkers, received her master's degree in Statistics with focus on biometry from the Technische Universität Dortmund.
Maria Blanco and Tim Müller are working students at Staburo who completed their bachelor's degree in Statistics at University of São Paulo and LMU, respectively. Both are pursuing their master's degree in Statistics and Data Science with specialization in biostatistics at LMU, and Tim is currently writing his master thesis at Staburo.
Armin Ott is also a LMU alumnus, where he has received his bachelor's and master's degree in Statistics. Currently, Armin works as Data Scientist and Biostatistician at Roche.