Please join the Biometric Colloquium

TIM MÜLLER UND HANNES BUCHNER

Staburo GmbH Munich

PERMUTATION-BASED MULTIPLE TESTING-CONTROLLED VARIABLE SELECTION USING RANDOM FORESTS

February 19th, 2025 at 9:00am

Seminarraum Center for Medical Data Science (previously CeMSIIS),
Spitalgasse 23, Room 88.03.513
Medical University of Vienna, 1090 Wien
Host: Franz König

Abstract:

Identifying relevant biomarkers is critical in clinical research and precision medicine, particularly when analysing high-dimensional data. Random forests (RFs) are promising for such settings due to their flexibility, ease of use and their ability to handle datasets with more variables than samples. RFs assess the importance of each variable in predicting the outcome using variable importance (VIMP) scores. However, the lack of a known statistical distribution of VIMP scores prevents standard statistical testing and associated multiple testing adjustment for the purpose of variable selection. To address this, we propose a novel method for multiple testing-controlled variable selection. Our approach, similar to permutation testing, involves generating permuted counterparts for each variable and comparing their VIMPs across iterations to calculate p-values. However, unlike competing methods, we preserve the correlation structure between the covariates in the permutations to guard against biases. With promising results, our method is evaluated against three competing RF variable selection approaches in simulations that involve high- and low-dimensional data, as well as correlated and categorical variables. Moreover, we apply it to a real dataset to demonstrate its practical use. The method's results integrate seamlessly into standard VIMP plots, providing a flexible and transparent way to interpret results in a familiar format.