Unbiasing split selection in forests with missing values

Nicolas Prost\textsuperscript{1}

\textsuperscript{1}Parietal, INRIA
\textsuperscript{1}CMAP

7\textsuperscript{th} May 2018

Abstract

Random forests are a powerful nonparametric tool to deal with complex relationships in the covariates. Their wide scope should make them a good candidate to deal with missing values without requiring prior imputation. The standard implementation of random forests is based on the CART algorithm (R package ‘rpart’), which features a missing value procedure. Unfortunately, the CART exhaustive search cause the split choice to be biased towards variables which either have many different categories or many missing values.

Conditional inference forests have been suggested to solve this issue (R package ‘party’). The idea is to separate variable selection and threshold selection. For the first step, they use linear statistics of association between each variable and the response, then compare the p-values based on the distribution of these statistics, conditionally to permutations of the response, under the null hypothesis of independence. This algorithm does remove the selection bias, but it also has flaws that we will discuss, as we explore alternative approaches.