Diversity forests: Using split sampling to enable innovative complex split procedures in random forests

Roman Hornung

Institute for Medical Information Processing, **B**iometry and **E**pidemiology,
University of Munich

DAGStat Conference 2022, Hamburg

March 28 to April 1, 2022

Introduction

Diversity Forests

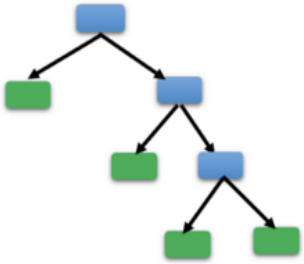
Roman Hornung

Introduction

Conventional split finding vs. the diversity forest algorithm

Empirical results

Examples of complex split procedures



- Conventional **random forests** (RFs) feature **strong predictive** performance. ✓
- However, when considering **complex split procedures** (e.g., multivariable splitting) the **split finding scheme** used in the tree construction is (much) **too expensive** computationally. ✗
- But **complex split procedures** can **solve practically important problems** using random forests (e.g., detection of interaction effects). 😐
- The **diversity forest** (DF) **algorithm** (Hornung, 2022) is a new split finding scheme that **allows** for using **complex split procedures** in random forests. 🙂 ✓

Conventional split finding vs. the diversity forest algorithm

- Conventional **candidate split set** sampling:
  For $l = 1, \ldots,$ *mtry*:
  1. Randomly select one of the **covariates**.
  2. Evaluate **all** possible binary **splits** in the sampled covariate.
- The **DF algorithm** (slightly simplified):
  For $l = 1, \ldots,$ *nsplits*:
  1. Randomly select one so-called **split problem**.
  2. Randomly select and evaluate **one or few splits** in the sampled split problem.
- The **structures** of the **split problems** depend on the split procedure used;
  examples: univariable, binary splitting: all binary splits in a covariate x_j, multivariable splitting: all splits that involve one or several of the covariates x_{j_1}, x_{j_2}, and x_{j_3}

Empirical results (obtained for univariable, binary splitting)

- In a large-scale comparison study using 220 datasets it was seen that the **DF** algorithm is associated with a very **similar** (slightly better) **predictive performance** as conventional **RFs**.

- In an analysis using 50 datasets the **predictive performance** of DFs was seen to be quite **insensitive** to the choice of ***nsplits***.

  $\Rightarrow$ The value of ***nsplits*** does not have to be optimized, but using a **fixed value** is **sufficient**.

Interaction Forests – first published DF method with a complex split procedure

Diversity Forests

Roman Hornung

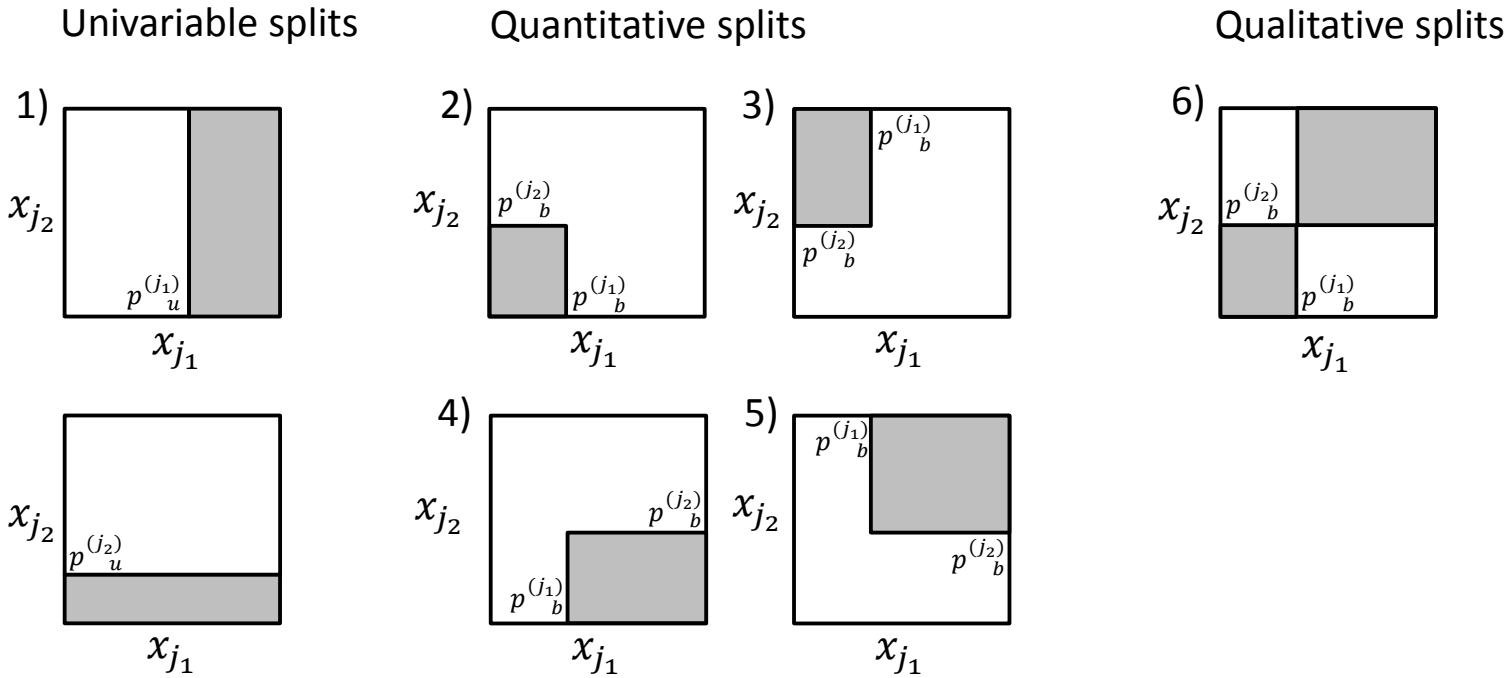
Introduction

Conventional split finding vs. the diversity forest algorithm

Empirical results

Examples of complex split procedures

- Interaction Forests (Hornung & Boulesteix, 2022) use **bivariable splitting** to model **interaction effects** (R package `diversityForest`)
- **Ranking of interactions** with respect to their **importance to prediction** via the **E**ffect **I**mportance **M**easure



Outlook: Further potential complex split procedures

- **Multi-way splitting**: Tackling K**-class outcomes** with K-way splitting. $\Rightarrow$ **Better variable importance** measure values for **covariates that differentiate well between all** K **classes** instead of only a subset; more flexible splits for two-class, continuous, or survival outcomes

- **Detecting predictive patterns** by generating **diffuse** (bivariable) **partitions** of the covariate space

References – **Thank you for your attention!**

Diversity Forests

Roman Hornung

Introduction

Conventional split finding vs. the diversity forest algorithm

Empirical results

Examples of complex split procedures

Breiman, L., 2001.

Random forests.

Machine Learning 45 (1), 5–32.

Hornung, R., 2022.

Diversity Forests: Using split sampling to enable innovative complex split procedures in random forests.

SN Computer Science 3, 1.

Hornung, R., Boulesteix, A.-L., 2022.

Interaction forests: Identifying and exploiting interpretable quantitative and qualitative interaction effects.

Computational Statistics & Data Analysis, 107460.