Multi forests: Variable importance for multi-class outcomes

Roman Hornung[1,2], Alexander Hapfelmeier[3]

CMStatistics 2024, London

December, 16th, 2024

[1] **I**nstitute for Medical Information Processing, **B**iometry and **E**pidemiology, LMU Munich, Munich, Germany

[2] Munich Center for Machine Learning (MCML), Munich, Germany

[3] Institute of AI and Informatics in Medicine, TUM School of Medicine and Health, Technical University of Munich, Munich, Germany

Introduction

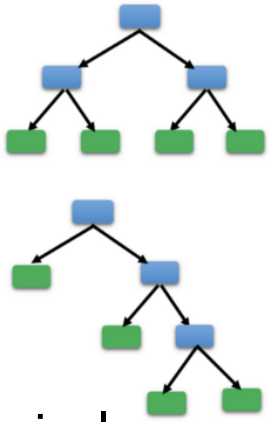
Multi forests
Roman Hornung
Introduction & background
MuF algorithm
MuF construction
Multi-class VIM and discriminatory VIM
Simulation study
Real data analysis
Further points

- In **multi-class prediction**, it can be important to **identify** covariates specifically associated with one or several individual classes, termed "**class-associated covariates**".
- The **influence** of these covariates is **well interpretable**; for instance, a low value might indicate class *a*, a moderate value class *b*, and a high value class *c*.
- Variable importance measures (**VIMs**) in random forests (**RFs**) **rank covariates by their impact** on the outcome.
- However, **conventional VIMs likely rank** both **class-associated** and **class-group-differentiating** covariates **similarly**.

Introduction

- This is because **conventional VIMs** like permutation VIM or (corrected) Gini importance **rank** covariates **based on overall performance**, without class differentiation.
- We introduce the **multi-class VIM**, **tailored** to identify **class-associated covariates**.
- It is **computed based on** the RF variant multi forests (**MuFs**), which perform **multi-way splits** into C **class-specific child nodes** (C: **# outcome classes**).
- Additionally, the **discriminatory VIM** is designed to **rank all influential covariates highly**, similar to conventional VIMs, without focussing on class association.

Background: From conventional random forests to multi forests



- RFs are **ensembles of decision trees**, employing **randomization** to enhance tree diversity and in consequence improve ensemble predictive performance.

- In **conventional RFs**, nodes are **split binarily** and recursively based on **covariates**:

  if covariate $x \leq const$, assign to child node 1; if $x > const$, to child node 2.

- **MuFs** have the same structure as RFs but employ **multi-way splitting**:

  if covariate $x \in]l_{c-1}, l_c]$, assign to child node c, with thresholds $l_0 := -\infty < l_1 < \cdots < l_{C-1} < l_C := \infty$.

Multi forests: Sketch of the split selection procedure

Multi forests
Roman Hornung
Introduction & background
MuF algorithm
MuF construction
Multi-class VIM and discriminatory VIM
Simulation study
Real data analysis
Further points

1. Select ***mtry* covariates** at random; **for each**, sample ***npervar* potential multi-way splits**.

2. **Randomly decide** whether the split is to be **multi-way or binary**.

3. **For each** selected covariate:

   If the split is to be **multi-way**:

   Evaluate **each multi-way split** using a split criterion **measuring how well** the **classes** are represented in their **respective child nodes**.

   If the split is to be **binary**:

   Evaluate **each split point** within the multi-way splits using a split criterion **measuring** the **purity** of the resulting **child nodes**.

4. **Select the split** that maximizes the respective **split criterion**.

Multi Forests: Calculation of the multi-way split criterion

1. The **classes** $1, \ldots, C$ are **assigned to** the **child nodes** $\mathcal{I}_1^{as}, \ldots, \mathcal{I}_C^{as}$ by selecting the **permutation that maximizes**:

$$\sum_{c=1}^{C} p_{c,\mathcal{I}_c^{as}}^2,$$

where $p_{c,\mathcal{I}_c^{as}}$ is the proportion of class c in child node $\mathcal{I}_c^{as}$.

2. The **final split criterion** is calculated as:

$$\sum_{c=1}^{C} p_{c,\mathcal{I}_c^{as}}^2 \frac{n_{l,\mathcal{I}_c^{as}}}{n_l},$$

where $n_{l,\mathcal{I}_c^{as}}$ is the count of observations in child node $\mathcal{I}_c^{as}$, and n_l is the total in the current node.

Multi-class VIM and discriminatory VIM

Multi forests
Roman Hornung
Introduction & background
MuF algorithm
MuF construction
Multi-class VIM and discriminatory VIM
Simulation study
Real data analysis
Further points

- The **multi-class VIM** value for a covariate s is the **average of tree-specific contributions**.
- These contributions are sums of the **differences in split criterion values** calculated **before and after permuting** s within nodes that use s for **multi-way splits**.
- Only the **out-of-bag observations** (not used for tree construction) are used here, **avoiding biases** similar to those seen with the Gini importance.
- The **discriminatory VIM** is computed **very similarly** to the multi-class VIM, but considers the **binary splits** instead of the multi-way splits.

Simulation study comparing multi-class VIM and discriminatory VIM with conventional VIMs: Study design

Multi forests

Roman Hornung

Introduction & background

MuF algorithm

MuF construction

Multi-class VIM and discriminatory VIM

Simulation study

Real data analysis

Further points

- **Compared methods**: multi-class VIM, discriminatory VIM, permutation VIM, corrected Gini importance (Nembrini et al., 2018)
- **Versions of multi-class/discriminatory VIM** evaluated, based on the following variations in the MuF algorithm:
  - **Squaring or not squaring proportions** $p_{c,\mathcal{I}_c^{as}}$ in the multi-way split criterion
  - **Split criterion for binary splits**: conventional Gini versus criterion similar to multi-way split criterion
- **Simulation design**:
  - Types of covariates: class-associated, distinguishing between class groups, noise covariates
  - Gaussian covariates, $C = 4, 6, 10$, varying sample sizes
  - AUC-based evaluation to compare discriminatory abilities of VIMs across covariate types

Simulation study: Results

Multi forests
Roman Hornung
Introduction & background
MuF algorithm
MuF construction
Multi-class VIM and discriminatory VIM
Simulation study
Real data analysis
Further points

- **Multi-class VIM** typically **ranked** the **class-associated** covariates **highest**. ✓
- The **conventional VIMs** ranked informative but **non-class-specific** covariates **similarly high or higher**. ✗
- **Best results** were achieved using **squared proportions** $p_{c,\mathcal{I}_c^{as}}$ and the **Gini criterion**; squaring proportions was more impactful.
- **Discriminatory VIM** rankings **mirrored** closely those of the **conventional VIMs**.
- **Comparisons between multi-class VIM** and **discriminatory VIM** rankings can provide **additional insights**, but comparing their absolute values requires **caution**.

Real data analysis comparing predictive performance of MuFs and RFs: Study design

Multi forests
Roman Hornung
Introduction & background
MuF algorithm
MuF construction
Multi-class VIM and discriminatory VIM
Simulation study
Real data analysis
Further points

- **MuFs** were **primarily** designed **to realize the multi-class VIM**.
- Nevertheless, we evaluated their predictive performance, as **users are likely** to **employ MuFs for prediction** as well.
- **Study design**:
  - **Compared methods**: four versions of the MuF algorithm (as in the simulation study) and conventional RFs
  - **Datasets**: 121 open-source multi-class datasets with $3 \leq C \leq 20$
  - **Performance metrics**: Brier score, accuracy, two multi-class versions of the AUC (AUNU and AUNP)
  - **Validation**: stratified 5-fold cross-validation, repeated five times

Real data analysis: Results

Multi forests

Roman Hornung

Introduction & background

MuF algorithm

MuF construction

Multi-class VIM and discriminatory VIM

Simulation study

Real data analysis

Further points

- **RFs most often ranked first** but also **frequently** took the **last** position.
- **Performance differences** between RFs and MuF versions were **generally minor**, where:
  - MuFs tended to perform **worse for large numbers of outcome classes**.
  - Differences were **more pronounced for the Brier score** than for the other metrics.
- The MuF version using **squared proportions** $p_{c,\mathcal{I}_c^{as}}$ and the **Gini criterion again performed best**, hence it is the **recommended version**.

Further points

Multi forests
Roman Hornung
Introduction & background
MuF algorithm
MuF construction
Multi-class VIM and discriminatory VIM
Simulation study
Real data analysis
Further points

- MuF algorithm **implemented in** the R package '**diversityForest**' using C++ code (closely based on 'ranger' package (Wright and Ziegler, 2017))
- **Categorical covariates** can be included, but the **multi-class VIM** can **only** be computed for **covariates with at least *C* categories**.
- **MuFs use** the **diversity forest algorithm** (Hornung, 2020), which allows for complex split procedures.

References

Multi forests

Roman Hornung

Introduction & background

MuF algorithm

MuF construction

Multi-class VIM and discriminatory VIM

Simulation study

Real data analysis

Further points

Hornung, R., Hapfelmeier, A., 2024.
Multi forests: Variable importance for multi-class outcomes.
arXiv Preprint, arXiv:2409.08925.

Breiman, L., 2001.
Random forests.
Mach Learn, 45:5–32.

Coppersmith, D., Hong, S. J., Hosking, J. R., 1999.
Partitioning nominal attributes in decision trees.
Data Min Knowl Discov, 3:197–217.

Hornung, R., 2022.
Diversity forests: Using split sampling to enable innovative complex split procedures in random forests.
SN Comput Sci, 3:1–16.

Kuhn, H. W., 1955.
The Hungarian method for the assignment problem.
Nav Res Logist Q, 2(1-2):83–97.

Nembrini, S., König, I. R., Wright, M. N., 2018.
The revival of the Gini importance?
Bioinformatics, 34 (21):3711–3718.

Wright, M. N., Ziegler, A., 2017.
ranger: A fast implementation of random forests for high dimensional data in C++ and R.
J Stat Softw, 77:1–17.

Thank you for your attention!

Multi forests

Roman Hornung

Introduction & background

MuF algorithm

MuF construction

Multi-class VIM and discriminatory VIM

Simulation study

Real data analysis

Further points

Recommended reading

Hornung, R., Hapfelmeier, A., 2024.
Multi forests: Variable importance for multi-class outcomes.
arXiv:2409.08925.