Interpretation of black box models using tree-based surrogate models

- Disputation -

Sofia Loibl

March 31$^{\text{th}}$, 2023

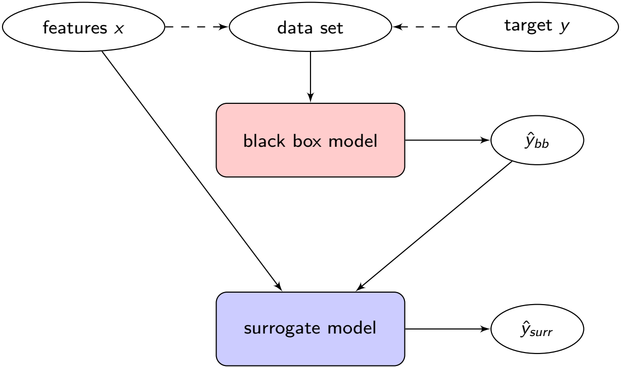
Department of Statistics
Ludwig-Maximilians-Universität München

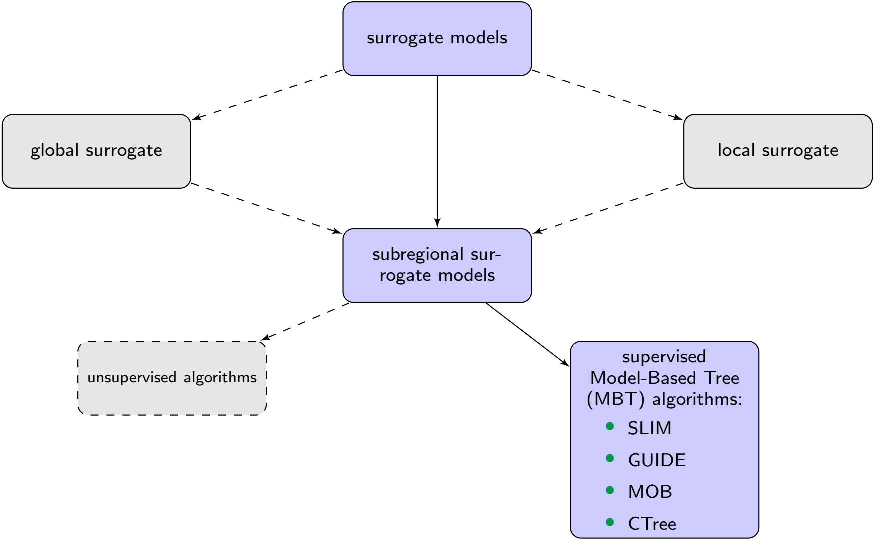
Supervised by Dr. Giuseppe Casalicchio

Table of contents

1. Introduction
2. Model-based trees
3. Selection bias
4. Performance, Interpretability and Stability
5. Conclusion

Surrogate models

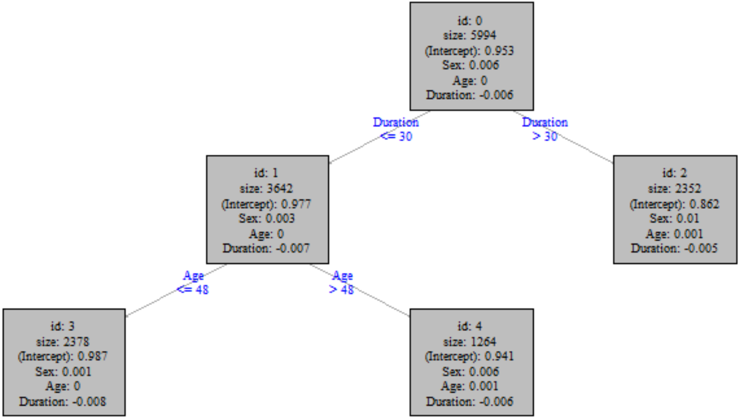



surrogate models
global surrogate
local surrogate
subregional surrogate models
unsupervised algorithms
supervised Model-Based Tree (MBT) algorithms:
SLIM
GUIDE
MOB
CTree

Goals

- Comparison of four different algorithms for the generation of MBTs with regard to:
  - Selection bias
  - Performance
  - Interpretability
  - Stability
- Comparison of interpretability and performance for different model classes fitted in the subregions
- Investigate the suitability of MBTs as surrogate model

Model-based trees



Requirements

Requirements for MBTs in this thesis:

- split by interactions
- main effect models in the subregions (nodes)
- flexible choice of main effect models (i.e. different objectives, regularized Regression, GAMs, ...)
- use features as potential splitting and regressor variables

Comparison of the algorithms

	SLIM	MOB	CTree	GUIDE
Split point selection	exhaustive search	two-step	two-step	two-step
Test	-	score-based M-fluctuation test	score-based permutation test	residual-based χ^2 test
Flexibility	high	low	low	high
Distinction between regressor and splitting variables	no	partly	partly	partly
Prepruning	improvement	alpha	alpha	improvement
Implementation	-	R package	R package	binary executable

Table: Comparison of MBT algorithms - Methodology

Selection bias

Selection bias - Independence

Definition unbiased for an independent target:
According to (Hothorn et al., 2006) an algorithm for recursive partitioning is called unbiased when, under the conditions of the null hypothesis of independence between a response y and feature $\mathbf{x}_1, ...\mathbf{x}_p$ the probability of selecting feature $\mathbf{x}_j$ is $1/p$ for all $j = 1, ..., p$ regardless of the measurement scales or number of missing values.

Problem: if an algorithm is biased in the case of independence, there is also a higher risk or bias if main effects of interactions are present

Simulation independence

	SLIM	MOB	CTree	GUIDE
Designation	biased	so called "unbiased"		
simulation numerical - numerical	biased	unbiased	unbiased	unbiased
simulation numerical - binary	biased	unbiased	unbiased	unbiased
simulation numerical - categorical	biased	biased	biased	biased

Table: Comparison of MBT algorithms - Selection bias independence

Simulation interactions - selection bias vs. splitting strategy

scenario	$\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_4$	$\mathbf{x}_3$	$f(\mathbf{x})$
numerical vs numerical	$[0, 1]$	$\{0, 0.1, \ldots, 0.9, 1\}$	$\mathbb{1}_{(\mathbf{x}_1 \leq mean(\mathbf{x}_1))}\mathbf{x}_2 + \mathbb{1}_{(\mathbf{x}_3 \leq mean(\mathbf{x}_3))}\mathbf{x}_4$
numerical vs binary	$[0, 1]$	$\{0, 1\}$	$\mathbb{1}_{(\mathbf{x}_1 \leq mean(\mathbf{x}_1))}\mathbf{x}_2 + \mathbb{1}_{(\mathbf{x}_3 = 0)}\mathbf{x}_4$

Table: scenarios selection bias interaction

Simulation interactions - selection bias vs. splitting strategy

scenario	$\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_4$	$\mathbf{x}_3$	$f(\mathbf{x})$
numerical vs numerical	$[0, 1]$	$\{0, 0.1, \ldots, 0.9, 1\}$	$\mathbb{1}_{(\mathbf{x}_1 \leq mean(\mathbf{x}_1))}\mathbf{x}_2 + \mathbb{1}_{(\mathbf{x}_3 \leq mean(\mathbf{x}_3))}\mathbf{x}_4$
numerical vs binary	$[0, 1]$	$\{0, 1\}$	$\mathbb{1}_{(\mathbf{x}_1 \leq mean(\mathbf{x}_1))}\mathbf{x}_2 + \mathbb{1}_{(\mathbf{x}_3 = 0)}\mathbf{x}_4$

Table: scenarios selection bias interaction

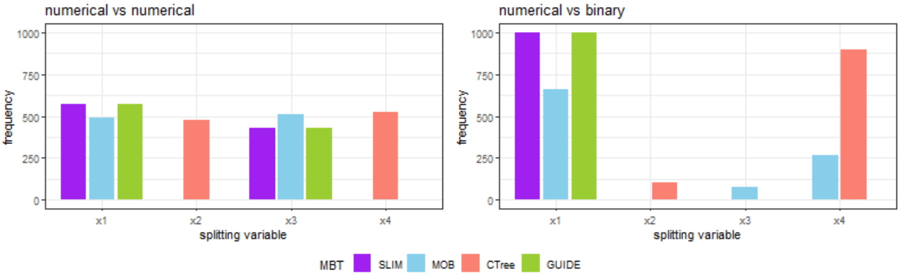


Figure: Simulated frequencies of first selected splitting features for the four interaction scenarios

Comparison of performance, interpretability and stability

Simulation - Comparison of Algorithms

How do the algorithms differ in terms of

- performance: R^2
- interpretability: number of leafnodes
- stability: variability of number of leafnodes and Rand Index

in different simulation scenarios?

Main difference between scenarios: smooth interactions vs. subgroup depending effects

Additional Variations:

- include correlation between features
- add noisy features

Main results

- Superiority of SLIM and GUIDE regarding interpretability and performance in subgroup detection tasks
- MOB and CTree show higher stability and slightly higher performance in scenarios with smooth interactions than SLIM and GUIDE
- Pruning with SLIM and GUIDE not optimal, as strongly asymmetrical trees are sometimes generated (scenario linear smooth)
- Stability tends to be higher when MBTs are used as surrogate models
- SLIM and GUIDE more frequently split by wrong variables when correlated or noisy features are added
- Fundamental problem: modelling of smooth interactions with high performance only possible through many binary splits
  $\Longrightarrow$ strong decrease in interpretability

Simulation - Comparison of SLIM MBTs with leaf node models of different complexity

Question: How do SLIM trees with models of different complexity in the leafnodes differ in terms of interpretability if non-linearities are present in the data?

	linear regression	polynomial lasso regression	GAM
Number of leafnodes	high	medium	low
Interpretability of parameter estimates	yes	partly	no
Separation of interactions and maineffects	no	medium	yes

Table: Interpretability results of SLIM MBTs with different models based on simulation; Prepruning by R^2

Conclusion

- SLIM and GUIDE are promising surrogate models, especially when subgroups are present
  $\Longrightarrow$ R-package
- Improve pruning for SLIM and GUIDE
- For very deep MBTs the results of different MBT algorithms move closer together
- Use models that can capture non-linearities
- Beware of the risk of selection bias

Bibliography

Hothorn, T., Hornik, K. and Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework, *Journal of Computational and Graphical Statistics* **15**(3): 651–674.

Scenario linear categorical

Numerical and binary features with linear effects and subgroup specific linear effects:

- $\mathbf{x}_1, \mathbf{x}_2 \sim U(-1, 1)$, $\mathbf{x}_3 \sim Bern(0.5)$,
- $f(x) = \mathbf{x}_1 - 8\mathbf{x}_2 + 16\mathbf{x}_2 \mathbb{1}_{(\mathbf{x}_3=0)} + 8\mathbf{x}_2 \mathbb{1}_{(\mathbf{x}_1 > mean(\mathbf{x}_1))}$
- $\epsilon \sim N(0, 0.1 \cdot sd(f(\mathbf{x}))$
- $y = f(\mathbf{x}) + \epsilon$

MBT	*impr*	n leaves	n leaves min	n leaves max	R^2_{test}	sd R^2_{test}	share $\mathbf{x}_2$
SLIM	0.15	2.00	2	2	0.8323	0.0118	0.0000
SLIM	0.10	4.00	4	4	0.9870	0.0029	0.0000
SLIM	0.05	4.00	4	4	0.9870	0.0029	0.0000
GUIDE	0.15	2.00	2	2	0.8323	0.0118	0.0000
GUIDE	0.10	4.00	4	4	0.9870	0.0029	0.0000
GUIDE	0.05	4.00	4	4	0.9870	0.0029	0.0000
	alpha						
MOB	0.001	13.45	11	16	0.9729	0.0069	0.8865
MOB	0.010	14.38	13	16	0.9765	0.0066	0.8656
MOB	0.050	14.63	13	16	0.9771	0.0062	0.8614
CTree	0.001	11.96	10	14	0.9545	0.0049	0.9914
CTree	0.010	12.76	10	15	0.9550	0.0050	0.9897
CTree	0.050	13.46	10	16	0.9558	0.0052	0.9838
xgboost					0.9778	0.9778	

Table: Mean simulation results on 100 simulation runs as surrogate models for XGBoost predictions; n = 1500