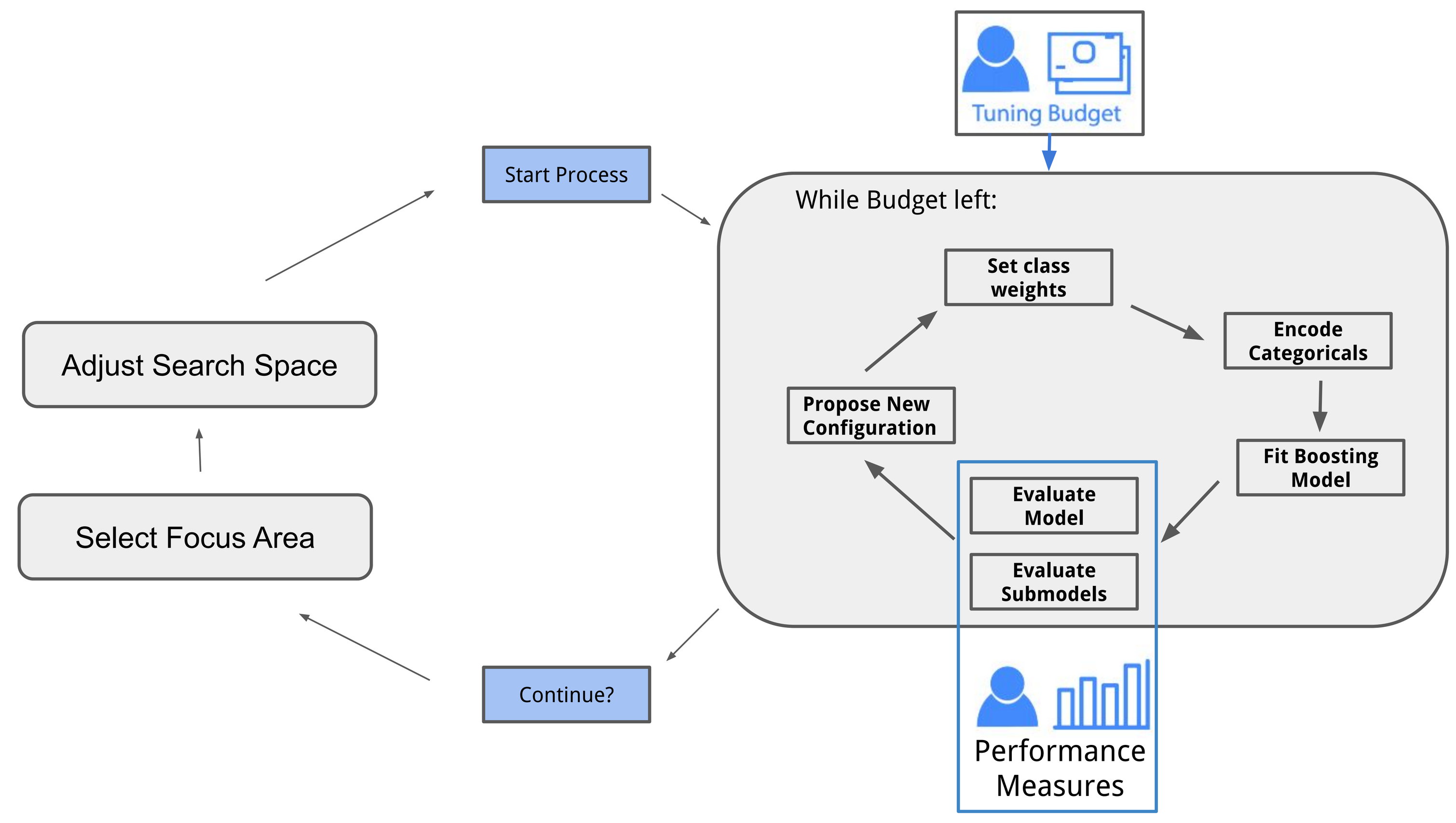
Multi-Objective AutoML

Florian Pfisterer, Janek Thomas, Stefan Coors, Bernd Bischl
LMU Munich

Abstract

AutoML systems are currently rising in popularity, as they can build powerful models without human oversight. They often combine techniques from many different sub-fields of machine learning in order to find a model or set of models that optimize a user-supplied criterion, such as predictive performance. The ultimate goal of such systems is to reduce the amount of time spent on menial tasks, or tasks that can be solved better by algorithms, while leaving decisions that require human intelligence to the end-user. In recent years, the importance of other criteria, such as fairness and interpretability, and many others has become more and more apparent. Current AutoML frameworks either do not allow to optimize such secondary criteria, or only do so by limiting the system's choice of models and preprocessing steps. We propose to optimize additional criteria defined by the user directly to guide the search towards an optimal machine learning pipeline. In order to demonstrate the need and usefulness of our approach, we provide a simple multi-criteria AutoML system and showcase an exemplary application.

TL;DR



Problem

- Some ML applications require optimizing multiple objectives at the same time.
- This is currently not possible in existing AutoML frameworks
- Awareness for multiple objectives often does not exist

Contribution

- We provide several interesting measures
- We propose a minimalistic implementation of a multi-objective AutoML system
- We demonstrate our implementation in a use-case

Measures

Fairness

- Independence

$$Pr\{Y_b = 1 | A = 0, Y = 1\} = Pr\{Y_b = 1 | A = 1, Y = 1\}$$

→ Minimize the absolute difference in FPR, FNR, ... between two sub-populations.

Interpretability

- **Complexity of main effects** [4]: Determine shape complexity of ALE main effects by number of parameters needed to approximate the curve with linear segments.
- **Interaction Strength** Impact of interaction effects is relevant when explanations are required, as interpretability techniques often use linear relationships to obtain explanations. Measures the fraction of variance that can not be explained by main effects.
- **Sparsity** Using fewer features simplifies interpreting models.

Robustness

- **Perturbations** Measure classifiers' robustness to perturbations in the input data.
- **Adversarial Examples** A variety of robustness measures can be derived from the different types of Adversarial Attacks [5] proposed.
- **Distribution shift** We might want models that are robust or work despite distribution shift. How to achieve this is an open research field.

Method

Pipeline

Based on autoxgboost [6], we create a **single-learner** AutoML system. We optimize hyperparameters of xgboost [2] along with class weights, and encoding for categorical variables.

Optimizer

We use Bayesian Optimization with **parEgo** [3], as it is a simple method, and it naturally lends itself to focussing on regions of the pareto front. **parEgo** is a rather simple extension, which scalarizes the set of target functions by using the augmented Tchebycheff norm

$$\max_{i=1,\dots,k}(w_i f_i(\theta)) + \rho \sum_{i=i}^{k} w_i f_i(\theta),$$

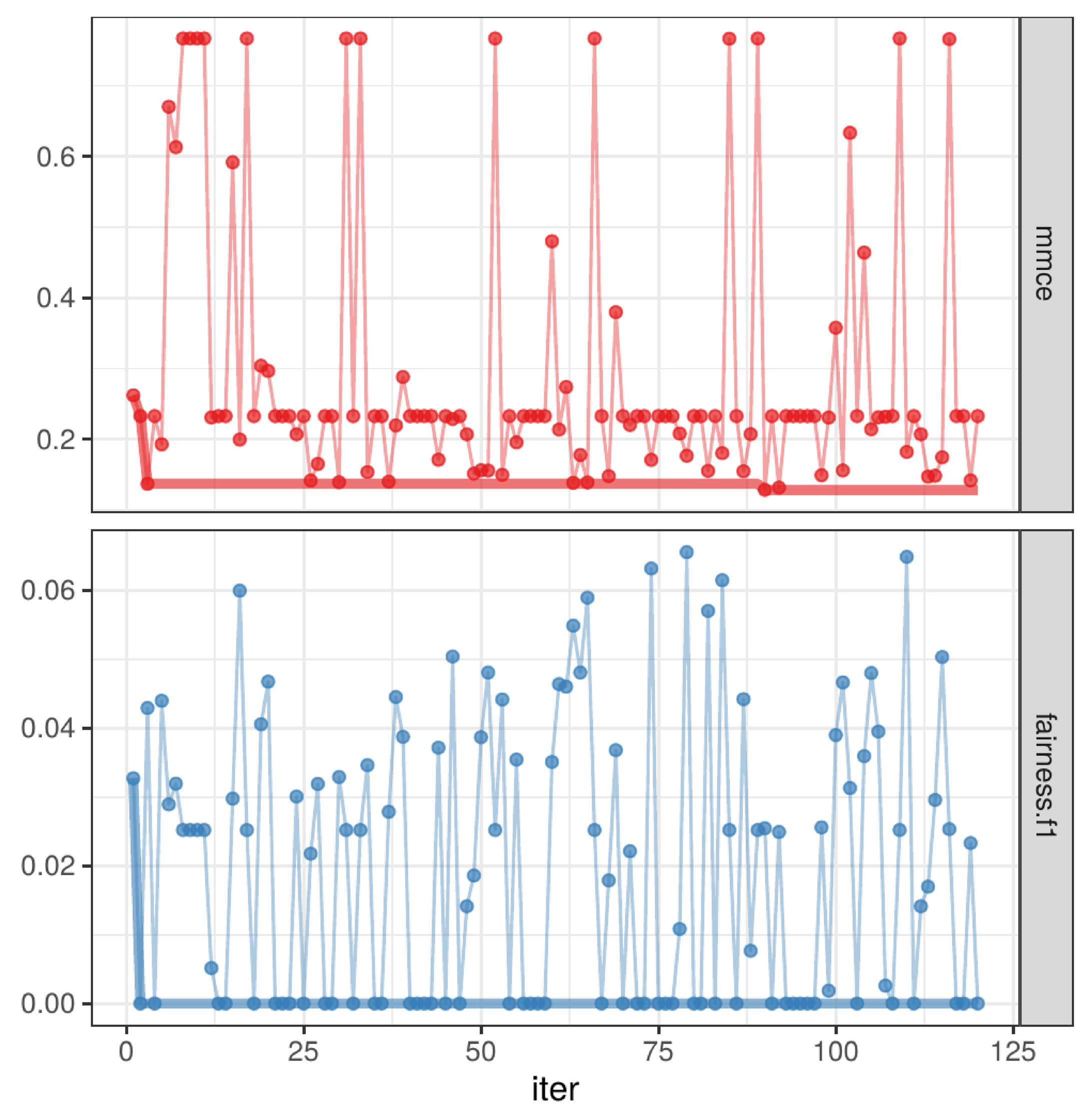
with a different uniformly sampled weight vector w such that $\sum_{i=1}^{k} w_i = 1$ in each iteration.

Sub-evaluations

As threshold-tuning and early-stopping are no longer trivial, after each iteration, we add different models with $n < nrounds$ and different classification thresholds to the pareto front. These can be obtained cheaply for a fixed model.

Application

The Adult dataset contains 48842 observations of 14 features, including 6 numeric and 8 factor features describing age, sex, education, marital status, race and more. The aim is to predict whether a person's income is below or above 50 k$ a year. We use the mean missclassification error as a performance measure for predictive performance. Additionally, we aim to assure fairness regarding a person's sex, i.e. we would like to obtain a model that performs equally good regardless of which sub-population we are considering. As a fairness measure we use the absolute difference in F1-Scores between two sub-populations *male* and *female*.



Tuning progress of the AutoML system on internal validation data.

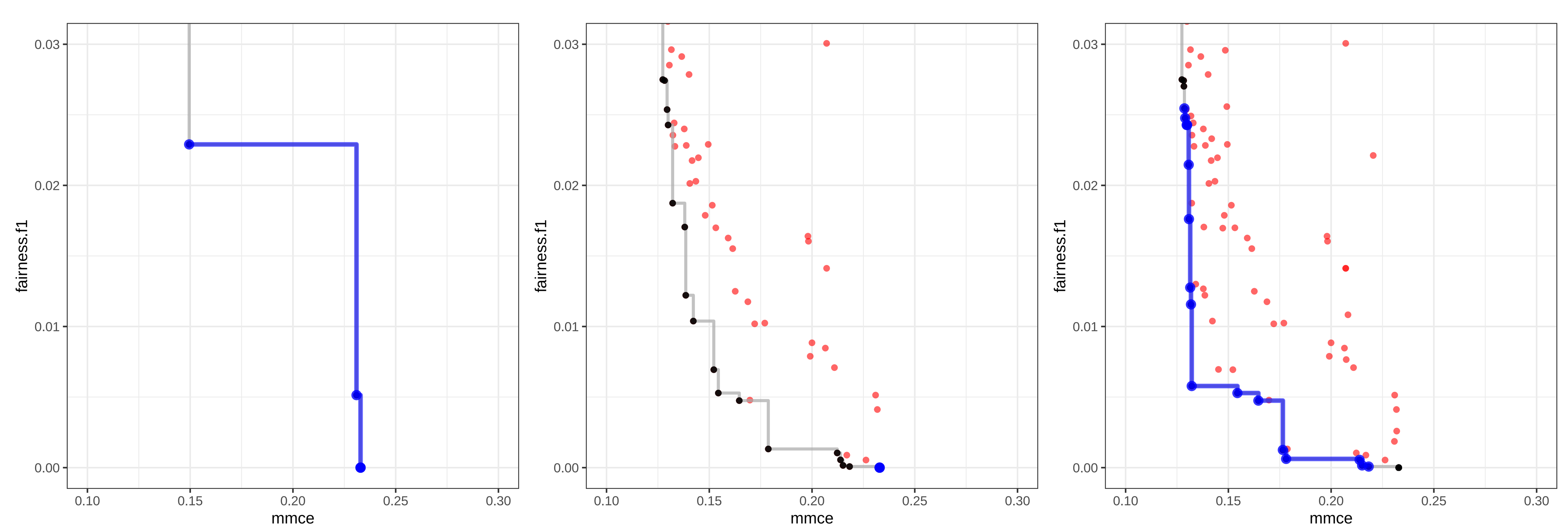
Outlook

- Implement pre- and post-processing methods tailored to measures.
- Add different optimizers (MO-Hyperband, more advanced multi-objective Bayesian Optimization).
- Explore functionality in different use-cases.

Github

`https://github.com/pfistfl/autoxgboostMC`

Use Case: Pareto Set



Pareto Front after 20, 70 and 120 iterations. Grey line: Pareto front; blue: Focus area

References

[1] Bernd Bischl, Michel Lang, Lars Kotthoff, Julia Schiffner, Jakob Richter, Erich Studerus, Giuseppe Casalicchio, and Zachary M. Jones. mlr: Machine learning in R. *Journal of Machine Learning Research*, 17(170):1–5, 2016.

[2] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, 2016. ACM.

[3] Joshua Knowles. Parego: A hybrid algorithm with on-line landscape approximation for expensive multiobjective optimization problems. Technical Report TR-COMPSYSBIO-2004-01, University of Manchester, September 2004.

[4] Christoph Molnar, Giuseppe Casalicchio, and Bernd Bischl. Quantifying interpretability of arbitrary machine learning models through functional decomposition. *arXiv preprint arXiv:1904.03867*, 2019.

[5] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[6] Janek Thomas, Stefan Coors, and Bernd Bischl. Automatic gradient boosting. In *International Workshop on Automatic Machine Learning at ICML*, 2018.

LaTeX TikZposter