Learning Optimal Fair Classification Trees

Nathanael Jo[1 [0000−0003−2295−9952]], Sina Aghaei[1 [0000−0002−3394−8864]], Andres Gomez[2 [0000−0003−3668−0653]], and Phebe Vayanos[1 [0000−0001−7800−7235]]

[1] USC Center for Artificial Intelligence
[2] University of Southern California

There is growing interest in using machine learning (ML) to make decisions in high-stakes domains, such as predicting a criminal's risk of recidivism [2], determining the best course of action for homeless individuals [3], to diagnose and treat various illnesses [4], and many more. In these contexts, it is necessary for such models to be both *accurate* (in order to minimize erroneous predictions that negatively affect stakeholders) and *interpetable* (so that decisions are transparent and hence accountable). Another crucial consideration in these high-impact situations is fairness; after all, an algorithm that affects people's well-being should be aware of the particular historical and/or social contexts that surround the learning problem. With these needs in mind, we focus our attention to the problem of learning optimal and fair classification (decision) trees.

Our approach and main contributions are:

- We present a mixed-integer optimization (MIO) formulation to learn optimal decision trees initially proposed in [1]. The original paper introduced the formulation without an emphasis on fairness, but in this work we will be using the formulation as a building block to which we add various fairness constraints. From hereon, we will refer to this approach as FairOCT.
- We conduct a comprehensive evaluation of its performance on several popular datasets, incorporating five popular notions of fairness where applicable: statistical parity, conditional statistical parity, predictive equality, equal opportunity, and equalized odds.
- We also compare our method to one of the most popular (heuristic-based) algorithms for learning fair trees proposed in [5] (DADT) and show that our optimal tree does, in fact, lead to significant performance improvements. FairOCT also has more flexible modeling power; it can incorporate arbitrary fairness constraints, solve for settings with more than two sensitive groups, and produces results that allow decision-makers to finely tune the accuracy-discrimination trade-off. In contrast, DADT only considers statistical parity and can only solve for settings with two sensitive groups.

We now briefly discuss our experimental results. In Figure 1, for any given discrimination level, FairOCT consistently has better in-sample and out-of-sample accuracies. This is expected since FairOCT finds an optimal solution whereas DADT relies on a heuristic. Given a fixed discrimination threshold, FairOCT improves out-of-sample (OOS) accuracy by 2.3 percentage points on average and obtains a higher OOS accuracy in 88.9% of the experiments.

One distinct advantage FairOCT has over DADT is the ability to fine-tune the accuracy-discrimination tradeoff. While both methods were trained on the

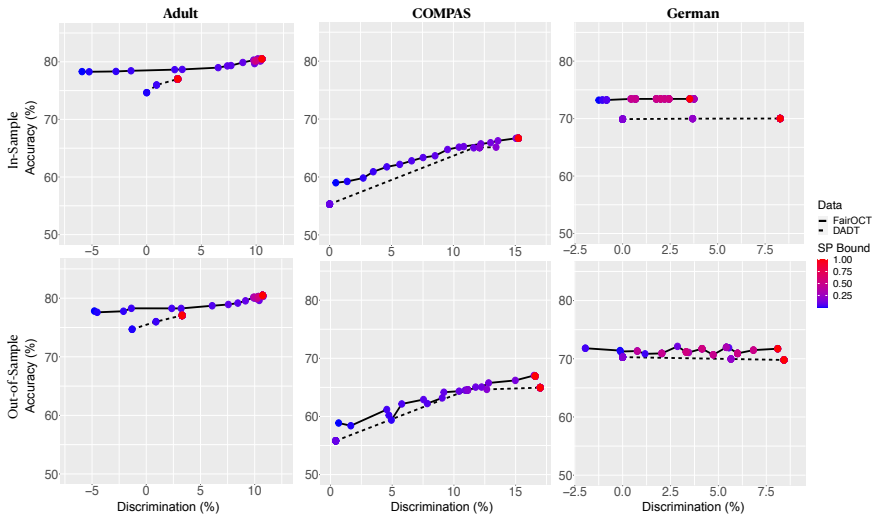

Fig. 1. Comparison of the accuracy and statistical disparity of FairOCT and DADT for trees of depth $d = 2$ when the discrimination bounds (δ) are varied on three datasets: Adult (left column), COMPAS (middle column), and German (right column) – averaged over 5 random train-test splits.

same fairness bounds, FairOCT produced more distinct results, which gives one more freedom to choose a policy that best suits their needs. This is because our fairness requirement is ingrained within the optimization problem, so a slight change in δ may change both the branching and labelling decisions.

To further showcase our approach's modelling power, we conducted experiments incorporating fairness notions other than statistical parity and included more than two sensitive groups in the data – neither of which DADT can accommodate. Our full paper has been submitted to the journal *Management Science*.

References

1. Aghaei, S., Gómez, A., Vayanos, P.: Strong optimal classification trees (2021)
2. Angwin, J., Larson, J., Mattu, S., Kirchner, L.: Machine bias (May 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
3. Azizi, M.J., Vayanos, P., Wilder, B., Rice, E., Tambe, M.: Designing fair, efficient, and interpretable policies for prioritizing homeless youth for housing resources. Lecture Notes in Computer Science **10848 LNCS**, 35–51 (2018). https://doi.org/10.1007/978-3-319-93031-23
4. Fatima, M., Pasha, M., et al.: Survey of machine learning algorithms for disease diagnostic. Journal of Intelligent Learning Systems and Applications **9**(01), 1 (2017)
5. Kamiran, F., Calders, T., Pechenizkiy, M.: Discrimination aware decision tree learning. In: Proceedings - IEEE International Conference on Data Mining, ICDM. pp. 869–874. IEEE, Sydney, Australia (2010). https://doi.org/10.1109/ICDM.2010.50