Extending Decision Tree to Handle Multiple Fairness Criteria

Alessandro Castelnovo[1,2]
[1]Intesa Sanpaolo S.p.A, Italy
[2]Dept. of Informatics, Systems & Communication, Univ. of Milan-Bicocca, Milan, Italy
a.castelnovo5@campus.unimib.it

Abstract

The demand for machine learning systems that can provide both transparency and fairness is constantly growing. Since the concept of fairness depends on the context, studies in the literature have proposed various formalisation and mitigation strategies. In this work, we propose a novel, flexible, discrimination-aware classifier that allows the user to: (i) select and mitigate the desired fairness criterion from a set of available options; (ii) implement more than one fairness criterion; (iii) handle more than one sensitive attribute; and (iv) specify the desired level of fairness to meet specific business needs or regulatory requirements. Our approach is based on an optimised extension to the decision-tree classifier, and aims to provide transparent and fair rules to the final users.

1 Introduction

AI-based decision systems interact with humans in many domains of application, and in some cases have replaced them. AI has spread widely because it is a powerful tool for making optimised decisions based on massive amounts of structured and unstructured data. However, the understanding and adoption of AI depends on criteria such as transparency, explainability and fairness. Explainable AI (XAI) is an immense help to companies in terms of producing AI systems that are trustworthy. However, XAI alone does not solve all of the possible ethics-related issues. AI systems are vulnerable to biases in the data that might render their decisions “unfair” to certain population subgroups based on gender or race. The different preferences and outlooks in different cultures involve different ways of looking at fairness, which makes it harder to come up with a single definition that is acceptable to everyone in a given situation.

Table 1 summarises three families of observational[1] fairness criteria typically adopted (for details see [Barocas *et al.*, 2017]), where Y, $\hat{Y}$ and S are the target variable, the model outcome and the sensitive attribute, respectively.

Independence	**Separation**	**Sufficiency**
$\hat{Y} \perp\!\!\!\perp S$	$\hat{Y} \perp\!\!\!\perp S \mid Y$	$Y \perp\!\!\!\perp A \mid \hat{Y}$

Table 1: Observational fairness criteria.

As detailed in [Zafar *et al.*, 2019], the mitigation strategies proposed in prior studies typically lack flexibility with respect to one or more of the following aspects: (i) They are specifically designed for only one fairness criterion, and as a consequence cannot accommodate more than one simultaneously; (ii) they cannot ensure fairness with respect to multiple sensitive features simultaneously; (iii) the proper mitigation model depends on the chosen definition of fairness; (iv) they are designed as a black box, i.e. they are not directly interpretable. To overcome these limitations we are developing `FFTree`, a new transparent, flexible and fairness-aware classifier.

2 Related Works

Our work is related to the *discrimination-aware decision tree*, which is an intersection between *fairness* and *interpretability* in supervised machine learning. A decision tree classifier [Brieman *et al.*,] is one of the most popular algorithms in machine learning given its intelligibility. In a discrimination-aware decision tree, fairness criteria are taken into consideration with the performance optimisation in the evaluation at each internal node. The first *”fair-tree”* was introduced in [Kamiran *et al.*, 2010], while the authors of [Zhang and Ntoutsi, 2019] have proposed important developments. All of these works handle the trade-off between accuracy and fairness by combining performance and fairness criteria through an operation, such as addition or multiplication. However, the different metrics could be incomparable. Moreover, they permit to mitigate a limited number of fairness criterion.

3 Contribution

We present `FFTree`, a new transparent, flexible and fairness-aware classifier. As a novelty, `FFTree` enchances the classical approach introduced in [Brieman *et al.*,] with a new approach to find a ”fair” split to: (i) satisfy a fairness constraint selected from a wide range of possible definitions of fairness; (ii) provide more than one fairness formalisation simultane-

[1] A criteria is called ”observational” if by knowing the joint distribution $(\hat{Y}, S, Y)$ is possible determine its value without ambiguity.

ously (when this is algebraically possible); (iii) handle more than one sensitive attributes at the same time; and (iv) set the required level of fairness as an input parameter to meet different business needs or regulatory requirements. FFTree can support users in many different application domains, including promoting gender equality and reducing inequalities within ethnic groups.

4 How FFTree Works

Given the domain $dom(X_i)$ of the input features $X = X_i,\ i = 1,..,n$ and the target label Y, the goal of a splitting criterion is to find a rule to separate a node D into *"purer"* branches D_j. The concept of purity is related to the target label Y, whereas the splitting criterion is determined by iteratively observing the variables X and their domains. Formally, an increase in purity is called *Information gain* (IG), and can be expressed as the differences in the entropy H of Y between the node and the average for the branches created by the splitting criterion a:

$$IG(a) = H_Y(D) - \sum_{j=1}^{m} \frac{|D_j|a|}{|D|} H_Y(D_i|a) \qquad (1)$$

Among all the splitting criterion under consideration, the one that locally (at the node) optimises the IG is chosen. In our approach, the best splitting criterion is the one that locally optimizes the IG, from among the options that satisfy the imposed fairness constraints FC.

$$\max IG(a) \quad \text{s.t.} \quad FC(a|S) \leq \delta \qquad (2)$$

Where FC is a fairness constraint, S is a binary sensitive attribute and δ is the maximum level of discrimination locally permitted. Currently, FFTree allows the users to choose FC among the observational fairness criteria summarized in Table 1, expressed as the absolute value of the difference between the two probability terms. FFTree permits to choose more than one fairness metric and more than one sensitive attribute S. All the selected criteria must be satisfied by each split; if no solutions are found the node becomes a leaf. To calculate the fairness metrics we compute the local prediction $\hat{Y}$ of each split assigning a value of one to the branch with the higher percentage of Y and zero to the other branch.

Demo. FFTree demonstration, using the open dataset Adult, is available at tinyurl.com/DemoFFTree.

4.1 Properties and Limitations

Properties. When FFTree is trained to provide perfect independence in each local branch ($\delta = 0$), the resulting global tree is *fairness-pruning-invariant* and *fairness-threshold-invariant*, since by pruning FFTree or changing the classification threshold, the output continues to provide perfect independence. Moreover, the optimisation at local level implies that all the rules shown by FFTree are consistent with the chosen fairness constraint, while other fair trees, optimised at global level, could be composed of a series of rules which, if taken individually, could be unfair. We suggest to tune the δ hyperparameter to identify at the global level the best performance-fairness trade-off for the application domain.

Limitations. Meeting different fairness criteria at the same time is not always possible. These criteria constrain the joint distribution in non-trivial ways. We should therefore suspect that imposing any two of them simultaneously will overconstrain the space to the point where only degenerate solutions remain. Moreover, under certain assumptions, different fairness criteria become mutually exclusive (see [Barocas *et al.*, 2017]).

5 Experiments and Future Directions

We evaluate FFTree on both (i) a well-known benchmark dataset from the literature, and (ii) a real-world private dataset owned by Intesa Sanpaolo containing about 250,000 records related to the granting of personal loans. We prove that FFTree is competitive with state-of-the-art fair trees in the common features, and thanks to its novel functionality, it can support users in very different application domains.

To make a more relevant contribution to the field of discrimination-aware data mining, we intend to release the code as open source, to allow the community to utilise it and improve it. A natural evolution of FFTree would be the addition of new fairness criteria or different formalisation of the current criteria. To improve performance, we are currently studying how to use FFTree for tree-based machine learning algorithms, such as Random Forest. FFTree can also be used in place of classic decision trees in XAI tools that act on surrogates such as ContrXT [Malandri *et al.*, 2022] or to monitor fairness through time [Castelnovo *et al.*, 2021].

Acknowledgements

Andrea Cosentini, Lorenzo Malandri, Fabio Mercorio and Mario Mezzanzanica contributed to this work. This research was supported by the outstanding Fairness team of the Data Science & AI Office at Intesa Sanpaolo.

References

[Barocas *et al.*, 2017] Barocas, Hardt, and Narayanan. Fairness in machine learning. *Nips tutorial*, 1:2, 2017.

[Brieman *et al.*,] Brieman, Friedman, Olshen, and Stone. Classification and regression trees. *Wadsworth Inc*.

[Castelnovo *et al.*, 2021] Castelnovo, Malandri, Mercorio, Mezzanzanica, and Cosentini. Towards fairness through time. In *ECML PKDD*, pages 647–663. Springer, 2021.

[Kamiran *et al.*, 2010] Kamiran, Calders, and Pechenizkiy. Discrimination aware decision tree learning. IEEE, 2010.

[Malandri *et al.*, 2022] Malandri, Mercorio, Mezzanzanica, Nobani, and Seveso. ContrXT: Generating contrastive explanations from any text classifier. *IF*, 2022.

[Zafar *et al.*, 2019] Zafar, Valera, Gomez-Rodriguez, and Gummadi. Fairness constraints: A flexible approach for fair classification. *JMLR*, 20(1):2737–2778, 2019.

[Zhang and Ntoutsi, 2019] Zhang and Ntoutsi. Faht: an adaptive fairness-aware decision tree classifier. In *AAAI*, pages 1480–1486, 2019.