ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

Boundary-Based Fairness Constraints in Decision Trees and Random Forests

Géraldin Nanfack*, Valentin Delchevalerie* and Benoît Frénay

University of Namur - NaDI/naXys - Faculty of Computer Science - PReCISE
Rue Grandgagnage 21, 5000 Namur - Belgium

Abstract. Decision Trees (DTs) and Random Forests (RFs) are popular models in Machine Learning (ML) thanks to their interpretability and efficiency to solve real-world problems. However, DTs may sometimes learn rules that treat different groups of people unfairly, by paying attention to sensitive features like for example gender, age, income, language, etc. Even if several solutions have been proposed to reduce the unfairness for different ML algorithms, few of them apply to DTs. This work aims to transpose a successful method proposed by Zafar et al. [1] to reduce the unfairness in boundary based ML models to DTs.

1 Introduction

Machine learning (ML) takes an increasingly important place in our society. It is used in many areas, even when it can directly affect citizens such as in hiring processes [2], health care [3], etc. However, ML algorithms may sometimes learn rules that treat different groups of people unfairly due to (un)intentional bias in data. For example, models that rely on sensitive features such as gender, age, income or language to take decisions should probably be rejected. Furthermore, some works [4, 5] show that simply removing these features from data is not sufficient since it ignores, sometimes complex, correlations with other features.

This concern led to a new area of research that aims to reduce the unfairness in ML models [6]. However, few of the existing methods propose a flexible way to reduce unfairness in Decision Trees (DTs) and Random Forests (RFs), while these models remain very popular models in the ML community.

A few years ago, Zafar et al. [1] proposed a powerful and "flexible constraint-based framework to enable the design of fair margin-based classifiers". They show how straightforward it is to use their framework for convex boundary-based classifiers (like logistic regression, SVM, etc.) thanks to the simple expression of the distance to the decision boundary in these cases. We propose then to extend this framework to DTs where, unlike in Zafar et al. [1], there is no such analytical expression of the distance to the decision boundary. To do so, we use the distance to the decision boundary as defined by Alvarez [7]. This leads to DTs that benefit from a powerful and flexible way to design fair models.

Section 2 describes the notion of unfairness in ML. Section 3 presents the state of the art in order to build fair DTs as well as the work of Zafar et al. [1]. Section 4 presents our work. Section 5 and Section 6 present the experiments and the discussion respectively, before concluding in Section 7.

*The first two authors contributed equally.

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

2 Background: On the Notion of Unfairness in ML

Several statistical measures have been proposed to define the *fairness* property [6] for ML models. We rely on the work of Zafar et al. [1], which revisits unfairness according to popular notions that include disparate impact and disparate mistreatment. First, disparate impact consists in providing outputs that benefit *disproportionately* to a group of people sharing the same value of the sensitive feature than other groups. Therefore, a classifier does not suffer from disparate impact if $p(\hat{y}|z=0) = p(\hat{y}|z=1)$ or equivalently $p(\hat{y}|z) = p(\hat{y})$, where z is a binary sensitive feature. Second, the disparate mistreatment refers to the *equality of opportunity* and a binary classifier is said to be fair according to disparate mistreatment if the misclassification rates, for different groups having different values of the sensitive feature, are equal.

3 Related Work

While several works have been proposed to learn ML models under fairness constraints (usually on differentiable models), very few exist on models such as DTs. Kamiran et al. [5] propose data pre-processing by relabelling and reweighting instances so as to reduce bias in data. However, as pre-processing (i) cannot eliminate discrimination that may come from the learning algorithm, and (ii) may miss complex correlations with other features, the same authors later propose the first discrimination-aware DT algorithm [8]. They introduce the information gain sensitivity, which measures the level of discrimination induced by a split. Similarly, Raff et al. [9] develop a fair version of the impurity score based on the Gini index. Instead of using impurity measures, Xhang et al. [10] propose a *fairness* gain based on the disparate impact measure. However, none of them are flexible, in the sense that they do not present a framework able at the same time to handle (i) different types of unfairness, (ii) m-ary classification and (iii) multiple sensitive features. Another notable work is the one of Aghaei et al. [11] who propose a mixed integer programming (MIP) formulation to learn optimal fair DTs. Nonetheless, MIP problems pose a scalability concern and, as highlighted by authors, their approach is computationally expensive.

While these previous works in DTs are usually limited to disparate impact, Zafar et al. [1] recently proposed a flexible approach to enforce fairness constraints through the decision boundary. They introduce a tractable proxy measure of unfairness, which is the covariance between the sensitive feature and the signed distance of instances to the decision boundary of the model. For disparate impact and disparate mistreatment, their problem can be formalized as

$$\text{minimize } L(\theta) \text{ s.t. either } \Big|\mathrm{Cov}\big(z, d_\theta(\mathbf{x})\big)\Big| \leq \gamma \text{ or } \Big|\mathrm{Cov}\big(z, \min\big(0, y\, d_\theta(\mathbf{x})\big)\big)\Big| \leq \gamma,$$

where $d_\theta(\mathbf{x})$ is the signed distance of the instance $\mathbf{x}$ to the decision boundary of the classifier and γ is a positive threshold. Notice that only one of the constraints is used, depending on the goal. The first constraint aims to enforce (but does not guarantee) independence between the decision boundary of the model and the sensitive feature, *i.e.*, $p\big(d_\theta(\mathbf{x}) \geq 0|z\big) \approx p\big(d_\theta(\mathbf{x}) \geq 0\big)$, thus vanishing

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

disparate impact. Similarly, the second constraint seeks independence between the sensitive feature and the misclassification rate, thus vanishing disparate mistreatment. Despite being attractive and flexible, this framework requires to explicitly know $d_\theta(\mathbf{x})$, which is unfortunately not the case for DTs.

4 A Boundary-Based Method to Learn Fairer DTs

This section shows how to use the developments presented in Section 3 in DTs. To do so, one should (i) evaluate the distance of an instance to the decision boundary and (ii) introduce the unfairness as a constraint during the training.

4.1 Evaluating the Distance to the Decision Boundary

Alvarez et al. [7] proposed a method to compute the distance of an instance $\mathbf{x} \in \mathbb{R}^d$ to the decision boundary in DTs. For each decision path in the tree that would lead to another label than the one of $\mathbf{x}$, and for all conditions $x_u \gtrless b$ that are false along that path, the value of the feature u is changed by b to produce a new instance $\mathbf{x}'$. Otherwise, $x'_u = x_u$. The distance to the decision boundary is then defined as $d_\theta(\mathbf{x}) = \min d(\mathbf{x})$ for all leaves where

$$d(\mathbf{x}) = \sqrt{\sum_{u=1}^{d} (x_u - x'_u)^2}. \tag{1}$$

4.2 Learning Fairer DTs

Once an expression of the distance to the decision boundary $d_\theta(\mathbf{x})$ is obtained, it is possible to transpose the idea in Section 3 to DTs. Here, in order to illustrate our methodology, we choose to focus on the first unfairness notion introduced in Section 2: Disparate Impact (DI). This will be extended to other unfairness notions in further work. In order to be integrated into splitting criteria and to constrain learning, unfairness in terms of DI can be assessed with the correlation

$$\mathrm{Corr}_{DI} = \frac{1}{N\sigma_z\sigma_{d_\theta}} \sum_{(\mathbf{x}_i, z_i)\in\mathcal{D}} (z_i - \bar{z})\, d_\theta(\mathbf{x}_i), \tag{2}$$

where z is the sensitive feature, $\bar{z}$ its mean value and σ_z and σ_{d_θ} are the standard deviations of z and d_θ. Notice that, contrarily to what is performed by Zafar et al. [1], we used a correlation instead of a covariance to assess the unfairness. Correlation $|\mathrm{Corr}_{DI}|$ has the advantage to be included in $[0, 1]$, as for the entropy in binary classification. Thus, it will be easier to manage the trade-off between accuracy and unfairness. Indeed, unfairness is introduced as a penalty term when considering a particular split by updating the definition of the gain to

$$\text{gain} = I_0 - sI_1 - (1-s)\, I_2 - \lambda\, |\mathrm{Corr}_{DI}|, \tag{3}$$

where I_0 is the impurity of the parent node, I_1 and I_2 are the impurity of the left and right child respectively, and s is the fraction of instances that will be assigned to the left child. Then, using the correlation instead of the covariance makes it easier to choose the meta-parameter λ that trades off between accuracy

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

and unfairness. Thanks to Equation (3), if $\lambda > 0$, a split may be rejected if it increases too much the DI (i.e., if the split is too much correlated with z).

It is important to note that the above developments are presented in the particular case of a binary classification task with only one binary sensitive feature, and where the DI is used to assess the unfairness. However, as presented by Zafar et al. [1], it is easy to extend to m-ary classification, other types of unfairness (see Section 2) and/or multiple categorical sensitive features.

5 Experiments

We implemented a version of the CART algorithm with the heuristic presented in Equation (3) to learn DTs in a top-down fashion. Among all possible splits, the one that leads to the greater gain is retained, at each level of the tree, until a stopping criterion is reached or none of the splits leads to a positive gain. The first dataset is an artificial dataset for binary classification that we called *synthetic*. It is built similarly to what is proposed by Zafar et al. [1] by drawing 3,000 instances from two different 2-d multivariate normal distributions, and a binary feature z correlated to the class label using a Bernouilli distribution. We also used the well-known *German Credit* and *Adult Income* datasets to test our methods on real-world problems. Table 1 summarizes these three datasets, along with the associated size, dimension, sensitive feature and classification task.

Name	N	d	Sensitive feature	Task
Synthetic	3,000	3	$z = 0$ / $z = 1$	$y = -1$ / $y = 1$
German Credit	1,000	20	Age ≥ 25	Good / Bad Credit
Adult Income	45,222	14	Male / Female	Income $<$ / $\geq$ 50k

Table 1: Details of the datasets used for our experiments.

For each dataset, we trained DTs and RFs made of 50 DTs. In both cases the maximum number of leaves is fixed to 15 to avoid overfitting. The entropy is chosen as the splitting criterion. During the training process, Corr_{DI} is computed by Equation (2) and used in Equation (3) to assess the unfairness. 40 values for λ between 0 and 10 are tested. When training RFs, bootstrapping is used to randomly select 500, 150 and 5,000 instances for *synthetic*, *German Credit* and *Adult Income*, respectively. Meta-parameters are fixed to these acceptable values in order to perform fair comparisons of the models for different values of λ. All the experiments are performed on 30 independent runs to compensate for the randomness when splitting the data into training and testing sets.

Figure 1 shows the results obtained for the three datasets and averaged over the 30 runs. It shows the evolution of the accuracy and the DI with respect to λ for both simple DTs and RFs. In accordance with previous work [1], the DI of each model is assessed during testing by computing $|p(\hat{y}|z = 0) - p(\hat{y}|z = 1)|$ as explained in Section 2.

6 Discussion

In the case of *synthetic*, very similar results are obtained for simple DTs and RFs. When λ increases, splits that lead to bigger DI are more and more penalized.

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

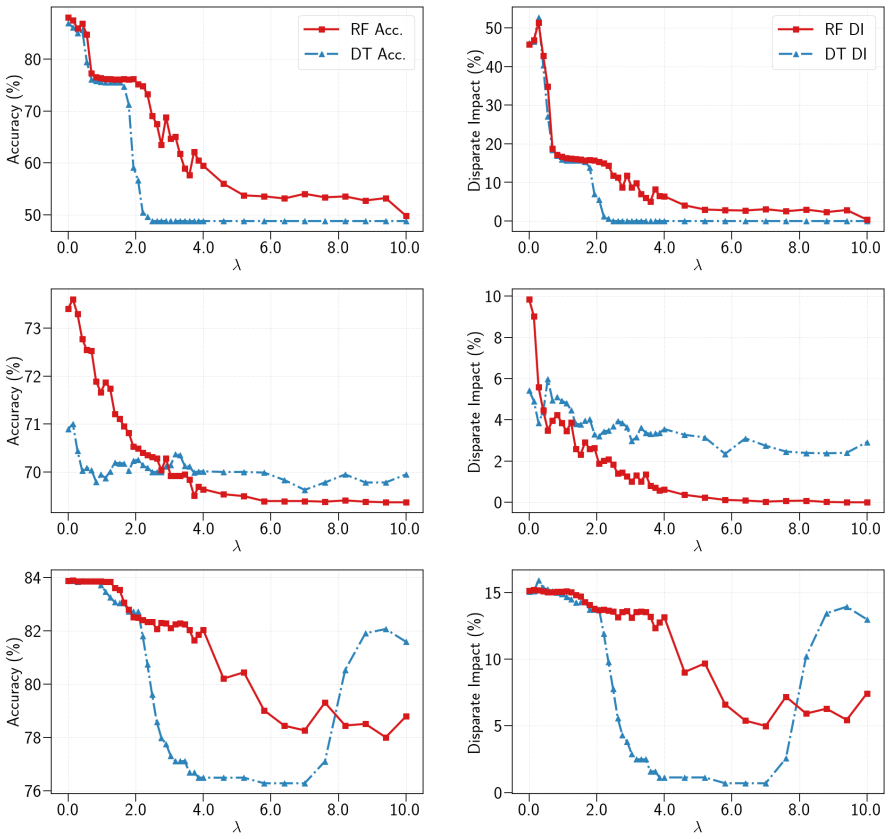

Fig. 1: Evolution of accuracy and DI with respect to λ on *synthetic* (1st row), *German Credit* (2nd row) and *Adult Income* (3rd row) for both DTs and RFs. Note that the DI is not assessed using Equation (2), but by computing instead the probabilities as explained in Section 2. Legend are shared by the columns.

They tend to be rejected and the DI consequently decreases. However, since the penalization increases, the change in selected splits also leads to a decreased accuracy. For both DTs and RFs, a good trade-off between accuracy and unfairness is achieved at the plateau for λ approximately in the range $[0.8, 1.9]$. For these values, the drop in accuracy is compensated by a significant DI reduction.

For *German Credit*, the DI is lower than for other two datasets. For DTs, our method is not able to further reduce the DI much. However, for RFs, a rapid drop in DI is observed for $\lambda < 0.5$, whereas the accuracy decreases more slowly.

For *Adult Income*, similar results to those obtained for *synthetic* are obtained, with a plateau for $\lambda \in [2, 4]$ for RFs. DI can be reduced for both simple DTs and RFs. The instability of DTs for large λ values may be explained by the fact that the CART algorithm performs a greedy search, i.e., a succession of local

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

optimizations that impact each other. Penalization may force CART to explore new directions and lead to more interesting solutions.

7 Conclusion

This work presents an adaptation for DT-based classifiers of the developments proposed by Zafar et al. [1] to reduce the unfairness in convex-boundary-based ML models. The proposed solution presents the advantage of being very flexible in its design of fair classifiers and could be easily extended in further work to handle multiple types of unfairness, m-ary classification and/or multiple categorical sensitive features. We tested it on different biased datasets, and it allows us to train DTs and RFs with an adjustable accuracy / unfairness trade-off. More extensive experiments should be performed to (i) better characterize its limitations, (ii) compare it with a baseline and (iii) investigate other choices to assess the unfairness than Equation (2) that would be more adapted for DTs.

Acknowledgments

G.N. and V.D. were supported by the EOS VeriLearn project n. 30992574. V.D. is supported by the Walloon region with a Ph.D. grant from FRIA (F.R.S.-FNRS). This research used resources of PTCI at UNamur, supported by the F.R.S.-FNRS under the convention n. 2.5020.11. The authors thank Adrien Bibal for the fruitful discussions on this paper.

References

[1] Muhammad Bilal Z., Isabel V., Manuel Gomez-Rodriguez, and Krishna P. G. Fairness Constraints: A Flexible Approach for Fair Classification. *JMLR*, 20(75):1–42, 2019.

[2] Miranda Bogen and Aaron Rieke. Help wanted: an examination of hiring algorithms, equity, and bias. Technical report, 2018.

[3] M. D McCradden, S. Joshi, J. A Anderson, M. Mazwi, A. Goldenberg, and R. Zlotnik Shaul. Patient safety and quality improvement: Ethical principles for a regulatory approach to bias in healthcare machine learning. *JAMIA*, 27(12):2024–2027, 2020.

[4] Faisal Kamiran and Toon Calders. Classifying without discriminating. In *Proc. of the International Conference on Computer, Control and Communication*, pages 1–6, 2009.

[5] T. Calders, F. Kamiran, and M. Pechenizkiy. Building classifiers with independency constraints. In *Proc. of ICDM Workshop on Domain Driven Data Mining*, pages 13–18, December 2009.

[6] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A Survey on Bias and Fairness in Machine Learning. *arXiv preprint arXiv:1908.09635*, September 2019.

[7] I. Alvarez, S. Bernard, and G. Deffuant. Keep the decision tree and estimate the class probabilities using its decision boundary. In *Proc. of IJCAI*, pages 654–659, 2007.

[8] Faisal Kamiran, Toon Calders, and Mykola Pechenizkiy. Discrimination Aware Decision Tree Learning. In *Proc. of ICDM*, pages 869–874, December 2010.

[9] Edward Raff, Jared Sylvester, and Steven Mills. Fair Forests: Regularized Tree Induction to Minimize Model Bias. In *Proc. of AIES*, pages 243–250, December 2018.

[10] Wenbin Zhang and Eirini Ntoutsi. Faht: An adaptive fairness-aware decision tree classifier. In *Proc. of IJCAI*, pages 1480–1486, July 2019.

[11] S. Aghaei, M. J. Azizi, and P. Vayanos. Learning Optimal and Fair Decision Trees for Non-Discriminative Decision-Making. In *Proc. of AAAI*, pages 1418–1426, July 2019.