Fair Augmentation of Decision Trees Through Selective Node Retraining

Coen Adler
Baskin School of Engineering
University of California, Santa Cruz
Santa Cruz, CA 95064
ctadler@ucsc.edu

Abstract

With modern machine learning models becoming ever more complex and embedded within society, there is a need for accurate, interpretable, and fair models that users can trust. Decision trees being a fully interpretable model, fill this role perfectly. Current research shows that algorithms exist that can train decision trees to be both accurate and fair. Despite this, decision trees are often trained solely on accuracy, resulting in biased or discriminative pre-trained trees. Frequently, the root of the bias in these pre-trained trees stems from a few select nodes or subtrees. In this paper, I propose a novel method of selective fair retraining of decision trees, modifying discriminative nodes to remove bias and retain a high accuracy. The experimental results indicate that the proposed tree modification method can result in fair decision trees with higher accuracy than those trained from scratch.

1 Introduction

The rise of machine learning in our society, such as the COMPAS tool being used to predict recidivism rates, raises the question, "are the models being implemented in our technology fair and non-discriminative?" [2] Innovations in the performance of modern models are constantly increasing, and so is the need for these models to be fair and interpretable. Decision trees (DT) are no exception to this statement. There have been consistent improvements on the already performant ID3 and CART DT training algorithms to be fair and non-discriminative [8, 6]. However, all these algorithms require training a DT from scratch, when often times one is given an already pretrained tree.

To modify a tree to become less discriminative, we must first define how to measure a DT's fairness. There are numerous metrics for fairness in machine learning, but Statistical Parity and Equalized Odds metrics are considered to be the most common [3]. A score of 0 with these metrics means the data is perfectly fair while a negative score shows discrimination against a minority class and positive score represents discrimination against a majority class. Modifying a DT to be perfectly fair is not always the right choice because of the fairness-accuracy tradeoff, where increasing fairness can decrease accuracy and vice versa [4].

ID3 and CART are the source algorithms that many fair decision tree training algorithms build off of [3]. ID3 uses the optimization of entropy (or gini for CART) as a splitting criterion; this optimization is known as information gain (IG).

$$\textsc{Entropy}(S) = -\sum P(I) \cdot \log_2(P(I)) \tag{1}$$

$$\textsc{InformationGain}(S, A) = \textsc{Entropy}(S) - \sum P(S|A) \cdot \textsc{Entropy}(S|A) \tag{2}$$

Training a tree on ID3 typically results in very accurate trees but at the cost of poor fairness. Many algorithms have been proposed to help tip the balance in favor of fairness such as the Fairness-Aware

Hoeffding Tree (FAHT) algorithm [9]. While this algorithm was tailored for streaming trees, its splitting criterion tended to work well for balancing accuracy and fairness in regular DTs. In addition to ID3's information gain, FAHT uses a metric called fairness gain (FG). A product of fairness gain and information gain is used to calculate the splitting criterion such that the tree finds the best split to optimize for both accuracy and discrimination.

$$\textsc{FairnessGain}(S, A) = \textsc{Disc}(S) - \sum P(S|A) \cdot \textsc{Disc}(S|A) \tag{3}$$

$$\text{FAHT} = \begin{cases} IG & \text{if } FG = 0 \\ IG \cdot FG & \text{if } FG \neq 0 \end{cases} \tag{4}$$

The concept of modifying fair trees is not an entirely new field. A study by Kamiran et al. found that it is possible to improve fairness by flipping the classification at select nodes [5]. However, there are a couple of stipulations that come with this method. The concept of flipping nodes, where doing so improves fairness, tends to have a very negative effect on the accuracy of the tree. They found out that by training the tree specifically to be very discriminative and accurate and then flipping the nodes, the results were better than when flipping nodes on a fair and accurate tree. Unfortunately, this method of tree augmentation also falls short on trees trained only for accuracy such as ID3 or CART.

With these methods in mind, I propose a dual-method approach (node selection and retraining methods) for editing pre-trained trees to balance both accuracy and fairness.

2 Methods

The modification of pre-trained decision trees consists of two main methods: node selection and subtree retraining. The node selection determines which nodes should be marked for retraining. Before marking nodes to retrain, the algorithm first calculates the FAHT splitting criterion of each node. A stream of ordered nodes is created that is sorted by increasing FAHT scores below a select threshold. The ordered nodes are pulled from the stream and subsequently retrained using the FAHT algorithm. The depth of the newly retrained subtree is determined by the size of the marked node's subtree. These retrained subtrees are compared with the marked node's subtree and if the new subtree improves the entire DT, it replaces the marked node. The algorithm iterates over the ordered nodes until either all marked nodes are retrained or the discrimination of the DT surpasses the termination threshold.

3 Results

Experiments using the proposed algorithm use Sklearn's[1] decision tree classifier method (optimized version of ID3) as the starting pre-trained tree. Additionally, the method was tested using the following two common datasets: the 1994 US Adult Census data [1] predicting whether annual income was above a $50k threshold and the Kaggle Credit Score dataset [7] classifying whether an individual has a strong credit score. Discrimination was calculated using statistical parity with sex as the sensitive attribute where *male* was the dominant attribute and *other* being the minority or discriminated attribute.

Table 1: Comparative Results

Dataset	Model	Accuracy	Discrimination
Adult	Sklearn	0.83	-0.226
	Modified	0.80	-0.022
	FAHT	0.76	0
Credit Score	Sklearn	0.86	-0.122
	Modified	0.79	-0.069
	FAHT	0.55	0.024

[1] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

When training a DT from scratch with Sklearn's ID3 algorithm, there was clear sense of bias shown through a statistical parity score of -0.226 on the Adult Census dataset. Retraining with the FAHT algorithm, removes the bias at the cost of performance, resulting in a 7% drop in accuracy. Additionally, the network generated by FAHT significantly differs from that of the default model. Using the proposed tree modification algorithm, we get the benefits of both models, seeing only a 3% decrease in accuracy, while bias is almost entirely eliminated. Similar results are obtained while using the Credit Score dataset.

4 Conclusion

While existing methods attempt to train a fair decision tree from scratch, the proposed approach starts with a pre-trained tree such that it only has to retrain smaller subtrees. This approach allows for comparable results to current algorithms. Additionally, this approach has the potential to generate a fair tree significantly quicker than training from scratch as it does not have to retrain every node in the tree. Future work on this problem includes the implementation of other fair splitting criteria and their comparison to using the FAHT splitting criterion.

References

[1] Barry Becker and Ronny Kohavi. Adult. UCI Machine Learning Repository, 1996. DOI: https://doi.org/10.24432/C5XW20.

[2] Julia Dressel and Hany Farid. The accuracy, fairness, and limits of predicting recidivism. *Science advances*, 4(1):eaao5580, 2018.

[3] Pratyush Garg, John Villasenor, and Virginia Foggo. Fairness metrics: A comparative analysis. In *2020 IEEE international conference on big data (Big Data)*, pages 3662–3666. IEEE, 2020.

[4] Nathanael Jo, Sina Aghaei, Jack Benson, Andres Gomez, and Phebe Vayanos. Learning optimal fair decision trees: Trade-offs between interpretability, fairness, and accuracy. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 181–192, 2023.

[5] Faisal Kamiran, Toon Calders, and Mykola Pechenizkiy. Discrimination aware decision tree learning. In *2010 IEEE international conference on data mining*, pages 869–874. IEEE, 2010.

[6] Wei-Yin Loh. Classification and regression trees. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, 1(1):14–23, 2011.

[7] Rohan Paris. Credit score classification. kaggle dataset, 2021.

[8] J. Ross Quinlan. Induction of decision trees. *Machine learning*, 1:81–106, 1986.

[9] Wenbin Zhang and Eirini Ntoutsi. Faht: an adaptive fairness-aware decision tree classifier. *arXiv preprint arXiv:1907.07237*, 2019.