Evolving Fair and Accurate Classification Models

[AUTHOR INFORMATION ELIDED FOR PEER REVIEW]

Machine learning models have traditionally been trained to maximize accuracy. Recently, concerns about fairness have motivated the development of multi-objective training methods, which optimize for both accuracy and fairness. Here, we present a multi-objective evolutionary algorithm that produces fair and accurate classification models. We compare our models to the models produced by SciKitLearn, and show that our models have much higher fairness at only a slight cost to accuracy. Furthermore, evolutionary algorithms produce the entire Pareto front, not just one model. Using this Pareto front, we explore the fairness-accuracy tradeoff in a way that would be impossible for many non-evolutionary methods.

ACM Reference Format:
[Author information elided for peer review]. 2020. Evolving Fair and Accurate Classification Models. 1, 1 (January 2020), 9 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Machine learning models play a role in decisions that alter the life prospects of their subjects. At the same time, these models are often biased [1], and because making biased decisions is unethical and illegal [2], it is important that machine learning practitioners actively reduce bias in their models [4]. Bias in algorithmic decision-making has many sources: a model can end up with bias if it is trained on historical data from humans who made biased decisions [5], a misrepresentative subset of data that paints a biased picture [16], or data with more datapoints for certain subgroups of the population [17].

Practitioners can identify biases in their models by applying metrics of fairness to their models and datasets. However, it can be difficult for practitioners to choose which metrics to use – the AIF360 project has over 70, none of which are intrinsically better than others [3]. Many of these metrics are contradictory, that is, an increase in one may necessitate a decrease in another [9]. For example, a definition of fairness could measure whether the model's accuracy is the same for different subgroups. However, this definition of fairness conflicts with accuracy when one subgroup

Author's address: [Author information elided for peer review].

Manuscript submitted to ACM

is easier to make predictions for than another. If you maximize accuracy overall, you cannot also have equal accuracy for each subgroup.

Practitioners can mitigate biases by pre-processing input datasets to remove latent bias, post-processing the model's predictions to ensure the trained model is fair even if the data is biased, or training models using algorithms that take fairness as well as accuracy into account [1]. We propose a new way to train models that are both fair and accurate.

Training a fair and accurate model is a multi-objective optimization problem, because one must optimize simultaneously for accuracy and some definition of fairness. Typical measures of the quality of classification models, such as specificity and AUC, do not capture the fairness in the distribution of benefits to different classes, so fairness requires a second metric, in addition to the metric being used to measure accuracy. We use a multi-objective evolutionary algorithm to train fair and accurate models. First, we show that an evolutionary algorithm, optimizing for accuracy alone, can produce models with similar accuracy to benchmarks. We then show that when optimizing for fairness as well as accuracy, the fair models are only slightly less accurate than the unfair models. The full source code for this analysis is available online [1].

2 FAIRNESS METRICS

A fairness metric is a function of a model and a dataset that produces a real-valued output representing the degree of unfairness of the given model on the given dataset. For simplicity, we limit our models to produce binary classifications: "positive outcome" or "negative outcome". We also limit our datasets to only have two groups: "privileged" and "unprivileged." Fairness metrics, then, are measures of how fairly the positive and negative outcomes are distributed across the privileged and unprivileged classes [4]. We will examine four fairness metrics:

(1) *Disparate impact,* the ratio between the ratio of positive to negative predictions for the privileged group and the ratio of positive to negative predictions for the unprivileged group. Traditionally, disparate impact is greater than one if the privileged group has a higher disparate impact score, and less than one if the unprivileged group does. We chose to represent the ratio as the higher ratio divided by the lower one, so that the minimum score is 1, and inequality favoring the priviliged or unpriviliged groups causes the score to be greater than 1. This step ensures that minimizing the adjusted score will lead to unbiased models, instead of models biased towards the unpriviliged group.
(2) *False negative rate ratio,* the ratio between the false negative rate for the privileged group and the unprivileged group. The same ratio adjustment as for disparate impact was applied.
(3) *True positive rate ratio,* the ratio between the true positive rate for the privileged group and the unprivileged group. The same ratio adjustment as for disparate impact was applied.

[1] [Link elided for peer review]

(4) *Between-group generalized entropy*, the amount of information required to represent the inequality in the distribution of benefits between groups by the model. This is the only one of our chosen metrics where a score of 0 represents perfect fairness. We choose specifically the between-group Theil index, a special case of between-group generalized entropy. For a more in-depth description, see [15].

3 EVOLVING FAIR MODELS

Training a classification model that is both accurate and fair is a multi-objective optimization problem. Because there are multiple objectives, some of which are not differentiable, standard methods of fitting models such as gradient descent and convex optimization—even the multi-objective version of gradient descent proposed by [7]—which rely on a differentiable loss function will not work [18]. Furthermore, we are not trying to find the optimal point in Euclidean space by some function, we are trying to find an optimal model. We must find the optimal model in model-space, not the optimal point in Euclidean space. Evolutionary optimization approaches have proven effective under these constraints, as in [19] where decision trees were evolved with false negative and false positive rate as objectives. Evolutionary algorithms can also support optimizing models for multiple fairness metrics at once, an advantage over parameterized fairness/accuracy calculations like the ones performed in [8].

We used the open-source Evvo[2] framework to implement the evolutionary algorithm which trains our models. The evolutionary algorithm implemented by Evvo is asynchonous and parallel. It runs as follows:

(1) Generate a starting population of random solutions.
(2) Starts threads which asynchronously copy and modify solutions in the population. The modified solutions are added to the population.
(3) Start theads which asynchronously select random samples from the population and delete the solutions that are bad. We chose to delete dominated solutions from the sample. Dominated solutions are those which score worse than another solution on every objective. In our case, this means that a model that is both less accurate and less fair than another would be dominated.
(4) When some amount of time has elapsed, stop the system.

To evolve a decision tree with Evvo, one needs an initial population, modification operators, and criteria for when to stop. Our initial population consists of randomly generated full depth-five decision trees, trained by splitting greedily to minimize entropy. We implemented the modification operators introduced in [10]. These operators change the feature that a node examines, change the threshold of a node, swap the class that a given leaf predicts, change a leaf to a node, or change a node to a leaf. Following [13], we employ a "crossover" operator, which takes two decision trees

[2]https://github.com/evvo-labs/evvo

and swaps a random subtree of one with a random subtree of another. After running for a specified amount of time, the system stops and returns the Pareto front.

Machine learning practitioners often evaluate models on accuracy alone, as increasing the accuracy of a model can directly produce value for their employer [12]. We show that large increases of fairness are obtainable without sacrificing much accuracy, hopefully laying the groundwork for increased adoption of fair modeling practices.

We evaluated our model training method on four datasets: the German credit dataset, Taiwan credit dataset, and Adult Income dataset from [6] and the COMPAS dataset from [11]. All four datasets contain numeric attributes and a binary prediction task. For the credit dataset, the task is to predict whether the person described by the data point will default on a loan or not. For the income dataset, the task is to predict whether a person's income is over $50,000. For the COMPAS dataset, the task is to predict whether a criminal will recidivate. For all of these tasks, there are multiple privileged classes with an intersectional effect. We have chosen to measure bias only across gender in the first three datasets and on race in the COMPAS dataset, for economy of presentation and the simplicity of the resulting fairness metrics. Gender and race, respectively, are not included as inputs to the models being trained.

On each dataset, we will have to define the success of a classification model at the prediction task. To do so, we split each dataset into a training set, consisting of a random sample of 70% of the data, and a test set, consisting of the remaining 30% of the data. During evolution, we optimize for accuracy and fairness on the training set. At the end, when we evaluate the model's performance, we measure its accuracy and fairness on the test set. To quantify how difficult each dataset is to model, we present in Table 1 the accuracy and fairness achieved by some of our evolved decision trees, and compare them to decision trees produced by SciKitLearn [14]. Table 1 shows that the models produced by Evvo have worse accuracy than those produced by SciKitLearn, but they are much more fair (as measured by disparate impact).

4 RESULTS

Traditional model-training methods only produce one model, while evolutionary algorithms produce a Pareto front. Traditional methods require manual human exploration of the potential models, based on an incomplete understanding of the fairness/accuracy tradeoff space. However, evolutionary algorithms allow the human in charge of developing the model to see the Pareto front of many possible solutions. When examining the Pareto front, it is immediately clear how much an increase in accuracy or fairness would cost in the other metric. The entire search space is presented to the practitioner choosing which model to deploy, allowing for decisions to be made based on more than just a few hand-picked data points.

On the COMPAS dataset, it is clear that improving the fairness of models decreases the model's accuracy. The points that are in the bottom-left-most parts of Figures 1 and 2, corresponding to

Dataset	Our Acc.	Benchmark Acc.	Our DI	Benchmark DI
German Credit	.752	.766	1.01	1.08
Taiwan Credit	.817	.827	1.22	1.33
Adult Income	.804	.859	1.99	2.64
COMPAS	.663	.685	1.75	2.05

Table 1. Accuracy and disparate impact obtained by evolving models using false positive rate and false negative rate as objectives, compared to a SciKitLearn DecisionTreeClassifier trained with a RandomizedSearchCV. Our reported accuracy is the accuracy on the test set of the decision tree with the highest accuracy on the training set. While our models are less accurate than the SciKitLearn models, they are more fair (as measured by disparate impact).

the highest accuracy, are the ones with the most unfairness. However, models with near-perfect fairness only slightly underperform the best models. The results were qualitatively similar for the German dataset, where fair models only slightly lagged unfair models in accuracy.

Disparate impact and true positive rate ratio are both measures of inequality in the allocation of the positive label. For both measures, an increase in the overall amount of positive predictions allows for a higher value of inequality. Figure 3 shows the opposite effect, because the false negative rate ratio is bounded by the number of total negative predictions.

The value of the overall Theil index must be lower than the maximum entropy for three classes ($\log_2(3) \approx 1.585$) as it is equivalent to the entropy of a series of observations with only three values. And, as noted in [15], the between-group entropy is a small amount (often less than one percent) of the overall entropy when you have few groups, so the small values in Figure 4 are to be expected.

Finally, we will examine the maximum accuracy achieved by fair models trained in each of these cases. Table 2 shows the accuracy of our models on the COMPAS dataset, when bounded by some threshold of a fitness metric. The unfair model, trained solely for accuracy, achieved an accuracy of 0.684. There are clearly diminishing returns to decreasing fairness, as most of the accuracy benefits can be obtained at a relatively high level of fairness (namely, 1.1). In fact, a disparate impact of 1.1 falls within the "four-fifths rule" established by Title VII [2], while having an accuracy only 2% lower than the highest accuracy of any model we trained. So, ensuring the legality of a model may require only a 2% decrease in accuracy.

5 FURTHER WORK

As mentioned when introducing the datasets, each dataset has multiple protected subgroups, and we debiased against only one. Adding objectives for each protected feature may present a different picture. Another straightforward extension would be the application of this framework to different

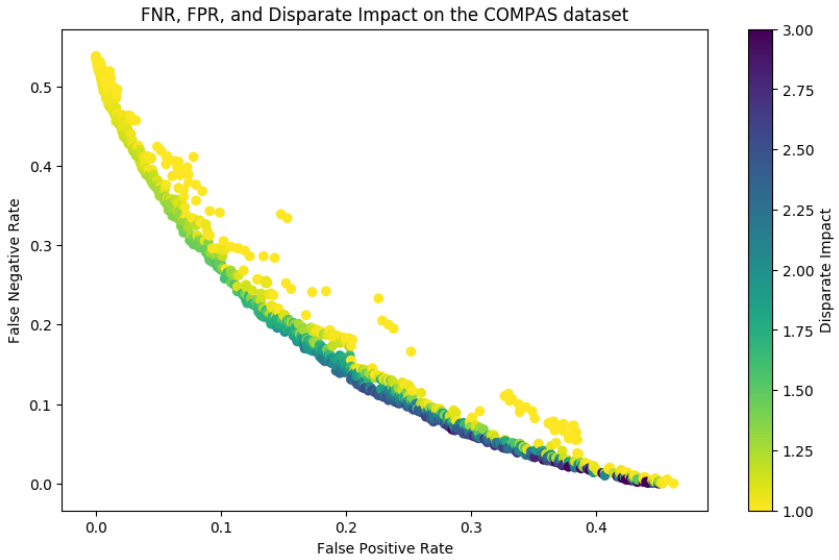

Fig. 1. False Negative Rate, False Positive Rate, and Disparate Impact on the COMPAS dataset. Higher false positive rates are correlated with higher disparate impact scores, and higher-accuracy models have worse disparate impact scores.

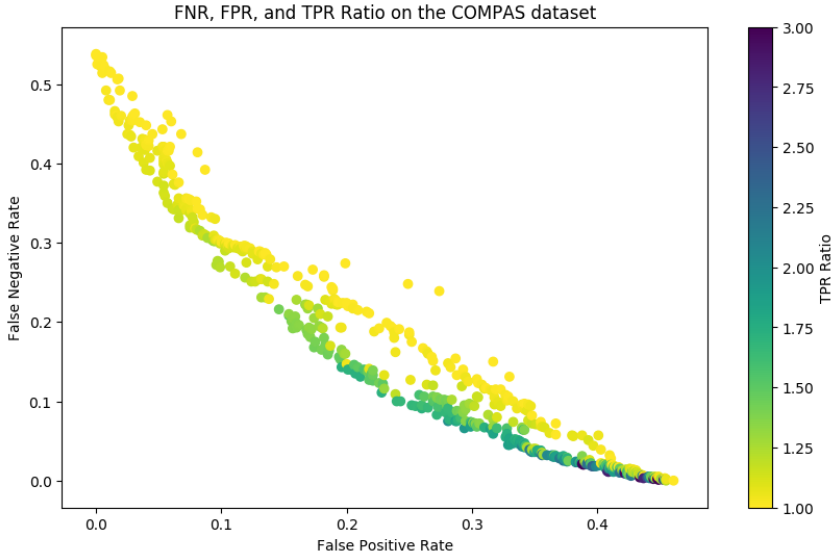

Fig. 2. False Negative Rate, False Positive Rate, and True Positive Rate Ratio on the COMPAS dataset. As above, the the true positive rate ratio is highest when the false positive rate is highest.

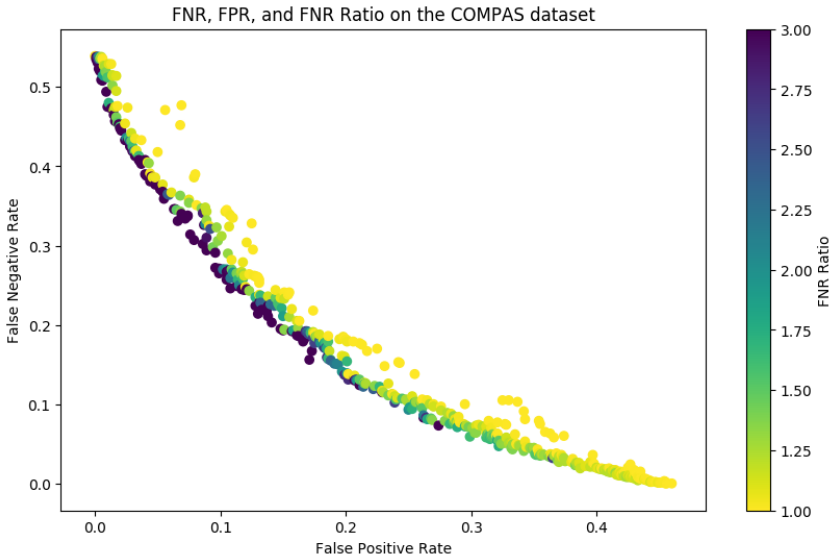

Fig. 3. False Negative Rate, False Positive Rate, and False Negative Rate Ratio on the COMPAS dataset. This figure shows the opposite trend as Figures 1 and 2, as false negative rate ratio is naturally correlated with false negative rate and not false positive rate.

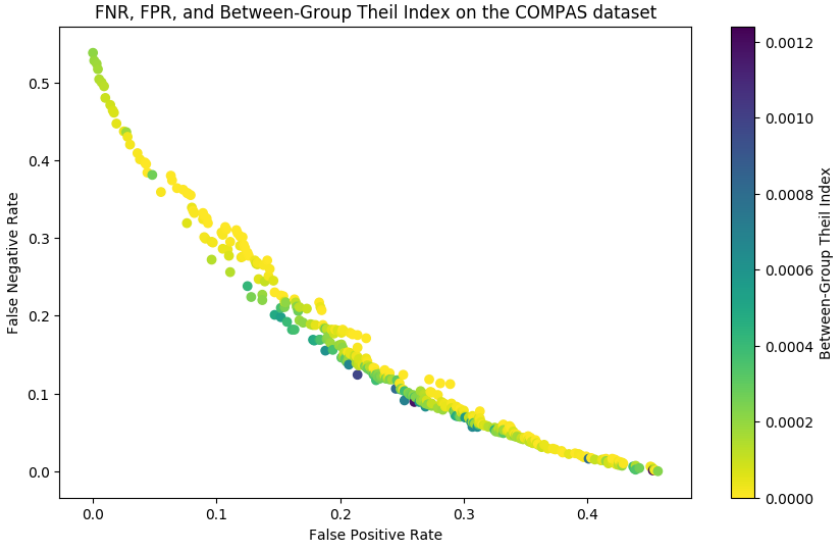

Fig. 4. False Negative Rate, False Positive Rate, and Between-Group Theil Index on the COMPAS dataset. Between-Group Theil Index is the only fairness metric that isn't correlated with overall false positive and negative rates.

Fairness Threshold	TPR Ratio	FNR Ratio	Disparate Impact
1.00	0.580	0.633	0.539
1.05	0.653	0.660	0.568
1.10	0.653	0.660	0.634
1.20	0.653	0.660	0.634
1.30	0.653	0.660	0.634
1.40	0.654	0.660	0.634
1.50	0.658	0.660	0.635
2.00	0.661	0.662	0.653
3.00	0.661	0.667	0.657

Table 2. Peak accuracy for models that achieved different levels of each fairness metric on the COMPAS dataset. Only models with fairness less than or equal to the fairness threshold are included in the accuracy calculation.

fairness metrics. In addition, the evolution of different types of models could be explored, to see if types other than decision trees may produce fairer, more accurate models. The AIF360 project provides pre- and post-processing steps for models, to reduce their bias. Evolution of models combine with pre- and post-processing might provide an additional way to reduce the cost of fairness.

6 CONCLUSION

We have shown that fair models can be trained with only a minor loss of accuracy. This result holds across multiple definitions of fairness and multiple datasets. In the face of ethical and legal concerns, this minimal tradeoff may motivate even purely profit-seeking companies to train and deploy fair models. Furthermore, we have used evolutionary computing not only as an optimization technique, but also to produce Pareto front for the exploration of different tradeoffs in fairness. We have explored the three-dimensional Pareto fronts between false negative, false positive, and different fairness metrics, as well as between pairs of fairness metrics, allowing an exploration of the entire accuracy/fairness tradeoff space.

REFERENCES

[1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks. (May 2016). https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[2] Solon Barocas and Andrew D. Selbst. 2016. Big Data's Disparate Impact. *California Law Review* (2016).

[3] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John T. Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. 2018. AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. *CoRR* abs/1810.01943 (2018). arXiv:1810.01943 http://arxiv.org/abs/1810.01943

[4] Reuben Binns. 2017. Fairness in Machine Learning: Lessons from Political Philosophy. *CoRR* abs/1712.03586 (2017). arXiv:1712.03586 http://arxiv.org/abs/1712.03586

[5] Toon Calders and Indre Zliobaite. 2013. *Why Unbiased Computational Processes Can Lead to Discriminative Decision Procedures*. 43–57. https://doi.org/10.1007/978-3-642-30487-3_3

[6] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository. (2017). http://archive.ics.uci.edu/ml

[7] Jean-Antoine Désidéri. 2012. Multiple-gradient descent algorithm (MGDA) for multiobjective optimization. *Comptes Rendus Mathematique* 350, 5 (2012), 313 – 318. https://doi.org/10.1016/j.crma.2012.03.014

[8] Sorelle A. Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P. Hamilton, and Derek Roth. 2019. A Comparative Study of Fairness-enhancing Interventions in Machine Learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. ACM, New York, NY, USA, 329–338. https://doi.org/10.1145/3287560.3287589

[9] Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent Trade-Offs in the Fair Determination of Risk Scores. *CoRR* abs/1609.05807 (2016). arXiv:1609.05807 http://arxiv.org/abs/1609.05807

[10] Marek Kretowski and Marek Grzes. 2005. Global learning of decision trees by an evolutionary algorithm. *Inform Process Security Syst* (01 2005), 401–410. https://doi.org/10.1007/0-387-26325-X_36

[11] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. 2016. *How We Analyzed the COMPAS Recidivism Algorithm*. Technical Report. ProPublica. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

[12] Ben Packer, M. Mitchell, Mario Guajardo-Céspedes, and Yoni Halpern. 2018. *Text Embeddings Contain Bias. Here's Why That Matters.* Technical Report. Google.

[13] A. Papagelis and D. Kalles. 2000. GA Tree: genetically evolved decision trees. In *Proceedings 12th IEEE Internationals Conference on Tools with Artificial Intelligence. ICTAI 2000*. 203–206. https://doi.org/10.1109/TAI.2000.889871

[14] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[15] Till Speicher, Hoda Heidari, Nina Grgic-Hlaca, Krishna P. Gummadi, Adish Singla, Adrian Weller, and Muhammad Bilal Zafar. 2018. A Unified Approach to Quantifying Algorithmic Unfairness: Measuring Individual & Group Unfairness via Inequality Indices. *CoRR* abs/1807.00787 (2018). arXiv:1807.00787 http://arxiv.org/abs/1807.00787

[16] Harini Suresh and John V. Guttag. 2019. A Framework for Understanding Unintended Consequences of Machine Learning. *CoRR* abs/1901.10002 (2019). arXiv:1901.10002 http://arxiv.org/abs/1901.10002

[17] Antonio Torralba and Alexei A. Efros. 2011. Unbiased Look at Dataset Bias. *Proc. of IEEE Computer Vision and Pattern Recognition, 2011* (2011), 1521–1528. https://ci.nii.ac.jp/naid/10030080158/en/

[18] Adrien Zerbinati, Jean-Antoine Desideri, and Régis Duvigneau. 2011. *Comparison between MGDA and PAES for Multi-Objective Optimization*. Research Report RR-7667. INRIA. 15 pages. https://hal.inria.fr/inria-00605423

[19] Huimin Zhao. 2007. A multi-objective genetic programming approach to developing Pareto optimal decision trees. *Decision Support Systems* 43, 3 (2007), 809 – 826. https://doi.org/10.1016/j.dss.2006.12.011 Integrated Decision Support.