Evolving Fair Models: Fair and Accurate Machine Learning Models with Multi-objective Evolutionary Computing

Julian F. Zucker julian.zucker@gmail.com
Khoury College, Northeastern University, Boston, MA

Clark Freifeld c.freifeld@northeastern.edu
Khoury College, Northeastern University, Boston, MA

Abstract

Machine learning practitioners must ensure that their models are fair, a difficult task in the face of many conflicting definitions of fairness. We present a multi-objective optimization technique for training classification models that allows practitioners to identify the tradeoff space between accuracy and fairness. We explore four different fairness metrics. After optimizing for false positive rate, false negative rate, and one fairness metric, we generate Pareto frontiers of possible models with varying accuracy/fairness tradeoffs. This Pareto frontier can be used by machine learning practitioners to identify how much accuracy must be sacrificed to achieve a specific level of fairness. After optimizing for false positive rate, false negative rate, and two fairness metrics, we look into the tradeoffs between different fairness metrics. Understanding which fairness metrics are contradictory to or synonymous with others reduces the number of metrics which must be examined to understand the overall fairness of a model. We examine the cost to accuracy of different levels of each fairness metric is computed, and see that the cost of greatly increasing fairness is small.

Keywords

Genetic algorithms, evolutionary computing, data mining, binary classification, decision tree, machine learning fairness.

1 Introduction

Machine learning models make decisions that alter the life prospects of their subjects. At the same time, these models are often biased, and because deploying biased models as decision-makers is unethical and illegal, it is important that machine learning practitioners take action to reduce bias in their models (Binns, 2017). Bias can be introduced when a model imitates the biases of the humans that created the data they are being trained on, is trained on a misrepresentative subset of data that paints a biased picture, or is given more data on subgroups of the population (Zerbinati et al., 2011). Practitioners can remove biases by preprocessing the input datasets to remove bias, building models taking fairness as well as accuracy in to account, and post-processing the model's predictions to ensure fairness. This paper will focus on the second method,

introducing a new way to train models that optimizes simultaneously for accuracy and fairness.

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 (Bellamy et al., 2018). Indeed, many of these metrics are contradictory, such that an increase in one may mathematically require a decrease in another (Kleinberg et al., 2016).

Training a fair and accurate model is a multi-objective optimization problem, where the objectives are accuracy and some definitions of fairness. Accuracy here refers to the rate of correct classifications in a hold-out dataset, with respect to the ground truth labels in the dataset. Fairness has multiple definitions. For example, a definition of fairness could measure how consistent the model's accuracy is for different subgroups. These two goals conflict when one subgroup is easier to make predictions for than another. If the easy subgroup is large, overall performance metrics will disregard weaknesses on smaller subgroups.

We chose evolutionary computing as our optimization approach. The results of evolutionary computing without adjusting for fairness are shown to be competitive with neural networks, and we show that, for our selected data sets, little accuracy is lost for large improvements in fairness. Furthermore, we show the tradeoffs required between each pair of fairness metrics, giving insight into which are correlated or anti-correlated in practice. The full source code for this analysis is available at [elided for anonymous peer review].

2 Fairness Metrics

A fairness metric is a function of a model and a dataset, that produces a real-valued output. For simplicity, we limit our models to produce binary classifications: "positive outcome" or "negative outcome," and our datasets to only have two groups: "privileged" and "unprivileged." Fairness metrics, then, are measures of how the positive and negative outcomes are distributed across the privileged and unprivileged classes. If you optimize for a fairness metric that doesn't properly reflect stakeholder values, your system will appear to be justified while being unjust. It is therefore crucial that the chosen fairness metrics capture the types of fairness that we value (Binns, 2017). We will explore four fairness metrics:

1. *Disparate impact,* the difference in the ratio of positive to negative predictions for the privileged group and the ratio of positive to negative predictions for the unprivileged group. While the ratio is normally conceived of as being greater than one if the privileged group has a higher true positive rate, and less than one is the unprivileged group does, we chose to represent the ratio as the higher true positive rate divided by the lower one. With this adjustment, the minimum score is 1, and inequality favoring the priviliged or unpriviliged groups causes the score to be greater than 1; this ensures that minimizing the adjusted score will lead to fair trees.

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 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 disparate impact was applied.

4. *Between-group generalized entropy,* the amount of inequality in the distribution of benefits between groups by the model, as measured by information-theoretic entropy. This is the only one of our chosen metrics where a score of 0 is perfect (representing equal distribution of outcomes). We choose specifically the between-group Theil index, a special case of between-group generalized entropy where α is 1 (Speicher et al., 2018).

3 Evolving Fair Models

Training a machine learning model that is both accurate and fair according to all the metrics listed above 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 (Désidéri, 2012) which rely on a differentiable loss function will not work (Zerbinati et al., 2011). 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 (Zhao, 2007) where decision trees were evolved with false negative and false positive rate as objectives. Evolutionary computing can also support optimizing models for multiple fairness metrics at once, an advantage over parameterized fairness/accuracy calculations like the ones performed in (Friedler et al., 2019).

We used the open-source Evvo framework[1] to perform our evolutionary computing. The core of generational evolutionary computation implemented by Evvo is:

1. Generate a starting population of random solutions.

2. Asynchronously, copy and modify solutions in the population.

3. Asynchronously, select samples from the population and delete the solutions that are bad. We chose to delete dominated solutions from the sample.

4. When some amount of time has elapsed, stop the system.

To evolve a decision tree with Evvo, you need an initial population, modification operators, and criteria for when to stop. Our initial population consists of randomly generated, full, depth-five decision trees. We implemented the mutation operators introduced in (Kretowski and Grzes, 2005). These operators change the feature that a node examines, change the threshold of a random node, swap the class that a given leaf predicts, change a leaf to a node, or change a node to a leaf. Following (Papagelis and Kalles, 2000), we employ a "crossover" operator, which takes two decision trees as input and produces a new tree, created by swapping 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 frontier. In contrast with running mutation operators

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

Dataset	Our Accuracy	Benchmark
German Credit	.760	0.792 (Abellán and Castellano, 2017) 0.792 (Bao et al., 2019) 0.787 (Shen et al., 2019)
COMPAS	.684	.71 (Larson et al., 2016)

Table 1: Accuracy obtained by evolving models using false positive rate and false negative rate as objectives. Our reported accuracy is the accuracy on the test set of the decision tree with the highest accuracy on the training set. While our accuracies are lower than the benchmarks, the purpose of this paper is not to evolve the most accurate decision trees, but to demonstrate the viability of evolving models for fairness.

a fixed number of times, time-based criteria ensure that increasing the computational complexity of our mutation operators is penalized, as those operators will take longer to run (Eiben and Smith, 2015).

4 Data

We evaluated our model training methods on two datasets: the German credit dataset (Dua and Graff, 2017) and the AIF360 pre-processed COMPAS dataset (Larson et al., 2016). Both datasets contain numeric attributes and a binary prediction task. For the German dataset, the task is to predict whether the person described by the data point will default on a loan or not. For the COMPAS dataset, the task is to predict whether a criminal will recidivise. For both of these tasks, there are multiple privileged classes with an intersectional effect. We have chosen to debias only gender in the German dataset and race in the COMPAS dataset, for economy of presentation and the simplicity of the resulting fairness metrics. Gender and race are two historical bases for systemic unfairness, though naturally there are many others.

On each dataset, we will have to define the success of a machine learning model at the prediction task. We do so by examining the model's accuracy and fairness on data it hasn't seen before. We first split each dataset into a training set, consisting of a random sample of 80% of the data, and a test set, consisting of the remaining 20% 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 their accuracy and fairness on the test set. To quantify how difficult each dataset is to model, we present in Table 1 accuracies achieved by evolving decision trees for accuracy alone. Table 1 shows that the models produced by evolutionary computing optimizing just for accuracy perform worse, but not too much worse, than those produced by alternative training regimes.

5 Results

Machine learning practitioners often evaluate models on accuracy alone, as increasing the accuracy of a model can directly produce profit (Packer et al., 2018). If we can show that large increases of fairness are obtainable without giving up much accuracy, perhaps we can increase the adoption of fair machine learning models. To further this

aim, we have provided a display of the Pareto frontier for each of our four metrics against accuracy in Figures 1 through 4. To generate these visualizations, models were evolved for five minutes, using false negative rate, false positive rate, and a fairness metric as the objectives. The value of each fairness metric is capped at three, to prevent extreme points from distorting the color scale. Raw data is available at [elided for anonymous peer review].

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 each figure, corresponding to 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. Note that the most accurate models are the ones closest to the line $y = x$, while the models that maximize accuracy with a given ratio of false negative to false positive are the ones along the FNR/FPR pareto frontier.

Figures 1 and 2 look quite similar, because 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. This effect is shown by the darker areas in the bottom right of the graph, where there are more positive predictions (and thus a higher false positive rate). 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 (Speicher et al., 2018), 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.

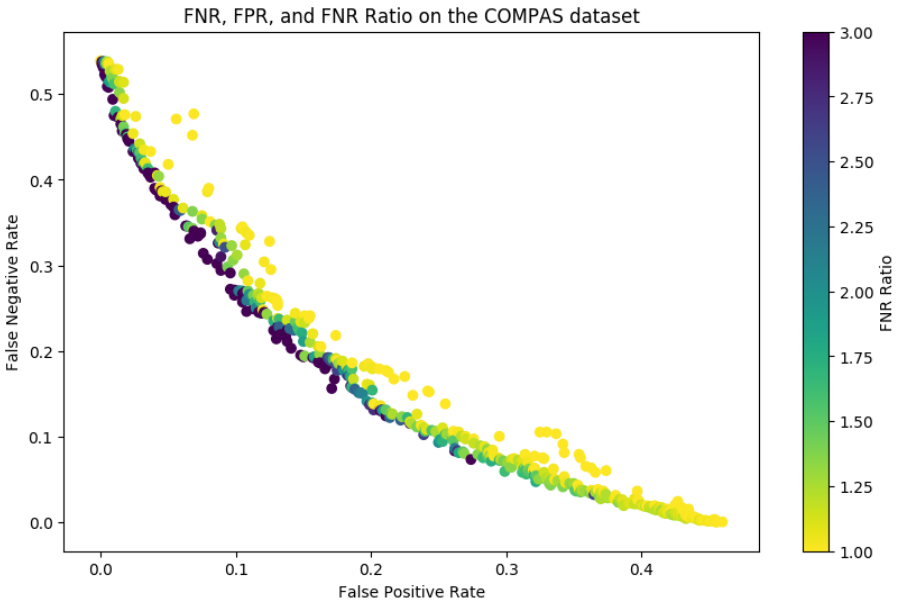

Figure 1: False Negative Rate, False Positive Rate, and False Negative Rate Ratio on the COMPAS dataset. Higher false negative rates are correlated with higher false negative rate ratios, and being closer to the Pareto frontier for false positive rate and false negative rate increases false negative rate ratio.

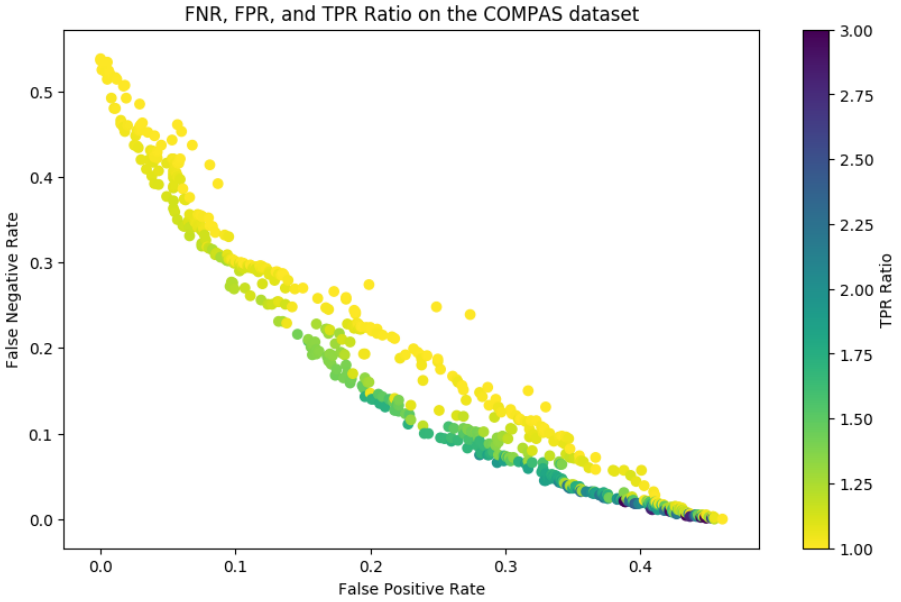

Figure 2: Similar to above, but the true positive rate ratio is highest when the false positive rate is highest.

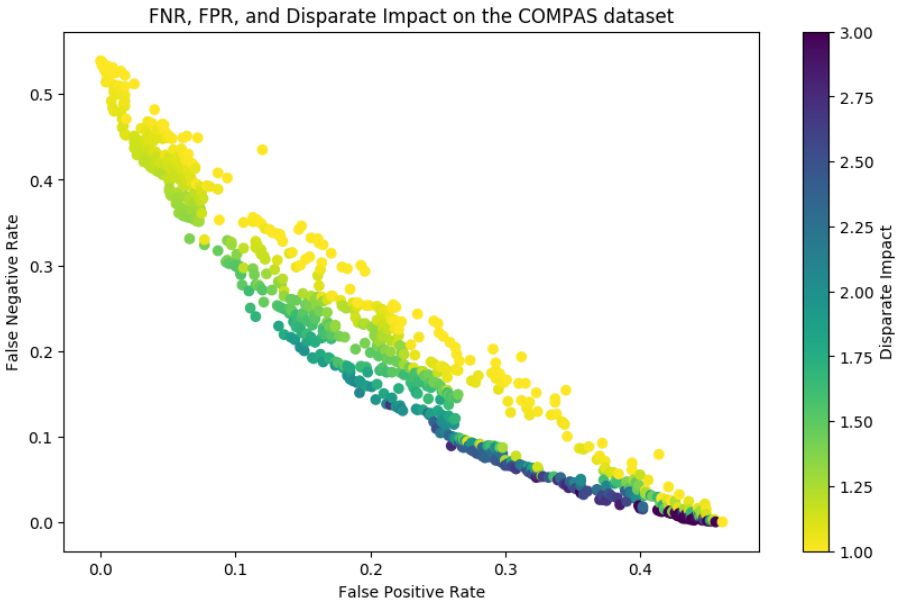

Figure 3: Similar to Figure 1. Note that the color gradient is similar to true positive rate ratio, as both fairness metrics are about inequality in allocation of positive labels. Disparate Impact has a much wider spread than TPR Ratio and FNR Ratio, because there are more positive label predictions than true positives or false negatives.

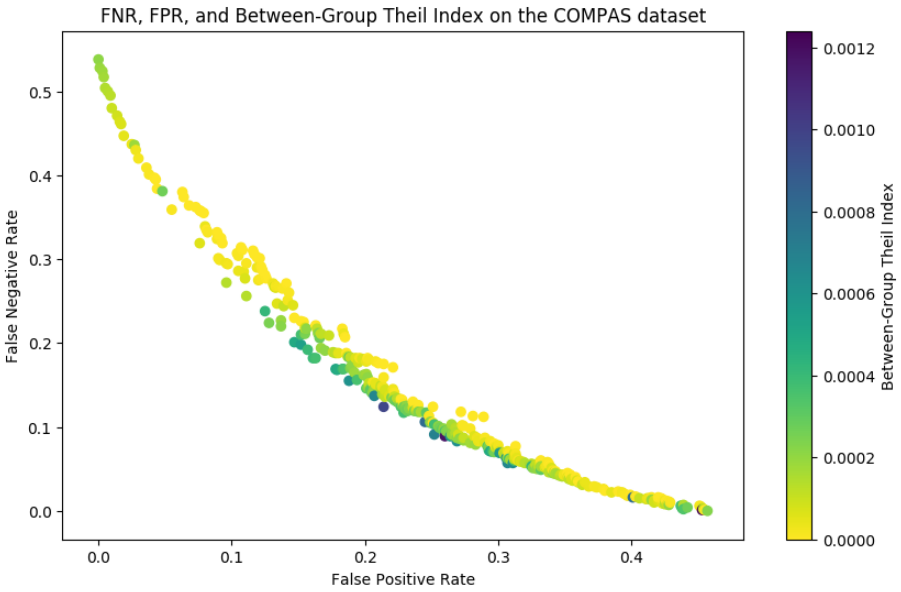

Figure 4: Between-Group Theil Index is the only fairness metric that isn't correlated with overall false positive and negative rates.

In Figures 5 and 6 we present the Pareto frontier for two pairs of objectives. The rest of the pairs of objectives are available in the Appendix. These results were obtained after five minutes of running Evvo, using false positive rate, false negative rate, and a given pair of metrics as the objectives. Examination of these Pareto frontiers reveals which metrics conflict with each other (for example, Between-Group Theil Index and Disparate Impact), and which are obtainable at the same time (for example, TPR Ratio and Disparate Impact). Between-Group Theil Index and Disparate Impact conflict because the Theil index requires that benefits be exactly evenly apportioned between groups, regardless of the labels on the data points in the test dataset, while the disparate impact measure requires the proportion of positive predictions for each group to scale with the proportion of positive labels in the test dataset for that group. True Positive Rate Ratio and Disparate Impact, on the other hand, are both measures of the appropriateness of benefit distribution, so they both tend to be satisfied at the same time.

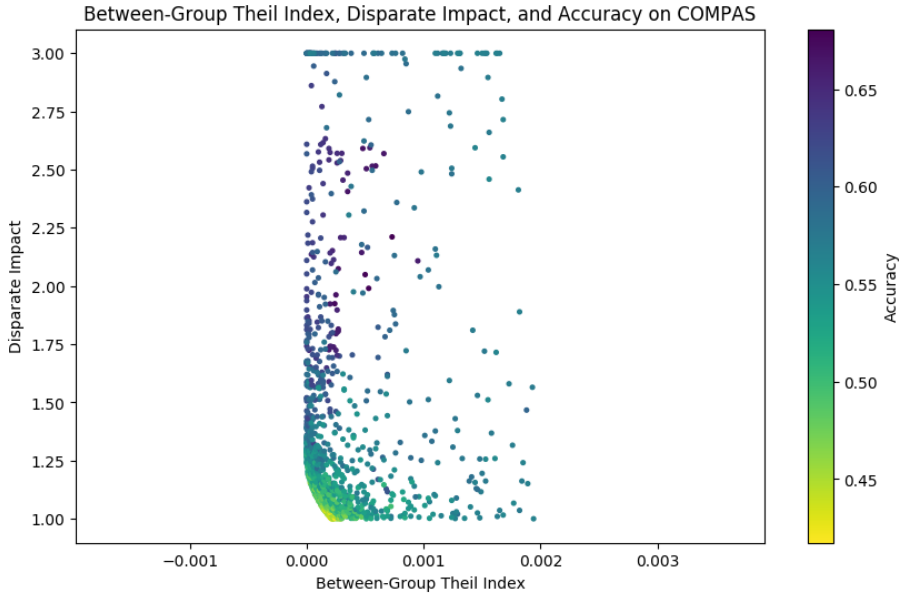

Figure 5: Theil Index and Disparate Impact are - anti-correlated, as evidenced by the lack of points in the bottom-left corner. It is also clear from this figure that accuracy suffers as these fairness metrics increase.

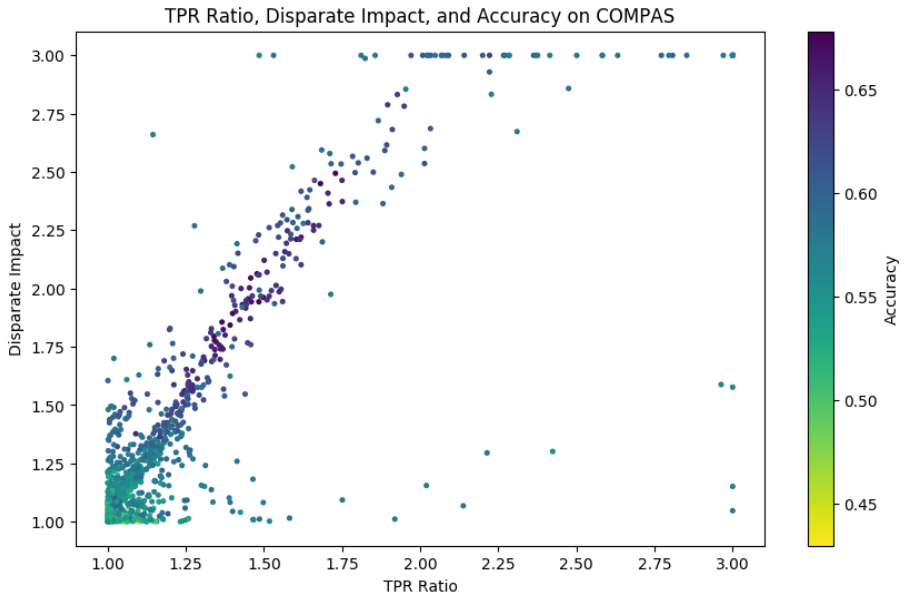

Figure 6: Disparate Impact and TPR Ratio are correlated, which makes intuitive sense as both measure inequality in distribution of positive labels. As with Figure 5, accuracy is lower when the models are more fair, but the effect is lesser here.

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. Recall that the unfair model, trained solely for accuracy, achieved an accuracy of 0.684. There are clearly diminishing returns to increasing unfairness, as most of the accuracy benefits can be obtained at a relatively low level of unfairness (1.1). In fact, a Disparate Impact score of 1.1 falls within the "four-fifths rule" established by Title VII, while having an accuracy only 2% lower than the accuracy of the highest accuracy model we trained (Barocas 2016).

Traditional methods of training models only produce one model, while evolutionary computing produces a Pareto frontier. So, instead of requiring manual human exploration of the potential models, based on an incomplete understanding of the fairness/accuracy tradeoff space, the human in charge of developing the model can see the Pareto frontier of many possible solutions. When examining the Pareto frontier, it is immediately obvious 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.

6 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 ap-

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 achieve 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.

plication of this framework to different 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 with pre- and post-processing might provide an additional way to reduce the cost of fairness.

7 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 choose fair models. Furthermore, we have used evolutionary computing not only as an optimization technique, but also to produce Pareto frontiers for the exploration of different tradeoffs in fairness. We have explored the three-dimensional Pareto front between false negative, false positive, and fairness, as well as between two different definitions of fairness, allowing an exploration of the entire accuracy/fairness tradeoff space.

References

Abellán, J. and Castellano, J. G. (2017). A comparative study on base classifiers in ensemble methods for credit scoring. *Expert Systems with Applications*, 73:1 – 10.

Bao, W., Lianju, N., and Yue, K. (2019). Integration of unsupervised and supervised machine learning algorithms for credit risk assessment. *Expert Systems with Applications*, 128:301 – 315.

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K. N., Richards, J. T.,

Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., and Zhang, Y. (2018). AI fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *CoRR*, abs/1810.01943.

Binns, R. (2017). Fairness in machine learning: Lessons from political philosophy. *CoRR*, abs/1712.03586.

Dua, D. and Graff, C. (2017). UCI machine learning repository.

Désidéri, J.-A. (2012). Multiple-gradient descent algorithm (mgda) for multiobjective optimization. *Comptes Rendus Mathematique*, 350(5):313 – 318.

Eiben, A. E. and Smith, J. E. (2015). *Introduction to Evolutionary Computing*. Springer Publishing Company, Incorporated, 2nd edition.

Friedler, S. A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E. P., and Roth, D. (2019). A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 329–338, New York, NY, USA. ACM.

Kleinberg, J. M., Mullainathan, S., and Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *CoRR*, abs/1609.05807.

Kretowski, M. and Grzes, M. (2005). Global learning of decision trees by an evolutionary algorithm. *Inform Process Security Syst*, pages 401–410.

Larson, J., Mattu, S., Kirchner, L., and Angwin, J. (2016). How We Analyzed the COMPAS Recidivism Algorithm. Technical report, ProPublica.

Packer, B., Mitchell, M., Guajardo-Céspedes, M., and Halpern, Y. (2018). Text embeddings contain bias. here's why that matters. Technical report, Google.

Papagelis, A. and Kalles, D. (2000). Ga tree: genetically evolved decision trees. In *Proceedings 12th IEEE Internationals Conference on Tools with Artificial Intelligence. ICTAI 2000*, pages 203–206.

Shen, F., Zhao, X., Li, Z., Li, K., and Meng, Z. (2019). A novel ensemble classification model based on neural networks and a classifier optimisation technique for imbalanced credit risk evaluation. *Physica A: Statistical Mechanics and its Applications*, 526:121073.

Speicher, T., Heidari, H., Grgic-Hlaca, N., Gummadi, K. P., Singla, A., Weller, A., and Zafar, M. B. (2018). A unified approach to quantifying algorithmic unfairness: Measuring individual & group unfairness via inequality indices. *CoRR*, abs/1807.00787.

Zerbinati, A., Desideri, J.-A., and Duvigneau, R. (2011). Comparison between MGDA and PAES for Multi-Objective Optimization. Research Report RR-7667, INRIA.

Zhao, H. (2007). A multi-objective genetic programming approach to developing pareto optimal decision trees. *Decision Support Systems*, 43(3):809 – 826. Integrated Decision Support.

8 Appendix

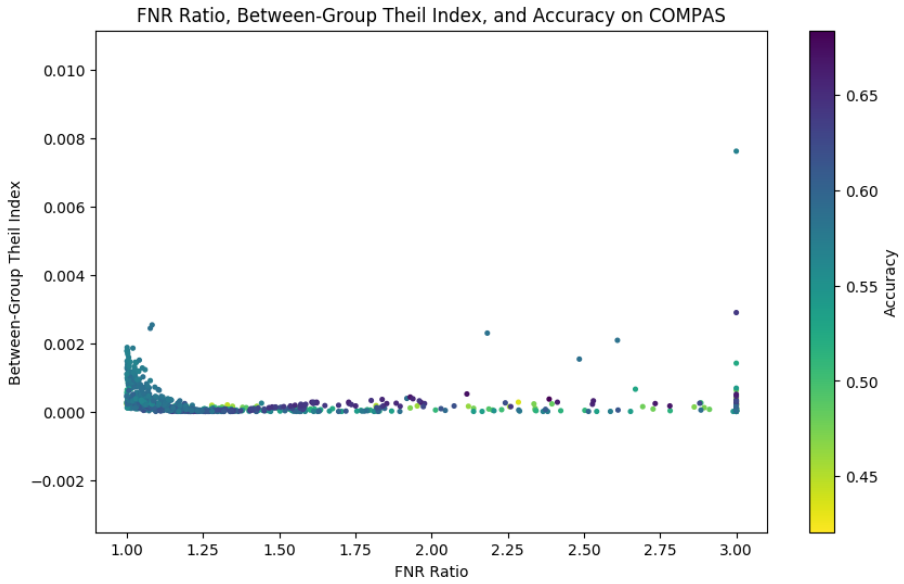

Figure 7: Between-Group Theil Index and False Negative Rate Ratio are can be optimized jointly, with little cost to predictive accuracy.

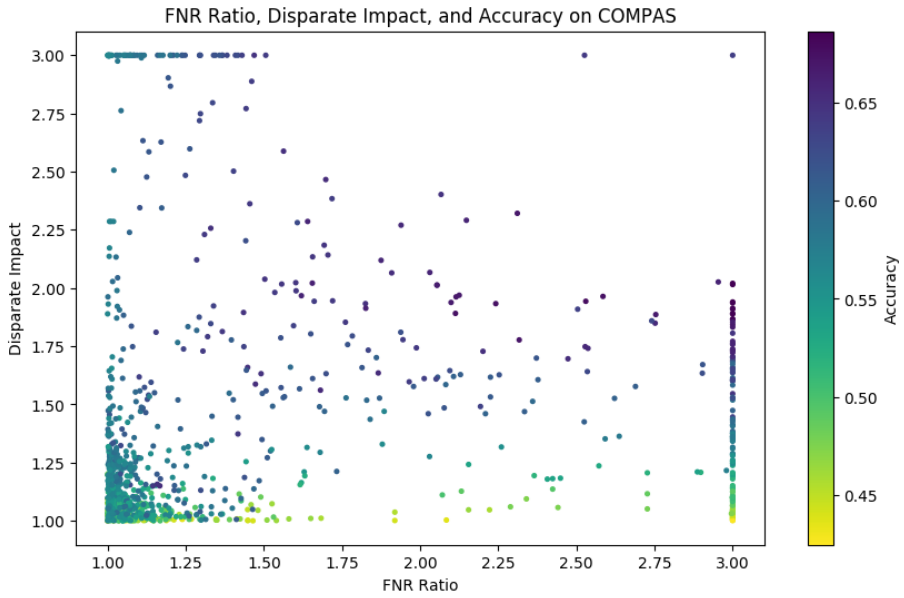

Figure 8: Decreasing Disparate Impact has a large influence on accuracy, while False Negative Rate Ratio has less of an affect.

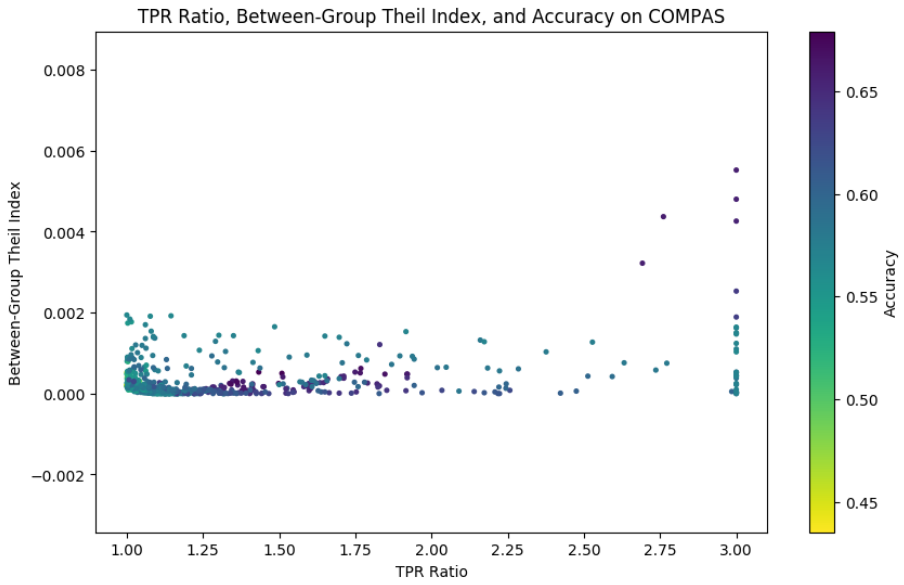

Figure 9: As with False Negative Rate Ratio in Figure 7, accuracy is relatively high, and models can be fair by both metrics.

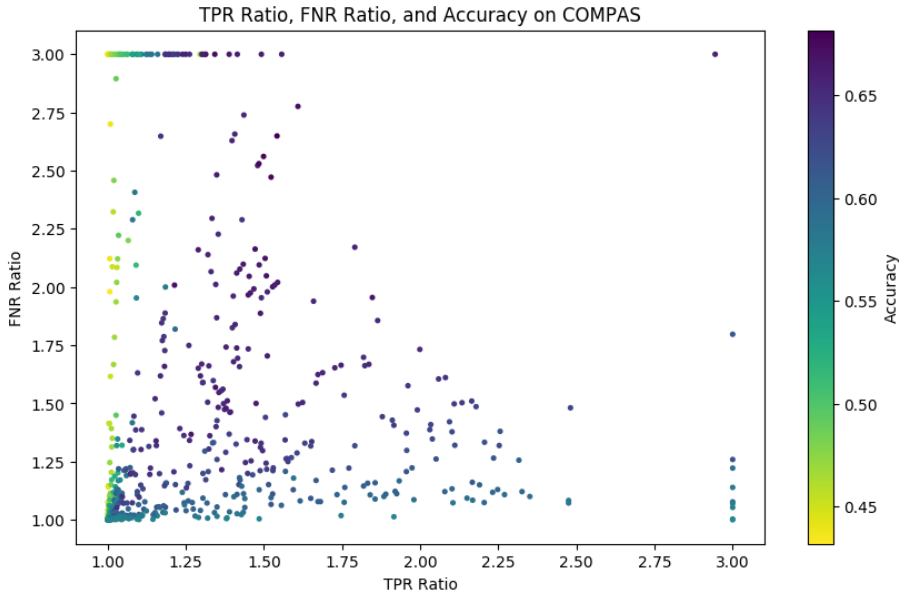

Figure 10: Here, the cost of True Positive Rate Ratio in accuracy is very high, although some points quite near (1,1) represent models that are incredibly fair, and still relatively accurate.