How to be Fair? A Discussion and Future Perspectives

Marco Favier[1], Toon Calders[1]

[1]University of Antwerp, Middelheimlaan 1, 2020 Antwerp, Belgium

Abstract

In our previous work titled "How to be fair? A study of label and selection bias," we discussed how the interaction between bias and fairness can yield fruitful insights for fairness research. We considered the scenario where an initially unobservable fair distribution becomes corrupted by bias, leading to an observable unfair distribution. By employing various fairness definitions and types of bias, we derived valuable mathematical conditions and properties that the observable distribution must adhere to. This would allow practitioners to better understand bias and mitigate its effect in their models. In this paper, we delve into the significance of these findings, address their limitations, and explore potential future research directions.

Keywords

Algorithmic fairness, Ethical AI, Classification,

1. Introduction

The fairness-accuracy trade-off [1, 2, 3] is frequently misunderstood in fairness literature. Rather than being viewed as a simple emergent phenomenon, as it should be, it is often portrayed as an insurmountable conflict between two ethical principles. While it is reasonable to expect a decrease in a classifier's accuracy when adding a fairness constraint, it's incorrect to assume that sacrificing accuracy is necessary to meet fairness requirements. Often, the trade-off is perceived as an unavoidable compromise, where the fairness of the classifier is exchanged for its accuracy, creating a false dichotomy between morality and performance quality. This line of thinking fosters the notion that fairness is a luxury reserved for situations where accuracy is not paramount. This dangerous perspective can lead to justifications for the absence of fairness in a classifier.

Furthermore, the ethical interpretation of the fairness-accuracy trade-off is inherently flawed, since it is self-contradictory in nature. If we acknowledge the necessity of implementing fairness interventions on our data, it implies that due to bias we have reason to believe that the data may contain artifacts that could hurt the performance of a classifier trained on them. We've already deemed our data flawed. It is then naive to expect that the accuracy of the classifier will not decrease, since the effort to make the data suitable for classification will necessarily shift the data distribution to a fairer one. If we pursue fairness, it is because we believe our data to be biased, and if we believe our data to be biased, we should also believe they cannot provide a good measure of the quality of the classifier. In reality, the trade-off should be considered as an

EWAF'24: European Workshop on Algorithmic Fairness, July 01–03, 2024, Mainz, Germany

*Corresponding author.

marco.favier@uantwerpen.be (M. Favier)

0000-0003-4676-5896 (M. Favier); 0000-0002-4943-6978 (T. Calders)

empty signifier: no meaning is conveyed by the trade-off, which should only be regarded as a numerical fact. In recent years, an increasing number of fairness researchers have endeavored to critically address the misconception surrounding the fairness-accuracy trade-off by highlighting that when bias is introduced into the data, the accuracy of a fair classifier on the unbiased data outperforms the accuracy of a fairness-agnostic classifier [4, 5]. In this regard, studying the effect of different biases on the probability distribution of the data is of primary importance, in order to develop effective fairness techniques capable of removing bias from the data.

Despite this, the literature still lacks a clear understanding of the relationship between bias and fairness. Current fairness measures only acknowledge the presence of bias, but they are unable to comprehend its nature or provide a clear path to remove it. This could pose potential challenges when attempting to enforce fairness in real-world applications.

For instance, when considering the fairness measure "Demographic Parity", it's clear that the objective of the metric is to ensure that the probability of being classified as positive is the same for all sensitive groups.

However, it remains unclear whether the distribution of positive labels achieved through a fairness intervention enforcing demographic parity is indicative of the successful elimination of societal bias or merely an instance of affirmative action. In other words, it is unclear if the new distribution is unbiased or if it is just a different kind of bias. We believe that studying the interaction between bias and fairness should be a priority for the fairness community, and our previous work is a step in this direction.

2. How to be fair? A study of label and selection bias

In our previous work "How to be fair? A study of label and selection bias" [1], we investigated how bias interacts with fair distributions and whether fair intervention techniques could effectively remove bias from data. Our investigation begins with the notion of an ideal world where the data collected from it are inherently fair by default. At the same time, for our available data, we assumed they are a distorted version of the data sampled from the ideal world. This distortion is caused by bias, which obfuscated the original data, resulting in two distinct distributions: an observable unfair distribution and a no-longer-observable fair distribution. Moreover, in our work, we assume that every classifier is able to learn the observable biased distribution from the available data and that the score assigned by the classifier corresponds to the unfair conditional probability of the label for a given data point. In technical terms, we assume no epistemic uncertainty.

Under these assumptions, asking whether fairness interventions are able to remove bias can be rephrased as whether fairness interventions manage to correct a classifier and align its score with the fair unobservable conditional probability. To mathematically formalize this, we consider a binary random variable B representing bias, independent of the non-sensitive features X, given the binary label Y and the sensitive attribute A. In layman's terms, for each data point in our initially fair dataset, we flip a weighted coin to decide whether to introduce bias, where the weight of the coin, i.e. the probability of bias, depends only on the sensitive attribute and the original label.

In this paper we analyze two types of biases: label bias and selection bias. Label bias arises when the assigned label for an individual does not accurately represent the label they should

	Demographic Parity	We're all equal
Label Bias	Necessary conditions on $P_U(Y\|X,A)$ $P_F(Y\|X,A)$ satisfies fairness	Necessary conditions on $P_U(Y\|X,A)$ $P_F(Y\|X,A)$ satisfies fairness
Selection Bias	Not detectable from $P_U(Y\|X,A)$ $P_F(Y\|X,A)$ doesn't satisfy fairness	Necessary conditions on $P_U(Y\|X,A)$ $P_F(Y\|X,A)$ satisfies fairness

Table 1
Combinations of bias and fairness, $P_U(Y|X,A)$ is the unfair observable distribution, $P_F(Y|X,A)$ is the fair unobservable distribution.

have been assigned, whereas selection bias occurs when certain individuals are not represented in the dataset, resulting in a non-representative sample of the population.

To study these biases, a mathematical model is needed. We adopted the following: for label bias, when a data point needs to be biased, i.e. when it loses the coin flip, we change the assigned binary label; for selection bias, we instead remove the data point entirely.

We then examined the impact of these biases on two specific fairness definitions: "Demographic Parity" and "We're all equal". Demographic parity, as previously mentioned, means that the distribution of the initial label is the same for both sensitive groups, formally $Y \perp\!\!\!\perp A$. "We're all equal", on the other hand, requires that the distribution of the initial label depends solely on the non-sensitive features, formally $Y \perp\!\!\!\perp A|X$. Based on these notions, our research question now is:

> Consider a distribution that satisfies either demographic parity or "we're all equal," which is transformed by either label or selection bias. Is it possible to detect or discern the bias based on a sample of the altered distribution? And if so, is there a fairness intervention that can counteract it?

Everything combined, it means there exist four possible combinations of bias and fairness we can study, as shown in Table 1.

Having a clear mathematical framework proved to be a fruitful choice for our research, as we were able to prove multiple statements about each of the four combinations. In particular, what became evident from our research is that in 3 out of 4 combinations, when bias has influenced the distribution, it is possible to confirm it purely by looking at the conditional distribution of the label, since there are strict conditions that the observable distribution must satisfy. Moreover, since different conditions are satisfied for different combinations, it is possible to discern the nature of the bias.

This result has positive implications for the field of fairness, as it suggests that, at least theoretically, the data itself could assist in determining the best course of action. This is because certain choices might already be excluded when they do not align with the data distribution, alleviating practitioners from the burden of blindly applying fairness techniques and helping them to take a more informed decision.

On the other hand, the remaining combination, selection bias combined with demographic parity, yields an even more interesting and somewhat counter-intuitive result. In this case, we demonstrated that the original fair distribution never satisfies demographic parity when computed on the available biased data. In other words, the correct fair distribution does not appear completely fair according to the used measure. This simple condition has a significant impact

on the application of many classifiers in terms of fairness. Indeed, many fairness techniques are designed to enforce demographic parity on the available data, which automatically fails to properly remove selection bias from the data as they cannot retrieve the original fair distribution. When bias influences the data, even the fairness measures themselves can become biased.

This also formally supports one of the most common criticisms of fairness, namely that we need a better understanding of the effect of fairness techniques on data.

3. Limitations, Future work and Conclusions

Despite the merits of our work, we acknowledge that it also has its limitations. Firstly, even though we utilized our theoretical framework to explain certain experimental results from the literature [4] and expanded upon them with our own, our work remains primarily theoretical in nature. Further exploration into practical applications is still necessary. Secondly, all our proofs rely on multiple, and often strong, assumptions, which would be interesting to generalize. In particular, the assumption that the bias is independent of the features is unrealistic.

In reality, it is often the case that bias may be influenced by both the features and the sensitive attributes. For instance, non-sensitive features that correlate with being part of a minority might inadvertently trigger confirmation bias in individuals with prejudices. This could subsequently exacerbate discrimination against certain groups of people, thereby perpetuating bias. A person with a foreign-sounding name might experience more rejections than a compatriot with a more common name when seeking employment. Police over-patrolling black neighborhoods gathers more data on residents of those neighborhoods than on black individuals residing in predominantly-white areas. In both cases, the bias is influenced by the features; label bias for the first example and selection bias for the second.

We believe this to be the most important point of improvement in our work. To overcome this limitation, in our future work we will explore what happens when the features are allowed to influence bias in our definitions. However, we also need to limit the extent to which the features influence the bias. This enables us to relax the independence assumption without altering the definitions of bias established thus far. We believe this to be a reasonable approach, as it allows us to still apply mathematical rigorousness when defining different kinds of biases.

Here we show three possible approaches: one based on differential privacy, the second on the Hirschfeld–Gebelein–Rényi coefficient, and the third is a direct chance-constraint on the probability distribution.

1. $P(B|X, A, Y) \leq e^{\varepsilon} P(B|A, Y)$
2. $\mathrm{HGR}(B, X|A, Y) := \sup \mathrm{corr}(f(B), g(X)|A, Y) < \varepsilon$ for functions f and g.
3. $P\left(|P(B = 1|X, A, Y) - P(B = 1|A, Y)| \geq \delta\right) \leq \varepsilon$

All these definitions are suitable, albeit with different levels of generality and ease of implementation. They can further improve our understanding of the relationship between bias and fairness and help us develop more effective fairness interventions. We believe our work is a step in the right direction: a new research line to connect bias types and data assumptions to measurable properties in observed data. The implications are both theoretical and practical, leading to a better understanding of fairness measures and better tools to guarantee fair AI.

References

[1] M. Favier, T. Calders, S. Pinxteren, J. Meyer, How to be fair? a study of label and selection bias, Machine Learning 112 (2023) 5081–5104.

[2] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, A. Huq, Algorithmic Decision Making and the Cost of Fairness, in: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '17, Association for Computing Machinery, New York, NY, USA, 2017, pp. 797–806. URL: https://doi.org/10.1145/3097983.3098095. doi:10.1145/3097983.3098095.

[3] A. K. Menon, R. C. Williamson, The cost of fairness in binary classification, in: Proceedings of the 1st Conference on Fairness, Accountability and Transparency, PMLR, 2018, pp. 107–118. URL: https://proceedings.mlr.press/v81/menon18a.html, iSSN: 2640-3498.

[4] M. Wick, J.-B. Tristan, et al., Unlocking fairness: a trade-off revisited, Advances in neural information processing systems 32 (2019).

[5] D. Lenders, T. Calders, Real-life performance of fairness interventions-introducing a new benchmarking dataset for fair ml, in: Proceedings of the 38th ACM/SIGAPP symposium on applied computing, 2023, pp. 350–357.