Technical Perspective: Database Repair Meets Algorithmic Fairness

Lise Getoor
UC Santa Cruz, USA

There has been an explosion of interest in fairness in machine learning. In large part, this has been motivated by societal issues highlighted in a string of well publicized cases such as gender biased job recommendation and racially biased criminal risk prediction algorithms. Both the recognition of the potential disparate impacts of machine learning due to historical bias in the data and the realization of how algorithmic decision making can exaggerate existing structural inequities has become increasingly well known.

This has spawned a growing body of work that examines fairness in ML [1]. From a theoretical perspective, it has opened a pandora's box of new fairness measures, impossibility results, and optimization strategies. However, this line of work has faced criticism. First, the notion that there is any one correct societal fairness definition, and the framing of ML fairness as a simple optimization problem, is suspect. Second, on technical grounds, unless one takes into account the underlying causal structure in the domain, there is no way to untangle, simply from data, whether the data is biased (and hence an algorithm trained on it is fair or not).

The paper "Database Repair Meets Algorithmic Fairness" by Babak Salimi, Bill Howe and Dan Suciu addresses this second criticism directly, and, I would argue, by generalizing the problem setting, they also address the first criticism. Furthermore, they introduce a refreshing database perspective on the problem. The lovely thing about this paper is that it tackles an important real-world issue, offers deep technical contributions, and includes convincing empirical results. Few papers are able to achieve all this, and none that I can think of do it as nicely and concisely.

First, it's useful to review Simpson's paradox, the well-known statistical phenomena that statistical correlations may reverse themselves depending on how data is aggregated. Within the fairness setting, we are interested in whether there is a dependence between a sensitive or (legally) protected attribute (such as gender, race or religion), and a decision outcome (such as admissions, hiring, credit or parole). If that dependence can reverse itself when we condition on another variable, such as age, then making conclusions about fairness will be difficult! Luckily if we have additional information about the underlying causal structure in the domain to reason about confounders we can unpack the correlations between protected attributes and outcomes. Pearl, in a long line of foundational work has developed a calculus of causation that enables one to translate between statistical statements and causal statements in a principled manner [2, 3]. With these tools in hand, when the full causal model is available, it allows us to determine whether there is an inappropriate dependency between a sensitive attribute and a decision.

However, this is a strong requirement. Salimi et al. use Pearl's causal framework as the foundation for a general and flexible construction for introducing admissible and inadmissible attributes while relaxing the requirement of having the full causal model available. Next, the authors draw an elegant connection between causal modeling and database theory to transform the problem of removing bias in data into a database repair problem. They show how to map causal interventions from statistical conditional independence constraints into multi-valued dependencies that should hold in the data. To ensure the statistical independencies required for fairness hold, they generate samples matching the empirical distribution as closely as possible and apply techniques from database repair to modify the data such that independences are satisfied. (Interestingly, this result can be used in *any* setting where one wishes to mix desired interventions and distributional constraints with empirical information.) The authors suggest that the repaired training data "can be seen as a sample from a hypothetical fair world".

To do this, they introduce a notion of justifiable fairness and prove that for a classifier to be justifiably fair, it is sufficient that the outcome variable is conditionally independent of the inadmissible attributes given the admissible attributes. Next, they show how to transform this requirement on a classifier into an integrity constraint on the training data! The paper's contributions include correctly setting up the theoretical machinery to make this translation between probability distributions and databases. While the high level intuition is simple, the details are quite non-trivial.

All in all, this is an important paper, and has something for everyone—real-world impact, theoretical results that bridge causal modeling and database theory, all in an elegant and well-written package.

1. REFERENCES

[1] S. Barocas, M. Hardt, and A. Narayanan. *Fairness and Machine Learning*. 2019. `http://www.fairmlbook.org`.
[2] J. Pearl. *Causality*. Cambridge University Press, 2009.
[3] J. Pearl and D. Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Basic Books, 2018.