Characterizing Fairness Over the Set of Good Models Under Selective Labels

Amanda Coston, Ashesh Rambachan, and Alexandra Chouldechova

International Conference on Machine Learning 2021

HARVARD UNIVERSITY DEPARTMENT OF ECONOMICS

Carnegie Mellon University Heinz College

ML MACHINE LEARNING DEPARTMENT

Algorithm-informed decisions may cause inequities

- May disproportionately affect different demographic groups
  - ➢ e.g., due to predictive disparities across groups
- Can we reduce disparities without affecting accuracy (too much)?

Can we reduce disparities without affecting accuracy?

Audit the business necessity defense of disparate impact[1,2]

Replace the model in use (the "benchmark" $\tilde{f}$) with a more equitable model that maintains performance

1. Civil Rights Act, 1964. 42 U.S.C. § 2000e
2. Equal Credit Opportunity Act, 1974. 15 U.S.C. § 1691

Can we reduce disparities without affecting accuracy?

Rashomon effect[1]

- Many models perform well but differ in their individual predictions

➢ May differ in terms of predictive disparities by demographic group

Over this set of good models,[2] what is the range of predictive disparities?

$$\min_{f \in \mathcal{F}} \text{disparities}(f) \;\; s.t. \underbrace{\text{loss}(f) \leq \text{loss}(\tilde{f}) + \epsilon}_{\text{set of good models}}$$

1. Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical science*, *16*(3), 199-231.
2. Dong, J., & Rudin, C. (2020). Exploring the cloud of variable importance for the set of all good models. *Nature Machine Intelligence*, *2*(12), 810-824.

Method

Over target population

$$\min_{f \in \mathcal{F}} \text{disparities}(f) \;\; s.t. \;\; \text{loss}(f) \leq \epsilon$$

- Solve via a reduction to cost-sensitive classification[1]
  - ➢ Applicable to a class of disparities, e.g., statistical parity, balance for +/- class
  - ➢ Applicable to any classification method that accepts weights

Theorem Under conditions on the function class complexity, this approach returns a randomized classifier that is approximately optimal wrt predictive disparities and that approximately satisfies the performance constraint

1. Agarwal, Alekh, et al. "A reductions approach to fair classification." *International Conference on Machine Learning*. PMLR, 2018.

Selective Labels

$$\min_{f \in \mathcal{F}} \underbrace{\text{disparities}(f)}_{\text{Over target population}} \; s.t. \; \underbrace{\text{loss}(f)}_{} \leq \epsilon$$

Lending

- Target population: all loan applicants
- Problem: Observe outcomes for *approved* applicants only
- Possible to achieve parity in approved applicants but still have disparities in target population[1,2]

Our solution:

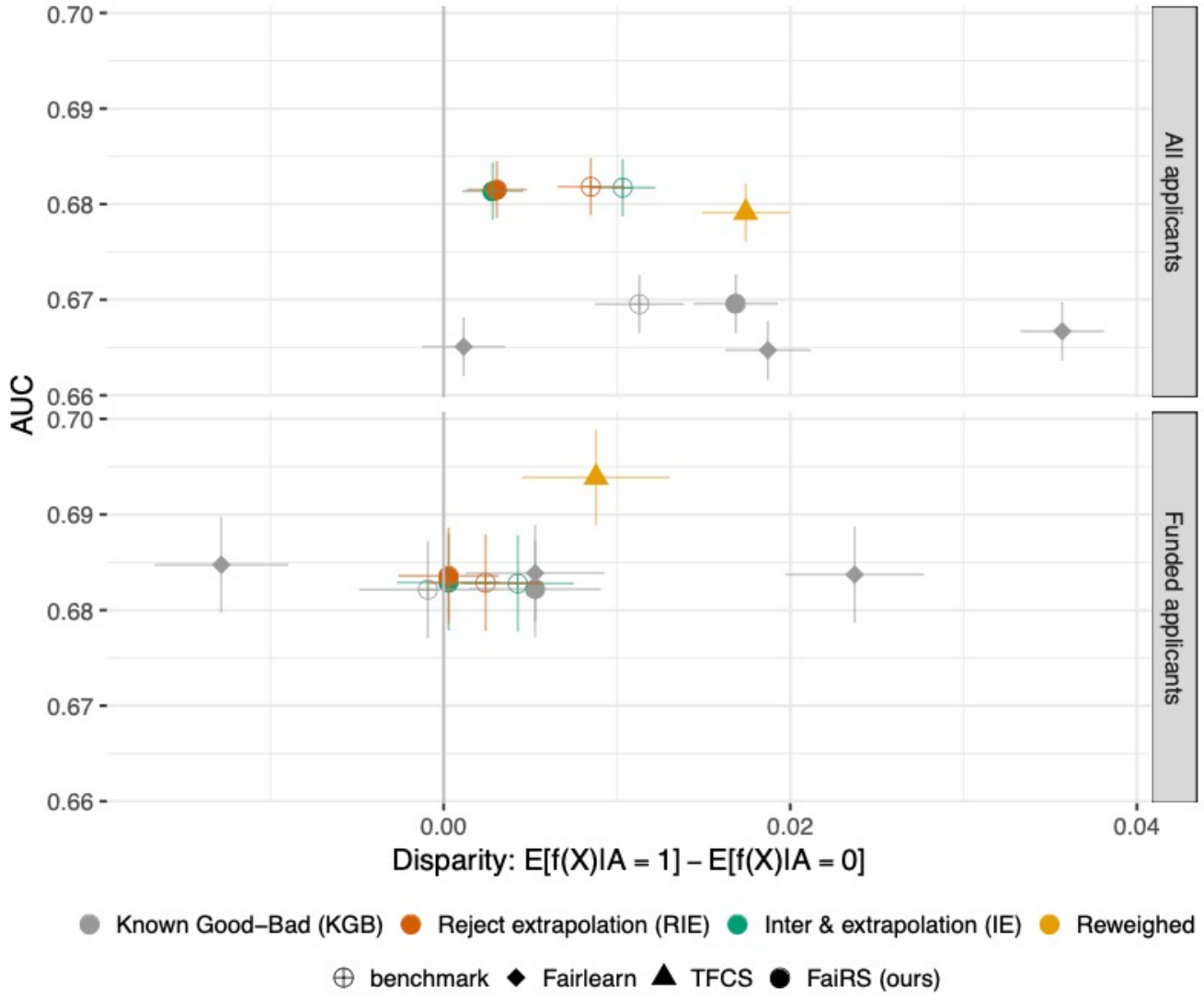
- Impute missing outcomes
- Assume that missing outcomes occur at random conditional on the observed features

1. Kallus, Nathan, and Angela Zhou. "Residual unfairness in fair machine learning from prejudiced data." *International Conference on Machine Learning*. PMLR, 2018.
2. Bechavod, Yahav, et al. "Equal opportunity in online classification with partial feedback." *Advances in Neural Information Processing Systems* 32 (2019).

Audit COMPAS for disparate impact

	Min. Disp.	Max. Disp.	COMPAS
Statistical parity	−0.060 (0.004)	0.120 (0.007)	0.194 (0.013)
Balance for + class	0.049 (0.005)	0.125 (0.012)	0.156 (0.016)
Balance for - class	0.044 (0.005)	0.117 (0.009)	0.174 (0.016)

Build a more equitable model than the benchmark



Thank you!

CommonwealthBank