An Anti-subordination Approach to Fair Classification

Vijay Keswani, L. Elisa Celis

Yale University

Fair classification

Biased data with demographics

Task-specific data, protected attributes

Objective function | **Fairness Metric**

Choose fairness parameters and accuracy tradeoff

Train Classifier

"Fair" Outcomes

Fairness metrics

- Used to audit classifiers and search for classifiers with low performance disparity
- Example - Statistical rate: difference b/w selection rate for majority group and minority group

Motivation for fair ML

- Fair classification addresses data bias and ensures similar performance for all groups
- **However, does it actively tackle the societal hierarchies that enable discrimination and corrupt data in the first place?**

The legal principle of anti-subordination

Anti-subordination principle "contends that guarantees of equal citizenship cannot be realized under conditions of pervasive social stratification and argues that law should reform institutions and practices that enforce the secondary social status of historically oppressed groups" (Balkin and Siegel, 2003).

A selection policy satisfies the anti-subordination principle when
- it does not have the effect of subordinating any group and
- does not propagate or enable any existing form of subordination

Anti-subordination principle is wider in scope than anti-classification principle

Anti-classification principle prohibits practices that discriminate based on protected attributes and favors rules that ensure that resource or opportunity access are independent of protected attributes of individuals.

Shortcomings of anti-classification principle:
- Similarity in performance does not lead to removal of structural inequalities
- Affirmative action harder to justify using anti-classification
- Anti-classification can be satisfied even when the application in question suffers from systemic issues leading to continued subordination

Why is anti-subordination important for fair ML?

Current fair ML methods focus on anti-classification and, either by design or use, can lead to continued subordination of historically-disadvantaged groups.

Along the axes listed below, improper design choices can lead to violation of anti-subordination principle. However, these choices still achieve anti-classification.

Dynamic and future impact

- Inappropriate choice of fairness metric can exacerbate feature disparities (Liu et al 2018, McCradden et al. 2020)
- Fair classifiers are often not robust to individuals' feature updates (Estornell et al 2022)

Protected attribute choice

- Most fair classifiers do not address the multi-faceted discrimination faced by intersectional groups (Sanchez-Monedero et al. 2020)
- Simplified gender and race categories ignore complex social identities (Scheuerman et al. 2019, Keyes 2018)

Fairness Parameters

- Proportional or equal representation is the norm but does not address historical lack of opportunities (Keswani, Celis 2022)
- Fairness-accuracy tradeoffs don't capture the true utility of fair classifiers (Dutta el. 2020)

Application issues

Besides design issues, the use of fair classification can legitimize systemic problems in many applications.
Examples:
- Recidivism risk assessment (Reisig 2007, Dressel and Farid 2018, Alikhademi et al. 2021)
- Predictive policing (Griffard 2019, Heaven 2020)

Recommendations

Anti-subordination can only be satisfied by ensuring that fair classifiers actively work towards dismantling bias-enabling hierarchies.

- Fair classifiers should take broader context of application into account
  - Model underlying social processes and inequalities (Mullainathan 2018)
  - Measure dynamic impact of fair classification (Liu et al. 2018)
- Protected attribute choice should represent all relevant demographic identities (Tomasev et al. 2021)
- Harms to intersectional groups should be considered independent of individual group memberships (Balashankar et al. 2019)
- Representation parameters should account for the existing stereotypes and historical lack of opportunities for minority groups (Noble 2018)
- Use of fair classification in applications with systemic issues should be discouraged (Eubanks 2018)

Full Paper available at:
https://bit.ly/3rgQuvt