An Empirical Study of Rich Subgroup Fairness for Machine Learning

Kyungseon Lee

April 1, 2025

Introduction

Main Contributions
Simplify the Kearns et al. [2018] algorithm to make it heuristically and test it on various datasets.

- **Problem in Kearns et al. [2018]:** Even if the algorithm guarantees perfect fairness in theory, it may fail in practice.
- We use **heuristic learner model and auditor model** to test the idea on real data.
- We study the trade-off between fairness and accuracy on different datasets.

Notation

- $x \in X$: Protected attribute vector.
- $x' \in X'$: Unprotected attribute vector.
- $y \in \{0, 1\}$: Binary label (e.g., 0 and 1).
- $\mathcal{X} = (x, x')$: Joint feature vector.
- P: Base probability distribution from which data is drawn.
- $D : \mathcal{X} \to \{0, 1\}$: Classifier that predicts a binary label given X.
- $\gamma \in [0, 1]$: Parameter for allowable fairness violation.
- $\mathcal{G}$: Set of indicator functions for subgroups defined by protected attributes ($\delta : X \to \{0, 1\}$).
- Each data point is given as a tuple $(x_i, c_{0,i}, c_{1,i})$:
  - $c_{0,i}$: Cost when predicting 0 for x_i.
  - $c_{1,i}$: Cost when predicting 1 for x_i.
- $\mathcal{H}$: Hypothesis space for classifiers.
- r_0, r_1: Linear regression models to predict costs for class 0 and class 1, respectively.

Related Work - Kearns et al. [2018]

Objective Function

- Fair metric: False Positive Subgroup Fairness

$$\alpha_F^P(\delta, P) \cdot \beta_F^P(\delta, D, P) \leq \gamma,$$

$$\alpha_F^P(\delta, P) = \Pr_P[\delta(x) = 1, \, y = 0], \quad \beta_F^P(\delta, D, P) = |\mathrm{FP}(D) - \mathrm{FP}(D, \delta)| \, .$$

$\mathrm{FP}(D) = \Pr_P[D(X) = 1 \mid y = 0]$: Overall FPR.

$\mathrm{FP}(D, \delta) = \Pr_P[D(X) = 1 \mid \delta(x) = 1, \, y = 0]$: FPR for subgroup δ.

- Fair ERM problem:

$$\begin{aligned} &\min_{D \in \Delta\mathcal{H}} \; \mathbb{E}_{h \sim D}[\mathrm{err}(h, \mathcal{P})] \\ &\text{s.t. } \forall g \in \mathcal{G} : \quad \alpha_{FP}(g, \mathcal{P}) \, \beta_{FP}(g, D, \mathcal{P}) \leq \gamma \end{aligned}$$

where $\mathrm{err}(h, \mathcal{P}) = \Pr_{\mathcal{P}}[h(x, x') \neq y]$ and D is a distribution over $\mathcal{H}$.

Related Work - Kearns et al. [2018]

Fictitious Play Algorithm

- Define models:
  - Learner: Linear classifier over all features.
  - Auditor: Linear classifier over protected features.
- Set up oracles:

$$h^* = \arg\min_{h \in \mathcal{H}} \sum_i \left[h(x_i) c_{1,i} + (1 - h(x_i)) c_{0,i} \right]$$

and

$$\delta_t = \arg\max_{\delta \in \mathcal{G}} \ \alpha_F^P(\delta, P) \cdot \beta_F^P(\delta, D, P).$$

- **Iterative Play (for each round t):**
  - **Auditor:** Compute and update δ_t using past plays.
  - **Learner:** Compute and update h_t via the CSC oracle.
  - Record strategies using a uniform distribution over past rounds.
- **Final Classifier:** Form the final classifier as a weighted average of all h_t's.

MainFramework

Heuristic algorithm

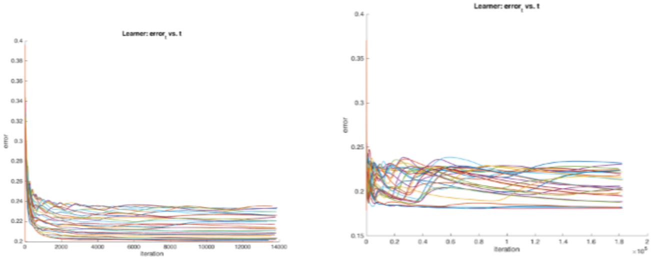
1. Learner: Predicts costs and finds a prediction model.

$$\hat{y} = \arg\min_{i \in \{0,1\}} r_i(x), \quad \hat{c}_i = r_i(x), \quad i = 1, 2.$$

2. Auditor: Evaluates unfairness for each subgroup $\rightarrow$ Selects the worst-off subgroup.
3. Learner: **Applies a cost penalty for that subgroup.**
4. Repeat.

Experiment

- We test the heuristic approach on real data.



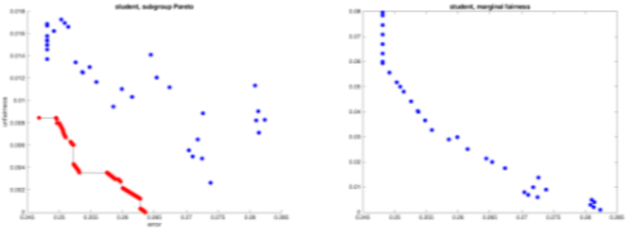
(a) (c)

Figure: Error graphs for Law School and Adult datasets

- The results show unstable error rates on some datasets.

Experiment

- Comparison between the SUBGROUP algorithm and the traditional fairness approach.

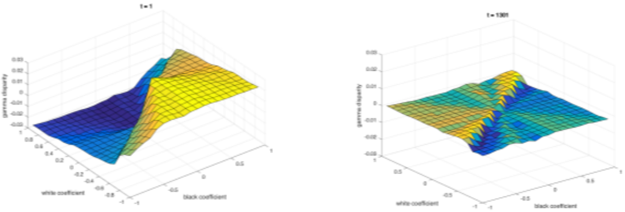


(g) (h)

Figure: Left: Points from SUBGROUP (red) and the traditional fairness algorithm (blue) on Student dataset. Right: Fairness of the traditional algorithm.

Experiment

- How racial bias changes in the Subgroup algorithm.



(a) (h)

Figure: Bias change graphs for white-black groups in the Communities and Crime dataset.

- The experiments show that the bias reduces well.

Conclusion

- This work shows a practical implementation of a rich subgroup fairness algorithm using heuristic learners and auditors.
- The algorithm converges fast on several datasets, achieving a large improvement in fairness with a small loss in accuracy.
- The study confirms that traditional fairness methods do not reduce subgroup unfairness enough.