Pareto Efficient Fairness for Skewed Subgroup Data

Ananth Balashankar, Alyssa Lees, Chris Welty, Lakshminaryanan Subramanian ananth@nyu.edu, {alyssalees, welty}@google.com, lakshmi@nyu.edu

ML can amplify societal biases

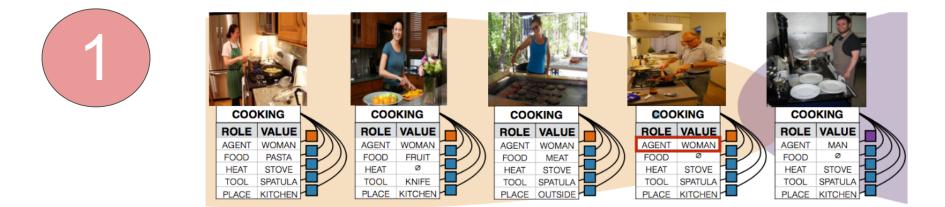
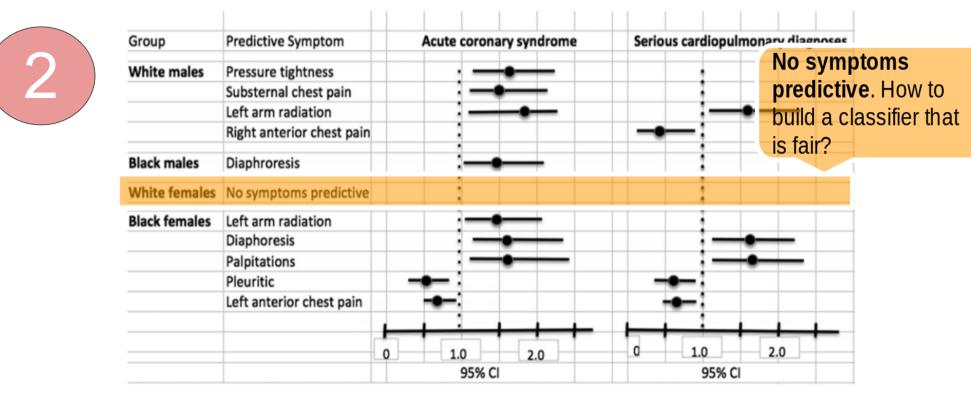


Figure 1: Five example images from the imSitu visual semantic role labeling (vSRL) dataset. Each in-33% male age is paired with a table describing a situation: the verb, cooking, its semantic roles, i.e agent, and oun values filling that role, i.e. woman. In the imSitu training set, 33% of cooking images have man is the agent role while the rest have woman. After training a Conditional Random Field (CRF), bias is amplified: man fills 16% of agent roles in cooking images. To reduce this bias amplification our calibration method adjusts weights of CRF potentials associated with biased predictions. After applying our methods, man appears in the agent role of 20% of cooking images, reducing the bias amplification by 25%, while keeping the CRF vSRL performance unchanged.

"Men Also Like Shopping Reducing Gender Bias Amplification using Corpuslevel Constraints" [Zhao et al., 2017]



Pareto-efficient fairness is better

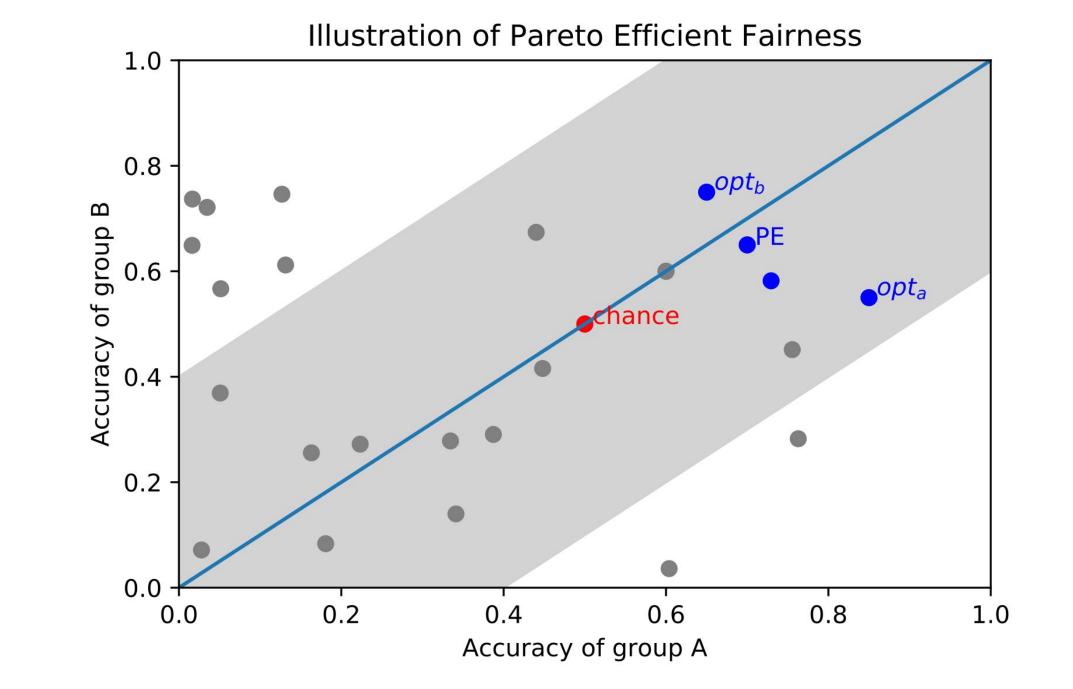


Figure 1 Significant chest pain symptoms associated with acute coronary syndrome (ACS) and serious cardio-pulmonary diagnoses (SCPD) patients in ED patients with chest pain according to race and gender (n=4162).

"Gender, race and the presentation of acute coronary syndrome and serious cardiopulmonary diagnosis in ED patients with chest pain" [Allabban et al., 2017]

Perfect Fairness and Accuracy are at odds

- Theorem: If the class probabilities of the target (Y) and sensitive features (S) are *aligned*, there is an unavoidable degradation in accuracy owing to the fairness requirement [Menon and Williamson, 2018]
- Examples of Perfect Fairness Requirements (assuming Y is binary):
 - Demographic Parity: $P(\hat{Y}=1|S=m) = P(\hat{Y}=1|S=n)$ [Calders and Verwer, 2010]
 - Equality of Opportunity: P(Ŷ=1IY=1, S=m) = P(Ŷ=1IY=1, S=n) [Hardt et al., 2016]
 - Equality of Odds: P(Ŷ=ŷIY=y, S=m) = P(Ŷ=ŷIY=y, S=n); ŷ, y ∈ {0,1},
 ∀m, n ∈ S, m ≠ n [Hardt et al., 2016]

- **Pareto-Efficient**: Set of points for which there does not exist another point, which is better performing across all sensitive groups
- Fairness: The deviation from each group's Pareto-optimal point is distributed equitably among groups

Pareto efficient bias mitigation $\mathcal{L}_p = \mathcal{L}_{ce} + \lambda(\alpha \|\mathcal{E}_G\|_1 + (1 - \alpha)\sigma_G^2(\mathcal{E}_G))$

 $\begin{array}{l}G: \mbox{ set of sensitive groups, } D: \mbox{ dataset, } D_g: \mbox{ data of group } g \in G \\ \mbox{ for } g \in G \mbox{ do } \\ M_g = \mbox{ arg min } \mathcal{L}_{ce}(D_g) \\ f_{opt-g} = \mbox{ eval}(M_g, D_g) \\ f_g = \emptyset \\ \mbox{ end for } \\ \mbox{ while } \exists g \in G, f_g = \emptyset \lor f_g > f_{opt-g} \mbox{ do } \\ f_{opt-g} = \mbox{ max}(f_g, f_{opt-g}), \forall g \in G \\ M = \mbox{ arg min } \mathcal{L}_p(D) \\ f_g = \mbox{ eval}(M, D_g), \forall g \in G \\ \mbox{ end while } \end{array}$

Evaluation on UCI Census

Existing Approaches

- Approximate Fairness with Lagrangian Constraints [Zhao et al., 2017]
 - Augments a fairness penalty term to the cross entropy loss
 - $\circ~$ The penalty factor λ is hard to fine-tune by the domain expert
- Adversarial Multi-Task Learning [Beutel et al., 2017]
 - Learn the target and ensure that the sensitive feature is not learnt through negative gradients of multi-task learning
 - No control given to domain expert to control the trade-off meaningfully

What we want in many cases...

- Policies like affirmative action and UN Sustainable Development goals aim to improve performance of protected groups to meet the levels of the highest performing groups [Foster and Vohra, 1992]
- In skewed subgroup datasets, there might be an opportunity to choose performance for all groups that are better than the best perfectly fair one
- We want to give the domain expert the ability to search for Pareto-Efficient performance within a Fairness bound

• data

- 14 demographic features like age, education, occupation, etc from 1994 census
- Predict if income is > 50K or not
- Sensitive variables assumed are race and gender
- Pareto Efficient loss is minimized, while achieving best overall accuracy

Table 1. UCI Adult dataset with bias mitigation algorithms

Model	Accuracy	FPR	FNR	Discrepancy	Pareto Loss
Baseline (no bias loss)	0.630	0.253	0.747	0.199	0.016
Minimize Discrepancy	0.619	0.283	0.712	0.167	0.133
Adversarial Loss	0.648	0.224	0.769	0.226	0.077
Pareto-Efficient Loss	0.678	0.165	0.830	0.250	0.000

• Pareto Efficient loss ensures best accuracy across all subgroups too!

Table 2. Subgroup performance on UCI Adult dataset

Model	Subgroup 1	2	3	4	Pareto Loss
Baseline (no bias loss)	0.890	0.883	0.818	0.784	0.016
Minimize Discrepancy	0.853	0.856	0.806	0.778	0.133
Adversarial Loss	0.882	0.872	0.824	0.780	0.077
Pareto-Efficient Loss	0.935	0.915	0.844	0.797	0.000
Subgroup Pareto Frontier	0.934	0.894	0.815	0.783	N/A



