AdaFair: Cumulative Fairness Adaptive Boosting

Vasileios Iosifidis and Eirini Ntoutsi

{iosifidis, ntoutsi}@L3S.de

1. Discrimination in Supervised Learning

- Discrimination is treatment or consideration of, or making a distinction towards, a person based on a protected attribute to which the person is perceived to belong.
- Protected attributes are considered to be: age, disability, race, religion, sex, sexual orientation, etc.
- Recent incidents of AI-based discrimination have raised concerns about implications of AI in our society.

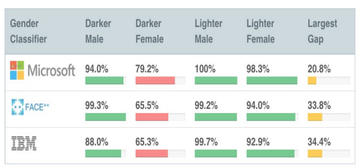


Image source[1]

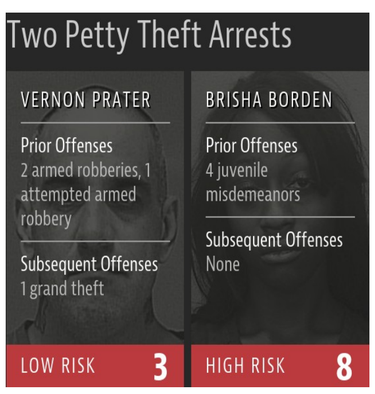
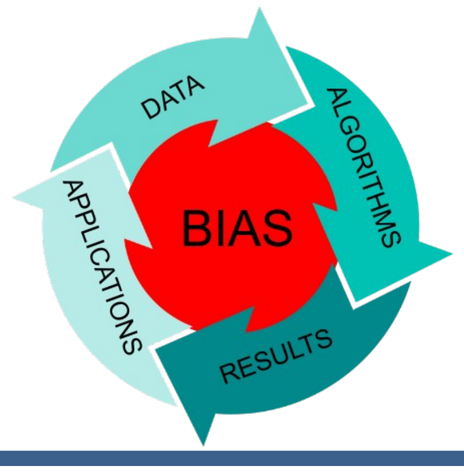


Image source[2]

2. Why Machine Learning can be Unfair?

- Data might encode existing societal biases.
- Data generated feedback loops e.g., predictive policing[3].
- Different characteristics for different populations e.g., men's vs women's height.
- There exist proxies e.g., zip codes can be proxies to race (Amazon services[4])



3. Notation and Problem definition

- Training dataset D drawn from a joint distribution P(F , S, y)
- We assume a binary class: y ∈ {+, −}
- F is the set of non-protected attributes and S is a binary protected attribute
- We define 4 different population segments:

	Class label	
Protected Attribute	Rejected	Granted
s (e.g., female)	s_-	s_+
š (e.g., male)	$š_-$	$š_+$

- **Fairness Notion:**
  - Equalized Odds = |δFPR| + |δFNR|
  - δFPR = $P(y \neq \hat{y} \mid š_-) - P(y \neq \hat{y} \mid s_-)$
  - δFNR = $P(y \neq \hat{y} \mid š_+) - P(y \neq \hat{y} \mid s_+)$

- **Goal:**
  Find a mapping function f(·) that minimizes Eq.Odds while maintaining good predictive performance for both classes (balanced error rate).

- **Balanced Error Rate:**

$$BER = 1 - \frac{1}{2} \cdot \left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right) = 1 - \frac{1}{2} \cdot (TPR + TNR)$$

4. AdaFair

- **Fair**ness-aware **boost**ing, deals with class-imbalance and unfair outcomes.
- Changes data distribution, using fairness-related weights, at each round based on the notion of cumulative fairness.
- Cumulative fairness evaluates long term/cumulative discrimination over the current sequent of learners.
- After the training phase, the best sequence of weak learners *(θ)* which achieve high performance and fairness is selected.

- **Cumulative fairness notion:**

$$\delta FNR^{1:j} = \frac{\sum_{i=1}^{|\bar{s}_+|} 1 \cdot \mathbb{I}[\sum_{k=1}^{j} a_k h_k(x_i^{\bar{s}_+}) \neq y_i]}{|\bar{s}_+|} - \frac{\sum_{i=1}^{|s_+|} 1 \cdot \mathbb{I}[\sum_{k=1}^{j} a_k h_k(x_i^{s_+}) \neq y_i]}{|s_+|}$$

$$\delta FPR^{1:j} = \frac{\sum_{i=1}^{|\bar{s}_-|} 1 \cdot \mathbb{I}[\sum_{k=1}^{j} a_k h_k(x_i^{\bar{s}_-}) \neq y_i]}{|\bar{s}_-|} - \frac{\sum_{i=1}^{|s_-|} 1 \cdot \mathbb{I}[\sum_{k=1}^{j} a_k h_k(x_i^{s_-}) \neq y_i]}{|s_-|}$$

- **Fairness-related weights:**

$$u_i = \begin{cases} |\delta FNR^{1:j}|, & if\ \mathbb{I}((y_i \neq h_j(x_i)) \wedge |\delta FNR^{1:j}| > \epsilon), x_i \in s_+, \delta FNR^{1:j} > 0 \\ |\delta FNR^{1:j}|, & if\ \mathbb{I}((y_i \neq h_j(x_i)) \wedge |\delta FNR^{1:j}| > \epsilon), x_i \in \bar{s}_+, \delta FNR^{1:j} < 0 \\ |\delta FPR^{1:j}|, & if\ \mathbb{I}((y_i \neq h_j(x_i)) \wedge |\delta FPR^{1:j}| > \epsilon), x_i \in s_-, \delta FPR^{1:j} > 0 \\ |\delta FPR^{1:j}|, & if\ \mathbb{I}((y_i \neq h_j(x_i)) \wedge |\delta FPR^{1:j}| > \epsilon), x_i \in \bar{s}_-, \delta FPR^{1:j} < 0 \\ 0, & otherwise \end{cases}$$

- **Data distribution update:**

$$w_i \leftarrow \frac{1}{Z_j} w_i \cdot e^{\alpha_j \cdot \hat{h}_j(x) \cdot \mathbb{I}(y_i \neq h_j(x_i))} \cdot (1 + u_i)$$

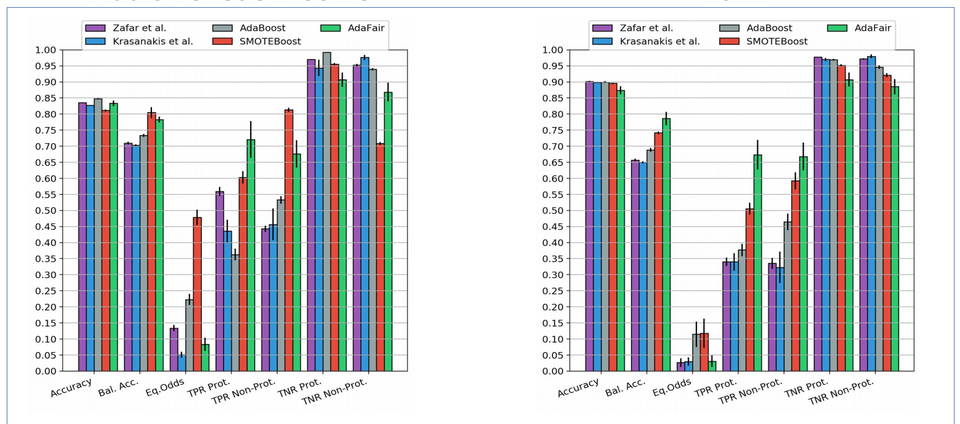
- **Objective function:**

$$\arg\min_{\theta} (c \cdot BER_\theta + (1 - c) \cdot ER_\theta + Eq.Odds_\theta)$$

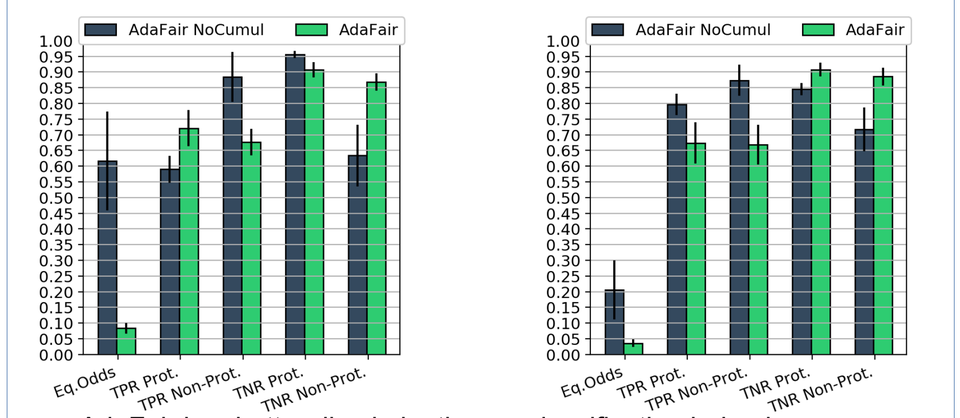
5. Results

AdaFair vs baselines (top), cumulative vs non-cumulative method (middle), and impact of parameter c (bottom).

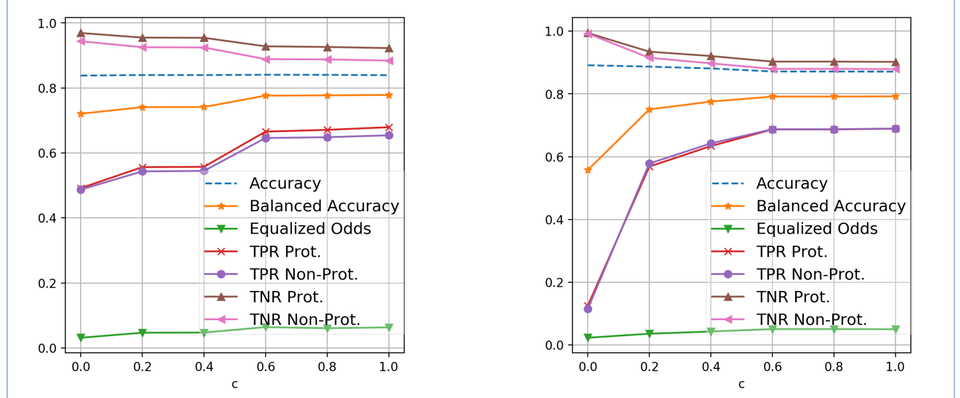
Adult Census Income **Bank**



- AdaBoost and SMOTEBoost do not consider fairness (high Eq.Odds).
- Krasanakis et al. and Zafar et al. produce low TPRs and high TNRs.



- AdaFair has better discrimination-vs-classification behavior.
- AdaFair is more stable.



- For c = 0, the error rate is optimized and c = 1 the balanced error rate.
- AdaFair can mitigate discrimination while tuned for BER or ER.

Sources:

1) https://www.media.mit.edu/projects/gender-shades/press-kit
2) https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
3) https://www.bloomberg.com/graphics/2016-amazon-same-day
4) https://www.theguardian.com/uk-news/2019/sep/16/predictive-policing-poses-discrimination-risk-thinktank-warns

AdaFair repository:

- https://iosifidisvasileios.github.io/AdaFair

ACKNOWLEDGMENTS

- This work was funded by the German Research Foundation (DFG) project OSCAR (Opinion Stream Classification with Ensembles and Active leaRners) and inspired by the Volkswagen Foundation project BIAS ("Bias and Discrimination in Big Data and Algorithmic Processing. Philosophical Assessments, Legal Dimensions, and Technical Solutions")

Leibniz Universität Hannover

Forschungszentrum · Research Center L3S