AdaFair: Cumulative Fairness Adaptive Boosting

V. Iosifidis[1] E. Ntoutsi[1]

[1]University of Leibniz
L3S Research Center
Hannover, Germany

November 11, 2019

Discrimination

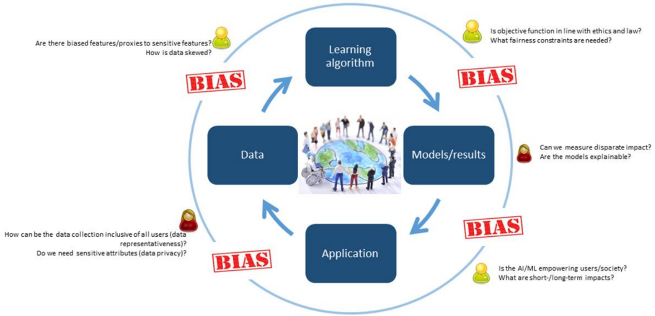
- Discrimination is treatment or consideration of, or making a distinction towards, a person based on a protected attribute to which the person is perceived to belong.
- Protected attributes are considered to be: age, disability, race, religion, sex, sexual orientation, etc.

Image source: https://www.hrmagazine.co.uk/article-details/discrimination-costs-the-uk-127-billion-a-year

Recent discrimination discoveries in machine learning applications

Why Machine Learning is Unfair?

- Data might encode existing biases e.g., bias caused by humans, features of minorities contain more noise.
- Data collection feedback loops.
- Different data distributions for different groups e.g., lack of observed examples.
- Proxies to protected attributes e.g., marital status *wife* or *husband* can reveal the gender.



Basic Notations

Notation

- Training dataset D drawn from a joint distribution $P(F, S, y)$
- We assume a binary class: $y \in \{+, -\}$
- F is the set of non-protected attributes and S is a binary protected attr.

	class label	
Protected Attribute	Rejected	Granted
s (Female)	s_-	s_+
$\bar{s}$ (Male)	$\bar{s}_-$	$\bar{s}_+$

Fairness notion [HPS+16]

$$Equalized\ Odds = |\delta FPR| + |\delta FNR|$$

$$\delta FPR = P(y \neq \hat{y}|\bar{s}_-) - P(y \neq \hat{y}|s_-)$$

$$\delta FNR = P(y \neq \hat{y}|\bar{s}_+) - P(y \neq \hat{y}|s_+)$$

The "trap" of *Equalized Odds*

Example

- Positive class << Negative class e.g., $|s^+| + |\bar{s}^+| = 5\%, |s^-| + |\bar{s}^-| = 95\%$
- Model classifies everything as negative.
- Accuracy is still high (95%) and model is "fair" i.e., $\delta FNR = 0, \delta FPR = 0$

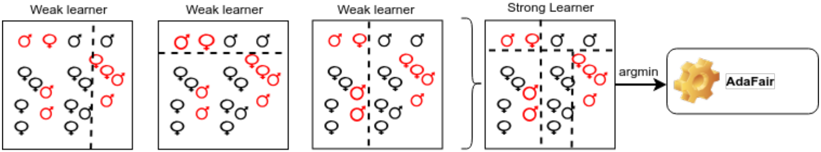
Goal

Find a mapping function $f(\cdot)$ that minimizes Eq.Odds while performing well for both classes.

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

AdaFair: Overview

- Fairness-aware boosting approach that deals with class-imbalance and unfair outcomes.
- Changes data distribution at each round based on the notion of cumulative fairness.
- After the training phase, the best sequence of weak learners which achieve high performance and fairness is selected.



Cumulative Fairness

- Let $j : 1 \ldots T$ be the current boosting round, T is user defined
- Let $H_{1:j} = \sum_{i=1}^{j} a_i h_i(x)$ be the ensemble model up to current round j.
- The cumulative fairness of the ensemble up to round j, is defined based on the parity in the predictions of weak learners $h1() \ldots hj()$ between protected and non-protected groups

Cumulative Fairness

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

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

Fairness weights

- Vanilla AdaBoost already boosts misclassified instances for the next round.
- Our weighting explicitly targets fairness by extra boosting discriminated groups for the next round.
- Instances $x_i \in D$ which belong to a group that is discriminated receive a fairness-related weight u_i

Weight calculation

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

AdaFair's pseudocode

Input: $D = (x_i, y_i)_1^N, T, \epsilon$
Output: Ensemble H

1. Initialize $w_i = 1/N$ and $u_i = 0$, for $i = 1, 2, \ldots, N$
2. For $j = 1$ to T:
    1. Train a classifier h_j to the training data using weights w_i.
    2. Compute the error rate $\text{err}_j = \frac{\sum_{i=1}^{N} w_i I(y_i \neq h_j(x_i))}{\sum_{i=1}^{N} w_i}$
    3. Compute the weight $\alpha_j = \frac{1}{2} \cdot \ln(\frac{1-\text{err}_j}{\text{err}_j})$
    4. Compute fairness-related $\delta FNR^{1:j}$
    5. Compute fairness-related $\delta FPR^{1:j}$
    6. Compute fairness-related weights u_i
    7. Update the distribution as $w_i \leftarrow \frac{1}{Z_j} w_i \cdot e^{\alpha_j \cdot \hat{h}_j(x) \cdot \mathbb{I}(y_i \neq h_j(x_i))} \cdot (1 + u_i)$
3. Output $H(x) = \sum_{j=1}^{T} \alpha_i h_j(x)$

Performance trade-off: error vs balanced error

- AdaFair optimizes for the balanced error rate.
- AdaFair selects the optimal number of weak learners $1 \cdots \theta, \theta \leq T$ that minimizes BER.
- AdaFair considers both ER and BER in the objective function as follows:

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

- Parameter c is user-defined and controls the impact of error and balanced error rate.

Baselines

- AdaBoost [Sch99]: vanilla AdaBoost.
- SMOTEBoost [CLHB03]: AdaBoost with SMOTE for imbalanced data.
- Krasanakis et al. [KXPK18]: Boosting method which minimizes Equalised Odds by approximating the underlying distribution of hidden correct labels.
- Zafar et al.[ZVGRG17]: Training logistic regression model with convex-concave constraints to minimize Equalised Odds.
- AdaFair NoCumul: Variation of AdaFair that computes the fairness weights based on individual weak learners.

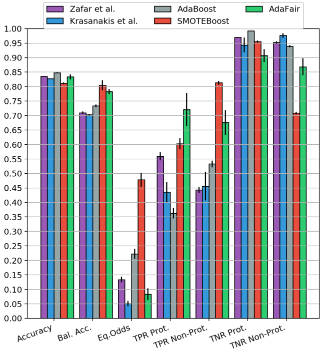
Datasets

	Adult Census	**Bank**	Compass	KDD Census
#Instances	45,175	40,004	5,278	299,285
#Attributes	14	16	9	41
Sen.Attr.	Gender	Marit. Status	Gender	Gender
Class ratio (+:−)	1:3.03	1:7.57	1:1.12	1:15.11
Positive class	*>50K*	*subscription*	*recidivism*	*>50K*

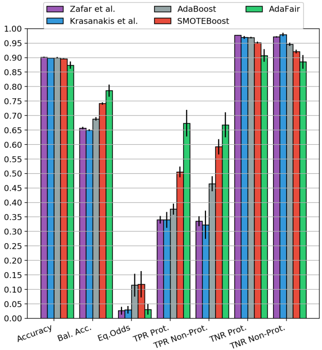
Employed datasets

We report on the average of 10 random splits [ZVGRG17], 50% training and 50% testing set.

AdaFair vs Baselines



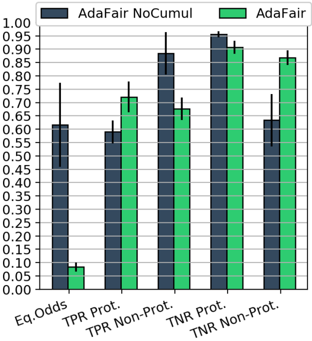
Adult Census



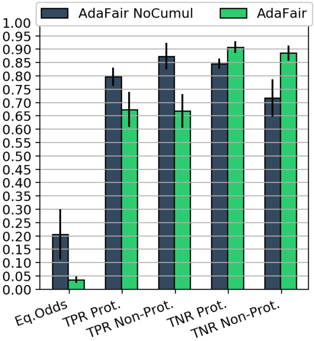
Bank

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

Cumulative vs Non Cumulative Overall Performance



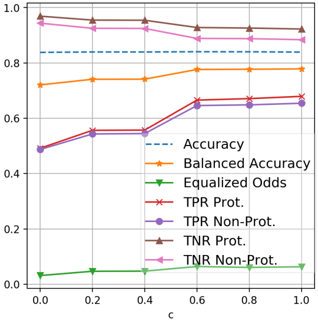
Adult Census



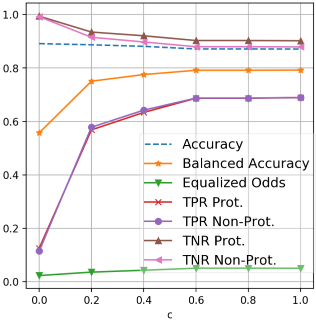
Bank

- AdaFair NoCumul has poor fairness performance.
- AdaFair NoCumul is very unstable.

Impact of parameter c



Adult Census



Bank

- For $c = 0$, the error rate is optimized and $c = 1$ the balanced error rate.

Conclusions

Conclusions

- AdaFair: fairness-aware boosting approach.
  - Data distributions alter based on *cumulative fairness*.
  - Deal with class-imbalance (indirectly).
- Substantial difference in performance compared to baselines.
- Cumulative fairness is superior to a non-cumulative approach.

Future Work

- Embed class-imbalance learning into training phase.
- Investigate theoretical properties e.g., convergence

Nitesh V Chawla, Aleksandar Lazarevic, Lawrence O Hall, and Kevin W Bowyer.
Smoteboost: Improving prediction of the minority class in boosting.
In *ECML PKDD*, 2003.

Moritz Hardt, Eric Price, Nati Srebro, et al.
Equality of opportunity in supervised learning.
In *NeurIPS*, 2016.

Emmanouil Krasanakis, Eleftherios Spyromitros Xioufis, Symeon Papadopoulos, and Yiannis Kompatsiaris.
Adaptive sensitive reweighting to mitigate bias in fairness-aware classification.
In *WWW*, 2018.

Robert E Schapire.
A brief introduction to boosting.
In *IJCAI*, 1999.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi.
Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment.

Thanks.

Questions?

Contact: {iosifidis,ntoutsi}@L3S.de