[Re] Exacerbating Algorithmic Bias through Fairness Attacks

UNIVERSITEIT VAN AMSTERDAM

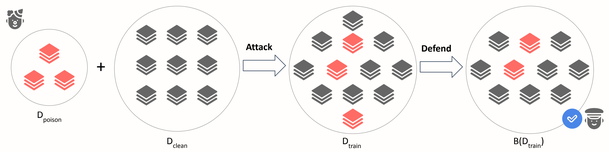
Angelos Nalmpantis – Apostolos Panagiotopoulos - Konstantinos Papakostas – John Gkountouras

Qualcomm

ellis European Laboratory for Learning and Intelligent Systems

Introduction

- Adversarial attacks typically aim to degrade the **performance** of a machine learning model.
- Following the same principles, we can deliberately influence the **fairness** of a model's output.
- Mehrabi et al. [1] proposed two novel **data poisoning** attacks, which we tried to replicate.
  - The **Anchoring Attack** (AA) places poisoned samples s.t. the decision boundary is skewed.
  - The **Influence Attack on Fairness** (IAF) performs gradient ascent on an adversarial fairness loss.



Attacks

Influence Attack on Fairness (IAF)

The **Influence Attack on Fairness** (IAF) generates poisoned points by starting with two datapoints $(\tilde{x}_1, \tilde{y}_1)$ and $(\tilde{x}_2, \tilde{y}_2)$ from each class of the clean dataset D_c and moving them in the direction that maximizes an adversarial loss $\mathcal{L}_{\text{adv}} = \mathcal{L}_{bc} + \lambda \cdot \mathcal{L}_f$

- $\mathcal{L}_{bc}$ is the BCE Loss
- $\mathcal{L}_f$ is the fairness loss is the one introduced in [3]: $\frac{1}{N}\sum_{i=1}^{N}(\mathbf{z}_i - \bar{\mathbf{z}})\, d_{\boldsymbol{\theta}}(\mathbf{x}_i)$
- In order to update the datapoints $(\tilde{x}_1, \tilde{y}_1)$ and $(\tilde{x}_2, \tilde{y}_2)$ we need to perform an EM scheme:
  - E-step: The model is trained on $B(D_c \cup D_p)$
  - M-step: The datapoints are updated using gradient ascent:

$$\mathbf{x}_{adv}^{(i+1)} = \mathbf{x}_{adv}^{(i)} + \eta \cdot \frac{\partial \mathcal{L}_{adv}}{\partial \mathbf{x}_{adv}}$$

Anchoring Attack on Fairness (AA)

The **Anchoring Attack** generates poisoned points by perturbing existing samples and flipping their target label, skewing the decision boundary in an unfair manner.

- In the Random Anchoring Attack (RAA), the samples are chosen randomly.
- In the Non-Random Anchoring Attack (NRAA), the samples are the most popular points w.r.t. their label and demographic group.

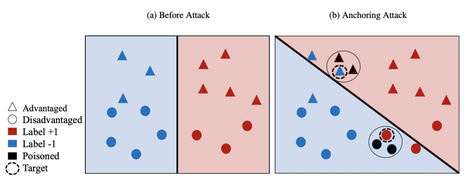


Figure by Mehrabi et al.

Defenses

- L2 defense: remove points that are too far from their class' centroid (in ℓ_2 distance).
- Slab defense: similar, but first project the points onto the line between the centroids.

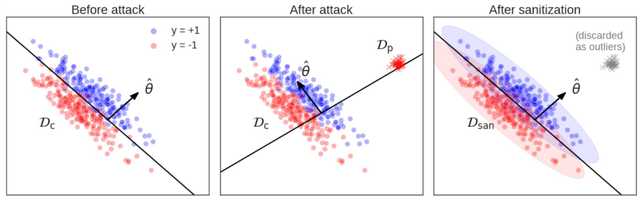


Figure by Koh et al. [2]

Adversarial Attacks on Fairness

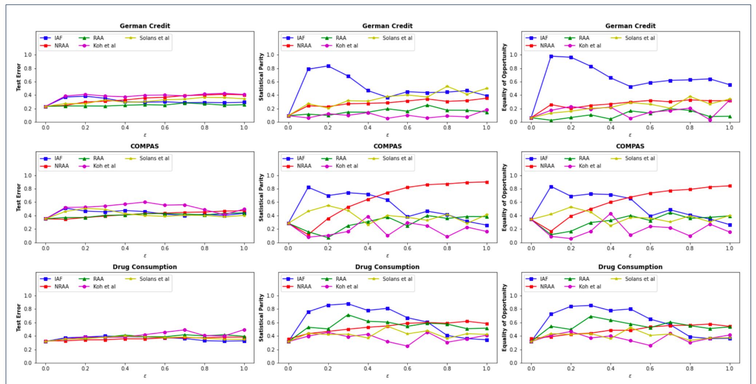


Figure 1: Impact on performance and fairness of a logistic regression classifier, using the attacks proposed in [11] and other state-of-the-art methods, for increasing ϵ values.

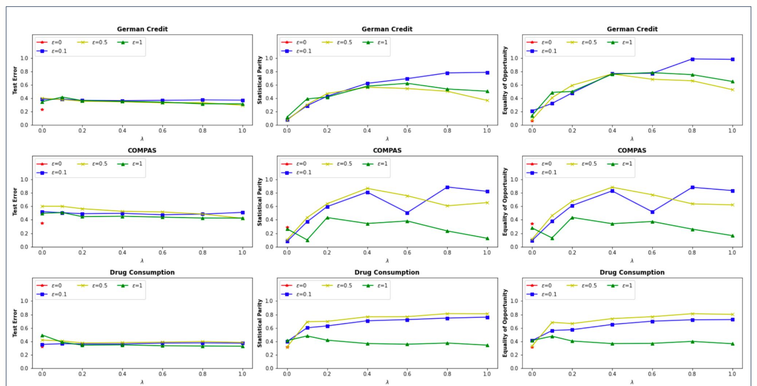


Figure 2: Impact on performance and fairness of a logistic regression model using the IAF, for increasing λ values.

Increasing Fairness via Data Augmentation

Using the principal idea behind IAF, we can **inverse the bias** present in a dataset.

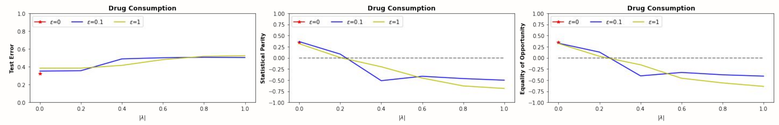
Steps:

i. Identify the direction of dataset's intrinsic bias.

ii. Change the sign of λ so that you reverse the bias.

iii. Apply the same methodology as in the Influence Attack.

As the method attempts to reverse the bias, there exists such a parameter λ that the bias will intersect with the line y=0. As a result, we can obtain a **perfectly fair** system.



Conclusion

Reproducibility

- Both of the proposed attacks impact fairness, surpassing other SoTA attacks.
- Using the Influence Attack on Fairness we can control the impact on performance by tuning the scale parameter λ accordingly.

Contributions

- Thoroughly reviewed & replicated the original paper, which resulted in a reproducibility report outlining all of the missing details needed for future reproductions.
- Open-sourced our implementation, with an improved structure and more extensible code.
- Provided a debiasing algorithm using the methods from the original paper.

Acknowledgements

The presentation of this paper was financially supported by the Amsterdam ELLIS Unit and Qualcomm.

This work was carried out on the Dutch national e-infrastructure with the support of SURF Cooperative.

We would like to thank Fernando P. Santos for his excellent initiative and setting up the UvA MSc FACT course.

References

[1] Mehrabi, Ninareh, et al. "Exacerbating algorithmic bias through fairness attacks." Proceedings of the AAAI Conference on Artificial Intelligence (AAAI 2021). Vol. 35. No. 10. 8930 – 8938

[2] Koh, Pang Wei, et al. "Stronger data poisoning attacks break data sanitization defenses." In Machine Learning (2022). Vol. 111. 1 – 47.

[3] Zafar, Muhammad Bilal, et al. "Fairness constraints: A flexible approach for fair classification." In the Journal of Machine Learning Research (JMLR 2019). Vol. 20. No. 1. 2737 – 2778