ALGORITHMIC FAIRNESS METRICS AND BIAS REDUCTION METHODS

Researchers: Chenruizhe Hu, Chris Chang   Mentors: Dr. Branka Hadji, Dr. Ali Hirsa, Dr. Joerg Osterrieder   Sponsor: European Cooperation in Science and Technology

Background and Introduction

•Bias in machine learning models increasingly impact various aspects of our lives, including hiring processes, loan approvals, and criminal justice decisions. Biases present in these models can perpetuate unfairness and discrimination, leading to negative social consequences. As a result, there is a growing need to develop techniques and metrics that can measure and mitigate bias in machine learning algorithms. Our research contributes to the field by introducing new metrics, evaluating debiasing methods, and advancing fairness-aware machine learning techniques.

•To provide an overview of our research, we will discuss the merits and limitations of **various bias metrics** and identify the **most suitable ones** for our study. Following that, we will conduct **empirical evaluations** to assess the effectiveness of different debiasing methods, aiming to **identify the most promising approaches**. In addition to the conventional bias analysis, we **propose a novel concept** called "**residual bias**." Finally, we explore **fairness-aware Gradient Boosting Decision Trees** and investigate their potential to incorporate fairness considerations during training.

State of Arts Fairness Metrics Review

•In this section, we will explore different fairness metrics and discuss their merits and limitations in addressing fairness concerns.

•**Unawareness** is a metric that excludes the protected feature from the prediction process. **Unawareness** assumes that removing the feature eliminates bias. However, this metric overlooks the potential influence of other features in predicting the sensitive attribute and may not fully address the bias issue.

•**Demographic Parity** aims to ensure that the probability of predicting a positive outcome remains consistent across different demographic groups when conditioned on the protected feature. However, achieving **Demographic Parity** can be challenging if there is **inherent bias in the population**, leading to different outcome probabilities among demographic groups. **Trivial classifiers** can also satisfy this metric without truly addressing the bias.

•**Accuracy Parity** focuses on maintaining consistent prediction accuracy across protected feature groups. However, this metric can be **easily satisfied by trivial classifiers** if the outcome cases are highly imbalanced, making it less effective in capturing and addressing bias.

•**Average Odds Difference** evaluates the disparity in false positive and false negative rates between different protected feature groups. The goal is to minimize the differences across groups, indicating equal treatment regardless of protected attributes. This metric requires the model's performance to be consistent across protected features, ensuring fairness in both positive and negative predictions. An unweighted **Average Odds Difference** assumes false positive and false negative cases are of equal importance.

Average Odds Difference as Preferred Fairness Metrics

•Our research argues that the **Average Odds Difference** metric emerges as the preferred measure for evaluating bias in predictive models due to its ability to account for the correlation between the protected feature and the outcome variable.

•This metric acknowledges the potential influence of the protected feature on the prediction outcome while prohibiting its direct use as a predictor.

•By focusing on disparities in false positive and false negative rates across different protected feature groups, **Average Odds Difference** offers a nuanced and comprehensive assessment of bias, promoting equitable treatment and unbiased decision-making.

•Our empirical evaluations and comparative analyses reinforce the value of this metric in capturing and addressing biases, while respecting the prohibition of using the protected feature directly as a predictor for the outcome.

State of Arts Bias Reduction Methods Review

•We provide a summary of popular bias reduction methods in the field. These methods can be classified into three categories based on their training procedures. **Pre-processing** methods focus on transforming the data prior to model training to mitigate bias. **In-processing methods** involve incorporating regularization techniques during model training to combat bias. **Post-processing methods** aim to address bias during the evaluation of the model on test sets.

•**Pre-processing**: **Reweighing**[1] generates weights for the training examples in each (group, label) combination differently to ensure fairness before classification. **Optimized preprocessing**[2] learns a probabilistic transformation that edits the features and labels in the data with group fairness, individual distortion, and data fidelity constraints and objectives.

•**In-processing**: **Adversarial debiasing**[3] learns a classifier to maximize prediction accuracy and simultaneously reduce an adversary's ability to determine the protected attribute from the predictions.

•**Post-processing**: **Equalized odds postprocessing**[4] solves a linear program to find probabilities with which to change output labels to optimize equalized odds. **Reject option classification**[1] gives favorable outcomes to unprivileged groups and unfavorable outcomes to privileged groups in a confidence band around the decision boundary with the highest uncertainty.

Empirical Evaluation of Bias Reduction Methods

•We conducted a comprehensive evaluation of reweighing, optimizing preprocessing, adversarial debiasing, and reject option classification methods on several population fairness machine learning datasets, including **German**[5], **COMPAS**[6], **Census Income**[7], and **Medical Expenditure Panel Survey**[8] **(MEPS)**.

•We compare the effectiveness of bias reduction using **Average Odds Difference** metric and the reduction in accuracy score after debiasing, accounting for the **bias and accuracy tradeoff**.

Empirical Evaluation Results

Dataset	Classifier	Bias Reduction Method	Debias Attribute	pre/in/post processing	Average Odds Difference after mitigation	Accuracy score after mitigation	reduction in accuracy
German	Logistic Regression	reweighing	sex	pre	0.07	78%	-3%
German	Logistic Regression	reweighing	age	pre	0.09	74%	1%
German	Logistic Regression	optimized preprocessing	sex	pre	-0.23	60%	15%
German	Logistic Regression	optimized preprocessing	age	pre	-0.21	61%	14%
German	Logistic Regression	adversarial debiasing	sex	in	-0.18	70%	5%
German	Logistic Regression	adversarial debiasing	age	in	-0.09	69%	6%
German	Logistic Regression	rejection option based classification	sex	post	0.09	77%	-2%
German	Logistic Regression	rejection option based classification	age	post	-0.12	75%	0%

Dataset	Classifier	Bias Reduction Method	Debias Attribute	pre/in/post processing	Average Odds Difference after mitigation	Accuracy score after mitigation	reduction in accuracy
COMPAS	Logistic Regression	reweighing	sex	pre	-0.02	66%	0%
COMPAS	Logistic Regression	reweighing	race	pre	0.03	66%	0%
COMPAS	Logistic Regression	optimized preprocessing	sex	pre	-0.17	65%	1%
COMPAS	Logistic Regression	optimized preprocessing	race	pre	0	67%	-1%
COMPAS	Logistic Regression	adversarial debiasing	sex	in	-0.02	66%	0%
COMPAS	Logistic Regression	adversarial debiasing	race	in	-0.13	65%	1%
COMPAS	Logistic Regression	rejection option based classification	sex	post	0.07	65%	1%
COMPAS	Logistic Regression	rejection option based classification	race	post	0	65%	1%

Dataset	Classifier	Bias Reduction Method	Debias Attribute	pre/in/post processing	Average Odds Difference after mitigation	Accuracy score after mitigation	reduction in accuracy
Census Income	Logistic Regression	reweighing	sex	pre	-0.04	81%	2%
Census Income	Logistic Regression	reweighing	race	pre	0	82%	1%
Census Income	Logistic Regression	optimized preprocessing	sex	pre	-0.11	71%	12%
Census Income	Logistic Regression	optimized preprocessing	race	pre	-0.1	74%	9%
Census Income	Logistic Regression	adversarial debiasing	sex	in	0.14	73%	10%
Census Income	Logistic Regression	adversarial debiasing	race	in	0.1	77%	6%
Census Income	Logistic Regression	rejection option based classificatio	sex	post	-0.07	82%	1%
Census Income	Logistic Regression	rejection option based classificatio	race	post	-0.04	82%	1%

Dataset	Classifier	Bias Reduction Method	Debias Attribute	pre/in/post processing	Average Odds Difference after mitigation	Accuracy score after mitigation	reduction in accuracy
Census Income	Logistic Regression	reweighing	sex	pre	-0.04	81%	2%
Census Income	Logistic Regression	reweighing	race	pre	0	82%	1%
Census Income	Logistic Regression	optimized preprocessing	sex	pre	-0.11	71%	12%
Census Income	Logistic Regression	optimized preprocessing	race	pre	-0.1	74%	9%
Census Income	Logistic Regression	adversarial debiasing	sex	in	0.14	73%	10%
Census Income	Logistic Regression	adversarial debiasing	race	in	0.1	77%	6%
Census Income	Logistic Regression	rejection option based classificatio	sex	post	-0.07	82%	1%
Census Income	Logistic Regression	rejection option based classificatio	race	post	-0.04	82%	1%

Method	classifier	training set	testing set	Debias Attribute	pre/in/post processing	difference after mitigation	Accuracy score after mitigation	reduction in accuracy
reweighing	Logistic Regression	MEPS Panel19	MEPS Panel19	race	pre	-0.015104	75.39%	0.02
reweighing	Random Forest	MEPS Panel19	MEPS Panel19	race	pre	-0.101374	76.44%	0.00
adversarial debiasing	Logistic Regression	MEPS Panel19	MEPS Panel19	race	in	0.052286	68.80%	-0.09
reweighing	Logistic Regression	MEPS Panel19	MEPS Panel19	race	pre	0.007135	73.11%	0.04
reweighing	Logistic Regression	MEPS Panel19	MEPS Panel21	race	pre	-0.01434	73.79%	0.04
reweighing	Logistic Regression	MEPS Panel20	MEPS Panel20	race	pre	0.044042	0.716856	0.06
reweighing	Logistic Regression	MEPS Panel20	MEPS Panel21	race	pre	-0.00177	0.730831	0.05

•**Reweighing** gives a good reduction of bias and results in few reduction in accuracy, although it is mathematically very simple.

•**Reweighing** solves sampling bias intuitively.

•Besides **reweighing**, **rejection option based classification**[1] is the second-best option.

•Bias reduction methods do not further decrease the model performance when the original model performance is poor, revealed in our evaluation on COMPAS dataset.

Proposing Residual Fairness[10]

•To deepen our understanding of bias origin, we introduce the concept of **residual fairness**.

•We posit that a classifier is unfair when it consistently under-predicts outcomes for one group and over-predicts for another group.

•We calculate the **average residual values** across different groups based on a protected attribute and conduct an **ANOVA** test to determine if there are statistically significant differences in residuals among the groups.

•**Reweighing** can lead to significant residual differences, despite improving average odds difference, according to our empirical tests. Different fairness metrics corresponds to different optimal bias reduction method.

•Residual fairness accounts for statistical significance and population bias comparing to **Demographic Parity.**

Fairness-aware Gradient Boosting Decision Trees

•A fairness-aware variant of gradient-boosting decision tree model is proposed, dubbed **FairGBM**, that focus on addressing **residual fairness**[9].

•**FairGBM,** an in-processing bias reduction method, leverages a novel **fairness constrained** optimization framework for **gradient-boosting**, where fairness metrics are transformed into **differentiable proxy Lagrangian duals** based on **cross-entropy**, enabling the integration of **fairness constraints** into the model training process.

•For different **accuracy and bias trade-off parameter** α, the figure[9] below shows **FairGBM** outperforms all other state of arts **Gradient Boosting Decision Tree** models.

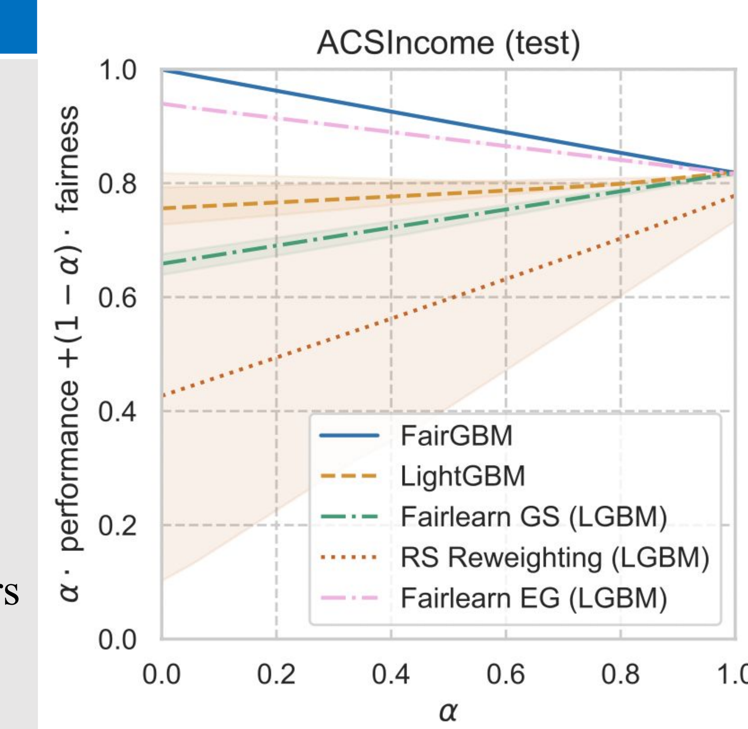
•The high-performance implementation of **FairGBM**[9]**:** *https://github.com/feedzai/fairgbm*

•**FairGBM** can be a potential solution to residual fairness.

Final Thoughts

•We cannot satisfy both **equalized odds difference** and **precision-to-negative predicted-value-fairness** in a realistic case[11]. We should realize the trade off between difference kinds of fairness and optimize over only one fairness metrics.

•**Reweighing method** trades one bias with another. Although it gives a good reduction of bias and results in few reduction in accuracy, the model considers the case of disadvantaged group with favorable outcome to be more important than that of advantaged group with favorable outcome. We should consider such fairness trade off before implementation.



(b) Best attainable trade-offs per algorithm.

Citations

1. Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. Knowledge and Information Systems, 33(1), 1-33.

2. Calmon, F., Weller, A., Wu, V., Sanchez, M., & Georgiou, T. (2017). Optimized pre-processing for discrimination prevention. In Advances in Neural Information Processing Systems (pp. 4114-4124).

3. Zhang, B., Lemoine, B., & Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (pp. 335-340).

4. Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In Advances in neural information processing systems (pp. 3315-3323).

5. UCI Machine Learning Repository. (1994). Statlog (German Credit Data). Retrieved from https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)

6. ProPublica. (2016). COMPAS dataset. Retrieved from https://www.propublica.org/datastore/dataset/compas-recidivism-risk-score-data-and-analysis

7. UCI Machine Learning Repository. (1996). Census Income Data Set. Retrieved from https://archive.ics.uci.edu/ml/datasets/census+income

8. Medical Expenditure Panel Survey (MEPS). Agency for Healthcare Research and Quality, Rockville, MD. https://meps.ahrq.gov/mepsweb/data_stats/download_data_files_detail.jsp?cboPufNumber=HC-181

9.Cruz, A. F., Belém, C., Jesus, S., Bravo, J., Saleiro, P., & Bizarro, P. (2023). FairGBM: Gradient Boosting with Fairness Constraints. In International Conference on Learning Representations.

10. Consulted Dimitri Bianco at Princeton Fintech and Quant Conference.

11. Berk, R., Heidari, H., Jabbari, S., Kearns, M., Roth, A., (2021). Fairness in Criminal Justice Risk Assessments: The State of the Art. In Sociology Methods & Research.