INFLUENCE OF DATA PROCESSING ON THE ALGORITHM FAIRNESS VS. ACCURACY TRADE-OFF: BUILDING PARETO FRONTS FOR EQUITABLE ALGORITHMIC DECISIONS

How can combining pre- and post-processing techniques optimize the fairness vs. accuracy trade-off within a Pareto front framework?

Name: Andres David Salvi
Email: andresdsalvi@gmail.com

Responsible Professor: Jie Yang
Supervisor: Sarah Carter

TU Delft

Introduction

Background

- Society dependent on algorithms (ML models, face recognition, ChatGPT, etc.)
- Algorithm bias common, but fairness methods can help, often with accuracy trade-off
- Weak research on tweaking trade-off; **especially with combining techniques**

Research Sub-Questions

1. How much could the selected fairness methods mitigate the group bias of a given (biased) dataset and system?
2. How could we use a Pareto-front to set a good accuracy vs. fairness trade-off?
3. How generalizable would the proposed fairness framework be with other datasets, models, and fairness methods?

Literature Review

Top Related Work

- Mortgage loan system biased against black people performed worse on both groups with off-the-shelf ML model & fairness methods [1]
- Framework for exploring the tradeoff while using pre- & post-processing fairness techniques individually and various metrics [2]

General Findings

- Many pre- & post-processing fairness methods; performance varies between datasets
- Different metrics influenced by different methods
- Pareto fronts show trade-offs well, one fairness & one accuracy metric

Experiment

Dataset

- Adult Income (predict if person's income >50k/year)
  - Protected attributes: race & gender
  - Selection & historical biases against females & black people

Outline

1. Clean up & prep dataset
2. Apply *Disparate Impact Remover (DIR);* 10 repair levels [3, 4]
3. Train *Logistic Regression* model
4. Apply *Reject Option Classiciation (ROC)* via SPD; binary [3, 4]
5. Build Pareto fronts using selected metrics
   - Test without methods applied, separately, and together

Evaluation

- Fairness: *Disparate Impact* & *Statistical Parity Diff. (SPD)* [3, 4]
- Accuracy: *Accuracy* & *Balanced Accuracy* (computed) [3, 4]
- Pareto fronts: trade-off between metrics
  - Select best point between pre- & post-ROC
  - Customizable fairness weight

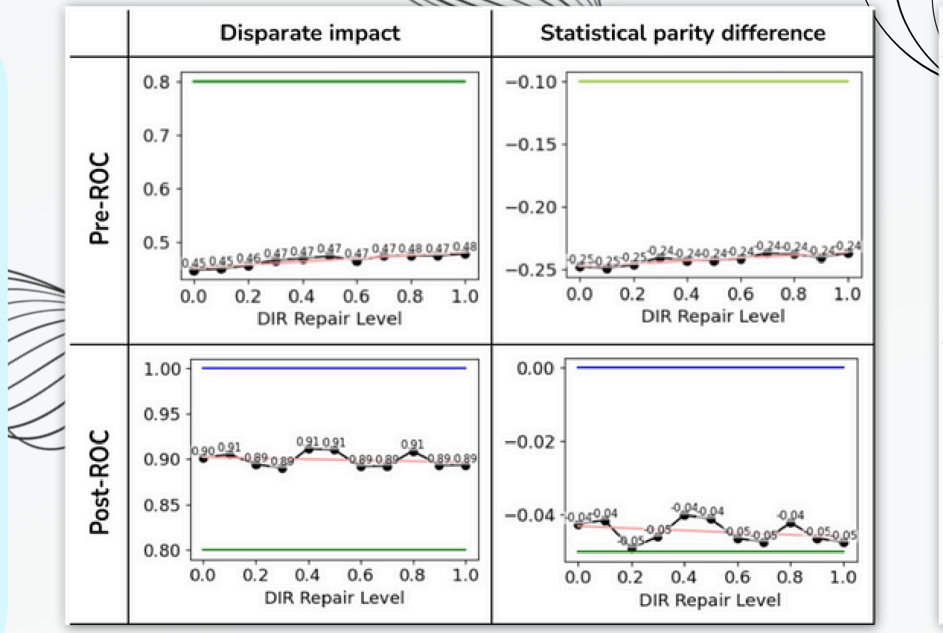

Figure 1: Fairness metrics for "gender" attribute

Discussion

Core Findings & Implications

- Non-linear trade-off typical; sometimes fairness methods did harm or accuracy improved
- Technique combination generally optimal, with exceptions (e.g., *Theil Index*)
- Pareto fronts can vary drastically between metric combos; make wide selection for solid overview and to reduce risk of missing strong trade-off or hefty cost

Limitations & Impacts

- Tested only DIR & ROC; other methods can yield different metric & front results
- Tested on one dataset (Adult); different dataset can yield different results
- Tested only on *Logistic Regression* model; other models may behave differently
- No error bar generated based on multiple model runs; increased margin of error of results

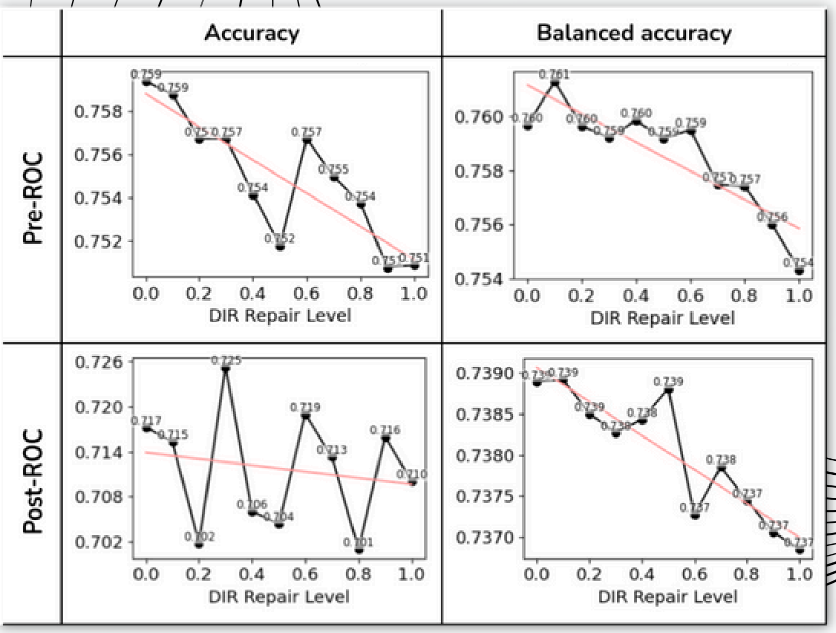

Figure 2: Accuracy metrics for "gender" attribute

Conclusions

Answering RQs

- **RQ 1:** Overall DIR gave modest improvement; ROC bigger improvements; optimal results typically with low DIR (0.1-0.5) & ROC applied
- **RQ 2**: Pareto fronts with trade-off scoring function for ranking to find optimal point
- **RQ 3**: If implementation tested with other datasets, models, methods, and many metrics, then likely yes!

Future Work

- Try other combinations of fairness methods, including in-processing
- Add more metrics & test on other datasets, also with cross-influential protected attributes
- Test on five other supported models (e.g., KNN or SVM), or add more
- Create error bars from multiple identical runs to reduce margin of error

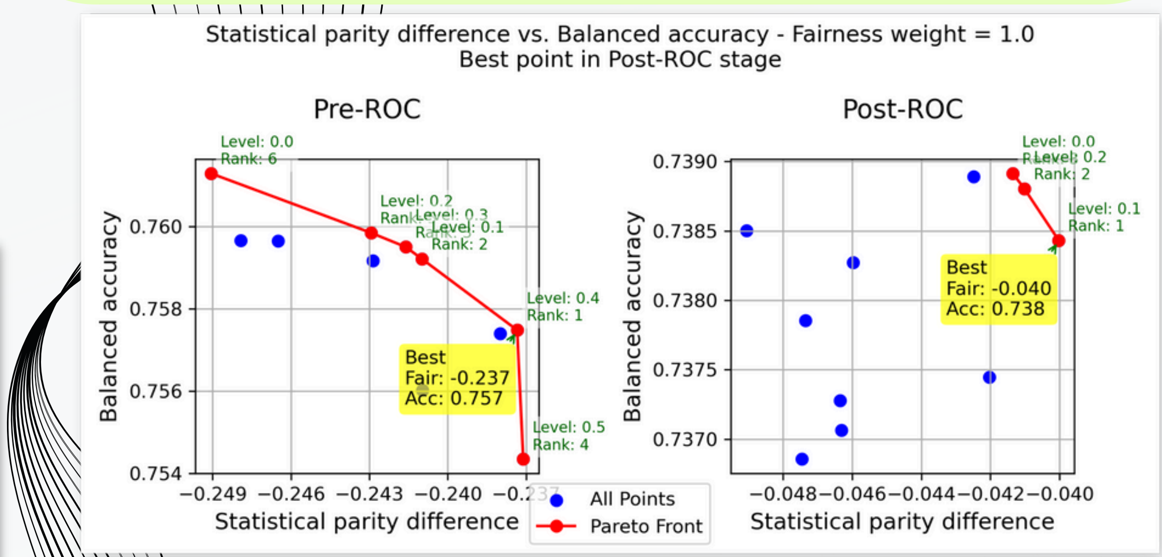

Figure 3: Pareto front pair between SPD & Balanced Accuracy for "gender" attribute

References

[1] Leying Zou and Warut Khern-am nuai. Ai and housing discrimination: the case of mortgage applications. AI and Ethics, 3(4):1271ã1281, November 2022
[2] Christian Haas. The price of fairness - a framework to explore trade-offs in algorithmic fairness, 2019
[3] Patrick Janssen and Bert M. Sadowski. Bias in algorithms: On the trade-off between accuracy and fairness, Jan 2021.
[4] Barry Becker. Adult, 1996.

GitHub Link: https://github.com/1Sulture/Fairness-vs.-Accuracy-Pareto-Front-Builder