Bursting the Burden Bubble?

An Assessment of Sharma et al.'s Counterfactual-Based Fairness Metric

Yochem van Rosmalen (*y.m.vanrosmalen@students.uu.nl*)
Florian van der Steen (*f.a.vandersteen@students.uu.nl*)
Sebastiaan Jans (*s.j.j.jans@students.uu.nl*)
Daan van der Weijden (*d.j.weijden@students.uu.nl*)

November 8, 2022

Utrecht University, The Netherlands

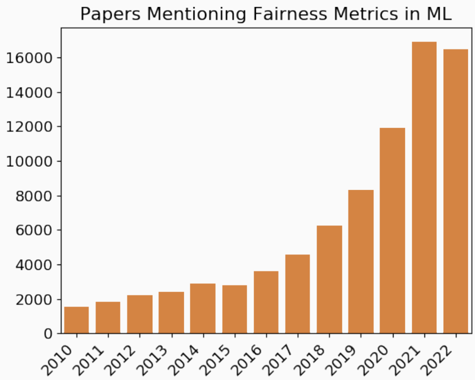
Outline

1. Fairness
   - Statistical Parity
   - Burden
2. Experiments
   - Synthetic Datasets
   - Taiwan Dataset
3. Conclusions
4. Future Work

Fairness

Fairness

- What is fairness?
- Many aspects of fairness metrics



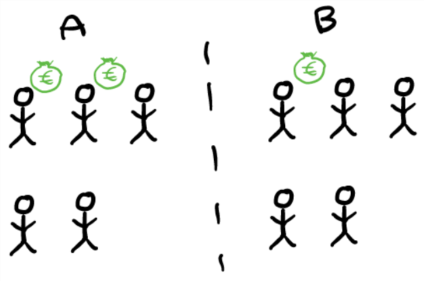
Statistical Parity

Statistical/Demographic Parity (SP_S) [2]:
Ratio of acceptance rates (AR_S).

$$SP_S = \frac{AR_{S=A}}{AR_{S=B}} = \frac{P(\hat{Y}=1|S=A)}{P(\hat{Y}=1|S=B)}$$

Perfect parity: $SP_S = 1$

80% rule [1]: $SP_S \geq 0.8$ is acceptable.



$\frac{1/5}{2/5} = 0.5 < 0.8$

No parity!

$\hat{Y}$: Model's prediction, S: Sensitive attribute/group

Burden

- Sharma et al.'s CERTIFAI framework [3] (2017)
- CognitiveScale
- Multiple domains
- Model Agnostic
- Counterfactuals
    - Not causal!
    - Generated with genetic algorithm

CERTIFAI: A Common Framework to Provide Explanations and Analyse the Fairness and Robustness of Black-box Models

Shubham Sharma
University of Texas at Austin
Austin, Texas
shubham_sharma@utexas,edu

Jette Henderson
CognitiveScale
Austin, Texas
jhenderson@cognitivescale.com

Joydeep Ghosh
CognitiveScale
Austin, Texas
jghosh@cognitivescale.com

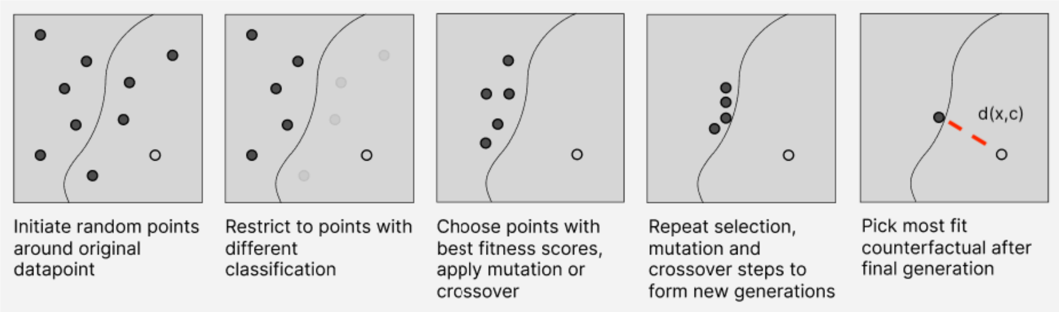
ABSTRACT

Concerns within the machine learning community and external pressures from regulators over the vulnerabilities of machine learning algorithms have spurred on the fields of explainability, robustness, and fairness. Often, issues in explainability, robustness, and fairness are confined to their specific sub-fields and few tools exist for model developers to use to simultaneously build their modeling pipelines in a transparent, accountable, and fair way. This can lead

ACM Reference Format:

Shubham Sharma, Jette Henderson, and Joydeep Ghosh. 2020. CERTIFAI: A Common Framework to Provide Explanations and Analyse the Fairness and Robustness of Black-box Models. In *Proceedings of the 2020 AAAI/ACM Conference on AI, Ethics, and Society (AIES '20), February 7–8, 2020, New York, NY, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3375627.3375812

Counterfactual generation



Evolutionary algorithm for the generation of realistic counterfactuals. Illustration adapted from [3].

Burden

- Distance to counterfactual →individual *recourse*.
- Average distance to counterfactual over instances in a group s, with $\mathbf{c}^*$ the found counterfactual(s):

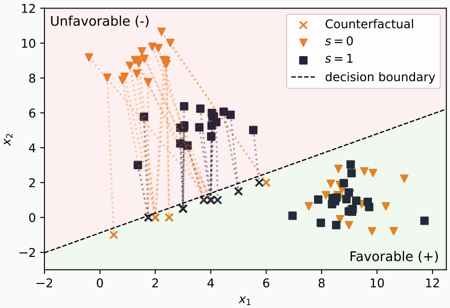
$$\text{Burden}_{S=s} = \mathbb{E}_{S=s}[d(\mathbf{x}, \mathbf{c}^*)]$$

Experiments

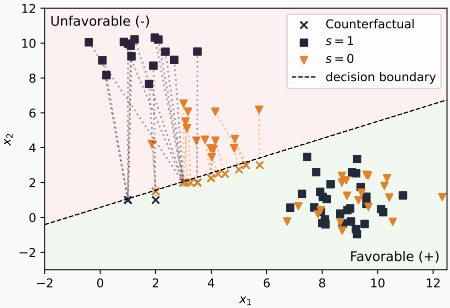
Synthetic datasets

- Goal: Highlighting theoretical difference between metrics
- What the metrics measure:
  - Burden: Estimated distance to counterfactual
  - SP: Rate of favorably classified
- Approach:
  - Dataset D_A: $AR_{S=0} = AR_{S=1}$, $\text{Burden}_{S=0} > \text{Burden}_{S=1}$
  - Dataset D_B: $AR_{S=0} > AR_{S=1}$, $\text{Burden}_{S=0} < \text{Burden}_{S=1}$

Results Synthetic Datasets



(a) D_A, where Burden and statpar disagree on the presence of unfairness.



(b) D_B, where Burden and statpar disagree on the direction of unfairness.

Real world dataset

- Default of Credit Card Clients Data Set, “Taiwan” [4]
    - Target: did the person default on loan?
    - 30,000 instances (1000 counterfactuals)
    - 4 Sensitive attributes (dropped for training)
    - All features concerning account balances
- Logistic Regression with 78% accuracy

Results

	Acceptance Rate		SP	Burden		
Dataset	$S = 0$	$S = 1$	$0/1$	$S = 0$	$S = 1$	$0/1$
D_A	0.500	0.500	1.00	**11.6**	4.65	2.49
D_B	**0.571**	0.667	0.857	3.31	**11.0**	0.302
Taiwan	0.967	**0.948**	1.02	**1.38**	0.940	1.47

Underprivileged group according to metric in **boldface**.

For Taiwan: 0 is women, 1 is men.

Conclusions

Conclusions

- Burden can provide more nuance than Statistical Parity
- Computational cost of Burden is high
- Burden can be used in addition to Statistical Parity

Concluding, be mindful when using a single fairness metric!

Future Work

Future Work

- Reduce computational complexity of Burden
- Find real-world datasets with SP and Burden disagree on the direction of unfairness

Questions?

References i

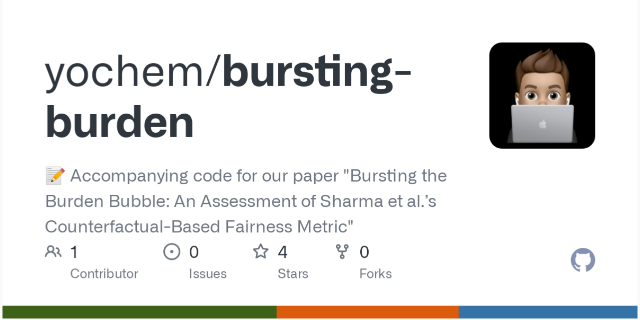
M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian.
Certifying and Removing Disparate Impact.
International Conference on Knowledge Discovery and Data Mining (KDD), 21(1):259––268, August 2015.

M. Hardt, E. Price, E. Price, and N. Srebro.
Equality of Opportunity in Supervised Learning.
Advances in Neural Information Processing Systems (NIPS), 29(1), 2016.

S. Sharma, J. Henderson, and J. Ghosh.
CERTIFAI: A Common Framework to Provide Explanations and Analyse the Fairness and Robustness of Black-Box Models.
In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, pages 166––172, New York, NY, 2020. Association for Computing Machinery.

I.-C. Yeh and C.-h. Lien.
The Comparisons of Data Mining Techniques for the Predictive Accuracy of Probability of Default of Credit Card Clients.
Expert Systems with Applications, 36(2):2473––2480, March 2009.

Find the slides, paper and code on GitHub!



github.com/yochem/bursting-burden