Measuring Fairness under Unawareness of Sensitive Attributes via Quantification

A. Fabris*, A. Esuli,
A. Moreo, F. Sebastiani

*fabrisal@dei.unipd.it
University of Padua

Goals

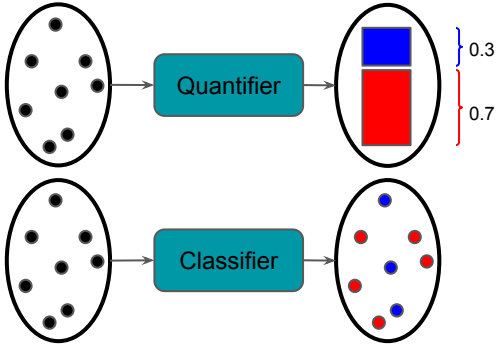
GOAL 1: Measuring fairness under unawareness of sensitive attributes.
GOAL 2: Decoupling group-level and individual-level inferences to avoid model misuse.

Background Summary

- Knowledge of sensitive attributes is necessary to measure group fairness.
- Sensitive attributes are often unavailable due to legislation, privacy requirements, data minimization, or prospect of negative media coverage [1,2].
- Fairness under unawareness is a setting of high practical interest which received little attention from the community [3].

Enter Quantification

- Quantifiers estimate class prevalence rather than individual membership [4]. They act on samples and output one value in [0,1].



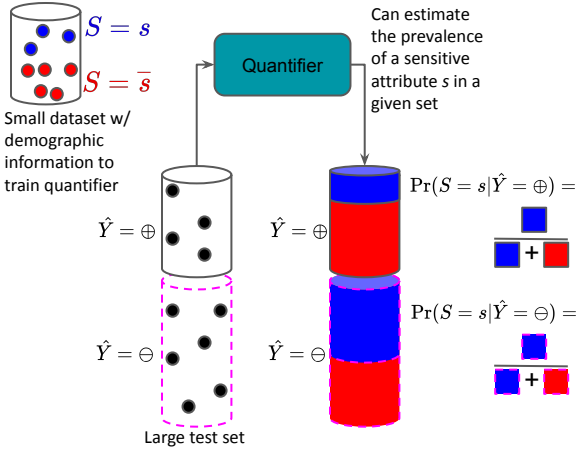
Key Proposition

Observational measures of algorithmic fairness, such as parity of acceptance rate, TP, TN, FP, and FN can be computed, under unawareness of sensitive attributes, by estimating the prevalence of the sensitive attribute in specific subsets of the test set.

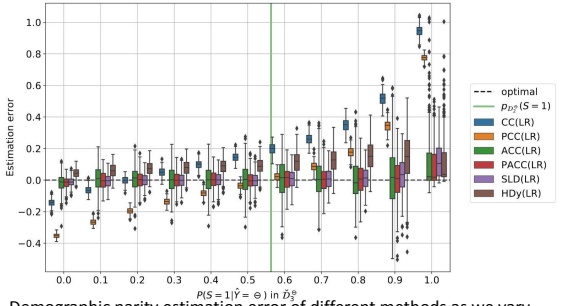
Proof. Sketch for demographic parity, i.e. parity of $\Pr(\hat{Y}=\oplus)$

$$\Pr(\hat{Y}=\oplus|S=s)=\overbrace{\Pr(S=s|\hat{Y}=\oplus)}^{\text{EST.}}\frac{\overbrace{\Pr(\hat{Y}=\oplus)}^{\text{KNOWN}}}{\underbrace{\Pr(S=s|\hat{Y}=\oplus)}_{\text{EST.}}\underbrace{\Pr(\hat{Y}=\oplus)}_{\text{KNOWN}}+\underbrace{\Pr(S=s|\hat{Y}=\ominus)}_{\text{EST.}}\underbrace{\Pr(\hat{Y}=\ominus)}_{\text{KNOWN}}}$$

Method



Results



Demographic parity estimation error of different methods as we vary the sensitive attribute prevalence in the test set.

- CC and PCC (prior art [3]) are outperformed by quantification methods ACC, PACC, SLD and HDy.

Conclusions

- Measuring fairness under unawareness can be cast as a prevalence estimation problem and effectively solved by methods of proven consistency from the quantification literature.
- Quantifiers can provide robust estimates even when trained on small auxiliary datasets with distribution drift.
- Quantifiers offer a path to decouple the (desirable) objective of estimating sensitive attributes at the group level from the (undesirable) side effect of inference at the individual level.

Essential Bibliography

[1] Bogen, M., Rieke, A., & Ahmed, S. (2020). Awareness in practice: Tensions in access to sensitive attribute data for antidiscrimination. (FAT* '20)
[2] Andrus, M., & Villeneuve, S. (2022). Demographic-reliant algorithmic fairness: Characterizing the risks of demographic data collection in the pursuit of fairness. (FAccT '22)
[3] Chen, J., Kallus, N., Mao, X., Svacha, G., & Udell, M. (2019). Fairness under unawareness: Assessing disparity when protected class is unobserved. (FAT* '19),
[4]Gonzalez, P., Castaño, A., Chawla, N. V., & del Coz, J. J. (2017). A review on quantification learning. (ACM Computing Surveys)