Parametrised Data Sampling for Fairness Optimisation

Vladimiro G. Zelaya
c.v.gonzalez-zelaya2@ncl.ac.uk

Paolo Missier
paolo.missier@ncl.ac.uk

Dennis Prangle
dennis.prangle@ncl.ac.uk

Introduction

Data preprocessing method to enforce fairness on machine learning classification tasks.

- Model and fairness-definition agnostic
- Correction level *tuned* for optimal fairness

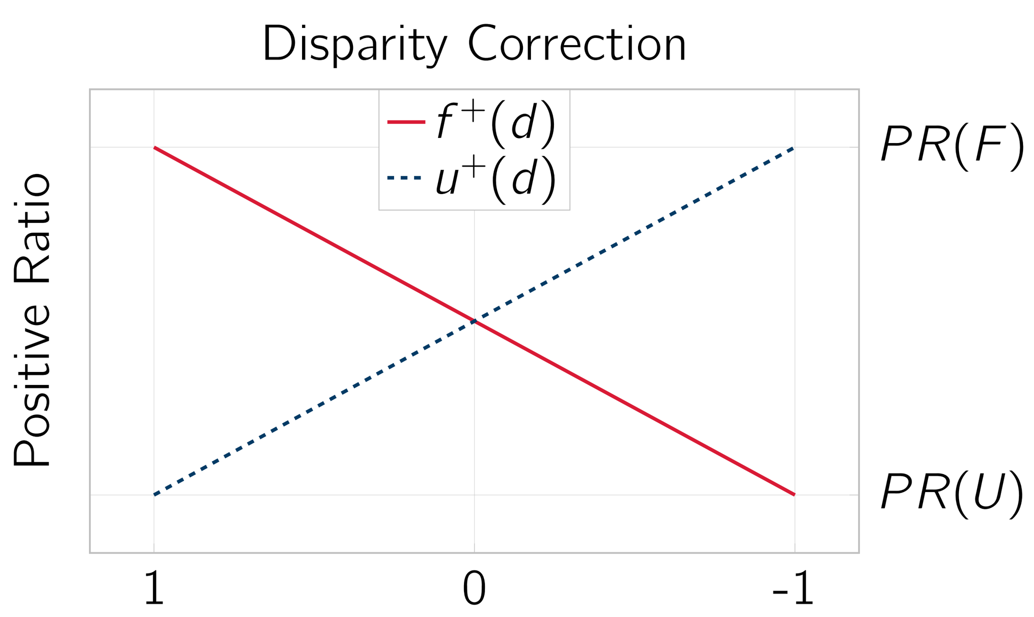
Population Subgroups

We split the train set into four groups:

By Protected Attribute:		By Class Label:	
Favoured	(F)	Positive	$(+)$
Unfavoured	(U)	Negative	$(-)$

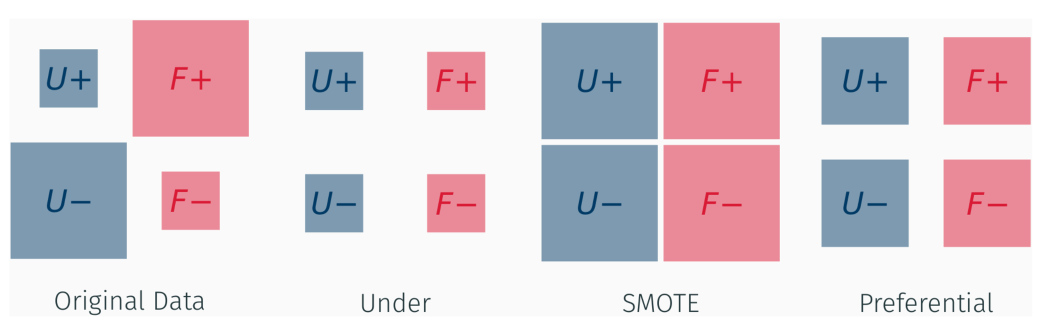
Train Set Correction

Subgroups are resampled to modify F and U *positive ratio* (PR), depending on $d \in [-1, 1]$.



Sampling Strategies

Resampling may be performed in different ways:



Biased data may lead to **unfair classification** of individuals.

We **restore fairness** through data preprocessing.

Scan for full paper, this poster and Jupyter Notebooks!

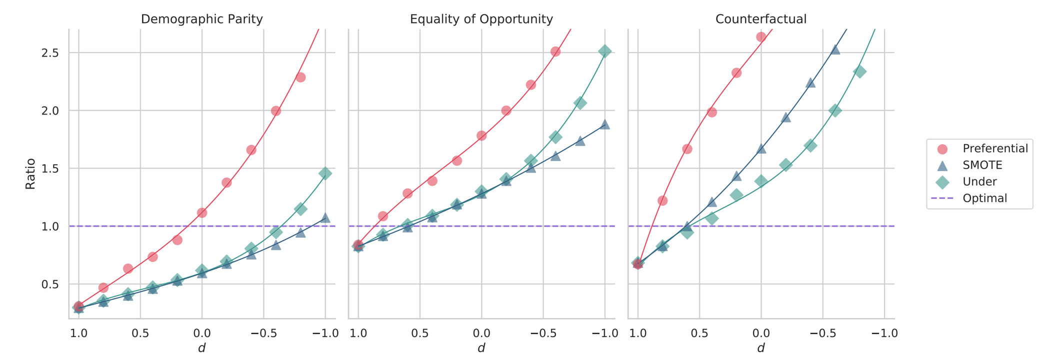
Fairness Definitions

Demographic Parity	$DPR = \dfrac{P(\hat{Y}=1 \mid PA=U)}{P(\hat{Y}=1 \mid PA=F)}$
Equality of Opportunity	$EOR = \dfrac{P(\hat{Y}=1 \mid PA=U, Y=1)}{P(\hat{Y}=1 \mid PA=F, Y=1)}$
Counterfactual (Proxy)	$CFR = \dfrac{PR(Test_{PA\leftarrow U})}{PR(Test_{PA\leftarrow F})}$

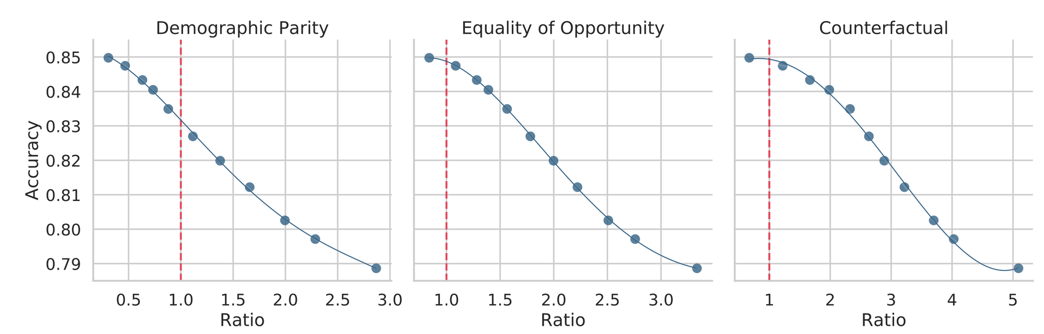
Experiments

Dataset	Protected	Favoured	Positive Class	Instances
COMPAS	Race	White	Won't reoffend	6907
Credit	Gender	Male	Will repay loan	1000
Income	Gender	Male	Income > $50k	48842

Fairness Correction



Accuracy Trade-off



Conclusion

Our method optimises classifier fairness with a small loss in accuracy.

EPSRC
Engineering and Physical Sciences Research Council

Newcastle University
Digital Institute