Parametrised Data Sampling for Fairness Optimisation

Vladimiro G. Zelaya, Paolo Missier and Dennis Prangle

Fairness, Transparency, Privacy, The Alan Turing Institute

4 October 2019

EPSRC
Engineering and Physical Sciences Research Council

Newcastle University
Digital Institute

What is This Talk About?

- Method for correcting *classifier fairness*
- *Model* and *definition* agnostic
- *Tune* correction level to optimise fairness

A Few Definitions

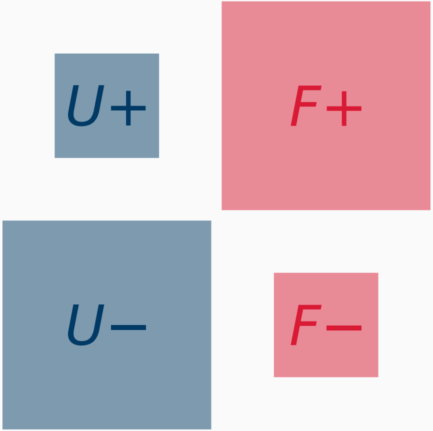
Protected Attribute (PA) Attribute on which unfairness is going to be corrected

Positive Ratio (PR) Proportion of positive labels in a data set

Favoured group (F) PA subgroup with *highest* PR

Unfavoured group (U) PA subgroup with *lowest* PR

Population Subgroups



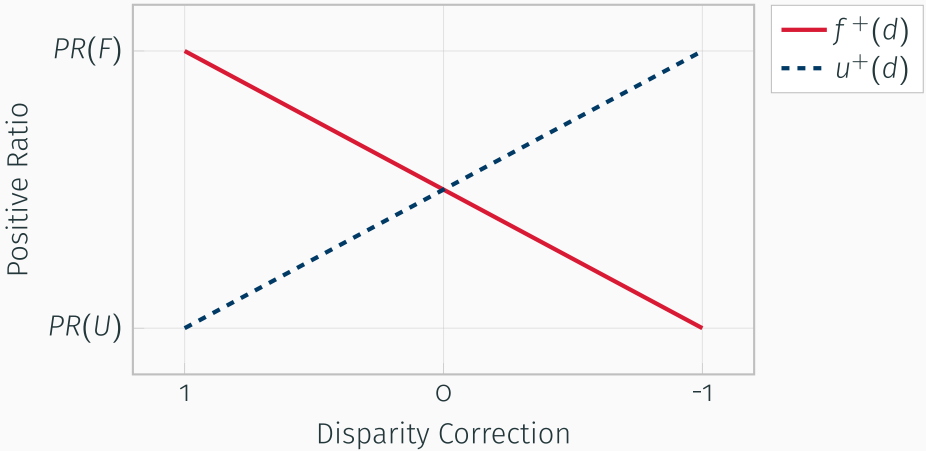
By Protected Attribute (PA):

Favoured	(F)
Unfavoured	(U)

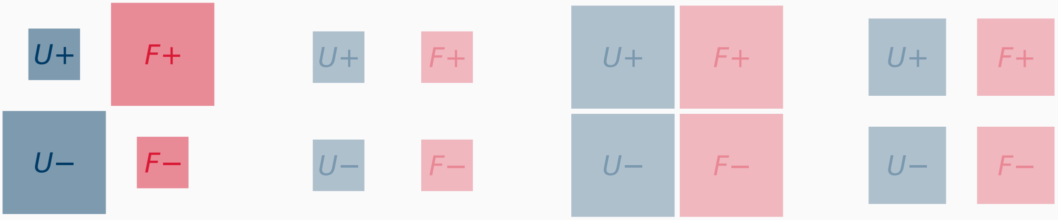
By Class Label:

Positive	($+$)
Negative	($-$)

Train Set Correction



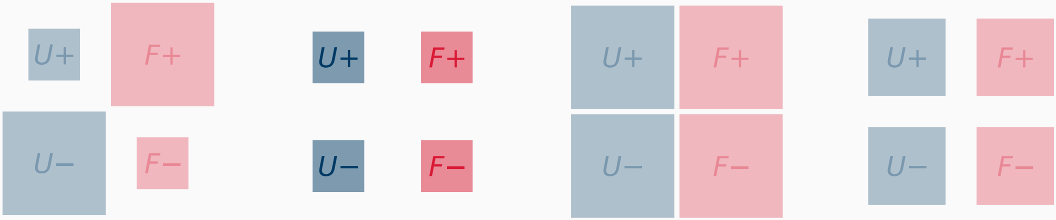
Sampling Strategies



Original Data   Under   SMOTE [1]   Preferential [2]

[1][Chawla et al., 2002]
[2][Kamiran and Calders, 2010]

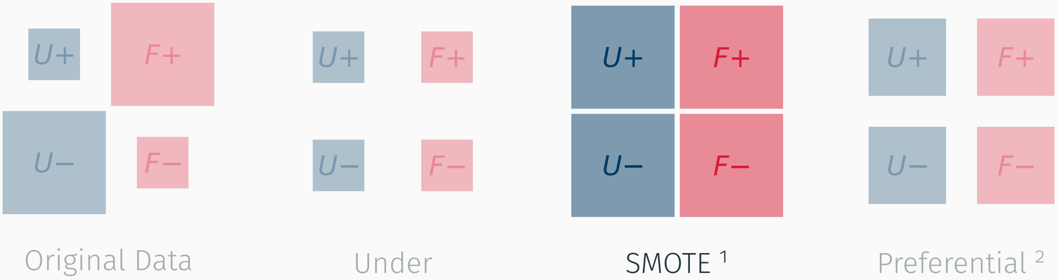
Sampling Strategies



Original Data | Under | SMOTE [1] | Preferential [2]

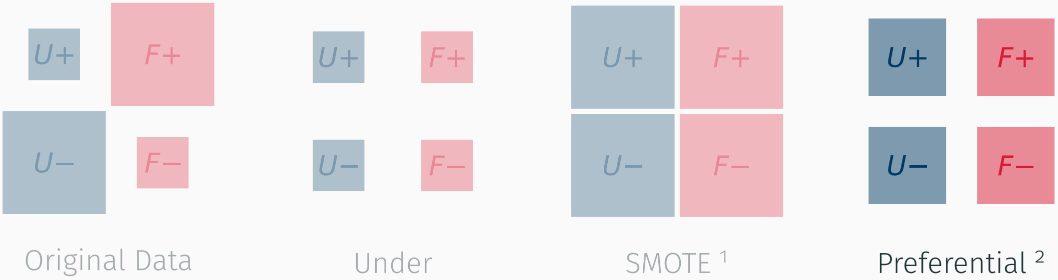
[1] [Chawla et al., 2002]
[2] [Kamiran and Calders, 2010]

Sampling Strategies



[1] [Chawla et al., 2002]
[2] [Kamiran and Calders, 2010]

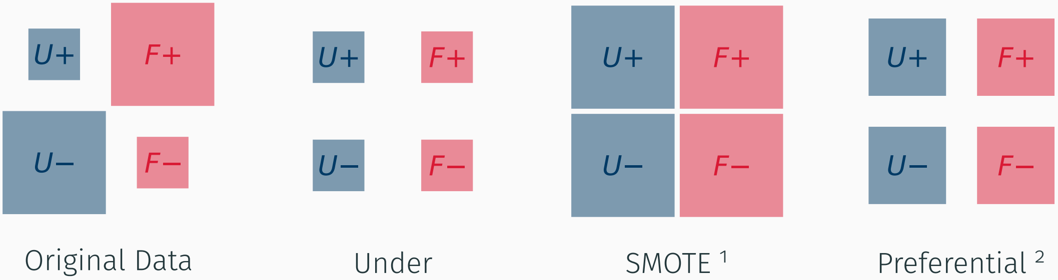
Sampling Strategies



[1] [Chawla et al., 2002]

[2] [Kamiran and Calders, 2010]

Sampling Strategies



[1] [Chawla et al., 2002]
[2] [Kamiran and Calders, 2010]

Ratio Form of Fairness Definitions

Equality Form

$$P\left(\hat{Y} = 1 \mid PA = U\right) = P\left(\hat{Y} = 1 \mid PA = F\right)$$

Ratio Form

$$\frac{P\left(\hat{Y} = 1 \mid PA = U\right)}{P\left(\hat{Y} = 1 \mid PA = F\right)} = 1$$

Ratio Form of Fairness Definitions

Equality Form $P\left(\hat{Y}=1 \mid PA=U\right) = P\left(\hat{Y}=1 \mid PA=F\right)$

Ratio Form $\dfrac{P\left(\hat{Y}=1 \mid PA=U\right)}{P\left(\hat{Y}=1 \mid PA=F\right)} = 1$

Some Fairness Ratios

Demographic Parity

$$DPR = \frac{P\left(\hat{Y} = 1 \mid PA = U\right)}{P\left(\hat{Y} = 1 \mid PA = F\right)}$$

Equality of Opportunity

$$EOR = \frac{P\left(\hat{Y} = 1 \mid PA = U,\ Y = 1\right)}{P\left(\hat{Y} = 1 \mid PA = F,\ Y = 1\right)}$$

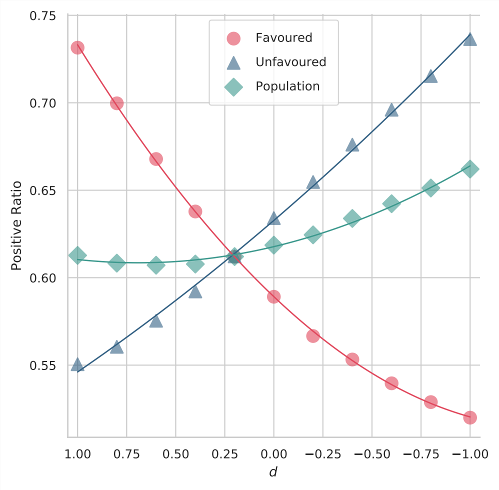
Counterfactual (Proxy)

$$CFR = \frac{PR\left(Test_{PA \leftarrow U}\right)}{PR\left(Test_{PA \leftarrow F}\right)}$$

Experiments

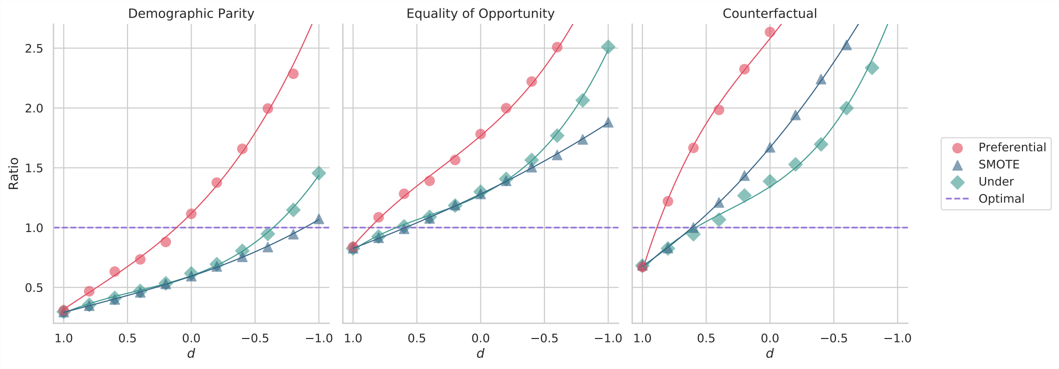
Dataset	Protected	Favoured	Positive Class	Instances
COMPAS	Race	White	Won’t reoffend	6907
Credit	Gender	Male	Will repay loan	1000
Income	Gender	Male	Income > $50k	48842

Effects on Test Set (COMPAS, Undersampling)



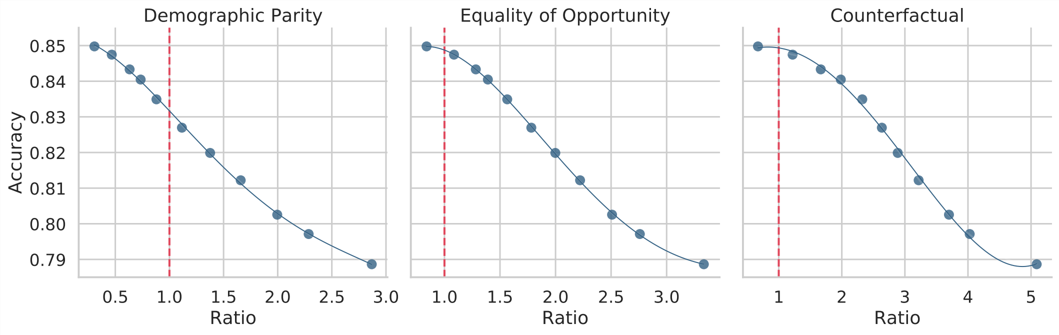
- Effect is correlated with correction
- But it occurs to a different extent
- Intersection is *not* at $d = 0$

Optimal Correction by Fairness and Sampling



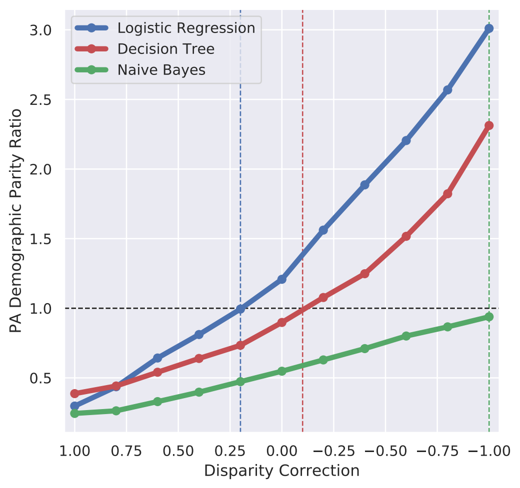
Plots for *Income* dataset

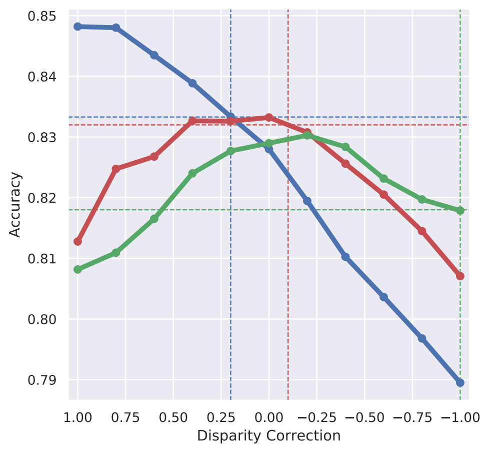
Accuracy vs Fairness Trade-off



Plots for *Income* dataset corrected by Preferential Sampling

Classifier Comparison





How to extend it?

- Make the PA multi-class
- Have more than one PA

Combined Protected Attribute

Age	Country	Gender	Race	Combined PA
(20-30]	Portugal	Male	Black	
(30-40]	France	Female	White	

Combined Protected Attribute

Age	Country	Gender	Race	Combined PA
(20-30]	Portugal	Male	Black	
(30-40]	France	Female	White	

Combined Protected Attribute

	Age	Country	Gender	Race	Combined PA
	(20-30]	Portugal	Male	Black	
	(30-40]	France	Female	White	
Subgroup PR	0.2	0.3	0.4	0.1	

Combined Protected Attribute

	Age	Country	Gender	Race	Combined PA
	(20-30]	Portugal	Male	Black	
	(30-40]	France	Female	White	
Subgroup PR	0.2	0.3	0.4	0.1	
Dataset PR	0.3	0.3	0.3	0.3	

Combined Protected Attribute

	Age	Country	Gender	Race	Combined PA
	(20-30]	Portugal	Male	Black	
	(30-40]	France	Female	White	
Subgroup PR	0.2	0.3	0.4	0.1	
Dataset PR	0.3	0.3	0.3	0.3	
Difference	−0.1	+0.0	+0.1	−0.2	

Combined Protected Attribute

Age	Country	Gender	Race	Combined PA
(20-30]	Portugal	Male	Black	
(30-40]	France	Female	White	

	Age	Country	Gender	Race	
Subgroup PR	0.2	0.3	0.4	0.1	
Dataset PR	0.3	0.3	0.3	0.3	
Difference	-0.1	$+0.0$	$+0.1$	-0.2	Sum $= -0.2$

Combined Protected Attribute

	Age	Country	Gender	Race	Combined PA
	(20-30]	Portugal	Male	Black	Unfavoured
	(30-40]	France	Female	White	
Subgroup PR	0.2	0.3	0.4	0.1	
Dataset PR	0.3	0.3	0.3	0.3	
Difference	-0.1	$+0.0$	$+0.1$	-0.2	Sum $= -0.2$

Combined Protected Attribute

	Age	Country	Gender	Race	Combined PA
	(20-30]	Portugal	Male	Black	Unfavoured
	(30-40]	France	Female	White	
Subgroup PR					
Dataset PR	0.3	0.3	0.3	0.3	
Difference					

Combined Protected Attribute

	Age	Country	Gender	Race	Combined PA
	(20-30]	Portugal	Male	Black	Unfavoured
	(30-40]	France	Female	White	
Subgroup PR	0.4	0.4	0.1	0.4	
Dataset PR	0.3	0.3	0.3	0.3	
Difference					

Combined Protected Attribute

	Age	Country	Gender	Race	Combined PA
	(20-30]	Portugal	Male	Black	Unfavoured
	(30-40]	France	Female	White	
Subgroup PR	0.4	0.4	0.1	0.4	
Dataset PR	0.3	0.3	0.3	0.3	
Difference	+0.1	+0.1	−0.2	+0.1	

Combined Protected Attribute

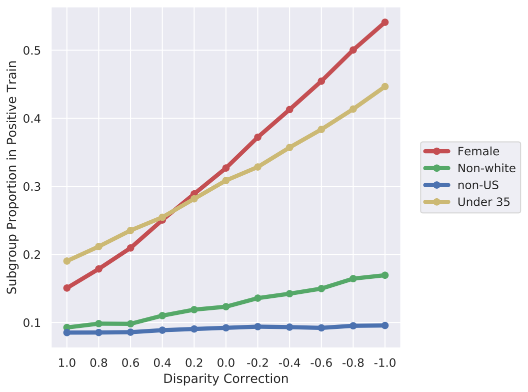
	Age	Country	Gender	Race	Combined PA
	(20-30]	Portugal	Male	Black	Unfavoured
	(30-40]	France	Female	White	
Subgroup PR	0.4	0.4	0.1	0.4	
Dataset PR	0.3	0.3	0.3	0.3	
Difference	$+0.1$	$+0.1$	-0.2	$+0.1$	Sum $= +0.1$

Combined Protected Attribute

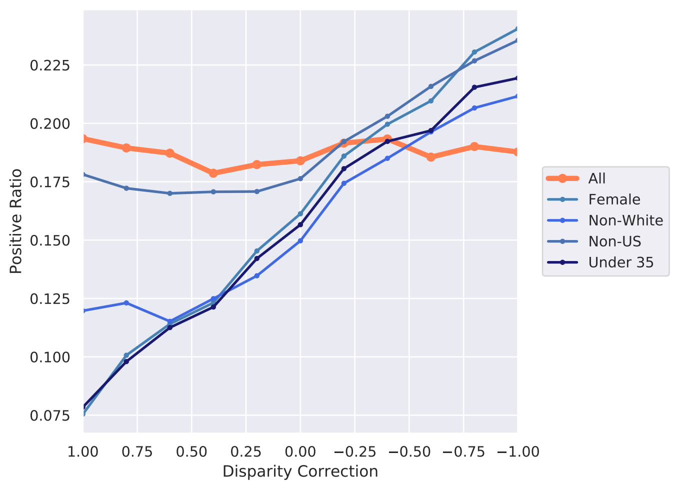
	Age	Country	Gender	Race	Combined PA
	(20-30]	Portugal	Male	Black	Unfavoured
	(30-40]	France	Female	White	Favoured
Subgroup PR	0.4	0.4	0.1	0.4	
Dataset PR	0.3	0.3	0.3	0.3	
Difference	$+0.1$	$+0.1$	-0.2	$+0.1$	Sum $= +0.1$

Unfavoured Subgroup Proportions in Positive Train Set

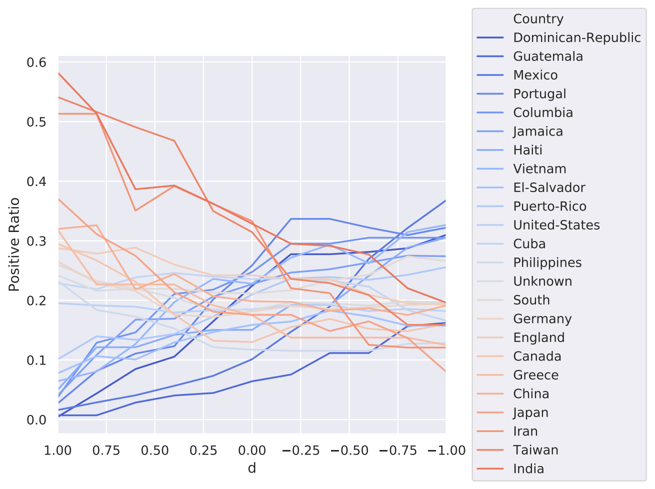
PA Subgroup	PR Difference
Female	−0.13
Non-US	−0.04
Non-White	−0.09
Under 35	−0.13



Multi-PA Correction



PR Correction for Original PAs



Conclusions

- Fairness-agnostic optimisation with a relatively small loss in accuracy
- Ideal correction level is definition dependant
- Different sampling strategies produced similar results

Future Work

- Optimise for more than one fairness definition
- Optimise for fairness and accuracy
- Worry about fairness *Gerrymandering* [3]

[3] [Kearns et al., 2019]

Thank You!

These slides, XAI paper and Jupyter Notebooks:

`https://github.com/vladoxNCL/fairCorrect`

c.v.gonzalez-zelaya2@ncl.ac.uk

For Further Reading

- Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002).
  SMOTE: Synthetic Minority Over-sampling Technique.
  Journal of Artificial Intelligence Research, 16:321–357.
- Kamiran, F. and Calders, T. (2010).
  Classification with no discrimination by preferential sampling.
  In *Proc. 19th Machine Learning Conf. Belgium and The Netherlands*, pages 1–6. Citeseer.
- Kearns, M., Neel, S., Roth, A., and Wu, Z. S. (2019).
  An empirical study of rich subgroup fairness for machine learning.
  In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 100–109. ACM.