Fairness Through the Lens of Proportional Equality

Extended Abstract

Arpita Biswas
Indian Institute of Science
Bangalore, Karnataka, India
arpitab@iisc.ac.in

Suvam Mukherjee
Microsoft Research
Bangalore, Karnataka, India
t-sumukh@microsoft.com

ABSTRACT

Today, automated algorithms, such as machine learning classifiers, are playing an increasingly pivotal role in important societal decisions such as hiring, loan allocation, and criminal risk assessment. This motivates the need to probe the outcomes of a prediction model for discriminatory traits towards specific groups of individuals. In this context, one of the crucial challenges is to formally define a satisfactory notion of *fairness*. Our contribution in this paper is to formalize *Proportional Equality* (PE) as a fairness notion. We additionally show that it is a more appropriate criterion than the existing popular notion called *Disparate Impact* (DI), which is used for evaluating the fairness of a classifier's outcomes.

KEYWORDS

Classification; Discrimination; Racial bias; Gender bias; Prior probability shifts; Fairness concepts

ACM Reference Format:
Arpita Biswas and Suvam Mukherjee. 2019. Fairness Through the Lens of Proportional Equality. In *Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019,* IFAAMAS, 3 pages.

1 INTRODUCTION

Recently, algorithmic fairness has been receiving significant attention from a wide spectrum of research communities like social science, computer science and statistics [3, 4, 6, 8, 11, 15, 18, 22, 23, 27, 30, 32]. Intelligent algorithms are playing an increasingly decisive role in real-world scenarios such as hiring, loan allocation, and criminal risk assessment. However, recent studies [1, 29] have shown that the predictions made by these algorithms often exhibit discrimination towards (one or more) population subgroups. Dependence on these unfair predictions for societal decisions is not only unethical, but also, in some cases, punishable by law [1, 7]. Thus, ensuring freedom from such discriminatory traits in the classifier predictions is of great importance. This motivates the need to define a fairness notion which appropriately captures such discriminatory traits in predictions.

This is precisely the focus of this paper—we formalize a fairness notion called *Proportional Equality* (PE). While proportional equality has been studied and discussed in philosophy [10, 20], to the best of our knowledge, this paper is the first to formalize PE as a measure of fairness in classification. We show that PE is a more appropriate fairness notion than the existing criterion called Disparate Impact (DI) [5, 9, 12, 21, 31]. We additionally show that PE satisfies some desirable properties of fairness notions. In particular, Theorem (3.1) establishes that a perfectly accurate classifier is always PE-fair.

Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), N. Agmon, M. E. Taylor, E. Elkind, M. Veloso (eds.), May 13–17, 2019, Montreal, Canada. © 2019 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). All rights reserved.

2 PRELIMINARIES AND NOTATIONS

We use $\mathcal{X} \subset \mathbb{R}^m$ to denote the feature space with m features, $\mathcal{Y} = \{0, 1\}$ to denote the label space and $Z = \{0, 1\}$ to denote the sensitive feature. The training data $D = \{(x_i, z_i, y_i)_{i=1}^N\}$ is assumed to be drawn from an unknown joint distribution $\mathcal{P}$ over $\mathcal{X} \times Z \times \mathcal{Y}$. Throughout the paper, the variable X denotes a feature vector[1] and the variable Y denotes a label.

2.1 Classification Problem

The goal of a classification problem is to learn a function $\hat{h} : \mathcal{X} \mapsto \mathcal{Y}$, defined in some hypothesis space $\mathcal{H}$, such that $\hat{h}$ minimizes some target loss function—say, misclassification error:

$$\hat{h} := \arg\min_{h \in \mathcal{H}} \mathbb{E}_{(X,Y)\sim\mathcal{P}}[\mathbb{I}\{h(X) \neq Y\}].$$

The performance of the classifier is then measured using a new set of data, *test dataset* $d = \{(x_j, y_j)_{j=1}^n\}$, by observing how accurate the predicted labels $\hat{h}(x_j)$'s, are with respect to the true labels y_j's.

The classification problem assumes that both the training and test datasets are independently and identically drawn from some unknown distribution $\mathcal{P}$ and, consequently, these two datasets represent the same probability distribution $\mathcal{P}(X, Y) = \mathcal{P}(Y) \cdot \mathcal{P}(X|Y)$. However, in practice, this assumption often does not hold true, i.e., $\mathcal{P}(X, Y)$ differs between the training and test datasets. This change is popularly known as *dataset shift*, which occurs when either the conditional probability $\mathcal{P}(X|Y)$ or the prior probability $\mathcal{P}(Y)$ changes. The phenomenon where the prior probability $\mathcal{P}(Y)$ changes between the training and test datasets, but the class conditional probability $\mathcal{P}(X|Y)$ remains unaltered, is known as *prior probability shift* [25, 26]. In this work, we focus on *prior probability shifts*, as such changes naturally occur in many real-world scenarios. A typical example of this phenomenon is medical diagnosis. For instance, the fraction of patients having a certain disease Y may vary over a period of time (i.e., $\mathcal{P}(Y = 1)$ may vary). However, the likelihood of the series of symptoms that appears when the disease occurs (i.e., $\mathcal{P}(X|Y = 1)$) remains constant. It is important to be aware of such phenomenon, ignoring which may result in drastic performance reduction of the classifiers that rely on the assumption that the distribution is unaltered between training and test data.

[1]Note that the sensitive feature Z is treated separately from the other d features only when there is a fairness concern with respect to that sensitive feature. In the absence of such fairness concerns, Z can be thought of as a dimension in the feature space $\mathcal{X}$. Thus, while defining problems without fairness constraints, we use X to denote *all* the features, including sensitive ones (if any).

We study this phenomenon to address such changes within groups, differentiated by a sensitive attribute Z, i.e., for each $z \in \{0, 1\}$, the conditional probability $\mathcal{P}(X|Y = 1, Z = z)$ remains constant but the prior probability $\mathcal{P}(Y = 1|Z = z)$ changes over a period of time. Empirically, prior probability shift leads to drastically different values of true prevalences among training dataset D and test dataset d (that is, $p_D^z \neq p_d^z$). Here, the *true prevalence* within each group $z \in \{0, 1\}$ for any dataset d is given as

$$p_d^z := \frac{|\{(x_i, z_i, y_i) \in d \mid y_i = 1, z_i = z\}|}{|\{(x_i, z_i, y_i) \in d \mid z_i = z\}|} \tag{1}$$

In such a situation, minimizing misclassification error may not be enough and the predictions need to be regulated using some fairness constraints. One of the popular metrics to identify such unfair decisions is *disparate impact* (DI), which we discuss next.

2.2 Disparate Impact as a Fairness Measure

A classifier's decisions are said to be free from *disparate impact* if, for $z, z' \in \{0, 1\}$, the prediction prevalence ratio $\hat{\varrho}_d^{z,z'} := \hat{\rho}^z / \hat{\rho}^{z'}$ (that is, the ratio of prediction prevalences on the test data of one group to another) is at least 0.8, where the prediction prevalence for a group z is given by $\hat{\rho}_d^z := \frac{\left|\{(x_i, y_i, z_i) \in d : \hat{h}(x_i) = 1, z_i = z\}\right|}{|\{(x_i, y_i, z_i) \in d : z_i = z\}|}$. The threshold 0.8 adheres to the EEOC's Uniform Guidelines [7], which suggest a 80% criterion for curbing adverse impact in age, ethnicity, race and gender discrimination cases.

Informally, a classifier is said to suffer from DI if the outcomes disproportionately benefit a group having certain sensitive attributes such as belonging to a particular gender, race, ethnicity. This undesirable effect may occur if the classifier had been trained using a dataset D that has extremely uneven fraction of positive labels among two groups (say, $p_D^1 = 0.2$ and $p_D^0 = 0.7$). Such a classifier may inherit the unjust patterns from training data and provide discriminatory predictions on the test dataset d, which may have almost equal prevalences ($p_d^1 = 0.6$ and $p_d^0 = 0.7$) due to prior probability shift. One way is to use $\hat{\varrho}_D^{z,z'} \geq 0.8$ as a constraint, while minimizing error, which may help reduce such disproportionate decisions.

The major drawback of using DI is the fixed threshold 0.8, since the prior probability shifts may not always lead to almost equal prevalences for two groups. For example, if the true prevalences are $p_d^1 = 0.6$ and $p_d^0 = 0.8$ respectively, (that is, $\varrho_d^{1,0} = 0.75$) then enforcing DI with threshold 0.8 may have adverse effect on $Z = 0$ group. This motivates the need to use a correct threshold depending on the extent of prior probability shift in the test dataset.

3 PROPORTIONAL EQUALITY

To address the intrinsic drawbacks associated with the definition of disparate impact, we need a fairness metric robust enough to deal with prior probability shifts. To this effect, we provide a novel formalization of the fairness notion called *proportional equality* (PE). The predictions $\hat{h}(x_i)$ on test dataset d is said to be PE-fair if the true prevalence ratio $\varrho_d^{z,z'} := p_d^z / p_d^{z'}$ and the prediction prevalence ratio $\hat{\varrho}_d^{z,z'} := \hat{p}_d^z / \hat{p}_d^{z'}$ satisfies the following:

$$\left|\varrho_d^{z,z'} - \hat{\varrho}_d^{z,z'}\right| \leq \epsilon \text{ for a small number } \epsilon > 0. \tag{2}$$

3.1 Desirable Properties of Fairness Notions

We state below some properties (by no means exhaustive) which are important for a fairness notion to satisfy:

P1 : The fairness concept should be agnostic to distributional changes within groups, for example, prior probability shifts.

P2 : The fairness concept holds true for a *perfect classifier* (a classifier with 100% accuracy).

The property **P1** is satisfied by PE-fairness since, by definition, it is aware of the prior probability shift between the training and the test dataset. Theorem 3.1 establishes an important connection between the accuracy of a classifier and the PE-fairness notion, and enables the PE concept to tick off the desirable property **P2**.

Theorem 3.1. *A perfect classifier is always* PE*-fair.*

3.2 Experimental Evaluation

On real-world datasets, COMPAS [28], Adult Income [24] and German Credit [19], we evaluate performances of two misclassification-minimizing classifiers, namely (1) PEC (Proportional Equality ensuring Classifier) which is aware of the prior probability shift and aims to ensure PE-fairness, and (2) BASE (a baseline classifier) which is unaware of such shifts and only minimizes misclassification.

The knowledge of prior probability shift during prediction seems difficult, since the true labels of the test dataset cannot be used. However, there exist quantification techniques [2, 13, 14, 16, 17] to *estimate* the true prevalences of test datasets in the presence of prior probability shifts. We employ such techniques to learn quantifiers for each group using training data and then use those to estimate the prevalences of each group in the test data. We use these estimates to ensure PE-fairness on the predictions of test dataset. Table 1 summarizes the fairness guarantees of PEC and BASE.

Dataset	True ratio $\varrho_d^{0,1}$	Estimation by quantifier [2]	PEC $\hat{\varrho}_d^{0,1}$	BASE $\hat{\varrho}_d^{0,1}$
COMPAS	0.7584	0.7679	**0.7573**	0.4617
Adult Income	0.3647	0.3855	**0.3759**	0.2033
German Credit	0.7463	0.7096	**0.7195**	0.6588

Table 1: This table shows that the true and estimated prevalence ratios are very close. Additionally, we observe that PEC **outperforms** BASE **in terms of fairness.**

Moreover, in terms of accuracy, PEC is close to that of BASE, with absolute worst case difference of 0.041 over all the datasets. For analysis, all the values are averaged over 20 iterations, where each iteration uses a random 70 − 30 train-test split of the datasets.

ACKNOWLEDGMENTS

This work was done while Arpita Biswas was an intern at Microsoft Research. She gratefully acknowledges the support of a Google Fellowship towards her PhD.

REFERENCES

[1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. *ProPublica* (2016). www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[2] Antonio Bella, Cesar Ferri, José Hernández-Orallo, and Maria Jose Ramirez-Quintana. 2010. Quantification via probability estimators. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*. IEEE, 737–742.

[3] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2018. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research* (2018).

[4] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. 'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions. In *CHI Conference on Human Factors in Computing Systems*. ACM, 377:1—377:14.

[5] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. 2009. Building classifiers with independency constraints. In *Data mining workshops, 2009. ICDMW'09. IEEE international conference on*. IEEE, 13–18.

[6] Alexandra Chouldechova and Aaron Roth. 2018. The Frontiers of Fairness in Machine Learning. *arXiv preprint arXiv:1810.08810* (2018).

[7] The U.S. Equal Employment Opportunity Commission. 1979. Uniform guidelines on employee selection procedures. *March 2* (1979).

[8] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 797–806.

[9] Andrew Cotter, Maya Gupta, Heinrich Jiang, Nathan Srebro, Karthik Sridharan, Serena Wang, Blake Woodworth, and Seungil You. 2018. Training Well-Generalizing Classifiers for Fairness Metrics and Other Data-Dependent Constraints. *arXiv preprint arXiv:1807.00028* (2018).

[10] Roger Crisp. 2014. *Aristotle: Nicomachean Ethics*. Cambridge University Press.

[11] Julia Dressel and Hany Farid. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science advances* 4, 1 (2018), eaao5580.

[12] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 259–268.

[13] George Forman. 2005. Counting positives accurately despite inaccurate classification. In *European Conference on Machine Learning*. Springer, 564–575.

[14] George Forman. 2006. Quantifying trends accurately despite classifier error and class imbalance. In *ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 157–166.

[15] Sorelle A Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P Hamilton, and Derek Roth. 2018. A comparative study of fairness-enhancing interventions in machine learning. *arXiv preprint arXiv:1802.04422* (2018).

[16] Pablo González, Jorge Díez, Nitesh Chawla, and Juan José del Coz. 2017. Why is quantification an interesting learning problem? *Progress in Artificial Intelligence* 6, 1 (2017), 53–58.

[17] Víctor González-Castro, Rocío Alaiz-Rodríguez, and Enrique Alegre. 2013. Class distribution estimation based on the Hellinger distance. *Information Sciences* 218 (2013), 146–164.

[18] Moritz Hardt, Eric Price, Nati Srebro, et al. 2016. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*. 3315–3323.

[19] Hans Hofmann. 1994. Statlog (German Credit Data) Data Set . https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data). [accessed 20-February-2019].

[20] Nan D Hunter. 2000. Proportional Equality: Readings of Romer. *Ky. LJ* 89 (2000), 885.

[21] Faisal Kamiran and Toon Calders. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems* 33, 1 (2012), 1–33.

[22] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2012. Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 35–50.

[23] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2017. Inherent trade-offs in the fair determination of risk scores. *Innovations in Theoretical Computer Science* (2017).

[24] Ronny Kohavi and Barry Becker. 1996. Adult Data Set . https://archive.ics.uci.edu/ml/datasets/adult. [accessed 20-February-2019].

[25] Meelis Kull and Peter Flach. 2014. Patterns of dataset shift. In *First International Workshop on Learning over Multiple Contexts (LMCE) at ECML-PKDD*.

[26] Jose G Moreno-Torres, Troy Raeder, RocíO Alaiz-RodríGuez, Nitesh V Chawla, and Francisco Herrera. 2012. A unifying view on dataset shift in classification. *Pattern Recognition* 45, 1 (2012), 521–530.

[27] Emma Pierson, Sam Corbett-Davies, and Sharad Goel. 2018. Fast Threshold Tests for Detecting Discrimination. In *International Conference on Artificial Intelligence and Statistics, AISTATS 2018*. 96–105.

[28] ProPublica. 2016. COMPAS Recidivism Risk Score Data & Analysis. github.com/propublica/compas-analysis.

[29] Camelia Simoiu, Sam Corbett-Davies, Sharad Goel, et al. 2017. The problem of infra-marginality in outcome tests for discrimination. *The Annals of Applied Statistics* 11, 3 (2017), 1193–1216.

[30] Blake Woodworth, Suriya Gunasekar, Mesrob I Ohannessian, and Nathan Srebro. 2017. Learning non-discriminatory predictors. *Preprint arXiv:1702.06081* (2017).

[31] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. 2017. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*. 962–970.

[32] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning fair representations. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*. 325–333.