A Generic Framework for Bias Evaluation and Mitigation for Fair AI

Jeanne Monnier[1,2,*], Thomas George[1], Christèle Tarnec[1], Frédéric Guyard[1] and Marios Kountouris[2]

[1] *Orange Innovation, Sophia Antipolis, France*
[2] *EURECOM, Sophia Antipolis, France*

Abstract

A plethora of fairness metrics have been proposed in the literature, some of them with the associated bias mitigation framework. Whether a metric is more adapted or relevant than another depends on the whole context of model construction and utilization. We propose a generic and comprehensive framework for bias evaluation and mitigation, which aims to cover all contexts of use. We define a *benefit function* making possible the integration of virtually any, existing or new, fairness notion. We then use*mutual information*, a concept from information theory, to estimate bias in different cases, including group fairness on several subgroups or multi-outputs models. Finally, we generalize an *in-processing* method of bias mitigation using a *regularization* term as a means to constrain the optimization problem. Our experiments highlight the utility, the efficiency, and the potential of our method.

Keywords

bias, fairness metrics, group fairness, mutual information, bias evaluation and mitigation, regularization

1. Introduction

Fairness and bias in Artificial Intelligence (AI) have recently attracted vivid interest in the research community, with more than 70 different types of biases identified in the whole data lifecycle [1]. AI models reproduce and magnify biases already present in data during model construction via for instance non-representative sampling methods or flawed learning objectives [2]. As such, bias mitigation is becoming imperative considering the expansion of AI in terms of both users and scopes. Nevertheless, most fairness metrics are based on a unique pattern, which only suits one kind of model training among all real-case scenarios. In this paper, we argue that a generic and comprehensive framework is needed to simplify and expand the use of fairness tools. For that, we generalize the *prejudice remover* [3], a regularization term initially designed for statistical parity, by introducing a generic framework of bias evaluation and mitigation leveraging tools from information theory.

Related Work. Fairness in Machine Learning (ML) aims at avoiding any type of discrimination in a model's prediction. However, this definition remains vast or unclear, and various interpretations have led to several different notions [4]. We focus on *group fairness*, which considers fair a model treating equally groups, called demographic groups, which differ by their *sensitive attributes* values. A *sensitive attribute* is a feature in the dataset that describes a

Aequitas 2024: Second Workshop on Fairness and Bias in AI | co-located with ECAI 2024, Santiago de Compostela, Spain

[*] Corresponding author.

jeanne.monnier@orange.com (J. Monnier)

characteristic that could generate discrimination. Sex, race, religion, and sexual orientation are typical examples of such attributes. Among these groups, some are characterized as *privileged*, e.g., *male*, and others as *unprivileged*, e.g., *female*, depending on whether they are likely to be discriminated against.

There are two major technical challenges in the study of bias in ML, namely bias evaluation and mitigation. Measuring bias requires the use of metrics. A *fairness metric* is a quantification of undesired bias in training data or models. Based on the notion of group fairness, a wide range of metrics have been proposed, categorized in three fundamental conditions of statistical independence: independence, sufficiency, and separation [5, 6], all corresponding to different versions of what is being considered *fair* or *ethical* towards all demographic groups. Some auditing tools have been provided, including the well-known AI Fairness 360 from IBM [7]. Once biases are estimated, it is desirable to reduce or mitigate them. Bias mitigation techniques can be used over all stages of model construction and ML lifecycle, i.e., *preprocessing*, *in-processing*, and *post-processing*. Preprocessing consists of modifying training data to learn debiased information. Despite being a relatively simple and flexible method to implement, particularly because it is model-agnostic, it does not prevent the formation of biases later on in the learning process. In-processing methods try to tackle biases that can emerge from the learning process, including both those originating from the data and those due to training mechanisms. Post-processing methods take the outputs of a model and apply some transformations to debias them. In-processing offers the advantage of taking all biases into account and avoiding the creation of a model that could be misused. In this paper, we focus on the latter type of method, and in particular, on prejudice remover [3], a fair regularization of the optimization objective.

Rationale and Problematic. A major roadblock in the field of fair AI is the dispersal across a plethora of fairness metrics/definitions and the associated notions of equity. What's more, most of these metrics are incompatible with each other [8], and methods proposed for bias mitigation are usually designed to mitigate a single specific metric [9][10][11]. Therefore, it is important to come up with a generic framework that supports and adapts to any desired fairness metric, whether an existing one or a new one specifically tailored to a use case. One of the criticisms of fairness is the so-called "fairness gerrymandering" [12], which claims that it is not suited to intersectionality. Society is made up of a multiplicity of demographic subgroups defined by the combination of diverse sensitive attributes and their values. For instance, *sex* and *race* can be both present in data, with *race* being a categorical feature with five different values, which would lead to ten demographic subgroups. Another limitation of these metrics is that they are constructed for models with binary outputs, forcing us to reduce all prediction cases to a duality, which can be impossible or lead to an information loss. For example, the prediction of customer satisfaction can be reduced to positive/negative, but nuances between 'satisfied' and 'very satisfied' are lost, as well as neutral opinions. Not to mention continuous targets, which do not fit at all into this framework. Popular fairness metrics in the literature include [7]: disparate impact (DI), statistical parity difference (SPD), equalized odds difference, equal opportunity difference (EOD), and average odds difference (AOD). They are all based on the same pattern, which only handles two demographic groups and a binary prediction.

The rationale of this paper lies in the crucial need for a more generic framework to overcome what makes fairness so difficult to adopt in real-life use cases at the moment and eventually make models aligned with society's and human needs.

Main Contributions. A key contribution of this work is the development of a generic and comprehensive framework for bias evaluation and mitigation, which tackles the aforementioned limitations of fair AI/ML techniques. In particular, the proposed group fairness metric can: (i) adapt to different notions of fairness, (ii) manage intersectionality, and (iii) support different types of model outputs, going beyond the limiting case of binary outputs.

2. Generic Fairness Metric and Bias Mitigation

We introduce here a generic framework for bias mitigation, which can be seen as a generalization of the prejudice remover [3]. Our framework consists of three steps: (i) introducing a *'benefit' function*, which in turn is designed depending on the nature of the prediction and on what is advantageous in the context of the model use, (ii) using mutual information to estimate the mutual dependence between the sensitive attributes and the 'benefit' of an instance, (iii) putting a regularization term in the loss function to mitigate bias during training. Steps (i) and (ii) taken together constitute the fairness metric that can be customized, while step (iii) is an in-processing bias mitigation method leveraging the proposed metric.

Let D be a dataset of p instances following a distribution $\mathcal{D}$. Each instance, represented by a tuple of $n+1$ values $(\mathbf{X}, \mathbf{A}, Y) = (x_1, \ldots, x_m, a_1, \ldots, a_{n-m}, y)$, is related to an individual. Each value x_i, a_i, and y belongs to a continuous or discrete space $\mathcal{X}_i$, $\mathcal{A}_i$, and $\mathcal{Y}$ respectively, and Y is the label that the model has to predict. The variables in $\mathcal{X} = \prod_{i=1}^{m} \mathcal{X}_i$ are called *non-sensitive attributes*, whereas the variables in $\mathcal{A} = \prod_{i=m+1}^{n} \mathcal{A}_i$ are called *sensitive attributes*, and are those for which no discrimination should be made.

Let M be a model and $\mathbf{w}$ its set of trainable parameters. Learning the task is written as an empirical risk minimization (ERM) problem as follows:

$$\min_{\mathbf{w} \in \mathbb{R}^h} F(D, \mathbf{w}) = \min_{\mathbf{w} \in \mathbb{R}^h} \sum_{d \in D} \ell \left(\mathbf{X}_d, \mathbf{A}_d, Y_d; \mathbf{w} \right) \tag{1}$$

where $\ell \left(\mathbf{X}_d, \mathbf{A}_d, Y_d; \mathbf{w} \right)$ is the loss of the prediction on instance $(\mathbf{X}_d, \mathbf{A}_d, Y_d)$ by the model of parameters $\mathbf{w}$. To implement in-processing fairness methods usually the ERM is modified by reweighting samples [13][14][15], but most of the time a regularization factor is employed. Such is the case with prejudice remover [3] that we propose to generalize. Specifically, the problem we consider here is that of a regularized optimization problem

$$\min_{\mathbf{w} \in \mathbb{R}^h} F(D, \mathbf{w}) + \eta R(D, \mathbf{w}) \tag{2}$$

where $R(D, \mathbf{w})$ is a regularization term and $\eta \in \mathbb{R}$ is the trade-off hyperparameter between ERM and regularizer.

2.1. Generalizing fairness metrics with 'benefit'

Observing the various metrics of group fairness, they all attempt to determine what should not depend on the sensitive attributes. They then use this as a basis to compare how different instances, from different demographic groups, are treated by the model. Consider, for example, the two worldviews in [16]: '*what you see is what you get*' supports that accurate prediction of training data should be independent of the sensitive attribute, while '*we are all equal*' supports that it is the rate of positive predictions, independent of the label, which should be independent of the sensitive attribute.

In order to generalize the fairness evaluation, we propose to introduce a function, coined *benefit function*:

$$f_b : (\hat{y}, y) \rightarrow b$$

where $\hat{y}$ denotes the prediction by the model, y is the initial label in the dataset, and $b \in \mathbb{R}$ (could be extended to $\mathbf{b} \in \mathbb{R}^p$). For a given data instance $(\mathbf{X}_d, \mathbf{A}_d) = (x_1^d, \ldots, x_m^d, a_1^d, ..., a_{n-m}^d)$, depending on what is considered advantageous or disadvantageous treatment, f_b quantifies how beneficial a prediction $\hat{Y}_d = M_{\mathbf{w}}(\mathbf{X}_d, \mathbf{A}_d)$ is. We obtain a new random variable $B = f_b(\hat{Y}, Y) = f_b(\mathbf{X}, \mathbf{A}, Y, \mathbf{w}) \in \mathcal{B} \subset \mathbb{R}$, where $\mathcal{B}$ can be a discrete or continuous space.

Particular cases: A key advantage of the *benefit function* is that most existing fairness metrics are particular cases. Once the corresponding 'benefit' is configured, its mutual dependence with the sensitive attributes $\mathbf{A}$ gives the final fairness metric. Below, we provide some examples of the *benefit function*, which recovers or is equivalent to widely used notions of fairness.

Table 1
Fairness notions, metrics for binary outputs, and corresponding definitions of the *benefit function*

Fairness Notion	Metric for binary outputs	f_b for equivalence
Statistical Parity (Independence)	$P(\hat{Y} \mid A = 0) - P(\hat{Y} \mid A = 1)$	$f_b : (\hat{Y}, Y) \rightarrow \hat{Y}$
Equal Opportunity (Separation)	$P(\hat{Y} = 1 \mid Y = 1, A = 0) - P(\hat{Y} = 1 \mid Y = 1, A = 1)$	$f_b : (\hat{Y}, Y) \rightarrow \hat{Y}Y$
Overall Accuracy Equality (Sufficiency)	$P(\hat{Y} = Y \mid A = 0) - P(\hat{Y} = Y \mid A = 1)$	$f_b : (\hat{Y}, Y) \rightarrow 1 - \lvert\hat{Y} - Y\rvert$

A group fairness metric consists of measuring how this 'benefit' is dependent on the sensitive attributes. Now that we have defined the 'benefit' of a prediction, we need to set up mathematical tools to assess the mutual dependence between the 'benefit' random variable B and the sensitive attributes $\mathbf{A}$.

2.2. Bias Evaluation using Mutual Information

Most existing, widely adopted fairness metrics measure the ratio or difference in some probability between two demographic groups. However, in practice, datasets contain more than one sensitive attribute, each taking on several values. Thus, a framework that can assess biases over several sub-groups at the same time is needed. To this end, rather than measuring a statistic on each sub-group and comparing them with each other, we propose to measure the mutual dependence between the 'benefit' variable $B = f_b(\hat{Y}, Y)$ and the random vector of sensitive attributes $\mathbf{A}$. For that, we use tools from information theory, namely mutual information. Let B and $\mathbf{A}$ be a discrete random variable and a vector, respectively, our final metric is defined as:

$$I(\mathbf{A}; B) = \sum_{\mathbf{a} \in \mathcal{A}} \sum_{b \in \mathcal{B}} P_{(\mathbf{A},B)}(\mathbf{a}, b) \log \left(\frac{P_{(\mathbf{A},B)}(\mathbf{a}, b)}{P_{\mathbf{A}}(\mathbf{a}) P_B(b)} \right) \tag{3}$$

where $P_{\mathbf{A},B}$ is the joint probability mass function of $\mathbf{A}$ and B, and $P_{\mathbf{A}}$ and P_B are the marginal probability mass functions of $\mathbf{A}$ and B respectively. If the real distributions are not known in advance, they can be approximated by the sample distributions. For the case of continuous variables, we have $I(\mathbf{A}; B) = \int_{\mathcal{A}} \int_{\mathcal{B}} P_{(\mathbf{A},B)}(\mathbf{a}, b) \log \left(\frac{P_{(\mathbf{A},B)}(\mathbf{a},b)}{P_{\mathbf{A}}(\mathbf{a}) P_B(b)} \right) \mathrm{d}\mathbf{a}\mathrm{d}b$. Measuring $I(\mathbf{A}; B)$ determines how much information about sensitive attributes is contained in B. In other words, how much benefit an instance derives from its prediction by the model that depends on the values of the sensitive attributes of that same instance, hence $I(\mathbf{A}; B)$ eventually enables the assessment of biases in the model outputs. Mutual information is non-negative and the lower its value, the lesser the dependence between the two variables. Note that $I(\mathbf{A}; B) = 0$ iff $\mathbf{A}$ and B are independent.

2.3. Mitigate bias during training

Once biases are identified and quantified using fairness metrics, they should be mitigated. We choose an in-processing mitigation technique because it takes the whole process into account, from biases emerging from data information to those introduced during training. We control biases in the model's outputs and constrain the training to mitigate them, ensuring a final model that is the least possible biased.
[3] proposes the prejudice remover, a regularizer based on the mutual information between outputs $\hat{Y}$ and sensitive attributes $\mathbf{A}$ of training data, hence minimizing statistical parity difference. Capitalizing on the same idea, we propose to add a well-designed regularizer to the loss function. An unconstrained ERM only focuses on accuracy, i.e., reproducing information contained in the training data, which also means reproducing all biases present in data. Here, regularizing the objective function aims at increasing accuracy while additionally paying attention to the distribution of the 'benefit' between different demographic subgroups.
Our approach differs from the prejudice remover [3] because we look at the mutual dependence between the 'benefit' $B = f_b(\hat{Y}, Y)$ and sensitive attributes $\mathbf{A}$. The flexibility in the *benefit function* definition enables the minimization of different types of biases.

As seen in Section 2.2, minimizing the mutual information of the variables B and $\mathbf{A}$ reduces their mutual dependence. The proposed regularizer derived from (3) takes on the form of

$$R_{MI}(\mathbf{A}; B) = \sum_{\mathbf{a} \in \mathcal{A}} \sum_{b \in \mathcal{B}} P_{(\mathbf{A},B)}(\mathbf{a}, b) \log \left(\frac{P_{(\mathbf{A},B)}(\mathbf{a}, b)}{P_{\mathbf{A}}(\mathbf{a}) P_B(b)} \right) \tag{4}$$

with $B = f_b(\hat{Y}, Y) = f_b(M_{\mathbf{w}}(\mathbf{X}, \mathbf{A}), Y)$ and the optimization problem (2) becomes

$$\min_{\mathbf{w} \in \mathbb{R}^h} F(D, \mathbf{w}) + \eta R_{MI}(D, \mathbf{w}). \tag{5}$$

3. Experiments

In this section, we provide experiments to illustrate the limitations of existing approaches and the potential of our proposal. Experiments are performed on UCI Adult dataset, containing information about whether the income of people exceeds $50K$/yr. All features except *Sex, Race, Age, Education* and *Income* are dropped and we take *Sex* as the sensitive attribute. The positive class is '*Income > 50K*' and the privileged sensitive attribute is *male* because the initial proportion of *male* data belonging to the positive class is higher than for *female* data. We train a simple logistic regression model. We implement the proposed regularizer (4) and vary the hyperparameter η in $\left[10^{-1}, 10^{2.5}\right]$. We experiment with two different *benefit functions*: experiment 1 uses the SPD fairness metrics recovering the prejudice remover [3], and experiment 2 uses the OAE fairness metrics resulting in different values of classification and fairness metrics (fig 1).

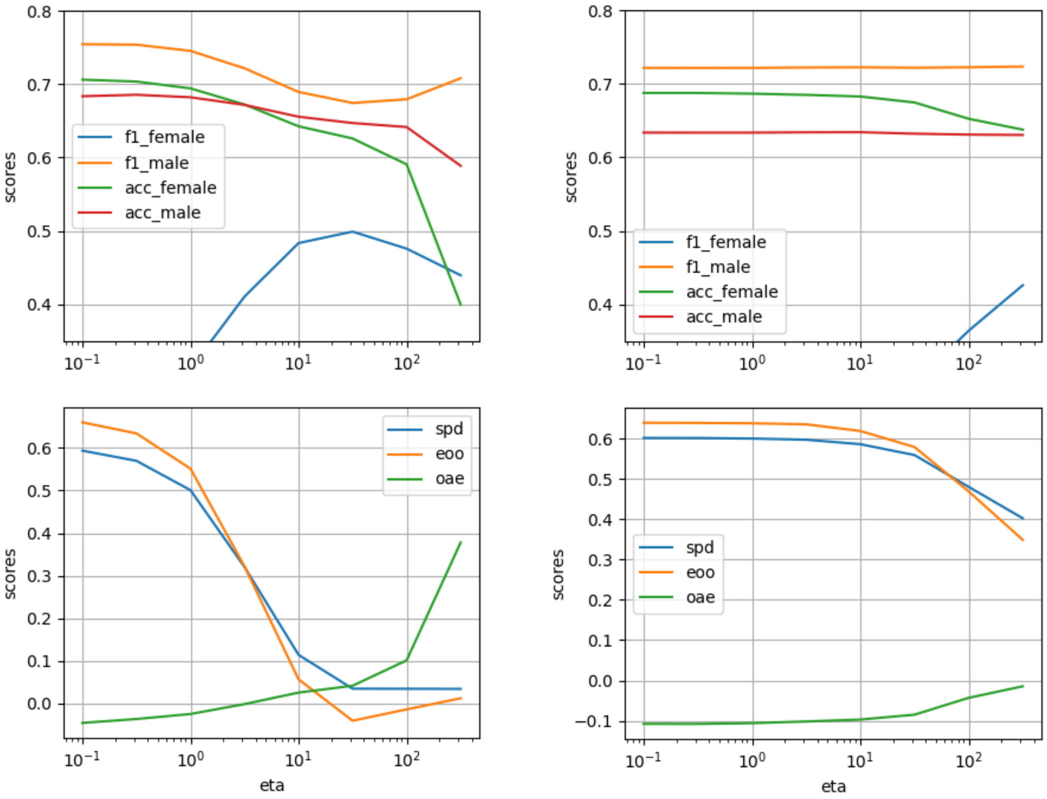

Figure 1: Performance metrics (*first row*) and fairness metrics (SPD, EOO and OAE) defined in Table 1 (*second row*) of models trained with specified fairness regularizers on Adult: R_{MI} with SPD (equivalent to prejudice remover) for experiment 1 (*left*) and R_{MI} with OAE for experiment 2 (*right*).

Results. Experiment 1 shows that the SPD metric tends towards zero as the strength of the regularizer (η) increases (as expected), but simultaneously, the OAE fairness metric gets worse (increases beyond zero). This guides us that having one prejudice remover is neither universal nor neutral, and enforces a single fairness notion. In cases where another metric is needed, the regularizer is useless, even counterproductive, which motivates our proposed generic bias mitigating strategy (section 2). In the second experiment, we implement our proposed regularizer R_{MI} by replacing the SPD with OAE while keeping all other experiment parameters fixed. This time, increasing the regularizer strength drives the OAE towards zero (as opposed to experiment 1). SPD slowly decreases but it is not a guaranteed knock-on effect of OAE minimization. This exemplifies the effectiveness of our generic bias mitigation strategy applied to other fairness metrics. Further experiments should explore the flexibility with new metrics, multiple subgroups, and multi-output models.

References

[1] A. Olteanu, C. Castillo, F. Diaz, E. Kıcıman, Social data: Biases, methodological pitfalls, and ethical boundaries, Frontiers in big data 2 (2019) 13.

[2] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A survey on bias and fairness in machine learning, ACM computing surveys (CSUR) 54 (2021) 1–35.

[3] T. Kamishima, S. Akaho, H. Asoh, J. Sakuma, Fairness-aware classifier with prejudice remover regularizer, in: Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2012, Bristol, UK, September 24-28, 2012. Proceedings, Part II 23, Springer, 2012, pp. 35–50.

[4] K. Makhlouf, S. Zhioua, C. Palamidessi, Machine learning fairness notions: Bridging the gap with real-world applications, Information Processing & Management 58 (2021) 102642.

[5] B. Ruf, M. Detyniecki, Towards the right kind of fairness in ai, arXiv preprint arXiv:2102.08453 (2021).

[6] S. Barocas, M. Hardt, A. Narayanan, Fairness and machine learning: Limitations and opportunities, MIT Press, 2023.

[7] R. K. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilović, et al., Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias, IBM Journal of Research and Development 63 (2019) 4–1.

[8] S. Mitchell, E. Potash, S. Barocas, A. D'Amour, K. Lum, Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions, arXiv preprint arXiv:1811.07867 (2018).

[9] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, R. Zemel, Fairness through awareness, in: Proceedings of the 3rd innovations in theoretical computer science conference, 2012, pp. 214–226.

[10] M. B. Zafar, I. Valera, M. Gomez Rodriguez, K. P. Gummadi, Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment, in: Proceedings of the 26th international conference on world wide web, 2017, pp. 1171–1180.

[11] M. Hardt, E. Price, N. Srebro, Equality of opportunity in supervised learning, Advances in neural information processing systems 29 (2016).

[12] M. Kearns, S. Neel, A. Roth, Z. S. Wu, Preventing fairness gerrymandering: Auditing and learning for subgroup fairness, in: International conference on machine learning, PMLR, 2018, pp. 2564–2572.

[13] V. Iosifidis, E. Ntoutsi, Adafair: Cumulative fairness adaptive boosting, in: Proceedings of the 28th ACM international conference on information and knowledge management, 2019, pp. 781–790.

[14] E. Krasanakis, E. Spyromitros-Xioufis, S. Papadopoulos, Y. Kompatsiaris, Adaptive sensitive reweighting to mitigate bias in fairness-aware classification, in: Proceedings of the 2018 world wide web conference, 2018, pp. 853–862.

[15] H. Jiang, O. Nachum, Identifying and correcting label bias in machine learning, in: International conference on artificial intelligence and statistics, PMLR, 2020, pp. 702–712.

[16] S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, The (im) possibility of fairness: Different value systems require different mechanisms for fair decision making, Communications of the ACM 64 (2021) 136–143.