Increasing Fairness via Combination with Learning Guarantees

Yijun Bian [1] Kun Zhang [2]

[1]University of Copenhagen [2]Hefei University of Technology

Funded by the European Union

Overview

We study the discrimination level of machine learning (ML) models, proposing
- a fairness measure (*discrimination risk*, ***DR***) from both individual- and group-fairness aspects
- the existence of a cancellation-of-discrimination effect in ensemble combination

Problem Statement and Motivation

Given a dataset S composed of instances including sensitive attributes (SAs)

$$S = \{(\underbrace{\breve{\boldsymbol{x}}_i}_{\text{non-sensitive / unprotected}}, \underbrace{\boldsymbol{a}_i}_{\text{sensitive / protected attributes}}, y_i)\}_{i=1}^n ,$$

- one instance $\boldsymbol{x}=(\breve{\boldsymbol{x}}, \boldsymbol{a})$, where the number of SAs $n_a \geqslant 1$, $n_a \in \mathbb{Z}_+$
- $\tilde{\boldsymbol{a}}$ is the slightly perturbed version of $\boldsymbol{a}$
- sensitive attributes $\boldsymbol{a}=[a_1, ..., a_{n_a}]^\mathsf{T}$ allows several attributes
  each $a_i \in \mathcal{Z}_+$ $(1 \leqslant i \leqslant n_a)$ is a finite set of values
- finite label space $y \in \mathcal{Y} = \{1, 2, ..., n_c\}$, where the number of labels $n_c \geqslant 2$

The *weighted voting* prediction by an ensemble

$$\mathbf{wv}_\rho(\boldsymbol{x}) = \operatorname{argmax}_{y \in \mathcal{Y}} \sum_{j=1}^m w_j \mathbb{I}(f_j(\boldsymbol{x}) = y)$$

- a function in the hypothesis space $f \in \mathcal{F} : \mathcal{X} \mapsto \mathcal{Y}$
- a set of m trained individual/discriminative classifiers $\{f_1(\cdot), ..., f_m(\cdot)\}$
- weight vector $\rho=[w_1, ..., w_m]^\mathsf{T} \in [0, 1]^m$, such that $\sum_{j=1}^m w_j = 1$

Proposed Fairness Measure: *DR*

Fairness quality from both individual and group fairness aspects

Following the principle of individual fairness,

the treatment/evaluation of one instance should not change solely due to minor changes in its sensitive attributes.

If it happens, this indicates the existence of underlying discriminative risks.

Naturally, the fairness quality of one hypothesis $f(\cdot)$ can be evaluated by

$$\ell_{\text{bias}}(f, \boldsymbol{x}) = \mathbb{I}(\overbrace{f(\breve{\boldsymbol{x}}, \boldsymbol{a}) \neq f(\breve{\boldsymbol{x}}, \tilde{\boldsymbol{a}})}^{f \text{ makes a discriminative decision}}) \tag{1a}$$

$$\hat{\mathcal{L}}_{\text{bias}}(f, S) = \frac{1}{n} \sum_{i=1}^n \ell_{\text{bias}}(f, \boldsymbol{x}_i) \tag{1b}$$

$$\mathcal{L}_{\text{bias}}(f) = \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}[\ell_{\text{bias}}(f, \boldsymbol{x})] \tag{1c}$$

- (1a) is evaluated on only one instance, from an individual aspect
- (1b), the empirical *DR* on S, describes this from a group aspect
- (1c), the true *DR* of the hypothesis, same as above in (1b)
- (1b) is an *unbiassed estimation* of (1c), & no restrictions apply on $f(\cdot)$ type

Distinctions of *DR* from the existing fairness measures

- *Two distinctions from individual fairness measures*
  - the latter relies on the choice of similarity/distance metric
  - instance pairs in comparison come from the original dataset
- *Two distinctions from group fairness measures*
  - the latter works for only one SA (usually with binary values)
  - needs to compute separately for each subgroup and then get the discrepancy
- *Five distinctions from causal fairness* (counterfactual fairness, proxy discrimination)
  - the latter works for only one SA (although possibly including multiple values)
  - based on causal models/graphs, and not a quantitative measure
  - non-sensitive attributes may be changed as well in counterfactual fairness
  - conditions for achieving them are stronger
  - *DR* can be proved to be bounded

vs. no such advantage

Similarities that *DR* shares with the existing fairness measures

- *Similarity with individual fairness measures*
  - the former follows the same principle (i.e., individual fairness)
  - sole change(s) in SAs indicates $(\breve{\boldsymbol{x}}, \boldsymbol{a})$ and $(\breve{\boldsymbol{x}}, \tilde{\boldsymbol{a}})$ are similar enough
- *Similarity with group fairness measures*
  - is calculated over a group of instances (like one dataset or a data distribution)
  - indicates the discrimination level from a statistical/demographic perspective —consistent with the idea of the latter
  - can be computed in the same way except that you don't have to

$$\mathcal{L}'_{\text{bias}}(f) = |\mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}|\boldsymbol{a}=1}[\ell_{\text{bias}}(f, \boldsymbol{x})] - \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}|\boldsymbol{a}=0}[\ell_{\text{bias}}(f, \boldsymbol{x})]| \tag{2}$$

Bounds Regarding Fairness for Weighted Vote

- tandem fairness quality of two hypotheses $f(\cdot)$ and $f'(\cdot)$

$$\ell_{\text{bias}}(f, f', \boldsymbol{x}) = \mathbb{I}(\underbrace{(f(\breve{\boldsymbol{x}}, \boldsymbol{a}) \neq f(\breve{\boldsymbol{x}}, \tilde{\boldsymbol{a}}))}_{\text{discriminative decision in } f} \wedge \underbrace{(f'(\breve{\boldsymbol{x}}, \boldsymbol{a}) \neq f'(\breve{\boldsymbol{x}}, \tilde{\boldsymbol{a}}))}_{\text{discriminative decision in } f'}) \tag{3}$$

- ρ-weighted majority vote (as the ensemble combination)

$$\mathbf{wv}_\rho(\boldsymbol{x}) = \operatorname{argmax}_{y \in \mathcal{Y}} \mathbb{E}_{f\sim\rho}[\mathbb{I}(f(\boldsymbol{x}) = y)]$$

- its fairness quality $\ell_{\text{bias}}(\mathbf{wv}_\rho, \boldsymbol{x}) = \mathbb{I}(\mathbf{wv}_\rho(\breve{\boldsymbol{x}}, \boldsymbol{a}) \neq \mathbf{wv}_\rho(\breve{\boldsymbol{x}}, \tilde{\boldsymbol{a}}))$
- for brevity, $\mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}[\cdot]$ and $\mathbb{E}_{f\sim\rho}[\cdot]$ may be abbreviated as $\mathbb{E}_{\mathcal{D}}[\cdot]$ and $\mathbb{E}_\rho[\cdot]$

If the weighted vote makes a discriminative decision, then at least a ρ-weighted half of the classifiers have made a discriminative decision and, therefore,

$$\ell_{bias}(\mathbf{wv}_\rho, \boldsymbol{x}) \leqslant \mathbb{I}(\mathbb{E}_\rho[\mathbb{I}(f(\breve{\boldsymbol{x}}, \boldsymbol{a}) \neq f(\breve{\boldsymbol{x}}, \tilde{\boldsymbol{a}}))] \geqslant 0.5)$$

Oracle bounds regarding fairness for weighted vote

- *Theorem 1. First-order oracle bound*

$$\mathcal{L}_{\text{bias}}(\mathbf{wv}_\rho) \leqslant 2\mathbb{E}_\rho[\mathcal{L}_{\text{bias}}(f)] .$$

- *Lemma 1.* In multi-class classification,

$$\mathbb{E}_{\mathcal{D}}[\mathbb{E}_\rho[\ell_{\text{bias}}(f, \boldsymbol{x})]^2] = \mathbb{E}_{\rho^2}[\mathcal{L}_{\text{bias}}(f, f')] .$$

- *Theorem 2. Second-order oracle bound*
  In multi-class classification,

$$\mathcal{L}_{\text{bias}}(\mathbf{wv}_\rho) \leqslant 4\mathbb{E}_{\rho^2}[\mathcal{L}_{\text{bias}}(f, f')] .$$

- *Theorem 3. C-tandem oracle bound*
  If $\mathbb{E}_\rho[\mathcal{L}_{\text{bias}}(f)] < 1/2$, then

$$\mathcal{L}_{\text{bias}}(\mathbf{wv}_\rho) \leqslant \frac{\mathbb{E}_{\rho^2}[\mathcal{L}_{\text{bias}}(f, f')] - \mathbb{E}_\rho[\mathcal{L}_{\text{bias}}(f)]^2}{\mathbb{E}_{\rho^2}[\mathcal{L}_{\text{bias}}(f, f')] - \mathbb{E}_\rho[\mathcal{L}_{\text{bias}}(f)] + \frac{1}{4}} .$$

A *prominent difference* between our work and the work of Masegosa *et al.* is that they investigate the expected risk or accuracy rather than fairness quality. In other words, their bounds are based on the 0/1 loss $\ell_{\text{err}}(f, \boldsymbol{x}) = \mathbb{I}(f(\boldsymbol{x}) \neq y)$, while ours are built upon $\ell_{\text{bias}}(f, \boldsymbol{x})$ in (1a).

PAC bounds regarding fairness for the weighted vote

- *Theorem 4.* For any $\delta \in (0, 1)$, with probability at least $(1-\delta)$ over a random draw of S with a size of n, for a single hypothesis $f(\cdot)$,

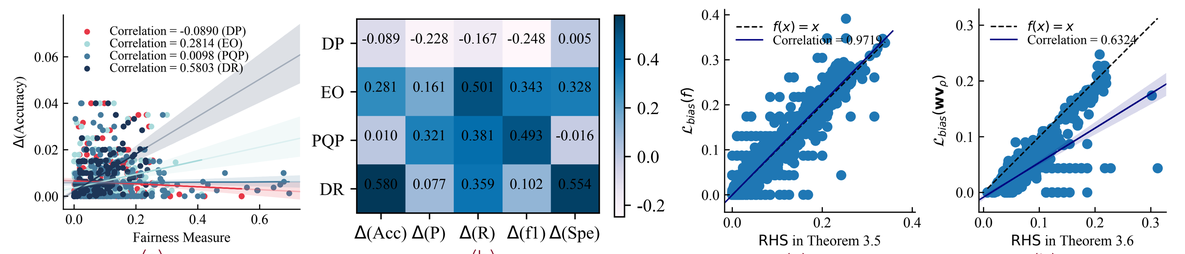
$$\mathcal{L}_{\text{bias}}(f) \leqslant \hat{\mathcal{L}}_{\text{bias}}(f, S) + \sqrt{\frac{1}{2n} \ln \frac{1}{\delta}} . \tag{4}$$

- *Theorem 5.* For any $\delta \in (0, 1)$, with probability at least $(1 - \delta)$ over a random draw of S with a size of n, for all distributions ρ on $\mathcal{F}$,

$$\mathcal{L}_{\text{bias}}(\mathbf{wv}_\rho) \leqslant \hat{\mathcal{L}}_{\text{bias}}(\mathbf{wv}_\rho) + \sqrt{\frac{1}{2n} \log \frac{|\mathcal{F}|}{\delta}} . \tag{5}$$

Generalisation bounds in (4) and (5) are derived from Hoeffding's inequality

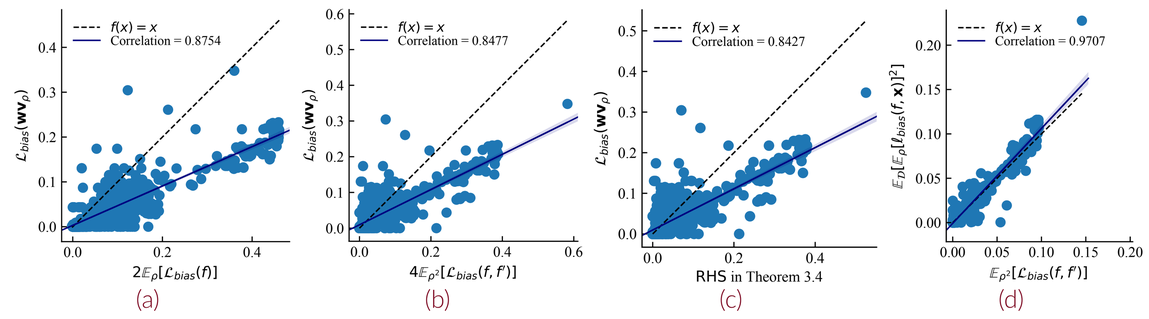
Empirical Results



(a) (b)

Figure 1. Comparison of the proposed *DR* with three group fairness measures.

Figure 2. Correlation for generalisation bounds.



(a) (b) (c) (d)

Figure 3. Correlation for oracle bounds.

Reference

[1] Andrés R Masegosa, Stephan Sloth Lorenzen, Christian Igel, and Yevgeny Seldin. Second order pac-bayesian bounds for the weighted majority vote. volume 33, pages 5263–5273. Curran Associates, Inc., 2020.

Acknowledgement

This research is funded by the European Union (MSCA, FairML, project no. 101106768). Views and opinions expressed are however those of the author(s) only.

SCAN ME

yibi@di.ku.dk