EasyChair Preprint

№ 6274

Resilience of Supervised Learning Algorithms to Discriminatory Poisoning of Training Data

Przemyslaw Grabowicz and Nicholas Perello

August 11, 2021

Resilience of Supervised Learning Algorithms to Discriminatory Poisoning of Training Data

Przemyslaw A. Grabowicz * Nicholas Perello*

Keywords: machine learning, discrimination, fairness, data poisoning, decision systems

Discrimination consists of treating somebody unfavorably because of their membership to a particular group, characterized by a *protected attribute*, such as race or gender. To prevent *disparate treatment*, the law often forbids the use of certain protected attributes in decision-making, e.g., in hiring, and dictates that decisions shall be based on relevant attributes. Historically, e.g., in the case of *redlining*, the prohibition of such direct discrimination was sometimes circumvented by the use of attributes correlated with the protected attribute as proxies. This is a particularly acute problem for machine learning data-rich systems, since they often find surprisingly accurate surrogates for protected attributes when a large set of legitimate-looking features is available, resulting in the *inducement* of discrimination via association. To prevent such inducements of discrimination, legal systems establish that the impact of a decision-making process should be the same across groups differing in protected attributes, unless relevant attributes justify it, according to a *business necessity clause* [1]. While multiple machine-learning fairness objectives have been developed to inhibit discrimination in models training on datasets potentially tainted by historical discrimination [2], they unfortunately give limited attention to the possibility of disparate impact justified via relevant attributes. To the best of our knowledge, this work is the first to define and inhibit the induction of discrimination via association in machine learning, while keeping the impact of relevant attributes.

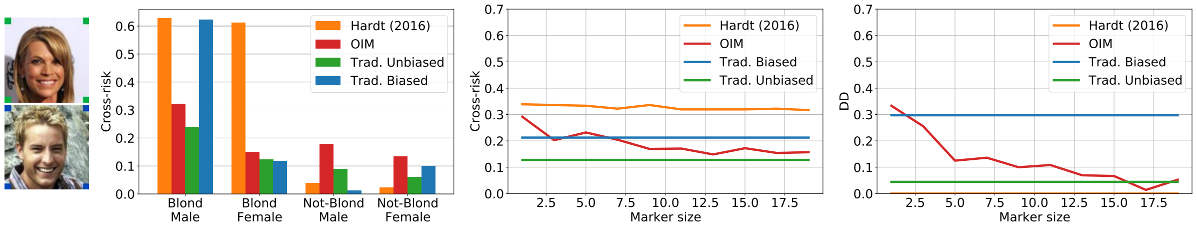
We develop a novel method for learning models of target decisions from data called the *optimal interventional mixture* (OIM) [3]. The OIM is a post-processing approach that eliminates the influence of the protected attribute by i) intervening probabilistically on the full model trained with all features, so that the protected attribute is set by the intervention to a random value, and ii) averaging the resulting post-interventional models via a mixing distribution that optimizes accuracy and is independent from the relevant features.

To evaluate this and other methods for inhibiting discrimination, we simulate a discriminatory poisoning (perturbations) of target decisions in training data, train fairness algorithms on this poisoned data, and evaluate them on the non-discriminatory test data. We refer to the mean error on the test data as *cross-risk*, i.e., the expected loss (risk) of this model w.r.t. the non-discriminatory data while training on the potentially discriminatory data (cross).

The distinction between non-discriminatory and discriminatory data allows for a causal definition of induced and direct discrimination as certain perturbations of the target decisions. We define a directly discriminatory perturbation as a transformation of non-discriminatory decisions into decisions that are directly causally influenced by the protected attribute. Conversely, an induced discriminatory perturbation is a transformation that modifies the direct causal impact of a relevant attribute associated with the protected attribute on decisions.

In addition to the introduced novel measure of cross-risk, we also compute well-known disparity measures such as *demographic disparity* and accuracy (mistreatment) disparity [2]. We test the OIM and several other methods addressing discrimination both on synthetic (not shown) and real-world datasets, such as Compas, German Credit, and CelebA [3]. Here, we show the result of the evaluation on the CelebA dataset, which is commonly used in computer

*U. of Massachusetts Amherst. Correspondence to: Przemyslaw A. Grabowicz, grabowicz@cs.umass.edu



(a) **(b)** Cross-risk per group **(c)** Cross-risk **(d)** Demographic disparity

Figure 1: (a) Image markings used for training the OIM. (b) The cross-risk by hair-gender group (10 pixel markers). (c) Overall cross-risk and (d) demographic disparity of learning algorithms as marking pixel size increases. "Trad." is traditional machine learning (ResNet-18). Lower values are better.

vision and deep learning literature. We use hair color classification as the target decisions and binary gender as the protected attribute. To avoid sampling bias, we balance the dataset based on the smallest group, i.e., blond males. To simulate a discriminatory perturbation, we swap the labels of 50% of blond males to not blond in the training data, resulting in a large error of traditional learning methods for blond males (blue in Figure 1). The methods train on this perturbed data, except for the traditional method trained on unbiased data (green in Figure 1).

As our base model architecture we use the popular ResNet-18 architecture. In addition to the OIM, we evaluate a post-processing method based on *equalized odds* [4]. The OIM requires the addition of the protected attribute to the feature set during training, therefore, we encode gender in the images via special markings placed in the corners of each image (Figure 1a). Other methods train without these markings.

Despite training on the perturbed data, the OIM reduces the cross-risk nearly to that of the traditional unbiased model trained without the data poisoning when using sufficiently large markers (blue, red, and green in Figures 1b & 1c). When the marking size decreases, the cross-risk of the OIM converges to that of the traditional biased model without the protected attribute encoded (Figures 1c). These results hold when we measure demographic disparity (Figure 1d) and other disparities (not shown). In real-world application domains where it is challenging to measure cross-risk due to potential unavailability of the non-discriminatory data, the size of markings for the OIM can be established based on the convergence of disparity measures.

Future works could measure discriminatory data poisonings via human subject experiments, enabling more realistic simulations of discriminatory data poisoning. For instance, *correspondence studies* in which resumes of fake applicants are sent to employers, such as the seminal work by Bertrand and Mullainathan, can be modified so that the first names of applicants are *not* correlated with race and the measurement of non-discriminatory decisions is possible.

In sum, we propose novel concepts and methods for studying discrimination, describe how state-of-the-art methods can induce discrimination via association, and develop a novel conservative learning method that deals with discriminatory poisoning of training data.

References

[1] Title VII of the Civil Rights Act, 1964. 7, 42 U.S.C., 2000e et seq.

[2] S. Barocas, M. Hardt, and A. Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019.

[3] P. Grabowicz, N. Perello, and K. Takatsu. Resilience of supervised learning algorithms to discriminatory data perturbations, 2021. arXiv:1912.08189.

[4] M. Hardt, E. Price, and N. Srebro. Equality of Opportunity in Supervised Learning. In *Adv. Neural Inf. Process. Syst. 29*, pages 3315–3323. Curran Assoc., Inc., 2016.