Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)

Achieving Non-Discrimination in Prediction

Lu Zhang, Yongkai Wu, and Xintao Wu
University of Arkansas
{lz006,yw009,xintaowu}@uark.edu

Abstract

In discrimination-aware classification, the pre-process methods for constructing a discrimination-free classifier first remove discrimination from the training data, and then learn the classifier from the cleaned data. However, they lack a theoretical guarantee for the potential discrimination when the classifier is deployed for prediction. In this paper, we fill this gap by mathematically bounding the discrimination in prediction. We adopt the causal model for modeling the data generation mechanism, and formally defining discrimination in population, in a dataset, and in prediction. We obtain two important theoretical results: (1) the discrimination in prediction can still exist even if the discrimination in the training data is completely removed; and (2) not all pre-process methods can ensure non-discrimination in prediction even though they can achieve non-discrimination in the modified training data. Based on the results, we develop a two-phase framework for constructing a discrimination-free classifier with a theoretical guarantee. The experiments demonstrate the theoretical results and show the effectiveness of our two-phase framework.

1 Introduction

Discrimination-aware classification is receiving an increasing attention in the data mining and machine learning fields. Many methods have been proposed for constructing discrimination-free classifiers, which can be broadly classified into three categories: the pre-process methods that modify the training data [Kamiran and Calders, 2009a; Feldman *et al.*, 2015; Zhang *et al.*, 2017; Calmon *et al.*, 2017; Nabi and Shpitser, 2017], the in-process methods that adjust the learning process of the classifier [Calders and Verwer, 2010; Kamishima *et al.*, 2011; 2012; Zafar *et al.*, 2017], and the post-process methods that directly change the predicted labels [Kamiran *et al.*, 2010; Hardt *et al.*, 2016]. All three categories of methods have their respective limitations. For the in-process methods, they usually perform some tweak or develop some regularizers for the classifier to correct or penalize discriminatory outcomes during the learning process. However, since the discrimination or fair constraints are generally not convex functions, surrogate functions are usually used for the minimization. As a result, additional bias might be introduced due to the approximation errors associated with the surrogate function. For the post-process methods, they are restricted to those who can modify the predicted label of each individual independently. Thus, methods that map the whole dataset or population to a new non-discriminatory one cannot be adopted for post-process, which means that a number of causal-based discrimination removal methods (e.g., [Zhang *et al.*, 2017; Nabi and Shpitser, 2017]) cannot be applied.

In our work, we target the pre-process methods that modify the training data, where some methods only modify the label, such as the *Massaging* [Kamiran and Calders, 2009a; Žliobaitė *et al.*, 2011] and the *Causal-Based Removal* [Zhang *et al.*, 2017], and some methods also modify the data attributes other than the label, such as the *Preferential Sampling* [Kamiran and Calders, 2012; Žliobaitė *et al.*, 2011], the *Reweighing* [Calders *et al.*, 2009], and the *Disparate Impact Removal* [Feldman *et al.*, 2015; Adler *et al.*, 2016]. The fundamental assumption of the pre-process methods is that, since the classifier is learned from a discrimination-free dataset, it is likely that the future prediction will also be discrimination-free [Kamiran and Calders, 2009b]. Although this assumption is plausible, however, there is no theoretical guarantee to show "how much likely" it is and "how discrimination-free" the prediction would be given a training data and a classifier. The lack of the theoretical guarantees places great uncertainty on the performance of all pre-process methods.

In this paper, we fill this gap by modeling the discrimination in prediction using the causal model [Pearl, 2009]. A causal model is a structural equation-based mathematical object that describes the causal mechanism of a system. We assume that there exists a fixed but unknown causal model that represents the underlying data generation mechanism of the population. Based on the causal model, we define the causal measure of discrimination in population as well as in prediction. We then formalize two problems, namely discovering and removing discrimination in prediction. Based on specific assumptions regarding the causal model and the causal measure of discrimination, we conduct concrete analysis of discrimination. We derive the formula for quantitatively measuring the discriminatory effect in population from

Notation	Definition
C	Protected attribute
$\mathbf{Z} = \{Z_1, \cdots, Z_m\}$	Non-protected attributes
L	Label of decision
$h : C \times \mathbf{Z} \to L$	Classifier
$\mathcal{M}$	Causal model of population
$\mathcal{M}_h$	Causal model of prediction
$\mathcal{D} = \{(c^{(j)}, \mathbf{z}^{(j)}, l^{(j)})\}$	Training data
$\mathcal{D}_h = \{(c^{(j)}, \mathbf{z}^{(j)}, h(c^{(j)}, \mathbf{z}^{(j)}))\}$	Training data w/ predicted labels

Table 1: Table of notations.

the observable probability distributions. We then derive the corresponding causal measure of the discrimination in prediction, as well as their approximations from the sample dataset. Finally, we link the discrimination in prediction with the discrimination in the training data by a probabilistic condition, which mathematically bounds the probability of the discrimination in prediction being within a given interval in terms of the training data and classifier.

As a consequence, we obtain two important theoretical results: (1) even if the discrimination in the training data is completely removed, the discrimination in prediction can still exist due to the bias in the classifier; and (2) for removing discrimination, different from the claims of many previous work, not all pre-process methods can ensure non-discrimination in prediction even though they can achieve non-discrimination in the modified training data. We show that to guarantee non-discrimination in prediction, the pre-process methods should only modify the label. Based on the results, we develop a two-phase framework for constructing a discrimination-free classifier with a theoretical guarantee, which provides a guideline for employing existing pre-process methods or designing new ones. The experiments demonstrate the theoretical results and show the effectiveness of our two-phase framework.

2 Problem Formulation

2.1 Notations and Preliminaries

We consider an attribute space which consists of some protected attributes, the label of certain decision attribute, and the non-protected attributes. Throughout the paper, we use an uppercase alphabet, e.g., X to represent an attribute; a bold uppercase alphabet, e.g., $\mathbf{X}$, to represent a subset of attributes. We use a lowercase alphabet, e.g., x, to represent a realization of attribute X; a bold lowercase alphabet, e.g., $\mathbf{x}$, to represent a realization of $\mathbf{X}$. For ease of representation, we assume that there is only one protected attribute, denoted by C, which is a binary attribute associated with the domain values of the non-protected group c^+ and the protected group c^-. We denote the label of the decision attribute by L, which is a binary attribute associated with the domain values of the positive label l^+ and negative label l^-. According to the convention in machine learning, we also define that $l^+ = 1$ and $l^- = 0$. The set of all the non-protected attributes is denoted by $\mathbf{Z} = \{Z_1, \cdots, Z_m\}$. Please refer to the notation table shown as Table 1.

A causal model is formally defined as follows.

Definition 1 (Causal Model)**.** *A causal model $\mathcal{M}$ is a triple $\mathcal{M} = \langle \mathbf{U}, \mathbf{V}, \mathbf{F} \rangle$ where*

1. $\mathbf{U}$ *is a set of hidden contextual variables that are determined by factors outside the model.*
2. $\mathbf{V}$ *is a set of observed variables that are determined by variables in* $\mathbf{U} \cup \mathbf{V}$.
3. $\mathbf{F}$ *is a set of equations mapping from* $\mathbf{U} \times \mathbf{V}$ *to* $\mathbf{V}$. *Specifically, for each* $V_i \in \mathbf{V}$, *there is an equation* f_i *mapping from* $\mathbf{U} \times (\mathbf{V} \backslash V_i)$ *to* V_i, *i.e.,*

$$v_i = f_i(pa_i, \mathbf{u}_i),$$

where pa_i *is a realization of a set of observed variables* $PA_i \subseteq \mathbf{V} \backslash V_i$ *called the parents of* V_i, *and* $\mathbf{u}_i$ *is a realization of a set of hidden variables* $\mathbf{U}_i \subseteq \mathbf{U}$.

The causal effect in the causal model is defined over the intervention that fixes the value of an observed variable(s) V to a constant(s) v while keeping the rest of the model unchanged. The intervention is mathematically formalized as $do(V = v)$ or simply $do(v)$. Then, for any variables $X, Y \in \mathbf{V}$, the distribution of Y after $do(x)$ is defined as [Pearl, 2009]

$$P(y|do(x)) \triangleq P(Y = y|do(X = x)) = \sum_{\{\mathbf{u}: Y_x(\mathbf{u}) = y\}} P(\mathbf{u}), \qquad (1)$$

where $Y_x(\mathbf{u})$ denotes the value of Y after intervention $do(x)$ under context $\mathbf{U} = \mathbf{u}$.

Note that $P(\mathbf{u})$ is an unknown joint distribution of all hidden variables. If the causal model satisfies the Markovian assumption: (1) the associated causal graph of the causal model is acyclic; and (2) all variables in $\mathbf{U}$ are mutually independent, $P(y|do(x))$ can be computed from the joint distribution of $\mathbf{V}$ according to the truncated factorization formula [Pearl, 2009]

$$P(y|do(x)) = \sum_{\mathbf{v}'} \prod_{V_i \neq X} P(v_i|pa_i)_{\delta_{X=x}}, \qquad (2)$$

where the summation is a marginalization that traverses all value combinations of $\mathbf{V}' = \mathbf{V} \backslash \{X, Y\}$, and $\delta_{X=x}$ means replacing X with x in each term.

2.2 Model Discrimination in Prediction

Assume that there exists a fixed population over the space $C \times \mathbf{Z} \times L$. The values of all the attributes in the population are determined by a causal model $\mathcal{M}$, which can be written as

$$\text{Causal Model } \mathcal{M} \qquad \begin{array}{l} c = f_C(pa_C, \mathbf{u}_C) \\ z_i = f_i(pa_i, \mathbf{u}_i) \quad i = 1, \cdots, m \\ l = f_L(pa_L, \mathbf{u}_L) \end{array}$$

where f_L can be considered as the decision making process in the real system. Without ambiguity, we can also use $\mathcal{M}$ to denote the population, and use the terms mechanism and population interchangeably. In practice, $\mathcal{M}$ is unknown and we can only observe a sample dataset $\mathcal{D} = \{(c^{(j)}, \mathbf{z}^{(j)}, l^{(j)}); j = 1, \cdots, n\}$ drawn from the population.

A classifier h is function mapping from $C \times \mathbf{Z}$ to L. A hypothesis space $\mathcal{H}$ is a set of candidate classifiers. A learning algorithm analyzes dataset $\mathcal{D}$ as the training data to find a classifier from $\mathcal{H}$ that minimizes the difference between the predicted labels $h(c^{(j)}, \mathbf{z}^{(j)})$ and the true labels $l^{(j)}$ $(j = 1, \cdots, n)$. Once training completes, the classifier is deployed to infer prediction on any new unlabeled data. It is

usually assumed that the unlabeled data is drawn from the same population as the training data, i.e., $\mathcal{M}$. Therefore, in prediction, the values of all the attributes other than the label are determined by the same mechanisms as those in $\mathcal{M}$, and meanwhile the classifier acts as a new mechanism for determining the value of the label. We consider $\mathcal{M}$ with function $f_L(\cdot)$ being replaced by classifier $h(\cdot)$ as the causal model of classifier h, denoted by $\mathcal{M}_h$, which is written as

$$\text{Causal Model } \mathcal{M}_h \qquad \begin{array}{l} c = f_C(pa_C, \mathbf{u}_C) \\ z_i = f_i(pa_i, \mathbf{u}_i) \quad i = 1, \cdots, m \\ l = h(c, \mathbf{z}) \end{array}$$

If we apply the classifier on $\mathcal{D}$, we can obtain a new dataset $\mathcal{D}_h$ by replacing the original labels with the predicted labels, i.e., $\mathcal{D}_h = \{(c^{(j)}, \mathbf{z}^{(j)}, h(c^{(j)}, \mathbf{z}^{(j)})); j = 1, \cdots, n\}$. Straightforwardly, $\mathcal{D}_h$ can be considered as a sample drawn from $\mathcal{M}_h$.

The discrimination in prediction made by classifier h can be naturally defined as the discrimination in $\mathcal{M}_h$. To this end, we first define a measure of discrimination in $\mathcal{M}$ based on the causal relationship specified by $\mathcal{M}$, denoted by $\text{DE}_{\mathcal{M}}$ called the *true discrimination*. By adopting the same measure, we denote the discrimination in $\mathcal{M}_h$ by $\text{DE}_{\mathcal{M}_h}$, called the *true discrimination in prediction*. Then, we denote the approximation of $\text{DE}_{\mathcal{M}}$ from dataset $\mathcal{D}$ by $\text{DE}_{\mathcal{D}}$, and similarly denote the approximation of $\text{DE}_{\mathcal{M}_h}$ from dataset $\mathcal{D}_h$ by $\text{DE}_{\mathcal{D}_h}$.

Our target is to discover and remove the true discrimination in prediction, i.e., $\text{DE}_{\mathcal{M}_h}$, based on certain causal measure of discrimination defined on $\mathcal{M}$, i.e., $\text{DE}_{\mathcal{M}}$. When calculating $\text{DE}_{\mathcal{M}_h}$ from dataset $\mathcal{D}$, we may encounter disturbances such as the sampling error of $\mathcal{D}$ and the misclassification of h. We then need to compute analytic approximation to $\text{DE}_{\mathcal{M}_h}$. Thus, we define the problem of discovering discrimination in prediction as follows.

Problem 1 (Discover Discrimination in Prediction)**.** *Given a causal measure of discrimination defined on $\mathcal{M}$, i.e., $\text{DE}_{\mathcal{M}}$, a sample dataset $\mathcal{D}$ and a classifier h trained on $\mathcal{D}$, compute analytic approximation to the true discrimination in prediction, i.e., $\text{DE}_{\mathcal{M}_h}$.*

If the true discrimination in prediction is detected according to the approximation, the next step is to remove the discrimination through tweaking the dataset and/or the classifier. Thus, we define the problem of removing discrimination in prediction as follows.

Problem 2 (Remove Discrimination in Prediction)**.** *Given $\text{DE}_{\mathcal{M}}$, $\mathcal{D}$ and h, tweak $\mathcal{D}$ and/or h in order to make $\text{DE}_{\mathcal{M}_h}$ be bounded by a user-defined threshold τ.*

3 Discover Discrimination in Prediction

In the above general problem definitions, $\text{DE}_{\mathcal{M}}$ can be any reasonable causal measure of discrimination defined on any causal model. However, a concrete analysis of discrimination must rely on specific assumptions regarding the causal measure of discrimination and the causal model. The remaining of the paper is based on following assumptions: (1) $\mathcal{M}$ satisfies the Markovian assumption; (2) we consider all causal effects (total effect) of C on L as discriminatory; (3) we assume that C has no parent and L has no child. The first assumption is necessary for computing the causal effect from the observable probability distributions. The second assumption is because the total causal effect is the causal relationship that is easiest to interpret and estimate. We will extend our results to other discrimination definitions such as those in [Zhang *et al.*, 2017; Bonchi *et al.*, 2017] in the future work. The last assumption is to make our theoretical results more concise and can be easily relaxed.

3.1 Causal Measure of Discrimination

We first derive the true discrimination in $\mathcal{M}$. The key of discrimination is a counterfactual question: whether the decision of an individual would be different had the individual been of a different protected/non-protected group (e.g., sex, race, age, religion, etc.)? To answer this question, we can perform an intervention on each individual to change the value of protected attribute C and see how label L will change. We consider the difference between the expectation of the labels when performing $do(c^+)$ for all individuals and the expectation of the labels when performing $do(c^-)$ for all individuals, and use it as the causal measure of discrimination.

To obtain the above difference, note that the causal model is completely specified at the individual level when context $\mathbf{U} = \mathbf{u}$ is specified. For each individual specified by $\mathbf{u}$, denote the label of individual $\mathbf{u}$ by $L_{c^+}(\mathbf{u})$ (resp. $L_{c^-}(\mathbf{u})$) when C is fixed according to $do(c^+)$ (resp. $do(c^-)$). Then, the difference in the label of individual $\mathbf{u}$ is given by $L_{c^+}(\mathbf{u}) - L_{c^-}(\mathbf{u})$. The expected difference of the labels over all individuals is hence given by $\mathbb{E}[L_{c^+}(\mathbf{u}) - L_{c^-}(\mathbf{u})]$. Based on this analysis, we obtain the following proposition.

Proposition 1. *Given a causal model $\mathcal{M}$, the true discrimination is given by*

$$\text{DE}_{\mathcal{M}} = P(l^+|c^+) - P(l^+|c^-).$$

Proof. The above expectations can be computed as

$$\begin{aligned} \mathbb{E}[L_{c^+}(\mathbf{u})] &= \sum_{\mathbf{u}} L_{c^+}(\mathbf{u}) P(\mathbf{u}) = \sum_{\{\mathbf{u}: L_{c^+}(\mathbf{u}) = l^+\}} l^+ P(\mathbf{u}) \\ &+ \sum_{\{\mathbf{u}: L_{c^+}(\mathbf{u}) = l^-\}} l^- P(\mathbf{u}) = \sum_{\{\mathbf{u}: L_{c^+}(\mathbf{u}) = l^+\}} P(\mathbf{u}) = P(l^+|do(c^+)), \end{aligned} \tag{3}$$

where the last step is according to Eq. (1). According to Eq. (2), we have

$$P(l^+|do(c^+)) = \sum_{\mathbf{z}} P(l^+|pa_L)_{\delta_{C=c^+}} \prod_{Z_i \in \mathbf{Z}} P(z_i|pa_i)_{\delta_{C=c^+}}.$$

On the other hand, since C has no parent, we have

$$P(l^+|c^+) = \frac{P(l^+, c^+)}{P(c^+)} = \frac{\sum_{\mathbf{z}} \cancel{P(c^+)} P(l^+|pa_L) \prod_{Z_i \in \mathbf{Z}} P(z_i|pa_i)}{\cancel{P(c^+)}},$$

Thus, we have $P(l^+|do(c^+)) = P(l^+|c^+)$, leading to $\mathbb{E}[L_{c^+}(\mathbf{u})] = P(l^+|c^+)$. Similarly we can prove $\mathbb{E}[L_{c^-}(\mathbf{u})] = P(l^+|c^-)$. Hence, the proposition is proven. □

Interestingly, the obtained discrimination causal measure is the same as the classic statistical discrimination metric *risk difference*, which is widely used as the non-discrimination

constraint in discrimination-aware learning [Romei and Ruggieri, 2014]. Our analysis can help understand the assumptions and scenarios in which the risk difference applies.

Given dataset $\mathcal{D}$, we approximate $\mathrm{DE}_{\mathcal{M}}$ using the maximum likelihood estimation, denoted by $\mathrm{DE}_{\mathcal{D}}$ as shown below.

Proposition 2. *Given a dataset $\mathcal{D}$, the discrimination in $\mathcal{D}$ is given by*

$$\mathrm{DE}(c^+, c^-)_{\mathcal{D}} = \hat{P}(l^+|c^+) - \hat{P}(l^+|c^-),$$

where

$$\hat{P}(l^+|c^+) = \sum_{\mathbf{z}} \hat{P}(l^+|c^+, \mathbf{z})\hat{P}(\mathbf{z}|c^+), \tag{4}$$

with $\hat{P}(\cdot)$ being the conditional frequency in $\mathcal{D}$.

Given a classifier $h : C \times \mathbf{Z} \to L$, denote the predicted labels by $\tilde{L}$. By adopting the same causal measure of discrimination of $\mathrm{DE}_{\mathcal{M}}$, we obtain $\mathrm{DE}_{\mathcal{M}_h}$ shown as follows.

Proposition 3. *Given a causal model $\mathcal{M}$ and a classifier h, the true discrimination in prediction is given by*

$$\mathrm{DE}_{\mathcal{M}_h} = P(\tilde{l}^+|c^+) - P(\tilde{l}^+|c^-),$$

where $P(\tilde{l}^+|c^+)$ (resp. $P(\tilde{l}^+|c^-)$) is the probability of the positive predicted labels for the data with $C = c^+$ (resp. $C = c^-$).

We similarly define $\mathrm{DE}_{\mathcal{D}_h}$ as the maximum likelihood estimation of $\mathrm{DE}_{\mathcal{M}_h}$.

Proposition 4. *Given a dataset $\mathcal{D}$ and a classifier h trained on $\mathcal{D}$, the discrimination in $\mathcal{D}_h$ is given by*

$$\mathrm{DE}_{\mathcal{D}_h} = \hat{P}(\tilde{l}^+|c^+) - \hat{P}(\tilde{l}^+|c^-),$$

where

$$\hat{P}(\tilde{l}^+|c^+) = \sum_{\mathbf{z}} \mathbb{I}_{[h(c^+,\mathbf{z})=l^+]}\hat{P}(\mathbf{z}|c^+). \tag{5}$$

with $\mathbb{I}_{[\cdot]}$ the indicator function.

3.2 Bounding Discrimination in Prediction

To approximate $\mathrm{DE}_{\mathcal{M}_h}$ from $\mathcal{D}$, sampling error cannot be avoided since $\mathcal{D}$ is only a subset of the entire population. Although exact measurement of sampling error is generally not feasible as M is unknown, it can be probabilistically bounded. In the following we first bound the difference between $\mathrm{DE}_{\mathcal{M}}$ and its approximation $\mathrm{DE}_{\mathcal{D}}$, and then extend the result to the difference between $\mathrm{DE}_{\mathcal{M}_h}$ and its approximation $\mathrm{DE}_{\mathcal{D}_h}$.

Proposition 5. *Given a causal model $\mathcal{M}$ and a sample dataset $\mathcal{D}$ with size of n, the probability of the difference between $\mathrm{DE}_{\mathcal{M}}$ and $\mathrm{DE}_{\mathcal{D}}$ no larger than t is bounded by*

$$P\left(|\mathrm{DE}_{\mathcal{M}} - \mathrm{DE}_{\mathcal{D}}| \le t\right) > 1 - 4e^{-\frac{n^+n^-}{n}t^2},$$

where n^+ and n^- ($n^+ + n^- = n$) are the numbers of individuals with c^+ and c^- in $\mathcal{D}$.

Proof. By definition of $\mathrm{DE}_{\mathcal{M}}$ and $\mathrm{DE}_{\mathcal{D}}$ we have

$$\mathrm{DE}_{\mathcal{M}} - \mathrm{DE}_{\mathcal{D}} = \left(P(l^+|c^+) - \hat{P}(l^+|c^+)\right) + \left(\hat{P}(l^+|c^-) - P(l^+|c^-)\right).$$

Denoting by $l^{(+j)}$ the label of the jth individual in $\mathcal{D}$ with $C = c^+$, we can write $\hat{P}(l^+|c^+)$ as

$$\hat{P}(l^+|c^+) = \frac{1}{n^+}\left(\mathbb{I}_{[l^{(+1)}=l^+]} + \cdots + \mathbb{I}_{[l^{(+n^+)}=l^+]}\right),$$

where indicators $\mathbb{I}_{[l^{(+j)}=l^+]}$ ($j = 1 \cdots n^+$) can be considered as independent random variables bounded by the interval $[0, 1]$. Note that $\mathbb{E}[\hat{P}(l^+|c^+)] = P(l^+|c^+)$. According to the Hoeffding's inequality [Murphy, 2012], we have

$$P\left(\left|P(l^+|c^+) - \hat{P}(l^+|c^+)\right| \ge t\right) \le 2e^{-2n^+t^2}.$$

Similarly, we have $P\left(\left|P(l^+|c^-) - \hat{P}(l^+|c^-)\right| \ge t\right) \le 2e^{-2n^-t^2}$. Therefore, we have

$$\begin{aligned}
&P\left(|\mathrm{DE}_{\mathcal{M}} - \mathrm{DE}_{\mathcal{D}}| \le t\right) \\
\ge& P\left(\left|P(l^+|c^+) - \hat{P}(l^+|c^+)\right| + \left|P(l^+|c^-) - \hat{P}(l^+|c^-)\right| \le t\right) \\
\ge& \max_{0\le x\le t} P\left(\left|P(l^+|c^+) - \hat{P}(l^+|c^+)\right| \le x\right) P\left(\left|P(l^+|c^-) - \hat{P}(l^+|c^-)\right| \le t - x\right) \\
\ge& \max_{0\le x\le t} (1 - 2e^{-2n^+x^2})(1 - 2e^{-2n^-(t-x)^2}) \\
>& 1 - 4e^{-\frac{n^+n^-}{n}t^2},
\end{aligned} \tag{6}$$

where the last line is by substituting x with $\frac{\sqrt{n^-}}{\sqrt{n^+}+\sqrt{n^-}}t$. □

For extending Proposition 5 to Proposition 6, the difference is that, since h is a classifier depending on training data $\mathcal{D}$, indicators $\mathbb{I}_{[h(c^{(+j)},\mathbf{z}^{(+j)})=l^+]}$ cannot be considered as independent. Thus, the Hoeffding's inequality cannot be directly applied and a uniform bound for all hypotheses in $\mathcal{H}$ is needed.

Proposition 6. *Given a causal model $\mathcal{M}$, a sample dataset $\mathcal{D}$, and a classifier $h : C \times \mathbf{Z} \to L$ from hypothesis space $\mathcal{H}$, the probability of the distance between $\mathrm{DE}_{\mathcal{M}_h}$ and $\mathrm{DE}_{\mathcal{D}_h}$ no larger than t is bounded by*

$$P\left(\left|\mathrm{DE}_{\mathcal{M}_h} - \mathrm{DE}_{\mathcal{D}_h}\right| \le t\right) \ge 1 - \delta(t),$$

where

$$\delta(t) = \begin{cases} 4|\mathcal{H}|^2 e^{-\frac{n^+n^-}{n}t^2} & \text{if } \mathcal{H} \text{ is finite,} \\ 4\dfrac{(2en^+)^d + (2en^-)^d}{d^d} e^{-\frac{n^+n^-}{n}t^2} & \text{if } \mathcal{H} \text{ is infinite,} \end{cases}$$

with d being the VC dimension of $\mathcal{H}$.

Proof. According to the definitions of $\mathrm{DE}_{\mathcal{M}_h}$ and $\mathrm{DE}_{\mathcal{D}_h}$,

$$\mathrm{DE}_{\mathcal{M}_h} - \mathrm{DE}_{\mathcal{D}_h} = \left(P(\tilde{l}^+|c^+) - \hat{P}(\tilde{l}^+|c^+)\right) + \left(P(\tilde{l}^+|c^-) - \hat{P}(\tilde{l}^+|c^-)\right).$$

Similar to the proof of Proposition 5, we treat $\hat{P}(\tilde{l}^+|c^+)$ as the mean of indicators $\mathbb{I}_{[h(c^{(+j)},\mathbf{z}^{(+j)})=l^+]}$ ($j = 1 \cdots n^+$). According to the generalization bound in statistical learning theory [Vapnik, 1998], if $\mathcal{H}$ is finite we have

$$P\left(\left|P(\tilde{l}^+|c^+) - \hat{P}(\tilde{l}^+|c^+)\right| \ge t\right) \le 2|\mathcal{H}|e^{-2n^+t^2},$$

where $|\mathcal{H}|$ is the size of $\mathcal{H}$. If $\mathcal{H}$ is infinite we have

$$P\left(\left|P(\tilde{l}^+|c^+) - \hat{P}(\tilde{l}^+|c^+)\right| \ge t\right) \le 4\left(\frac{2en^+}{d}\right)^d e^{-2n^+t^2},$$

where d is the VC dimension of $\mathcal{H}$. Then the proposition can be proven similarly to Eq. (6). □

Proposition 6 provides an approximation of $\mathrm{DE}_{\mathcal{M}_h}$ from $\mathrm{DE}_{\mathcal{D}_h}$. However, since pre-process methods deal with the training data, it is imperative to further link $\mathrm{DE}_{\mathcal{M}_h}$ with $\mathrm{DE}_{\mathcal{D}}$. Next, we give the relation between $\mathrm{DE}_{\mathcal{D}_h}$ and $\mathrm{DE}_{\mathcal{D}}$ in terms of a bias metric that we refer to as the the *error bias*.

Definition 2 (Error Bias)**.** *For any classifier h trained on a training dataset $\mathcal{D}$, the error bias is given by*

$$\varepsilon_{h,\mathcal{D}} = \varepsilon_1^+ - \varepsilon_2^+ - (\varepsilon_1^- - \varepsilon_2^-),$$

where $\varepsilon_1^+, \varepsilon_1^-$ are the percentages of false positives on data with $C = c^+$ and $C = c^-$ respectively, and $\varepsilon_2^+, \varepsilon_2^-$ are the percentages false negatives on data with $C = c^+$ and $C = c^-$ respectively.

Proposition 7. *For any classifier h that is trained on $\mathcal{D}$, we have*

$$\mathrm{DE}_{\mathcal{D}_h} - \mathrm{DE}_{\mathcal{D}} = \varepsilon_{h,\mathcal{D}}.$$

Proof. By definition, ε_1^+ is given by

$$\varepsilon_1^+ = \frac{1}{n^+} \sum_{\{j: c^{(j)}=c^+, l^{(j)}=l^-\}} \mathbb{I}_{[h(c^{(j)},\mathbf{z}^{(j)})=l^+]},$$

which can be rewritten as

$$\varepsilon_1^+ = \sum_{\mathbf{z}} \hat{P}(\mathbf{z}|c^+) \cdot \mathbb{I}_{[h(c^{(j)},\mathbf{z}^{(j)})=l^+]} \cdot (1 - \hat{P}(l^+|c^+, \mathbf{z})).$$

Similarly, ε_2^+ is given by

$$\varepsilon_2^+ = \sum_{\mathbf{z}} \hat{P}(\mathbf{z}|c^+) \cdot \mathbb{I}_{[h(c^{(j)},\mathbf{z}^{(j)})=l^-]} \cdot \hat{P}(l^+|c^+, \mathbf{z}).$$

Subtracting ε_2^+ from ε_1^+, we obtain

$$\varepsilon_1^+ - \varepsilon_2^+ = \sum_{\mathbf{z}} \hat{P}(\mathbf{z}|c^+)\Big(\mathbb{I}_{[h(c^+,\mathbf{z})=l^+]}(1-\hat{P}(l^+|c^+,\mathbf{z})) - \mathbb{I}_{[h(c^+,\mathbf{z})=l^-]}\hat{P}(l^+|c^+,\mathbf{z})\Big),$$

which is equivalent to

$$\varepsilon_1^+ - \varepsilon_2^+ = \sum_{\mathbf{z}} \hat{P}(\mathbf{z}|c^+) \cdot \Big(\mathbb{I}_{[h(c^+,\mathbf{z})=l^+]} - \hat{P}(l^+|c^+, \mathbf{z})\Big).$$

Similarly for data with $C = c^-$, we have

$$\varepsilon_1^- - \varepsilon_2^- = \sum_{\mathbf{z}} \hat{P}(\mathbf{z}|c^-) \cdot \Big(\mathbb{I}_{[h(c^-,\mathbf{z})=l^+]} - \hat{P}(l^+|c^-, \mathbf{z})\Big).$$

On the other hand, according to Eq. (4) and (5) we have

$$\mathrm{DE}_{\mathcal{D}_h} - \mathrm{DE}_{\mathcal{D}} = \sum_{\mathbf{z}} \hat{P}(\mathbf{z}|c^+)\Big(\mathbb{I}_{[h(c^+,\mathbf{z})=l^+]} - \hat{P}(l^+|c^+, \mathbf{z})\Big) - \sum_{\mathbf{z}} \hat{P}(\mathbf{z}|c^-)\Big(\mathbb{I}_{[h(c^-,\mathbf{z})=l^+]} - \hat{P}(l^+|c^-, \mathbf{z})\Big) = \varepsilon_1^+ - \varepsilon_2^+ - (\varepsilon_1^- - \varepsilon_2^-).$$

Letting $\varepsilon_{h,\mathcal{D}} = \varepsilon_1^+ - \varepsilon_2^+ - (\varepsilon_1^- - \varepsilon_2^-)$ completes the proof. □

Using Proposition 7, we rewrite Propositions 6 to Theorem 1 that is easier to interpret and utilize in practice.

Theorem 1. *Given a causal model $\mathcal{M}$, a sample dataset $\mathcal{D}$ and a classifier h trained on $\mathcal{D}$, $\mathrm{DE}_{\mathcal{M}_h}$ is bounded by*

$$P\left(\left|\mathrm{DE}_{\mathcal{M}_h}\right| \leq \left|\mathrm{DE}_{\mathcal{D}} + \varepsilon_{h,\mathcal{D}}\right| + t\right) \geq 1 - \delta(t),$$

where the meaning of $\delta(t)$ is same as that in Proposition 6.

Theorem 1 gives a criterion of non-discrimination in prediction that incorporates both the discrimination in the training data and the error bias of the classifier, i.e., $\left|\mathrm{DE}_{\mathcal{D}} + \varepsilon_{h,\mathcal{D}}\right|$ being bounded by a threshold τ. It shows that either given a discrimination-free dataset $\mathcal{D}$, i.e., $|\mathrm{DE}_{\mathcal{D}}| \leq \tau$, or a "balanced" classifier, i.e., $\left|\varepsilon_{h,\mathcal{D}}\right| \leq \tau$, we cannot guarantee non-discriminatory prediction. Instead, it requires to ensure that the sum of $\mathrm{DE}_{\mathcal{D}}$ and $\varepsilon_{h,\mathcal{D}}$ is within the threshold.

4 Remove Discrimination in Prediction

This section solves the problem of removing discrimination in prediction: if criterion $\left|\mathrm{DE}_{\mathcal{D}} + \varepsilon_{h,\mathcal{D}}\right| \leq \tau$ is not satisfied for a classifier, how can we meet the criterion through modifying the training data and tweaking the classifier? Denote by D^* a dataset obtained by modifying $\mathcal{D}$, and by h^* a new classifier trained on $\mathcal{D}^*$. Note that when the training data is modified, the error bias of the classifier built on it will also change. Thus, it is difficult to perform the training data modification and the classifier tweaking simultaneous. We propose a framework for modifying the training data and the classifier in two successive phases, as summarized in Algorithm 1.

Algorithm 1: Two-phase framework.

1. If $\left|\mathrm{DE}_{\mathcal{D}} + \varepsilon_{h,\mathcal{D}}\right| \leq \tau$, we are done. Otherwise, modify the labels in the training dataset $\mathcal{D}$ to obtain a modified dataset $\mathcal{D}^*$ such that $|\mathrm{DE}_{\mathcal{D}^*}| \leq \tau$.
2. Train a classifier h^* on $\mathcal{D}^*$. If $\left|\mathrm{DE}_{\mathcal{D}^*} + \varepsilon_{h^*,\mathcal{D}^*}\right| \leq \tau$, we are done. Otherwise, tweak classifier h^* to meet the above requirement.

In the first phase, we modify $\mathcal{D}$ to remove the discrimination it contains. The modified dataset $\mathcal{D}^*$ can be considered as being generated by a causal model $\mathcal{M}^*$ that is different from $\mathcal{M}$ with respect to the modification. Note that if $\left|\mathrm{DE}_{\mathcal{D}^*} + \varepsilon_{h^*,\mathcal{D}^*}\right| \leq \tau$ is achieved, Theorem 1 ensures the bound of discrimination for $\mathcal{M}^*_{h^*}$, i.e., the discrimination of h^* performed on $\mathcal{M}^*$, but not for $\mathcal{M}_{h^*}$, i.e., the discrimination of h^* performed on the original population. If we only modify the label of $\mathcal{D}$, $\mathcal{M}^*$ can be written as

$$\text{Causal Model } \mathcal{M}^* \qquad \begin{array}{l} c = f_C(pa_C, u_C) \\ z_i = f_i(pa_i, u_i) \quad i = 1, \cdots, m \\ l = f_L^*(pa_L^*, u_L^*) \end{array}$$

Then, the causal model of any classifier h^* trained on D^* and performed on $\mathcal{M}^*$ is given by

$$\text{Causal Model } \mathcal{M}^*_{h^*} \qquad \begin{array}{l} c = f_C(pa_C, u_C) \\ z_i = f_i(pa_i, u_i) \quad i = 1, \cdots, m \\ l = h^*(c, \mathbf{z}) \end{array}$$

which is equivalent to $\mathcal{M}_{h^*}$. Thus, non-discrimination in $\mathcal{M}^*_{h^*}$ also means non-discrimination in $\mathcal{M}_{h^*}$. On the other hand, if we modify attributes other than L, since the new unlabeled data is drawn from the original population, $\mathcal{M}_{h^*}$ is inconsistent with $\mathcal{M}^*_{h^*}$. As a result, the non-discrimination result of the training data cannot be applied to the prediction of the new data. Therefore, we have the following corollary derived from Theorem 1.

Size	$\text{DE}_{\mathcal{M}}$	$\text{DE}_{\mathcal{D}}$	$\text{DE}_{\mathcal{D}_h}$		$\text{DE}_{\mathcal{M}_h}$	
			DT	SVM	DT	SVM
500	**0.130**	**0.131** ± 1.6E−3	**0.145** ± 4.1E−3	**0.132** ± 8.2E−3	**0.139** ± 3.5E−3	**0.125** ± 6.8E−3
2000		**0.131** ± 4.8E−4	**0.129** ± 1.1E−3	**0.121** ± 7.4E−3	**0.130** ± 9.4E−4	**0.120** ± 7.1E−3
10000		**0.129** ± 8.0E−5	**0.138** ± 4.0E−4	**0.150** ± 4.3E−3	**0.138** ± 3.8E−4	**0.150** ± 4.3E−3

Table 2: Measured discrimination before discrimination removal (values larger than threshold are highlighted as bold).

Size	Two-phase framework (MSG)			DI	
	$\text{DE}_{\mathcal{D}^*}$	$\text{DE}_{\mathcal{M}_{h^*}}$	$\text{DE}_{\mathcal{M}_{h^*}}$ (w/o classifier tweaking)	$\text{DE}_{\mathcal{D}^*}$	$\text{DE}_{\mathcal{M}_{h^*}}$
500	0.004 ± 3.7E−6	0.015 ± 1.0E−3	**0.068** ± 4.6E−3	2E−4 ± 1.4E−3	**0.092** ± 6.1E−3
2000	0.001 ± 1.7E−7	0.016 ± 5.3E−4	**0.067** ± 4.3E−3	0.001 ± 3.4E−4	**0.095** ± 1.6E−3
10000	2E−4 ± 9.7E−9	0.013 ± 3.3E−4	**0.061** ± 3.3E−3	0.001 ± 6.8E−5	**0.107** ± 5.4E−4

Table 3: Measured discrimination after discrimination removal (decision tree as the classifier).

Corollary 1. *Let $\mathcal{D}^*$ be a modified dataset from $\mathcal{D}$, and h^* be a new classifier trained on $\mathcal{D}^*$. If $\mathcal{D}^*$ only modifies the labels, then $\left|\text{DE}_{\mathcal{D}^*} + \varepsilon_{h^*,\mathcal{D}^*}\right| \leq \tau$ is a sufficient condition to achieve*

$$P\left(\left|\text{DE}_{\mathcal{M}_{h^*}}\right| \leq \tau + t\right) \geq 1 - \delta(t),$$

where the meaning of $\delta(t)$ is same as that in Proposition 6.

In the second phase, we make modifications to h^* to reduce the error bias. Although dealing with a different fairness criterion, existing methods for balancing the misclassification rates (e.g., [Hardt *et al.*, 2016]) can be easily extended for solving this problem. For the purpose of evaluating the correctness of our theoretical results, here we use a simple algorithm *RandomFlip* for reducing the error bias that can be applied to any classifier. After the classifier makes a prediction, *RandomFlip* randomly flips the predicted label with certain probability p^+ (resp. p^-) if the individual has $C = c^+$ (resp. $C = c^-$) to achieve $\left|\text{DE}_{\mathcal{D}^*} + \varepsilon_{h^*,\mathcal{D}^*}\right| \leq \tau$, where p^+ can be computed according to the prediction of h^* over $\mathcal{D}^*$. Denoting $\sigma = \tau - |\text{DE}_{\mathcal{D}^*}|$, it suffices to make $|\varepsilon_1^+ - \varepsilon_2^+| \leq \sigma/2$ and $|\varepsilon_1^- - \varepsilon_2^-)| \leq \sigma/2$. Assume that $\varepsilon_1^+ - \varepsilon_2^+ > \sigma/2$, then it can be easily shown that p^+ should satisfy $(\varepsilon_1^+ - \varepsilon_2^+ - \sigma/2)\left(\frac{n^+}{fp+tp}\right) \leq p^+ \leq (\varepsilon_1^+ - \varepsilon_2^+)\left(\frac{n^+}{fp+tp}\right)$. Similar result can be obtained if $\varepsilon_1^+ - \varepsilon_2^+ < -\sigma/2$.

5 Empirical Evaluation

5.1 Experimental Setup

In this section, we conduct experiments to evaluate our theoretical results. For simulating a population, we adopt a semi-synthetic data generation method. We first learn a causal model $\mathcal{M}$ for a real dataset, the Adult dataset [Lichman, 2013], and treat it as the ground-truth. We then generate the training data $\mathcal{D}$ based on $\mathcal{M}$. The causal model is built using the open-source software Tetrad [Glymour and others, 2004].

The Adult dataset consists of 11 attributes including `age`, `education`, `sex`, `occupation`, `income`, etc. Due to the sparse data issue, we binarize each attribute's domain values into two classes to reduce the domain sizes. We treat `sex` as C and `income` as L. The discrimination is measured as 0.13 in $\mathcal{M}$, i.e., $\text{DE}_{\mathcal{M}} = 0.13$. Based on the underlying distribution of $\mathcal{M}$, we generate a number of training data sets with different sample sizes.

When constructing discrimination-free classifiers using the two-phase framework, we select one representative data modifying algorithm that only modifies L, the *Massaging* (MSG) algorithm [Kamiran and Calders, 2009a]. For other algorithms, we will evaluate their performance in preserving data utility in the future work. For comparison, we also include an algorithm that modifies $\mathbf{Z}$, the *Disparate Impact Removal* (DI) algorithm [Adler *et al.*, 2016]. The proposed *RandomFlip* algorithm is used for tweaking the classifier. We assume a discrimination threshold $\tau = 0.05$, i.e., we want to ensure that the discrimination in prediction is not larger than 0.05.

5.2 Experiment Results

We first measure the discrimination in each training data set, i.e., $\text{DE}_{\mathcal{D}}$. Then, we learn the classifier h from the training data where two classifiers, decision tree (DT) and support vector machine (SVM) are used. By replacing the labels in the training data with the labels predicted by the classifier, we obtain $\mathcal{D}_h$ whose discrimination is measured as $\text{DE}_{\mathcal{D}_h}$. Finally, we measure the discrimination in prediction, i.e., $\text{DE}_{\mathcal{M}_h}$ according to Proposition 3. For each sample size, we repeat the experiments 100 times by randomly generating 100 sets of training data. We report the average and variance of each measured discrimination.

The results are shown in Table 2. As expected, with the increase of the sample size, the difference between $\text{DE}_{\mathcal{M}}$ and $\text{DE}_{\mathcal{D}}$ decreases as shown by the variance. Since $\text{DE}_{\mathcal{D}} > 0.05$, the training data contains discrimination. As a result, both the training data with predicted labels, i.e., $\mathcal{D}_h$, and the prediction, i.e., $\mathcal{M}_h$, also contain discrimination.

To show the effectiveness of the two-phase framework, we first apply MSG to completely remove the discrimination in the above training data, obtaining the modified training data $\mathcal{D}^*$. Then, a decision tree h^* is built on $\mathcal{D}^*$, and the *RandomFlip* algorithm is executed to tweak the classifier so that the error bias is less than 0.05, i.e., $\left|\varepsilon_{h^*,\mathcal{D}^*}\right| \leq 0.05$. Finally, we measure the discrimination in $\mathcal{M}_{h^*}$. For comparison, the same process is also performed for DI. The results are shown in Table 3. By using the two-phase framework, discrimination is removed from the training data as shown by $\text{DE}_{\mathcal{D}^*}$, and more importantly, removed from the prediction as shown by $\text{DE}_{\mathcal{M}_{h^*}}$. We also see that, if the classifier tweaking is not performed, the prediction still contains discrimination. How-

ever, for DI, even when the discrimination is removed from the training data, and the error bias in the classifier is also removed, there still exists discrimination in prediction. These results are consistent with our theoretical conclusions.

6 Conclusions

In this paper, we addressed the limitation of the pre-process methods that there is no guarantee about the discrimination in prediction. Our theoretical results show that: (1) only removing discrimination from the training data cannot ensure non-discrimination in prediction for any classifier; and (2) when removing discrimination from the training data, one should only modify the labels in order to obtain a non-discrimination guarantee. Based on the results, we developed a two-phase framework for constructing a discrimination-free classifier with a theoretical guarantee. The experiments adopting a semi-synthetic data generation method demonstrate the theoretical results and show the effectiveness of our two-phase framework.

Acknowledgments

This work was supported in part by NSF 1646654.

References

[Adler *et al.*, 2016] Philip Adler, Casey Falk, Sorelle A Friedler, Gabriel Rybeck, Carlos Scheidegger, Brandon Smith, and Suresh Venkatasubramanian. Auditing black-box models for indirect influence. In *Proceedings of ICDM 2016*, 2016.

[Bonchi *et al.*, 2017] Francesco Bonchi, Sara Hajian, Bud Mishra, and Daniele Ramazzotti. Exposing the probabilistic causal structure of discrimination. *International Journal of Data Science and Analytics*, 3(1):1–21, 2017.

[Calders and Verwer, 2010] Toon Calders and Sicco Verwer. Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21(2):277–292, 2010.

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

[Calmon *et al.*, 2017] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems*, pages 3995–4004, 2017.

[Feldman *et al.*, 2015] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *KDD*, pages 259–268. ACM, 2015.

[Glymour and others, 2004] Clark Glymour et al. The TETRAD project. `http://www.phil.cmu.edu/tetrad`, 2004.

[Hardt *et al.*, 2016] Moritz Hardt, Eric Price, Nati Srebro, et al. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, pages 3315–3323, 2016.

[Kamiran and Calders, 2009a] Faisal Kamiran and Toon Calders. Classifying without discriminating. In *Computer, Control and Communication, 2009. IC4 2009. 2nd International Conference on*, pages 1–6. IEEE, 2009.

[Kamiran and Calders, 2009b] Faisal Kamiran and Toon Calders. Discrimination-aware classification. In *21st Benelux Conference on Artificial Intelligence (BNAIC)*, pages 333–334, 2009.

[Kamiran and Calders, 2012] Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *KAIS*, 33(1):1–33, 2012.

[Kamiran *et al.*, 2010] Faisal Kamiran, Toon Calders, and Mykola Pechenizkiy. Discrimination aware decision tree learning. In *ICDM*, pages 869–874. IEEE, 2010.

[Kamishima *et al.*, 2011] Toshihiro Kamishima, Shotaro Akaho, and Jun Sakuma. Fairness-aware learning through regularization approach. In *ICDMW*, pages 643–650. IEEE, 2011.

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

[Lichman, 2013] M. Lichman. UCI machine learning repository. `http://archive.ics.uci.edu/ml`, 2013.

[Murphy, 2012] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[Nabi and Shpitser, 2017] Razieh Nabi and Ilya Shpitser. Fair inference on outcomes. *arXiv preprint arXiv:1705.10378*, 2017.

[Pearl, 2009] Judea Pearl. *Causality: models, reasoning and inference*. Cambridge university press, 2009.

[Romei and Ruggieri, 2014] Andrea Romei and Salvatore Ruggieri. A multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review*, 29(05):582–638, 2014.

[Vapnik, 1998] Vlamimir Vapnik. *Statistical learning theory*, volume 1. Wiley New York, 1998.

[Žliobaitė *et al.*, 2011] Indre Žliobaitė, Faisal Kamiran, and Toon Calders. Handling conditional discrimination. In *ICDM*, pages 992–1001. IEEE, 2011.

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

[Zhang *et al.*, 2017] Lu Zhang, Yongkai Wu, and Xintao Wu. A causal framework for discovering and removing direct and indirect discrimination. In *IJCAI*, 2017.