New Proportion Measures of Discrimination Based on Natural Direct and Indirect Effects

Ryusei Shingaki, Manabu Kuroki

Graduate School of Engineering Science, Yokohama National University
79-5 Tokiwadai, Hodogaya-ku, Yokohama 240-8501 JAPAN
shingaki-ryusei-kw@ynu.jp, kuroki-manabu-zm@ynu.ac.jp

Abstract

Discrimination-aware data mining is expected to play an important role in data-driven decision making, as "BIG data" can be obtained from the actual society. To build the appropriate decision making system, AI researchers and practitioners have proposed various discrimination measures. However, most of the existing discrimination measures cannot be interpreted as the "proportion" and thus may not provide the comparable evaluation of the discrimination level. To evaluate how much of discrimination is based on a sensitive feature directly, indirectly, or totally, we propose three proportion measures of discrimination using natural direct and indirect effects (Pearl, 2009). The effectiveness of the proposed discrimination measures is confirmed on Adult Census Data (Becker and Kohavi, 1996).

Introduction

Current society is complicated and increasingly relies on decision-making systems based on observed data. In this situation, discrimination-aware data mining is expected to play an important role in data-driven decision making, as "BIG data" can be obtained from the society (Žliobaitė 2017; Toreini et al. 2020). In recent years, AI researchers and practitioners have developed a number of discrimination-aware data mining algorithms, using various measures to evaluate how much of discrimination is based on a sensitive (or protected) feature (Žliobaitė 2017). Such measures are called discrimination measures in this paper.

Regarding the definition of discrimination, Žliobaitė (2017) stated that "In the context of data mining and machine learning, non-discrimination can be defined as follows: (1) people that are similar in terms of non-protected characteristics should receive similar predictions, and (2) differences in predictions across groups of people can only be as large as justified by their non-protected characteristics." The first condition is related to direct discrimination or disparate treatment (Zafar et al. 2017), and the second condition ensures that there is no indirect discrimination or no disparate impact (Zafar et al. 2017). In addition, Žliobaitė (2017) pointed out, for example, that "(The first condition) can be illustrated by so called twin test: if gender is the protected attribute and we have two identical twins that share all characteristics, but gender, they should receive identical predictions." The statement "two identical twins that share *all* characteristics, but gender" is a counterfactual situation implying that statistical discrimination measures based on observed data cannot evaluate direct or indirect discrimination; the idea from causal inference is necessary to develop discrimination-aware data mining algorithms. Nevertheless, most of AI researchers and practitioners have proposed various discrimination measures from statistical viewpoints. Such discrimination measures have drawbacks. First, most of the existing discrimination measures are not designed to have a value on the range $[0, 1]$ without any assumptions: they cannot be interpreted as the 'proportion' in the mathematical meaning and thus may not provide the comparable evaluation of the discrimination level, because they are formulated based on the different target populations. Thus, it is required to interpret them based on not only the values taken by these measures but also the characteristics of the relevant parameters used in the discrimination measures, such as the values, the signs, and the interactions. Second, it is not obvious how they reflect the dependences between sensitive and non-sensitive features. Third, it is too strict and unlikely to satisfy no discrimination completely in actual situations, and thus it is necessary to formulate new discrimination measures that describe how much of discrimination is based on a sensitive feature directly, indirectly or totally.

To solve these problems, we propose causal discrimination criteria based on natural direct and indirect effects (Pearl 2009). Causal discrimination criteria are useful to clarify causal aspects of the discrimination mechanism and highlight the importance of causal inference in the field of discrimination-aware data mining. In addition, under the assumption that there is no confounder, we clarify the difference between the proposed discrimination criteria and the statistical discrimination criteria. Furthermore, to evaluate the discrimination level, referring to Zhang and Bareinboim

(2018); Plecko and Bareinboim (2022), we formulate three novel proportion measures of discrimination based on natural direct and indirect effects: the proportion of the total variation explained by direct discrimination ($\text{PTV}^{\text{direct}}$), the proportion of total discrimination explained by direct discrimination ($\text{PTD}^{\text{direct}}$), and the proportion of the total variation explained by total discrimination ($\text{PTV}^{\text{total}}$). Finally, the effectiveness of the proposed discrimination measures is confirmed on Adult Census Data (Becker and Kohavi 1996). Unlike the existing discrimination-aware data mining based on causal inference (e.g., Kilbertus et al. 2017; Kusner et al. 2017), the results of this paper contribute to the reliable evaluation of how much of discrimination is based on a sensitive feature directly, indirectly, or totally, and thus are also applicable to evaluating the degree of a sensitive feature appeared in discrimination when we wish to judge whether or not discrimination-unaware data mining is appropriate to analyze observed data.

Structral causal model

In this paper, we assume that the readers are familiar with the language of counterfactuals from the semantics of structural causal models, as given in Pearl (2009).

Given a set $\boldsymbol{V} = \{V_1, V_2, \ldots, V_m\}$ of random variables, let v_i and $\boldsymbol{v}$ represent the values taken by V_i and $\boldsymbol{V}$, respectively. In addition, for $V_i, V_j \in \boldsymbol{V}$, let $P(V_i = v_i) = P(v_i)$, $P(\boldsymbol{V} = \boldsymbol{v}) = P(\boldsymbol{v})$, and $P(V_i = v_i \mid V_j = v_j) = P(v_i|v_j)$ denote the (marginal) probability of $V_i = v_i$, the joint probability of $\boldsymbol{V} = \boldsymbol{v}$, and the conditional probability of $V_i = v_i$ given $V_j = v_j$, respectively. Furthermore, for $V_i, V_j \in \boldsymbol{V}$, let $E_{V_i}[V_i]$ and $E_{V_i|v_j}[V_i \mid V_j = v_j] = E_{V_i|v_j}[V_i|v_j]$ denote the (marginal) expectation of V_i and the conditional expectation of V_i given $V_j = v_j$, respectively. Similar notation is used for other probabilities and expectations. Here, it is noted that the discussion of this paper is mainly based on survival functions such as $P(V_i > v_i)$, because we have

$$E_{V_i}[V_i] = \sum_{v_i=0}^{\infty} \{P(V_i > v_i) - P(V_i \leq -v_i)\} \quad (1)$$

for $E_{V_i}[|V_i|] < \infty$, where the symbol $|V_i|$ indicates the absolute value of V_i. In addition, summation symbols are replaced by integrals whenever the summed variables are continuous unless noted otherwise.

The structural causal model is then defined as follows:

Definition 1 (Structural Causal Model)**.** *A structural causal model is the four-tuple $\langle \boldsymbol{U}, \boldsymbol{V}, \boldsymbol{F}, P(\boldsymbol{u}) \rangle$, where*

(1) $\boldsymbol{U}$ is a set of exogenous variables, determined by factors outside the model;

(2) $\boldsymbol{V}$ is a set of endogenous variables, determined by variables in $\boldsymbol{U} \cup \boldsymbol{V}$;

(3) $\boldsymbol{F}$ is a set of functions, where each $f_{v_i} \in \boldsymbol{F}$ is a data generating process that assigns a value

$$v_i = f_{v_i}(\boldsymbol{v} \setminus \{v_i\}, \boldsymbol{u}) \quad (2)$$

to $V_i \in \boldsymbol{V}$ in response to the values $\boldsymbol{u} \cup \boldsymbol{v} \setminus \{v_i\}$ taken by $\boldsymbol{U} \cup \boldsymbol{V} \setminus \{V_i\}$; and

(4) $P(\boldsymbol{u})$ is the probability of $\boldsymbol{U} = \boldsymbol{u}$.

Through this paper, we assume that the structural causal model does not have feedback, i.e., the output of one variable in the data generating process (2) is not returned as the input of the same variable.

In the framework of structural causal models, the external intervention of $X = x$, denoted by $\text{do}(X = x)$, represents a model manipulation that X is set to some fixed value x, regardless of how the value is ordinarily determined by the function f_x. The probability of $Y > y$ when the external intervention $\text{do}(X = x)$ is conducted, denoted by $P(Y > y \mid \text{do}(X = x))$, is called a causal risk of $X = x$ on $Y > y$ in this paper. When causal quantities such as $P(Y > y \mid \text{do}(X = x))$ are given as a non-trivial function of both x and y, it is said that X has an effect on $Y > y$, or Y is affected by X.

Using the semantics of structural causal models, $P(Y > y \mid \text{do}(X = x))$ can be translated into the probability of the potential outcome variable Y_x. Here, the potential outcome $Y_x = y$ represents the counterfactual sentence "Y would have the value y, had X been x." Similar notation is applied to other potential outcome variables. Then, we have

$$P(Y > y \mid \text{do}(X = x)) = P(Y_x > y).$$

When a randomized experiment is conducted, X is independent of Y_x for any x, denoted as

$$Y_x \perp\!\!\!\perp X$$

for any x. This condition is often called exogeneity. Under the assumption of exogeneity, by the consistency property (Pearl 2009), i.e.,

$$X = x \Longrightarrow Y_x = Y, \quad (3)$$

$P(Y_x > y)$ is identifiable and is given by

$$P(Y_x > y) = P(Y > y \mid x).$$

Here, 'identifiable' means that causal quantities such as $P(Y_x > y)$ can be estimated consistently from a joint probability of observed variables. When a randomized experiment is difficult to conduct, $P(Y_x > y)$ can still be identified in accordance with conditional ignorability, or graphically, the back door criterion (Pearl 2009): when there exists a set $\boldsymbol{Z}$ of variables such that X is conditionally independent of Y_x given $\boldsymbol{Z}$ for any x, denoted as

$$Y_x \perp\!\!\!\perp X \mid \boldsymbol{Z}$$

for any x, and $P(x \mid \boldsymbol{z}) > 0$ for any x and $\boldsymbol{z}$, we say that $\boldsymbol{Z}$ satisfies the conditional ignorability condition relative to (X, Y). Then, by the consistency property,

$P(Y_x > y)$ can be estimated using a set $\boldsymbol{Z}$ of observed variables as follows:

$$P(Y_x > y) = E_{\boldsymbol{z}}\left[P(Y > y \mid x, \boldsymbol{Z})\right],$$

where $E_{\boldsymbol{z}}\left[P(Y > y \mid x, \boldsymbol{Z})\right]$ stands for the expectation of $P(Y > y \mid x, \boldsymbol{Z})$ regarding $\boldsymbol{Z}$. The other identification conditions of causal risks are given in Pearl (2009).

Effect measures

To propose new discrimination measures, we let X be a sensitive feature to be protected and Y be an outcome variable. A variable that is not sensitive is called a non-sensitive feature. In addition, we let S stand for an intermediate variable that would be affected by X and could have an effect on Y. Note that our discussion is based on a single intermediate variable, but the extension of our results to multiple intermediate variables is straightforward. Let $P(X = x) = P(x)$ and $P(Y > y \mid X = x) = P(Y > y \mid x)$ denote the (marginal) probability of $X = x$ and the conditional probability of $Y > y$ given $X = x$, respectively. Similar notation is used for other probabilities.

One of representative discrimination measures is the total variation (TV) (Žliobaitė 2017; Plecko and Bareinboim 2022):

Definition 2 (Total Variation)**.** *The total variation (TV) on $Y > y$ comparing $X = x_1$ to $X = x_0$, denoted by $\mathrm{TV}_y(x_1, x_0)$, is defined as*

$$\mathrm{TV}_y(x_1, x_0) = P(Y > y \mid x_1) - P(Y > y \mid x_0)$$

for the risk difference scale, and

$$\mathrm{TV}_y(x_1, x_0) = P(Y > y \mid x_1)/P(Y > y \mid x_0)$$

for the risk ratio scale assuming $P(Y > y \mid x_0) \neq 0$.

The TV is nothing more than a statistical measure, since it is the comparison between the probabilities of Y in the passively observed groups of $X = x_1$ and $X = x_0$. Thus, the TV is identifiable without causal knowledge if observed probabilities $P(y \mid x)$ are available.

Let Y_x be the potential outcome variable that represents the counterfactual sentence "Y would have the value y, had X been x." Similar notation is applied to other potential outcome variables. Then, another representative discrimination measure is the total effect (TE):

Definition 3 (Total Effect)**.** *The total effect (TE) on $Y > y$ comparing $X = x_1$ to $X = x_0$, denoted by $\mathrm{TE}_y(x_1, x_0)$, is defined as*

$$\mathrm{TE}_y(x_1, x_0) = P(Y_{x_1} > y) - P(Y_{x_0} > y)$$

for the risk difference scale, and

$$\mathrm{TE}_y(x_1, x_0) = P(Y_{x_1} > y)/P(Y_{x_0} > y)$$

for the risk ratio scale assuming $P(Y_{x_0} > y) \neq 0$.

Differently from the TV, the TE measures the change on the probability of Y when X changes from $X = x_0$ to $X = x_1$ by the external intervention while S is allowed to track the change in X.

Referring to Pearl (2009), we define the NDE and NIE as follows:

Definition 4 (Natural Direct and Indirect Effects)**.** *The natural direct effect (NDE) on $Y > y$ comparing $X = x_1$ to $X = x_0$ when S is set to S_{x_1}, denoted by $\mathrm{NDE}_y(x_1, x_0|S)$, is defined as*

$$\mathrm{NDE}_y(x_1, x_0|S) = P(Y_{x_1, S_{x_1}} > y) - P(Y_{x_0, S_{x_1}} > y)$$

for the risk difference scale, and

$$\mathrm{NDE}_y(x_1, x_0|S) = P(Y_{x_1, S_{x_1}} > y)/P(Y_{x_0, S_{x_1}} > y)$$

for the risk ratio scale assuming $P(Y_{x_0, S_{x_1}} > y) \neq 0$. The natural indirect effect (NIE) on $Y > y$ comparing S_{x_1} to S_{x_0} when X is set to x_0, denoted by $\mathrm{NIE}_y(x_1, x_0|S)$, is defined as

$$\mathrm{NIE}_y(x_1, x_0|S) = P(Y_{x_0, S_{x_1}} > y) - P(Y_{x_0, S_{x_0}} > y)$$

for the risk difference scale, and

$$\mathrm{NIE}_y(x_1, x_0|S) = P(Y_{x_0, S_{x_1}} > y)/P(Y_{x_0, S_{x_0}} > y)$$

for the risk ratio scale assuming $P(Y_{x_0, S_{x_0}} > y) \neq 0$.

The NDE measures the change on the probability of Y as X changes from $X = x_0$ to $X = x_1$ by the external intervention while setting S to whatever value it would have obtained under $X = x_1$. Contrary, the NIE measures the change on the probability of Y when the X is held constant at $X = x_0$, and S changes to whatever value it would have attained under $X = x_1$. Here, when we focus on direct and indirect discrimination based on S, note that the idea of the 'twin test' stated in Žliobaitė (2017) is reflected in the NDE and NIE. In fact, in an employment-discrimination case known as Carson versus Bethlehem Steel Corp. (70 FEP Cases 921, 7th Cir. (1996)), which was introduced by Pearl (2001) to discuss the NDE and NIE, the court wrote

> The central question in any employment-discrimination case is whether the employer would have taken the *same action* had the employee been of a *different race* (age, sex, religion, national origin, etc.) and *everything else had been the same*,

which implies that the twin test is required to evaluate the discrimination level, and such a court ruling is taken into account as the unit-level NDE and NIE.

If we have

$$P(Y_x > y) = P(Y > y \mid x) \tag{4}$$

for all x and y, then it is said that X and Y are not confounded (Pearl 2009). Based on this situation, we define the spurious effect (SE) as follows:

Definition 5 (Spurious Effect)**.** *The spurious effect (SE) on $Y > y$ of comparing $X = x_1$ to $X = x_0$, denoted by $\mathrm{SE}_y(x_1, x_0)$, is defined as*

$$\mathrm{SE}_y(x_1, x_0) = \mathrm{TV}_y(x_1, x_0) - \mathrm{TE}_y(x_1, x_0)$$

for the risk difference scale, and

$$\mathrm{SE}_y(x_1, x_0) = \mathrm{TV}_y(x_1, x_0)/\mathrm{TE}_y(x_1, x_0)$$

for the risk ratio scale assuming $\mathrm{TE}_y(x_1, x_0) \neq 0$.

Intuitively, equation (4) states that X and Y are not confounded whenever the observationally witnessed association between them is the same as the association that would be measured in a randomized experiment. Note that the SE does not depend on the selection of intermediate variables.

Referring to Zhang and Bareinboim (2018); Plecko and Bareinboim (2022), the following theorem is straightforward:

Theorem 1. *The TV, NDE, NIE, and SE obey the following relationships:*

$$\begin{aligned}\mathrm{TV}_y(x_1,x_0) = \mathrm{NDE}_y(x_1,x_0|S) + \mathrm{NIE}_y(x_1,x_0|S) \\ + \mathrm{SE}_y(x_1,x_0)\end{aligned} \tag{5}$$

for the risk difference scale, and

$$\begin{aligned}\mathrm{TV}_y(x_1,x_0) = \mathrm{NDE}_y(x_1,x_0|S) \times \mathrm{NIE}_y(x_1,x_0|S) \\ \times \mathrm{SE}_y(x_1,x_0)\end{aligned} \tag{6}$$

for both the risk ratio.

Hereafter, we focus on the risk difference scale because the risk ratio scale can be written as the risk difference scale through the logarithm transformation of the TV, i.e.,

$$\begin{aligned}\log \mathrm{TV}_y(x_1,x_0) = \log \mathrm{NDE}_y(x_1,x_0|S) \\ + \log \mathrm{NIE}_y(x_1,x_0|S) + \log \mathrm{SE}_y(x_1,x_0)\end{aligned}$$

from equation (6). From equation (5), if the NDE, NIE, and SE (or the TE and SE) are zero simultaneously then TV is also zero. This fact would be useful to detect the possibility of discrimination from observed data because $\mathrm{TV} \neq 0$ implies that at least one of NDE, NIE, and SE is non-zero. On the contrary, $\mathrm{TV} = 0$ does not imply that the NDE, NIE, and SE are zero simultaneously, because of the parametric cancellation. In addition, if a set $\{X, Y, S\} \cup \boldsymbol{Z}$ of observed variables satisfies the sequential ignorability condition, i.e.,

$$\begin{aligned}\{Y_{x,s}, S_{x'}\} \perp\!\!\!\perp X \mid \boldsymbol{Z}, \quad Y_{x,s} \perp\!\!\!\perp S_{x'} \mid \boldsymbol{Z}, \\ Y_{x,s} \perp\!\!\!\perp S \mid \{X\} \cup \boldsymbol{Z},\end{aligned}$$

$P(x \mid \boldsymbol{z}) > 0$, and $P(s \mid x, \boldsymbol{z}) > 0$ for $x, x' \in \{x_1, x_0\}$, then $P(Y_x > y)$ and $P(Y_{x,S_{x'}} > y)$ are identifiable (Imai et al. 2011).

Discrimination criteria

In this section, we propose the discrimination criteria based on the semantics of structural causal models as follows:

Definition 6 (Causal Discrimination Criteria)**.** *Letting X, S, and Y be a sensitive feature, an intermediate variable, and an outcome variable, respectively, we say that*

(1) there is no causal direct discrimination not via S if $\mathrm{NDE}_y(x_1,x_0|S) = 0$ holds for any y,

(2) there is no causal indirect discrimination via S if $\mathrm{NIE}_y(x_1,x_0|S) = 0$ holds for any y,

(3) there is no spurious discrimination if $\mathrm{SE}_y(x_1,x_0) = 0$ holds for any y.

Here, if the assumption of (1) does not hold, we say that there is causal direct discrimination not via S. Similarly, if the assumption of (2) does not hold, we say that there is causal indirect discrimination via S. In addition, if the assumption of (3) does not hold, there is spurious discrimination. Especially, there is no causal total discrimination if $\mathrm{TE}_y(x_1,x_0) = 0$ holds; otherwise, we say that there is causal total discrimination.

The sequential ignorability condition plays an important role in clarifying the difference between causal discrimination criteria and the existing statistical discrimination criteria. To see this, assuming that $\boldsymbol{Z}$ is empty in the sequential ignorability condition, note that $\mathrm{NDE}_y(x_1,x_0|S)$ is zero if $X \perp\!\!\!\perp Y \mid S$ holds and $\mathrm{NIE}_y(x_1,x_0|S)$ is zero if $X \perp\!\!\!\perp S$ or $S \perp\!\!\!\perp Y \mid X$ holds. Based on the consideration, we propose statistical discrimination criteria as follows:

Definition 7 (Statistical Discrimination Criteria (I))**.** *Letting X, $\boldsymbol{W}$ and Y be a sensitive feature, a set of non-sensitive features, and an outcome variable, respectively, we say that*

(1) there is no statistical direct discrimination given $\boldsymbol{W}$ if $X \perp\!\!\!\perp Y \mid \boldsymbol{W}$, $X \not\perp\!\!\!\perp \boldsymbol{W}$, and $\boldsymbol{W} \not\perp\!\!\!\perp Y \mid X$ hold,

(2) there is no statistical indirect discrimination given $\boldsymbol{W}$ if $X \perp\!\!\!\perp \boldsymbol{W}$ or $\boldsymbol{W} \perp\!\!\!\perp Y \mid X$ holds.

Statistical discrimination criteria (I) is similar to Wang and Taylor's criteria for validating surrogate endpoints (Wang and Taylor 2002) in the context of randomized clinical trials (RCTs). Especially, Kamishima et al. (2012) introduced the condition of $\boldsymbol{W} \not\perp\!\!\!\perp Y \mid X$ as the statistical concept of "direct prejudice" into discrimination-aware data mining.

Statistical discrimination criteria (I) can be considered to reflect statistical aspect of causal discrimination criteria through the sequential ignorability condition. Contrary, to derive statistical discrimination criteria which refer to the existing discrimination criteria, according to Kamishima et al. (2012) and Žliobaitė (2017), consider the following conditions:

(1) The probabilities of the outcome variable are equal for all possible values taken by the sensitive feature. Statistically, this is interpreted as $X \perp\!\!\!\perp Y$.

(2) The probabilities of the outcome variable are equal for all possible values taken by the sensitive feature given a specific value of a non-sensitive feature. Statistically, this is interpreted as $X \perp\!\!\!\perp Y \mid \boldsymbol{W}$.

The condition of $X \perp\!\!\!\perp Y$ is called the independence criterion in the sense that the information on X is not necessary to predict Y. Meanwhile, the condition of $X \perp\!\!\!\perp Y \mid \boldsymbol{W}$ is called the sufficiency criterion in the sense that we do not need to see X when we know $\boldsymbol{W}$ to predict Y. Here, note that these discrimination criteria do not consider the statistical dependence between X and $\boldsymbol{W}$. To solve the problem, consider statistical conditions of $X \not\perp\!\!\!\perp \boldsymbol{W}$ and $Y \not\perp\!\!\!\perp \boldsymbol{W}$ which are introduced

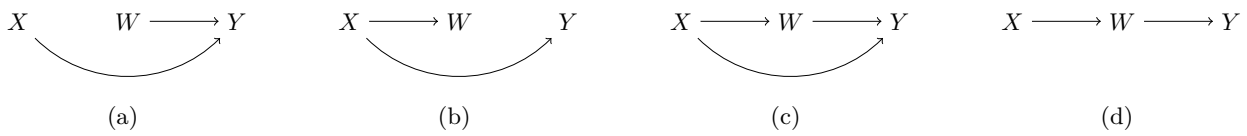

Figure 1: Graphical representations of statistical discrimination criteria (I) and (II): (a) no indirect discrimination; (b) no indirect discrimination by statistical discrimination criteria (I), but inseparable discrimination by statistical discrimination criteria (II); (c) inseparable discrimination; and (d) no direct discrimination.

as the concepts of "latent prejudice" and "indirect prejudice", respectively, by Kamishima et al. (2012). Then, we re-define the existing statistical discrimination criteria as follows:

Definition 8 (Statistical Discrimination Criteria (II))**.** *Letting X, $\boldsymbol{W}$, and Y be a sensitive feature, a set of non-sensitive features, and an outcome variable, respectively, we say that*

(1) there is no statistical direct discrimination given $\boldsymbol{W}$, if $X \perp\!\!\!\perp Y \mid \boldsymbol{W}$, $X \not\perp\!\!\!\perp \boldsymbol{W}$, and $\boldsymbol{W} \not\perp\!\!\!\perp Y$ hold,

(2) there is no statistical indirect discrimination given $\boldsymbol{W}$, if $X \perp\!\!\!\perp \boldsymbol{W}$ or $\boldsymbol{W} \perp\!\!\!\perp Y$ holds.

Statistical discrimination criteria (II) is similar to Prentice's criteria for validating surrogate endpoints (Prentice 1989) in RCTs.

When W is an intermediate variable, the difference between statistical discrimination criteria (I) and (II) can be clarified in a causal diagram (Pearl 2009) that illustrates the direct and indirect effects via different paths between X and Y (see Figure 1). In this setting, using either statistical discrimination criteria, Figure 1a and 1d show the situations judged as no indirect and no direct discrimination, respectively. In contrast, Figure 1b shows the situation judged as no indirect discrimination from statistical discrimination criteria (I), but as both direct and indirect discrimination from statistical discrimination criteria (II), because $W \not\perp\!\!\!\perp Y$, $X \not\perp\!\!\!\perp Y \mid W$, and $W \perp\!\!\!\perp Y \mid X$ hold. This consideration implies that causal discrimination criteria and the existing statistical discrimination criteria are derived from the different motivations for a sensitive feature, even if no unmeasured confounders exist.

Contrary, regarding the equivalence between statistical discrimination criteria (I) and (II), the following theorem is derived:

Theorem 2. *When the probabilities of X, Y, and $\boldsymbol{W}$ are strictly positive and $X \not\perp\!\!\!\perp Y$ and $X \perp\!\!\!\perp Y \mid \boldsymbol{W}$ hold, there is no statistical direct discrimination given $\boldsymbol{W}$ in the sense of both statistical discrimination criteria (I) and (II).*

Proof. Assuming that the probabilities of X, Y, and $\boldsymbol{W}$ are strictly positive and $X \perp\!\!\!\perp Y \mid \boldsymbol{W}$ holds, then $\boldsymbol{W} \perp\!\!\!\perp Y \mid X$ and $X \perp\!\!\!\perp Y \mid \boldsymbol{W}$ imply that $\{X, \boldsymbol{W}\} \perp\!\!\!\perp Y$ by the intersection property (Pearl 1988), which induces both $X \perp\!\!\!\perp Y$ and $\boldsymbol{W} \perp\!\!\!\perp Y$ from the decomposition property (Pearl 1988). However, as both $X \not\perp\!\!\!\perp Y$ and $X \perp\!\!\!\perp Y \mid \boldsymbol{W}$ are assumed, $\boldsymbol{W} \not\perp\!\!\!\perp Y \mid X$ holds by contraposition.

	$X \perp\!\!\!\perp Y \mid \boldsymbol{W}$		$X \not\perp\!\!\!\perp Y \mid \boldsymbol{W}$	
	$\boldsymbol{W} \perp\!\!\!\perp Y \mid X$	$\boldsymbol{W} \not\perp\!\!\!\perp Y \mid X$	$\boldsymbol{W} \perp\!\!\!\perp Y \mid X$	$\boldsymbol{W} \not\perp\!\!\!\perp Y \mid X$
$X \perp\!\!\!\perp \boldsymbol{W}$	—	—	NI	NI
$X \not\perp\!\!\!\perp \boldsymbol{W}$	—	ND	NI	ID

(a) Statistical discrimination criteria (I).

	$X \perp\!\!\!\perp Y \mid \boldsymbol{W}$		$X \not\perp\!\!\!\perp Y \mid \boldsymbol{W}$	
	$\boldsymbol{W} \perp\!\!\!\perp Y$	$\boldsymbol{W} \not\perp\!\!\!\perp Y$	$\boldsymbol{W} \perp\!\!\!\perp Y$	$\boldsymbol{W} \not\perp\!\!\!\perp Y$
$X \perp\!\!\!\perp \boldsymbol{W}$	—	—	NI	NI
$X \not\perp\!\!\!\perp \boldsymbol{W}$	—	ND	NI	ID

(b) Statistical discrimination criteria (II).

Table 1: Comparison between statistical discrimination criteria (I) and (II). "ND", "NI", "ID", and "—" mean no direct discrimination, no indirect discrimination, both direct and indirect discrimination, and contradiction against $X \not\perp\!\!\!\perp Y$, respectively.

Similarly, $\boldsymbol{W} \perp\!\!\!\perp Y$ and $X \perp\!\!\!\perp Y \mid \boldsymbol{W}$ imply $\{X, \boldsymbol{W}\} \perp\!\!\!\perp Y$ by the contraction property (Pearl 1988). As both $X \not\perp\!\!\!\perp Y$ and $X \perp\!\!\!\perp Y \mid \boldsymbol{W}$ are assumed, $\boldsymbol{W} \not\perp\!\!\!\perp Y$ also holds by contraposition. Finally, if both $X \perp\!\!\!\perp Y \mid \boldsymbol{W}$ and $X \perp\!\!\!\perp \boldsymbol{W}$ imply $X \perp\!\!\!\perp \{Y, \boldsymbol{W}\}$ by the contraction property (Pearl 1988). However, from the assumption of both $X \not\perp\!\!\!\perp Y$ and $X \perp\!\!\!\perp Y \mid \boldsymbol{W}$, $X \not\perp\!\!\!\perp \boldsymbol{W}$ holds by contraposition. □

Under the assumption of $X \not\perp\!\!\!\perp Y$, the relationships between statistical dependencies and statistical discrimination criteria (I) and (II) are shown in Table 1a and 1b, respectively. Table 1 demonstrates that the statistical judgment of direct, indirect, or both direct and indirect discrimination depends on which criteria are used to detect or explain how discrimination occurs.

Theorem 2 does not mean that there are both direct and indirect discrimination in the sense of statistical discrimination criteria (I) if and only if there are both direct and indirect discrimination in the sense of statistical discrimination criteria (II). Based on the consideration, by introducing the separation criterion $X \perp\!\!\!\perp \boldsymbol{W} \mid Y$ (Zafar et al. 2017), the equivalence condition between statistical discrimination criteria (I) and (II) is derived as follows:

Theorem 3. *When the probabilities of X, Y, and $\boldsymbol{W}$ are strictly positive and $X \perp\!\!\!\perp \boldsymbol{W} \mid Y$ holds under the*

assumption $X \not\perp\!\!\!\perp Y$, statistical discrimination criteria (I) and (II) are equivalent.

Proof. The combination of $\boldsymbol{W} \perp\!\!\!\perp Y \mid X$ and $X \perp\!\!\!\perp \boldsymbol{W} \mid Y$ induces $\boldsymbol{W} \perp\!\!\!\perp \{X, Y\}$ by the intersection property (Pearl 1988). Similarly, the combination of $\boldsymbol{W} \perp\!\!\!\perp Y$ and $X \perp\!\!\!\perp \boldsymbol{W} \mid Y$ induces $\boldsymbol{W} \perp\!\!\!\perp \{X, Y\}$ by the contraction property (Pearl 1988). In addition, the combination of $X \perp\!\!\!\perp Y \mid \boldsymbol{W}$ and $X \perp\!\!\!\perp \boldsymbol{W} \mid Y$ induces $X \perp\!\!\!\perp \{\boldsymbol{W}, Y\}$ by the intersection property (Pearl 1988). However, as $X \not\perp\!\!\!\perp Y$ holds, $X \not\perp\!\!\!\perp Y \mid \boldsymbol{W}$ can be derived by the contradiction. This implies that there is direct discrimination. Thus, under the assumption that both $X \not\perp\!\!\!\perp Y$ and $X \perp\!\!\!\perp \boldsymbol{W} \mid Y$ hold, $X \perp\!\!\!\perp \boldsymbol{W}$ implies that there is no indirect discrimination in the sense of both statistical discrimination criteria (I) and (II), and $X \not\perp\!\!\!\perp \boldsymbol{W}$ implies that there are both direct and indirect discrimination in the sense of both statistical discrimination criteria (I) and (II). □

Proportion measures of discrimination

In actual situations, it would be rare to strictly satisfy causal discrimination criteria so that it is reasonable to evaluate the discrimination level representing how much of discrimination is based on the sensitive feature directly, indirectly or totally. However, most of the existing discrimination measures cannot be interpreted as the "proportion" in the mathematical meaning and thus may not provide the comparable evaluation of the discrimination level, because they are formulated based on the different target populations. To solve the problem, we propose three types of novel proportion measures of discrimination: the proportion of the TV explained by direct discrimination ($\mathrm{PTV}^{\mathrm{direct}}$), the proportion of total discrimination explained by direct discrimination ($\mathrm{PTD}^{\mathrm{direct}}$), and the proportion of the TV explained by total discrimination ($\mathrm{PTV}^{\mathrm{total}}$). These measures are defined as follows:

$$\mathrm{PTV}_y^{\mathrm{direct}}(x_1, x_0|S) = \frac{\mathrm{NDE}_y(x_1, x_0|S)^2}{\mathrm{SE}_y(x_1, x_0)^2 + \mathrm{NIE}_y(x_1, x_0|S)^2 + \mathrm{NDE}_y(x_1, x_0|S)^2} \tag{7}$$

$$\mathrm{PTD}_y^{\mathrm{direct}}(x_1, x_0|S) = \frac{\mathrm{NDE}_y(x_1, x_0|S)^2}{\mathrm{NIE}_y(x_1, x_0|S)^2 + \mathrm{NDE}_y(x_1, x_0|S)^2} \tag{8}$$

$$\mathrm{PTV}_y^{\mathrm{total}}(x_1, x_0) = \frac{\mathrm{TE}_y(x_1, x_0)^2}{\mathrm{SE}_y(x_1, x_0)^2 + \mathrm{TE}_y(x_1, x_0)^2}, \tag{9}$$

where 0/0 is defined as 0 in this paper. As seen from equations (7), (8), and (9), the proposed discrimination measures are defined based on a single target population, and always fall inside the range $[0, 1]$ without any assumptions.

The higher values of the proposed discrimination measures show a more severe situation in the sense that most part of discrimination is attributed to the sensitive feature alone and thus may not be removable by adjusting the non-sensitive features (in this paper, "severe" does not always mean the degree of social seriousness based on the sensitive feature). In this sense, the proposed discrimination measures help us to classify the severity of discrimination, as shown in the last part of this section. In addition, the proposed discrimination measures are applicable to evaluating how much of discrimination is explained by causal direct or total discrimination, in order to judge whether or not the discrimination-unaware data mining algorithm is appropriate to analyze observed data.

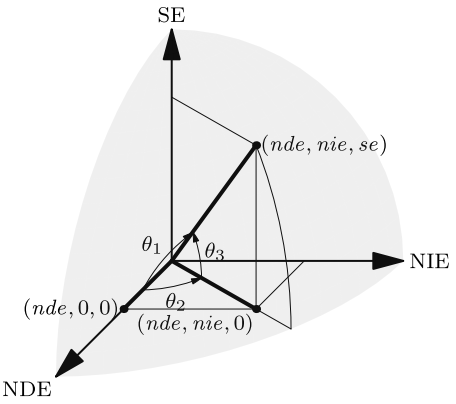

Figure 2: Motivating idea of the proposed discrimination measures.

Here, it would be worthwhile to state that the motivating idea behind the proposed discrimination measures comes from the similarity measure used in cluster analysis, i.e., the cosine similarities between the TV and the NDE and between the TE and the NDE, as shown in Figure 2. Letting TV, TE, and NDE correspond to the vectors (nde, nie, se), $(nde, nie, 0)$, and $(nde, 0, 0)$, respectively, the cosine similarities between the TV and the NDE and between the TE and the NDE correspond to $\mathrm{PTV}_y^{\mathrm{direct}}(x_1, x_0|S)$ and $\mathrm{PTD}_y^{\mathrm{direct}}(x_1, x_0|S)$, respectively. Indeed, denoting the angle between the TV and the NDE as θ_1, which is the angle between the vectors (nde, nie, se) and $(nde, 0, 0)$ and the angle between the TE and the NDE as θ_2, which is the angle between the vectors $(nde, nie, 0)$ and $(nde, 0, 0)$, we have

$$\mathrm{PTV}_y^{\mathrm{direct}}(x_1, x_0|S) = \cos^2\theta_1,$$
$$\mathrm{PTD}_y^{\mathrm{direct}}(x_1, x_0|S) = \cos^2\theta_2.$$

Then, letting θ_3 be the angle between (nde, nie, se) and $(nde,\ nie,\ 0.0)$, which is interpreted as the angle between the TV and the TE, we have $\cos^2\theta_1 = \cos^2\theta_2 \cos^2\theta_3$, i.e.,

$$\mathrm{PTV}_y^{\mathrm{direct}}(x_1, x_0|S) = \mathrm{PTD}_y^{\mathrm{direct}}(x_1, x_0|S)\cos^2\theta_3$$

from "Theorem of Three Perpendiculars". This shows that these discrimination measures are not sufficient for

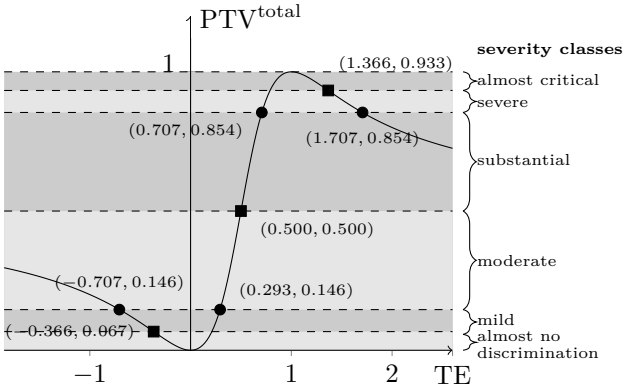

Figure 3: $\mathrm{PTV}^{\mathrm{total}}$ versus TE graph for TV = 1.000 and TE = $(-1.500, 2.500)$. ■ and • indicate inflection and jerk points, respectively.

determining each other, but are not entirely independent. In addition, $\cos^2\theta_3$ can be considered as another discrimination measure representing the proportion of the TV explained by total discrimination.

From a mathematical viewpoint, unlike the existing discrimination measures, the proposed discrimination measures can provide cut-off values for judging the severity class of discrimination based on the derivatives of $\mathrm{PTD}^{\mathrm{direct}}$ and $\mathrm{PTV}^{\mathrm{total}}$. To see this, we consider that $\mathrm{PTV}^{\mathrm{total}}$ is a function of the $\mathrm{TE}_y(x_1, x_0)(=\mathrm{TE})$ for a given value of the $\mathrm{TV}_y(x_1, x_0)(=\mathrm{TV})$. Then, the inflection points of the function are obtained when $\mathrm{TE} = \frac{1}{2}\mathrm{TV}$, $\frac{1 \pm \sqrt{3}}{2}\mathrm{TV}$ by taking the second derivative of $\mathrm{PTV}^{\mathrm{total}}$ with respect to TE. The jerk points of the function are obtained when $\mathrm{TE} = \frac{1}{\sqrt{2}}\mathrm{TV}$, $\frac{2 \pm \sqrt{2}}{2}\mathrm{TV}$ by taking the third derivative of $\mathrm{PTV}^{\mathrm{total}}$ with respect to TE. With this consideration, we suggest several severity classes of discrimination shown in Figure 3. As the ranges of "substantial" and "moderate" may be considered practically too wide, noting that the fourth derivative of the $\mathrm{PTV}^{\mathrm{total}}$ does not provide the value in the range $[0.500, 0.854)$ or $[0.146, 0.500)$, we divide the range $[0.500, 0.854)$ into two parts, $[0.500, 0.634)$ and $[0.634, 0.854)$ based on the fifth derivative of the $\mathrm{PTV}^{\mathrm{total}}$. Similarly, we divide the range $[0.146, 0.500)$ into two parts, $[0.146, 0.366)$ and $[0.366, 0.500)$. Considering the sampling variability, it would be better to judge the severity class of discrimination by comparing the upper/lower limits of the confidence intervals of the discrimination measures with Figure 3, rather than the point estimates.

Application

We apply the discrimination measures to the Adult Census Dataset, available from the UCI Repository of Machine Learning Databases (Becker and Kohavi 1996). The data consists of 1994 census information in the US. The Adult Census Dataset contains 48,842 instances with 6 numerical and 8 categorical features. A predictive model developed on this data was expected to determine whether a person's income makes over 50K a year.

The individual's income Y is a dichotomous outcome variable indicating whether a person's income makes over 50,000 a year, i.e., his/her income above 50,000 ($Y = y_1$) or below 50,000 ($Y = y_0$). Following Hamilton (2017), "sex" is considered as a sensitive feature X. Here, $X = x_1$ indicates the disadvantaged group (female for "sex") and $X = x_0$ indicates the advantaged group (male for "sex"). For details, refer to Hamilton (2017).

In order to evaluate the discrimination measures, under the sequential ignorability condition, we assume the logistic regression model of Y on X and $\boldsymbol{W}$ as a predictive model:

$$P(Y = y_1 \mid X = x, \boldsymbol{W} = \boldsymbol{w}) = \frac{1}{1 + \exp\{-(\theta_0 + \theta_x x + \boldsymbol{\theta}_w^\top \boldsymbol{w})\}},$$

where $(\theta_0, \theta_x, \boldsymbol{\theta}_w^\top)^\top$ is a coefficients vector of the logistic regression model and $\boldsymbol{W}$ includes 13 features (excluding the sensitive feature X). In this scenario, the performances of the existing and proposed discrimination measures are listed in Table 2. Here, the normalized mean difference (NMD) is a representative discrimination measure, whereas "slift" and "elift" are known as the impact (risk) ratio and the ratio of additive interactions, respectively (Žliobaitė 2017). In addition, Table 2 shows the sample estimates from the original data (denoted by "estimate"), the standard errors (denoted by "s.e.") and the 95% bootstrap confidence intervals (CIs) (denoted by "lcl" for the lower confidence limits, and "ucl" for the upper confidence limits) evaluated by 2,000 bootstrap replications.

From Table 2, the 95% CIs of the TV and NMD do not include zero, and the 95% CIs of the slift and elift do not include one. It seems that the association between the sensitive feature and the outcome is statistically significant, and thus we can consider that $X \not\perp\!\!\!\perp Y$ holds. However, from the existing discrimination measures, it is uncertain how much of discrimination is based on the sensitive feature not via non-sensitive features. On the contrary, the estimates of the $\mathrm{PTD}^{\mathrm{direct}}$, $\mathrm{PTV}^{\mathrm{direct}}$, and $\mathrm{PTV}^{\mathrm{total}}$ show that most of total discrimination and total variation are based on the sensitive feature directly and totally. In addition, the 95% lower confidence limits of $\mathrm{PTD}^{\mathrm{direct}}$, $\mathrm{PTV}^{\mathrm{direct}}$, and $\mathrm{PTV}^{\mathrm{total}}$ are above 0.854. This shows that "sex" may be judged as "severe", "almost critical" or "critical" sensitive feature from the viewpoint of the proposed measures.

Conclusion

Most of the existing measures proposed for discrimination-aware data mining have the deficiencies stated in

	TV	NMD	slift	elift	$\text{PTD}^{\text{direct}}$	$\text{PTV}^{\text{direct}}$	$\text{PTD}^{\text{total}}$
estimate	−0.196	0.545	0.358	1.270	0.935	0.934	1.000
s.e.	0.004	0.011	0.010	0.006	0.038	0.042	0.029
lcl	−0.197	0.545	0.358	1.269	0.922	0.914	0.977
mean	−0.196	0.545	0.358	1.270	0.924	0.916	0.978
ucl	−0.196	0.546	0.359	1.270	0.926	0.917	0.980

Table 2: Summary statistics of the discrimination measures from Adult Census Dataset.

Section . To overcome these deficiencies, we proposed causal discrimination measures based on the natural direct and indirect effects. The proposed discrimination measures are not estimable from observed data without causal knowledge, but the bounding formulas for the causal quantities (e.g., Balke and Pearl 1997; Cai et al. 2008) would play an important role in evaluating the discrimination level. In addition, although Zhang and Bareinboim (2018) introduced the idea of the effect decomposition based on the "effect of treatment on the treated" into discrimination-aware data mining, the application of our results to their framework is straightforward.

Acknowledgments

This research was partially funded by JFE Engineering Corporation and Japan Society for the Promotion of Science (JSPS), Grant Number 19K11856 and 21H03504.

References

Balke, A.; and Pearl, J. 1997. Bounds on Treatment Effects From Studies With Imperfect Compliance. *J. Amer. Statist. Assoc.*, 92(439): 1171–1176.

Becker, B.; and Kohavi, R. 1996. Adult. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5XW20.

Cai, Z.; Kuroki, M.; Pearl, J.; and Tian, J. 2008. Bounds on Direct Effects in the Presence of Confounded Intermediate Variables. *Biometrics*, 64(3): 695–701.

Hamilton, E. 2017. *Benchmarking Four Approaches to Fairness-Aware Machine Learning*. Ph.D. thesis, Haverford College. Department of Computer Science.

Imai, K.; Keele, L.; Tingley, D.; and Yamamoto, T. 2011. Unpacking the Black Box of Causality: Learning about Causal Mechanisms from Experimental and Observational Studies. *American Political Science Review*, 105(4): 765–789.

Kamishima, T.; Akaho, S.; Asoh, H.; and Sakuma, J. 2012. Fairness-Aware Classifier with Prejudice Remover Regularizer. In Flach, P. A.; De Bie, T.; and Cristianini, N., eds., *Machine Learning and Knowledge Discovery in Databases*, 35–50. Berlin, Heidelberg: Springer Berlin Heidelberg.

Kilbertus, N.; Rojas-Carulla, M.; Parascandolo, G.; Hardt, M.; Janzing, D.; and Schölkopf, B. 2017. Avoiding Discrimination through Causal Reasoning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, 656–666. Red Hook, NY, USA: Curran Associates Inc.

Kusner, M. J.; Loftus, J.; Russell, C.; and Silva, R. 2017. Counterfactual Fairness. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 30*, NIPS '17, 4066–4076. Curran Associates, Inc.

Pearl, J. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann.

Pearl, J. 2001. Direct and Indirect Effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, 411–420. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Pearl, J. 2009. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2nd edition.

Plecko, D.; and Bareinboim, E. 2022. Causal Fairness Analysis. Technical Report R-90, Causal Artificial Intelligence Lab, Columbia University.

Prentice, R. L. 1989. Surrogate endpoints in clinical trials: Definition and operational criteria. *Statistics in Medicine*, 8(4): 431–440.

Toreini, E.; Aitken, M.; Coopamootoo, K.; Elliott, K.; Zelaya, C. G.; and van Moorsel, A. 2020. The Relationship between Trust in AI and Trustworthy Machine Learning Technologies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, 272–283. New York, NY, USA: Association for Computing Machinery.

Wang, Y.; and Taylor, J. M. G. 2002. A Measure of the Proportion of Treatment Effect Explained by a Surrogate Marker. *Biometrics*, 58(4): 803–812.

Zafar, M. B.; Valera, I.; Gomez Rodriguez, M.; and Gummadi, K. P. 2017. Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, 1171–1180. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee.

Zhang, J.; and Bareinboim, E. 2018. Fairness in Decision-Making — The Causal Explanation Formula.

Proceedings of the AAAI Conference on Artificial Intelligence, 32(1).

Žliobaitė, I. 2017. Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery*, 31: 1060–1089.