Assessing the Fairness of Classifiers with Collider Bias

Zhenlong Xu
Univsersity of South Austrlia
zhenlong.xu@mymail.unisa.edu.au

Jixue Liu
Univsersity of South Austrlia
jixue.liu@unisa.edu.au

Debo Cheng
Univsersity of South Austrlia
debo.cheng@mymail.unisa.edu.au

Jiuyong Li
Univsersity of South Austrlia
jiuyong.li@unisa.edu.au

Lin Liu
Univsersity of South Austrlia
lin.liu@unisa.edu.au

Ke Wang
Simon Frasier University
ke_wang@sfu.ca

Abstract

The increasing maturity of machine learning technologies and their applications to decisions relate to everyday decision making have brought concerns about the fairness of the decisions. However, current fairness assessment systems often suffer from collider bias, which leads to a spurious association between the protected attribute and the outcomes. To achieve fairness evaluation on prediction models at the individual level, in this paper, we develop the causality-based theorems to support the use of direct causal effect estimation for fairness assessment on a given a classifier without access to original training data. Based on the theorems, an unbiased situation test method is presented to assess individual fairness of predictions by a classifier, through the elimination of the impact of collider bias of the classifier on the fairness assessment. Extensive experiments have been performed on synthetic and real-world data to evaluate the performance of the proposed method. The experimental results show that the proposed method reduces bias significantly.

CCS Concepts

• **Computing methodologies → Machine learning algorithms**.

Keywords

Individual Fairness, Collider Bias, Causal Graph, Causal Effect Estimation, Situation Test

Citation Format:
Zhenlong Xu, Jixue Liu, Debo Cheng, Jiuyong Li, Lin Liu, and Ke Wang. 2020. Assessing the Fairness of Classifiers with Collider Bias. In * ** *(***)*. 9 pages.

1 Introduction

With the widespread use of machine learning for decision making in various applications such as job hiring, credit scoring, and home loan, there are increasing concerns over the fairness of decisions made by machine learning algorithms. Achieving fairness in machine learning is a non-trivial and challenging task. Unfairness/discrimination can be inadvertently inserted into machine learning models in several ways. Moreover, machine learning systems have become more and more complex, and they are commonly used in a black-box fashion. All these make the investigation of the fairness of such systems very difficult.

We consider the problem of discriminating against some individuals by machine learning algorithms. Unfair or discriminatory machine learning based decisions come in many forms with different degrees of consequences. For example, it can vary from simple misattribution of female authorship to males by machine learning language translation solution to unfair judgments by counts [2] due to racially-biased recidivism predictions.

In this paper, we consider a new problem of detecting and assessing the fairness of a decision system (a classifier) without accessing the training data. The problem is practical since a regulatory organization may access the decision system of a private company but does not have access to the data used by the company for building the system due to the confidentiality of the data. To our best knowledge, there is no existing work addressing this problem.

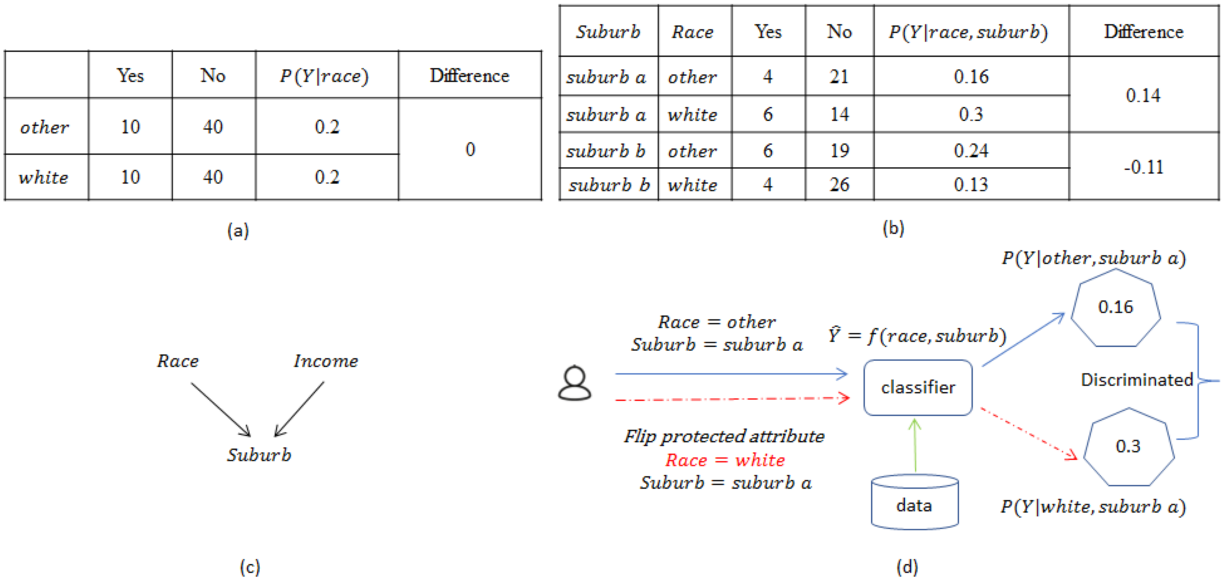
“Counterfactual reasoning" can be used for detecting discrimination against an individual with a protected value, e.g., female. One can flip the value to the opposite, e.g., male, then the same record, but with the flipped protected value is input to the classifier to obtain a decision. If two decisions are different, then there is possible discrimination. This is a typical situation test [3]. However, we will point out in the following that such a situation test on classifiers is unsound and may fail due to collider bias.

Let us consider a simple decision system used by a company to determine an employee's salary. We assume the causal relations between the three variables *Race*, *Suburb* and *Salary* as those shown in Figure 1(c), where *Race* is independent of *Salary*, indicating that there is no racial discrimination in salary payment. In the causal graph, *Suburb* is known as a *collider* since the two causal links from *Race* and *Salary* “collide" at *Suburb*. Other variables are not shown in the graph, as they are irrelevant to the discussion here. Figure 1 shows the probabilities provided by a classifier for a person to receive a high salary (Y= yes) and the corresponding summaries of data, without given suburbs (Figure1(a)) and given suburb (Figure1(b)), respectively.

Now we assess the fairness of the classifier, assuming that we only have access to the probabilities and have no access to the training data. From Figure 1(a), when suburbs are not given, if we flip the value of the protected attribute Race, the probability is the same. This indicates that the classifier is fair. However, from Figure 1(b), in a given suburb, when flipping the value of *Race*, the probability becomes different, indicating that the classifier is unfair. Figure 1(d) demonstrates the whole process.

The above example has demonstrated that collider bias brings spurious fairness evaluation. This is because, as seen from the causal graph in Figure 1(c), without given or conditioning on the collider, *Race* and *Salary* are independent, but when conditioning on *Suburb*,

Authors' addresses: Z. Xu, J. Liu, D. Cheng, J. Li and L. Liu, UniSA STEM, University of South Australia, Adelaide, 5095, Australia; K. Wang, Simon Frasier University, British Columbia, Canada.



	Yes	No	$P(Y\|race)$	Difference
other	10	40	0.2	0
white	10	40	0.2	

(a)

Suburb	*Race*	Yes	No	$P(Y\|race, suburb)$	Difference
suburb a	*other*	4	21	0.16	0.14
suburb a	*white*	6	14	0.3	
suburb b	*other*	6	19	0.24	-0.11
suburb b	*white*	4	26	0.13	

(b)

(c)

(d)

Figure 1: An example for collider bias on high salary prediction. 1(a): Data view on *Race* only. 1(b): Data view on *Race* and *Suburb*. 1(c): The relationships between *Race*, *Salary* and *Suburb*. 1(d): An example of biased situation test using a classifier.

a spurious association between *Race* and *Salary* is introduced. In practice, colliders are often used by a classifier due to their positive contribution to the accuracy of predictions. However, for fairness assessment, we must eliminate assessment bias caused by colliders. When we do not have access to training data, removing collider bias in fairness assessment is challenging. Currently, there is no work on removing collider bias for achieving sound fairness assessment.

In this paper, we propose a causality-based method UST (Unbiased situation test) for sound fairness assessment. The main contributions of the work are summarized as follows:

- We investigate the problem of potential collider biases in the situation test on individual fairness evaluation. To our best knowledge, this is the first work to reveal the impact of collider bias on fairness evaluation. Some previous work has discussed spurious associations in discrimination detection but not explicitly indicates the source of spurious associations.
- To achieve a fairness assessment at the individual level, we develop the causality-based theorems to support unbiased estimation of the direct causal effect of a protected attribute on the outcome.
- We propose an unbiased situation test method to correct collider biases based on the developed theorems. The extensive experiments on synthetic and real-world data show that the developed method can effectively reduce the impact of the collider bias in assessing fairness at the individual level.

The rest of the paper is organized as follows. Section 2 reviews the background of causal inference. In Section 3, we present a formal definition of collider bias in situation test. Section 4 gives the theoretical solutions to discrimination score estimation. Section 5 presents the developed UST (Unbiased situation test) algorithm. Section 6 provides the experimental results. In Section 7, we review related work , and we conclude the paper in Section 8.

2 Background for Causal Inference

In this section, we present the necessary background of causal inference. In the presentation, we use upper case letters to represent attributes and bold-faced upper case letters to denote sets of attributes. The values of attributes are represented using lower case letters.

Let $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ be a graph, where $\mathbf{V} = \{V_1, \ldots, V_p\}$ is the set of nodes and $\mathbf{E}$ is the set of edges between the nodes, i.e. $\mathbf{E} \subseteq \mathbf{V} \times \mathbf{V}$. A path π from V_s to V_e is a sequence of distinct nodes $< V_s, \ldots, V_e >$ such that every pair of successive nodes are adjacent in $\mathcal{G}$. A path π is a directed path if all edges along the path are directed edges. In $\mathcal{G}$, if there exists $V_i \longrightarrow V_j$, V_i is a parent of V_j and we use $\mathrm{PA}(V_j)$ to denote the set of all parents of V_j. In a directed path π, V_i is an ancestor of V_j and V_j is a descendant of V_i if all arrows point to V_j. The sets of ancestors and descendants of V_i are denoted as $An(V_i)$ and $De(V_i)$, respectively. Given a path π, V_k is a collider node on π if there are two edges incident like $V_i \longrightarrow V_k \longleftarrow V_j$.

A DAG is a directed graph without directed cycles. When a DAG satisfies the following Markov condition and faithfulness assumption, we can read the dependency or independency relationships of a distribution from the DAG.

Definition 2.1 (Markov condition [18]). Given a DAG $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ and $P(\mathbf{V})$, the joint probability distribution of $\mathbf{V}$, $\mathcal{G}$ satisfies the Markov condition if for $\forall V_i \in \mathbf{V}$, V_i probabilistically independent of all non-descendants of V_i, given the parents of V_i.

Under the Markov condition, the joint distribution of $P(\mathbf{V})$ can be factorized into: $P(\mathbf{V}) = \prod_i P(V_i | \mathrm{PA}(V_i))$. To conduct causal inference with graphical models, we often make the following faithfulness and causal sufficiency assumptions.

Definition 2.2 (Faithfulness [21]). A DAG $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ is faithful to $P(\mathbf{V})$ iff every independence presenting in $P(\mathbf{V})$ is entailed by $\mathcal{G}$ and fulfills the Markov condition. A distribution $P(\mathbf{V})$ is faithful to a DAG $\mathcal{G}$ iff there exists a DAG $\mathcal{G}$ which is faithful to $P(\mathbf{V})$.

Definition 2.3 (Causal sufficiency [21]). A dataset satisfies causal sufficiency if for every pair of variables (V_i, V_j) in $\mathbf{V}$, all their common causes are also in $\mathbf{V}$.

A causal DAG is a DAG in which a node's parents are interpreted as to its direct causes. The d-separation criterion [18] determines all independencies entailed by the Markov condition in a causal DAG.

Definition 2.4 (d-separation [18]). A path π in a causal DAG is said to be d-separated (or blocked) by a set of nodes $\mathbf{Z}$ if and only if (1) π contains a chain $V_i \rightarrow V_k \rightarrow V_j$ and a fork $V_i \leftarrow V_k \rightarrow V_j$ node such that the middle node V_k is in $\mathbf{Z}$, or (2) π contains a collider V_k such that V_k is not in $\mathbf{Z}$ and such that no descendant of V_k is in $\mathbf{Z}$.

Definition 2.5 (Total Average casual effect). The total Average Causal Effect of a treatment, denoted as A on the outcome of interest, denoted as Y is defined as $ACE(A, Y) = \mathbb{E}(Y|do\ (A = 1)) - \mathbb{E}(Y|do\ (A = 0))$, where $do()$ is the do-operator and $do(A = a)$ represents the manipulation of A by setting its value to a [18].

The do-operator can be interpreted as an intervention that modifies a select set of functions in the underlying model. The set of inference rules that emerge from this interpretation will be called as do-calculus. The inference rules of do-calculus are used in the proof of the theorems in section 4 and introduced as follows.

For a DAG $\mathcal{G}$ and a subset of nodes $\mathbf{X}$ in $\mathcal{G}$, $\mathcal{G}_{\overline{\mathbf{X}}}$ represents the DAG obtained by deleting from $\mathcal{G}$ all arrows pointing to nodes in $\mathbf{X}$. We let $\mathcal{G}_{\underline{\mathbf{X}}}$ denotes the DAG by deleting from $\mathcal{G}$ all arrows from nodes in $\mathbf{X}$. The following theorem states the inference rules of the do-calculus [18].

Theorem 2.6 (Rules of do-Calculus [18]). *Let $\mathcal{G}$ be the DAG associated with a causal model. For any disjoint subsets of nodes* $\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \mathbf{W}$, *the following rules hold, where* $\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{w}$ *are the shorthands of* $\mathbf{X} = \mathbf{x}, \mathbf{Y} = \mathbf{y}, \mathbf{Z} = \mathbf{z}$ *and* $\mathbf{W} = \mathbf{w}$ *respectively.*

Rule 1. (Insertion/deletion of observations):
$P(\mathbf{y}|do(\mathbf{x}), \mathbf{z}, \mathbf{w}) = P(\mathbf{y}|do(\mathbf{x}), \mathbf{w})$ if $(\mathbf{Y} \perp\!\!\!\perp \mathbf{Z}|\mathbf{X}, \mathbf{W})$ in $\mathcal{G}_{\overline{\mathbf{X}}}$.
Rule 2. (Action/observation exchange):
$P(\mathbf{y}|do(\mathbf{x}), do(\mathbf{z}), \mathbf{w}) = P(\mathbf{y}|do(\mathbf{x}), \mathbf{z}, \mathbf{w})$ if $(\mathbf{Y} \perp\!\!\!\perp \mathbf{Z}|\mathbf{X}, \mathbf{W})$ in $\mathcal{G}_{\overline{\mathbf{X}}\underline{\mathbf{Z}}}$.
Rule 3. (Insertion/deletion of actions):
$P(\mathbf{y}|do(\mathbf{x}), do(\mathbf{z}), \mathbf{w}) = P(\mathbf{y}|do(\mathbf{x}), \mathbf{w})$ if $(\mathbf{Y} \perp\!\!\!\perp \mathbf{Z}|\mathbf{X}, \mathbf{W})$ in $\mathcal{G}_{\overline{\mathbf{X}\mathbf{Z}(\mathbf{W})}}$,
where $\mathbf{Z}(\mathbf{W})$ is the nodes in $\mathbf{Z}$ that are not ancestors of any node in $\mathbf{W}$ in $\mathcal{G}_{\overline{\mathbf{X}}}$.

In graphical causal modeling, the back-door criterion is the main principle for identifying a set of adjustment variables (denoted as $\mathbf{Z}$) to obtain an unbiased estimation of causal effect from observational data.

Definition 2.7 (Back-door criterion in DAG). A set of variables $\mathbf{Z}$ satisfies the back-door criterion relative to (A, Y) in a causal DAG $\mathcal{G}$ if (1) $\mathbf{Z}$ does not contain only descendants of A; (2) $\mathbf{Z}$ blocks every back-door path between A and Y (i.e., the paths with an arrow pointing to A).

If we find an adjustment set $\mathbf{Z}$ from the observational data, the average causal effect $ACE(A, Y)$ can be calculated unbiasedly as follows:

$$\begin{aligned} ACE(A, Y) = \sum_{\mathbf{z}} [&\mathbb{E}(Y \mid a = 1, \mathbf{Z} = \mathbf{z}) \\ &- \mathbb{E}(Y \mid a = 0, \mathbf{Z} = \mathbf{z})] p(\mathbf{Z} = \mathbf{z}) \end{aligned} \tag{1}$$

3 Problem Definition

We first discuss an application scenario. A regulatory authority audits a decision support system of a company using a set of test instances collected from, e.g., complaints. The data contains a protected attribute A, other attributes $\mathbf{X}$, and a decision outcome (the response attribute) Y. The company has built the decision support system, a classifier $f()$, from the training data with the same attribute set as the authority. The authority can access the classifier $f()$ and the distribution of the training data set, but not the training data set itself because of confidentiality. The regulatory authority will use the classifier $f()$, data distribution, and the test data set to determine if $f()$ is fair.

To assess the fairness of the classifier, the regulatory authority needs regulations. We assume that the regulations are represented as a causal DAG, which determines which attributes can be used for predicting Y and shows the relationships between the attributes.

Assumption 1. *A causal DAG $\mathcal{G}$ represents the regulatory policy and relationships among variables.*

This is a reasonable assumption because, in practice, regulatory authorities have been using a DAG implicitly, if not explicitly. From the perspective of a regulatory authority, the eligible attributes for making a fair decision need to be the direct causes (i.e., parents in a causal DAG) of Y. Additionally, the effect of Y (i.e., children of Y in a causal DAG) is normally known by domain experts. The same knowledge exists for protected attribute A. Therefore a causal DAG can be easily constructed based on such knowledge. Also, a causal DAG can be built from the data [21], which can then be reviewed by domain experts and adopted by the authority.

We define the following discrimination score as the direct causal effect of A on Y [18].

Definition 3.1. (Discrimination Score for an Individual) Given a causal DAG $\mathcal{G}$, for a decision system $f()$ containing a binary protected attribute A, binary outcome Y and other attributes $\mathbf{X}$, given a causal DAG $\mathcal{G}$ representing the regulatory policy regarding $f()$ and the distribution of the data for training $f()$, the discrimination score of an individual for which $\mathbf{X} = \mathbf{x}_i$, denoted as D_i, is defined as follows where y denotes $Y = 1$.

$$\begin{aligned} DS_i =& |DCE(A, Y)_{\mathbf{X}=\mathbf{x}_i}| \\ =& |P(y|do(A = 1), do(\mathbf{X} = \mathbf{x}_i)) \\ &- P(y|do(A = 0), do(\mathbf{X} = \mathbf{x}_i))| \end{aligned} \tag{2}$$

Our discrimination score is defined by the direct causal effect. We assume that we conduct a data experiment while following the policy specified by the causal DAG. The protected attribute is manipulated (e.g., changing gender from female to male), and all other variables are set to be $\mathbf{x}_i$. The discrimination score is the change of Y as a result of manipulating A while holding the other variables' values to be $\mathbf{x}_i$.

We have the following problem definition for this paper.

Definition 3.2 (Problem definition). Given a causal DAG, $\mathcal{G}$, and the distribution of data for training a decision system, i.e. classier $f()$. Our goal is to determine if $f()$ is fair for an individual i by testing whether $DS_i > \alpha$ without access to the data, where α is a threshold specified by the regulation.

Unlike previous work (e.g., [7–9, 17]), our definition of discrimination score is based on the direct causal effect of the protected attribute A on Y, which uses the manipulated probabilities, instead of conditional probabilities. Thus the spurious association between A and Y caused by conditioning on colliders (as illustrated in Section 1) will be avoided.

However, as we do not have access to the training data of classifier$f()$, obtaining an unbiased estimation of the direct causal effect, i.e., the discrimination score is challenging. In the next section, we will develop the theorems to support unbiased estimation of the discrimination scores.

4 Estimating Discrimination Score

Before presenting our solution, we will analyze a problem with a Naïve solution. The problem is shared by most association based methods.

Definition 4.1 (Naive Situation Test (NST)). A naive situation test determines whether a decision on an individual is fair by testing a non-causal based discrimination score $NDS_i = P(y|A = 1, \mathbf{X} = \mathbf{x}_i) - P(y|A = 0, \mathbf{X} = \mathbf{x}_i)$ where y denotes $Y = 1$. The individual i has been discriminated if $NDS_i > \alpha$, where α is a threshold specified by the regulation.

The above solution is intuitively attractive in practice. However, as described below, NST is unsound when collider bias exists.

Definition 4.2 (Collider bias). Collider bias is the bias in a causal effect estimation due to the spurious association introduced by conditioning on a common child (or its descendant) of the cause and effect variables.

In a graphical term, C is a collider of A and Y in $\mathcal{G}$ if $A \longrightarrow C \longleftarrow Y$. In this case, A and Y are actually independent, but when given C, A and Y become associated. Hence the association is spurious. Conditioning on a descendant of C will produce a spurious association between A and Y too.

As the discrimination score used by NST is based on the conditional probabilities given $\mathbf{X}$, we have the

Observation 1. *NST is biased when* $\mathbf{X}$ *contains a collider of* A *and* Y *or descendant of a collider of* A *and* Y *in* $\mathcal{G}$.

The causality based discrimination score in Definition 3.1 can resolve the problem. The following theorem shows that the manipulated probability in our discrimination score is calculated based on the conditional probability given the parents of Y, thus avoiding conditioning on a collider or its descendants.

For the sake of fairness assessment, A is assumed to be a parent node of Y so we can use direct causal effect for the assessment.

Theorem 4.3. *Suppose that DAG* $\mathcal{G}$ *contains a protected attribute* A*, an outcome variable* Y *and a set of other observed variables* $\mathbf{X}$. *Causal sufficiency is satisfied by the data involved. We have* $P(y|do(A = a), do(\mathbf{X} = \mathbf{x})) = P(y|A = a, PA'(Y) = \mathbf{pa})$ *where* $\mathrm{PA}'(Y)$ *is the set of all parents of* Y *in* $\mathcal{G}$ *excluding* A.

Proof. Firstly, let $\mathbf{X} = \{\mathbf{C} \cup \mathbf{Q}\}$ where $\mathbf{C}$ contains descendant nodes of Y and $\mathbf{Q}$ contains non-descent nodes of Y. We have $P(y|do(A = a), do(\mathbf{C} = \mathbf{c}), do(\mathbf{Q} = \mathbf{q})) = P(y|do(A = a), do(\mathbf{Q} = \mathbf{q}))$. This is achieved by repeatedly using Rule 3 of Theorem 2.6. We show this by an example where $C \in \mathbf{C}$. $P(y|do(A = a), do(C = c), do(\mathbf{Q} = \mathbf{q})) = P(y|do(A = a), do(\mathbf{Q} = \mathbf{q}))$ because $Y \perp\!\!\!\perp C$ in DAG $\mathcal{G}_{\overline{A,C}}$ where the incoming edges to A and to C have been removed.

Secondly, we consider $P(y|do(A = a), do(\mathbf{Q} = \mathbf{q}))$ only. Based on the Markov condition 2.1, Y is independent of all its non-descendant nodes given its parents. Therefore, $P(y|do(A = a), do(\mathbf{Q} = \mathbf{q})) = P(y|do(A = a), do(\mathrm{PA}'(Y) = \mathbf{pa}))$.

Thirdly, we will prove $P(y|do(A = a), do(\mathrm{PA}'(Y) = \mathbf{pa})) = P(y|A = a, \mathrm{PA}'(Y) = \mathbf{pa})$. This can be achieved by repeatedly applying Rule 2 of Theorem 2.6.

Let $\mathrm{PA}(Y) = \{A, X_1, X_2, \ldots, X_k\}$.

$$
\begin{aligned}
& P(y|do(A = a), do(X_1 = x_1), do(X_2 = x_2), \ldots, do(X_k = x_k)) \\
& = P(y|A = a, do(X_1 = x_1), do(X_2 = x_2), \ldots, do(X_k = x_k)) \\
& \mathit{since}\ (Y \perp\!\!\!\perp A|X_1, X_2, \ldots, X_k)\ \mathit{in}\ \mathcal{G}_{\overline{X_1},\overline{X_2},\ldots,\overline{X_k},\underline{A}} \\
& = P(y|A = a, X_1 = x_1, do(X_2 = x_2), \ldots, do(X_k = x_k)) \\
& \mathit{since}\ (Y \perp\!\!\!\perp X_1|A, X_2, \ldots, X_k)\ \mathit{in}\ \mathcal{G}_{\overline{X_2},\ldots,\overline{X_k},\underline{X_1}} \\
& \mathit{repeat}\ (k-1)\ \mathit{times} \\
& = P(y|A = a, X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k) \\
& = P(y|A = a, \mathrm{PA}'(Y) = \mathbf{pa})
\end{aligned}
$$

Now, we get $P(y|do(A = a), do(\mathbf{X} = \mathbf{x})) = P(y|A = a, PA'(Y) = \mathbf{pa})$. □

Theorem 4.3 removes the descendant nodes of Y from the conditioning set in the conditional probabilities in discrimination score estimation, and this removes possible collider bias. Furthermore, it gives a succinct set of attributes for estimating discrimination scores.

For example, in Figure 2(a), $P(y|do(a), do(x_1, x_2, x_3, x_4)) = P(y|a, x_1, x_2)$ based on Theorem 4.3 where we use x_i for $X_i = x_i$. The discrimination score is determined by conditional probabilities on A, X_1 and X_2. Since X_3 is not used in a condition, there will be no collider bias. In our example in Figure 1 in the introduction, the parent node set is empty, and hence the discrimination score is calculated by $P(y|other) - P(y|white) = 0$. The collider bias is removed.

However, Theorem 4.3 is based on the causal sufficiency assumption, which assumes that there are no non-measured common causes in data. In real-world applications, hidden variables are unavoidable. When there are unobserved variables, when is the discrimination score estimation sound? The following theorem answers this question.

Theorem 4.4. *Suppose that DAG* $\mathcal{G}$ *contains a protected attribute* A*, an outcome variable* Y *and a set of other observed variables* $\mathbf{X}$. *Causal sufficiency is not satisfied by the data involved. Let* $\mathrm{CA}(Y)$

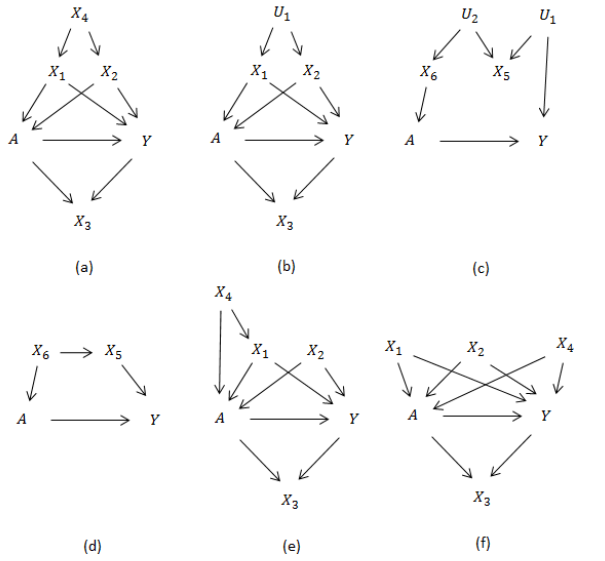


Figure 2: DAGs for the examples of theorems. X_1 **to** X_6 **are observed variables, and** U_1 **and** U_2 **are unobserved variables.**

include all the direct causes and only direct causes of Y*. We have* $P(y|do(A=a), do(\mathbf{X}=\mathbf{x})) = P(y|A=a, \mathrm{CA}(Y)=\mathbf{ca})$.

Proof. This theorem is an extension of Theorem 4.3. When there are hidden common causes and $\mathrm{PA}'(Y) = \mathrm{CA}(Y)$, 1) there is no new collider bias introduced; and 2) there will be no unblocked back-door paths.

Firstly, the edge from a direct cause into Y emits from the cause. There is not a possibility for a direct cause to be a collider when there are hidden variables. So, no collider bias will be introduced.

Secondly, there is not a possibility to form a back-door path by an edge-emitting from Y since this will violate the acyclic assumption of the DAG. Hence, all back-door door paths must run into Y and must pass the direct causes. Therefore, the set of all causes will block all back-door paths for nodes (A, Y), and no bias will be introduced when estimating the causal effect of A on Y.

Since there is no collider bias introduced, and the backdoor condition is satisfied, the estimated cause effect is unbiased. □

Theorem 4.4 indicates that discrimination detection is sound when the regulatory authority knows all the direct causes of Y and uses them as the conditioning set when calculating discrimination scores.

For example, in Figure 2(b), $P(y|do(a), do(x_1, x_2, x_3, U_1)) = P(y|a, x_1, x_2)$ based on Theorem 4.4 where we use x_i for $X_i = x_i$, if X_1 and X_2 are all direct causes of Y. All other observed or unobserved antecedent variables of Y are blocked off from Y by X_1 and X_2, and hence they do not affect the probability of Y. The unobservable variables can be in the descendant nodes of Y and A too, but they do not affect the discrimination score estimation since they will not be used anyway.

We will further explain why direct causes are necessary for Theorem 4.4. Let Figure 2(c) be a true DAG with two unobserved variables U_1 and U_2, and X_5 is not a direct cause of Y. Since U_1 and U_2 are unobserved and regulatory authority may have a DAG as Figure 2(d) where X_5 is perceived as a parent of Y. In this case, X_5 is a collider in the true DAG with the unobserved variables, and conditioning on X_5 may cause bias.

Both Theorems 4.3 and 4.4 each give a succinct conditional set for direct causal effect estimation. In fact, a superset will work as long as the superset contains the parents (or direct causes) and all or some other antecedent nodes of Y. In a causal DAG, the antecedent nodes are those having direct paths into Y. Semantically, antecedent nodes represent the direct causes and indirect causes of Y. We use set $\mathbf{B}$ to represent the antecedent nodes of Y.

Theorem 4.5. *Suppose that DAG* $\mathcal{G}$ *contains a protected attribute* A*, an outcome variable* Y*, and a set of other observed variables* $\mathbf{X}$*. Causal sufficiency is not satisfied with the data involved. Let* $\mathbf{B}$ *include all the direct causes and all or some indirect causes of* Y*. We have* $P(y|do(A=a), do(\mathbf{X}=\mathbf{x})) = P(y|A=a, \mathbf{B}=\mathbf{b})$.

Proof. The proof follows from Theorem 4.4. All direct causes have blocked the back-door paths between A and Y. When an indirect cause is considered as a parent node of Y (as a direct cause), does this lead to an unblocked path? The answer is no since the direction of the edge from the indirect cause to Y is outgoing from the indirect cause, and hence the indirect cause cannot form a collider to introduce a collider bias for (A, Y). All direct causes still block all back-door paths, and adding one or more indirect causes in the block set does not bias in direct causal effect estimation. □

Theorem 4.5 allows some redundancy in the conditional set by comparing to Theorem 4.4. In practice, the redundancy gives flexibility for users to choose the parent nodes in the DAG. Sometimes, a direct cause and an indirect cause are difficult to distinguish, and Theorem 4.5 indicates that including both does not bias the result.

For example, in Figure 2(e), $P(y|do(a), do(x_1, x_2, x_3, x_4)) = P(y|a, x_1, x_2) = P(y|a, x_1, x_2, x_4)$ based on Theorem 4.5 where we use x_i for $X_i = x_i$, if X_1 and X_2 are all direct causes of Y. Let us assume that Figure 2(e) is the true DAG, but a regulatory authority has a DAG as Figure 2(f) since they do not know which one of X_1 and X_4 is the direct cause. The discrimination score based on the imprecise DAG in Figure 2(f) is also correct as long as the all direct causes are included in the parent set of Y.

In our algorithm, we will use antecedent nodes as the conditional variable set instead of the parent nodes of Y only, and this gives some room of possible impreciseness of DAG specification in the assumption.

5 Implementing Unbiased Situation Test

Now we can summarise the discussion in Section 4 in the following unbiased situation test.

Definition 5.1. [Unbiased Situation Test (UST)] A unbiased situation test determines whether a decision on an individual i is fair by testing a discrimination score $DS_i = P(y|A=1, \mathbf{B}=\mathbf{b}_i) - P(y|A=0, \mathbf{B}=\mathbf{b}_i)$, where y denotes $Y=1$ and $\mathbf{B}$ is the set of antecedent nodes of Y in DAG $\mathcal{G}$. Individual i is discriminated if $DS_i > \alpha$, where α is a threshold specified by the regulation.

All variables in the problem can be categorized into four types $\mathbf{B}, \mathbf{C}, \mathbf{I}$ and $\mathbf{S}$, where $\mathbf{B}$ is the set of antecedent nodes of Y, $\mathbf{C}$ is the set of descendent nodes of Y, $\mathbf{I}$ are the irrelevant variables of Y, and

Algorithm 1 Unbiased Situation Test on a classifier (UST)

Input: DAG $\mathcal{G}$, classifier $f()$, training data distribution, test data set D_T, the discrimination score threshold α

Output: L, a list of discriminated individuals in D_T.

```
let $L = \emptyset$
for each $r_i \in D_T$ do
    let $r'_i$ be the record $r_i$ by flipping the value of $A$
    $P(y|r_i) \leftarrow f(r_i)$ and $P(y|r'_i) \leftarrow f(r'_i)$
    obtain $P(y|A = A(r_i), \mathbf{B} = B(r_i))$ and $P(y|A = A(r'_i), \mathbf{B} = B(r'_i))$ by Eqn 3 where $A()$ and $B()$ return values of $A$ and $\mathbf{B}$ in a record respectively
    Conduct situation test by Definition 5.1
    if $DS_i > \alpha$ then add $r_i$ to $L$
end for
return $L$
```

S are spouses of Y(A spouse of Y share a common child with Y). Irrelevant variables are independent of Y and are not considered by a classifier and are not used in discrimination detection, and hence are ignored. Spouses are associated with Y when conditioned on their common children. When the descendent nodes are removed, spouses are independent of Y and hence can be ignored.

In the following discussions, we consider $\mathbf{X} = \mathbf{B} \cup \mathbf{C}$. To conduct the unbiased situation test as defined in Definition 5.1, the problem is that we cannot get the conditional probability $P(y|a, \mathbf{B} = \mathbf{b}_i)$ directly since we do not have data. Instead, we have $P(y|a, \mathbf{B} = \mathbf{b}_i, \mathbf{C} = \mathbf{c}_i)$ from the classifier $f()$. Therefore we propose the following marginalization approach to obtain $P(y|a, \mathbf{B} = \mathbf{b}_i)$.

$$P(y|a, \mathbf{B} = \mathbf{b}_i) = \sum_{\mathbf{c}_i \in \mathbf{C}} P(y|a, \mathbf{B} = \mathbf{b}_i, \mathbf{C} = \mathbf{c}_i) * P(\mathbf{c}_i) \tag{3}$$

In Equation 3 above, while $P(y|a, \mathbf{B} = \mathbf{b}_i, \mathbf{C} = \mathbf{c}_i)$ is obtained from the classifier $f()$, $P(\mathbf{C} = \mathbf{c}_i)$ can be retrieved from the data distribution.

Then our algorithm for the proposed unbiased situation test method (known as the UST algorithm) is presented in Algorithm 1. The UST algorithm is fast with the complexity of $O(n)$, where n is the size of D_T, i.e. linear to the number of test samples, as can be seen from Algorithm 1.

6 Empirical Analysis

In this section, we evaluate the performance of the proposed UST (unbiased situation test) method on synthetic and real-world datasets, in comparison with the NST (naive situation test) method.

We use the naive situation test as the benchmark to make a fair comparative study since existing (published) fairness assessment methods require access to training data, but our UST method does not. Hence, we compare our method with the naive approach to show the bias reduction.

In the experiments, we use the popular classifiers (as $f()$ in our problem setting), including LR (Logistic Regression), SVM (Support Vector Machine), KNN (K Nearest Neighbors), NB (Naive Bayes), RF (Random Forest) and NN (Neural Network). RMSE (Root Mean Square Error) is used to measure the performance against the true measure of fairness. We also employ density plot and box plot to illustrate the difference between NST and UST.

6.1 Experiments on Synthetic Data

We create synthetic data in two steps. Firstly we follow the same procedure as in [10] to generate 500,000 records with 12 variables, including a protected attribute A, an outcome variable Y and 10 other variables, $X_1, X_2, \ldots, X_{10}$. Then in the next step, we insert a collider into the data set.

Specifically, the 10 other variables are generated according to the following specified distributions: $(X_1, X_2)(X_5, X_6) \sim \mathcal{N}((0,0)^T, ((1,0.5)^T(1,0.5)^T)$ in which $\mathcal{N}$ denotes the normal distribution and the correlation coefficients of (X_1, X_2) (X_5, X_6) are 0.5, $(X_7, X_8) \sim Bernoulli((0.5, 0.5)^T, ((1, 0.7)^T$ $(1, 0.7)^T))$ with the correlation coefficient 0.7, $X_3, X_{10} \sim Bernoulli(0.5)$ and $X_4, X_9 \sim \mathcal{N}(0, 1)$.

The true discrimination score τ is the true individual causal effect in our experiments and $\tau = \delta * (\eta_1 + \eta_2 + \eta_7)$, where $\eta_j = 2X_j$, if $X_j > 0$; otherwise $\eta_j = 0$, $j \in \{1, 2, 7\}$. And δ is a constant to control the scale of causal effect of A on Y. Then the outcome Y is generated based on τ, A and the other variables as follows: $Y_i(A) = [1 + exp((A-1) * \tau + f_Y)]^{-1}$ where $f_Y = 4 * X_1 + 2 * X_2$ $+2 * X_5 + 4 * X_6 + 4 * X_8$.

In the second step, we add collider variable into the original synthetic data, and we let $C = \alpha * Y + \beta * A + \sigma$. Here α and β are constants, and σ is random noise between $[0, 0.01]$.

For this evaluation experiment, we do not have the causal DAG specifying the regulatory policy, so we learn from the generated data (500,000 records) a causal DAG (as shown in Figure 3), by using the PC algorithm [21], a well-known causal structure learning implemented in the R package pcalg [11]. In this way, we assure the faithfulness of the learned causal DAG and the data.

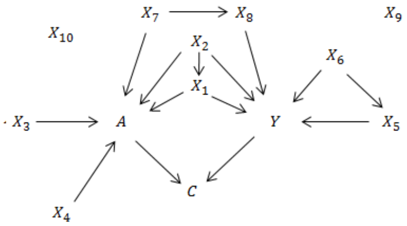


Figure 3: Causal relations in the synthetic data. A **is a protected attribute,** Y **is the outcome variableand the set of other variables** $\mathbf{X} = \{X_1, X_2, \ldots, X_{10}, C\}$ **, where** C **is a collider.**

We run our experiments with the generated datasets containing 10K, 100K and 500K samples, respectively, for each type of the classifiers, including LR, SVM, KNN, NB, RF, and NN. Each data set is split into with 70% records for training and 30% for testing.

The RMSE scores of UST and NST are shown in Table 1. Clearly, UST achieves lower RMSE than NSTconsistently across all datasets and for all classifiers.

Figure 4 presents the distributions of the ground truth discrimination scores, estimated discrimination scores by NST and UST. We see that UTS significantly outperform NTS as the distribution of discrimination score estimated by UTS is closer to the true discrimination score distribution, particularly with respect to the peak of the ground truth distribution at 0.4.

	10K		100K		500K	
	NST	UST	NST	UST	NST	UST
LR	0.404	**0.337**	0.407	**0.337**	0.407	**0.338**
SVM	0.399	**0.345**	0.462	**0.354**	0.479	**0.351**
KNN	0.399	**0.384**	0.438	**0.413**	0.435	**0.411**
NB	0.340	**0.336**	0.343	**0.339**	0.342	**0.338**
RF	0.437	**0.381**	0.454	**0.381**	0.470	**0.392**
NN	0.425	**0.357**	0.395	**0.353**	0.423	**0.362**

Table 1: Root Mean Square Error (RMSE) of of NST and UST with different classifiers on datasets with different numbers of records, 10K, 100K and 500K. The smaller error is highlighted.

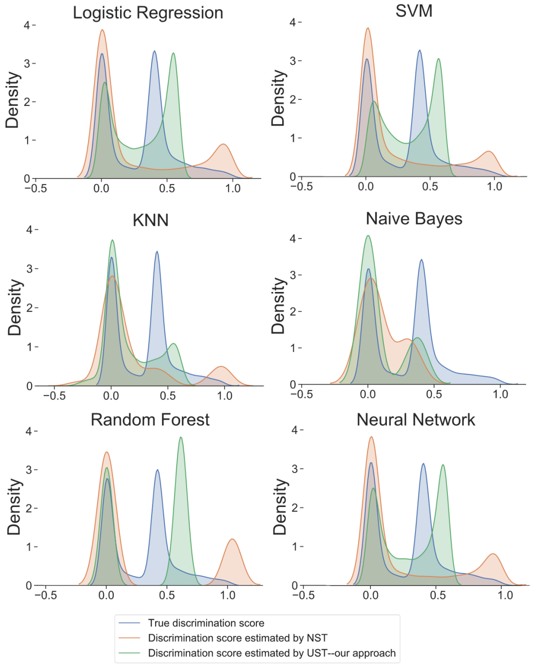


Figure 4: Comparison of the distributions of discrimination scores estimated by UST (unbiased situation test) and NST (naive situation test), w.r.t. different classifiers used, against the true discrimination score distributions. Here we use plots instead of box plots because the true discrimination score have dual peaks.

Moreover, the result in Figure 4 is consistent with the RMSE in Table 1. The closer the density distribution is, the smaller the value in the RMSE table.

6.2 Experiments on Real-world Data

We also evaluate UST using a real-world dataset: Census Adult [6], with a total of 48842 records. Because there is no causal graph given, as with the synthetic data generation, we use the PC algorithm (in the pcdag package) to learn a faithful DAG from the data, as shown in Figure 5. In the causal graph, we consider *occupation* as an outcome variable, *sex* as a protected attribute, and *salary* is a collider, and we have binarized this variable for straightforward interpretation. We note that most methods use *salary* as the outcome, but there is not a collider if the *salary* is considered as the outcome. So we use *occupation* as the outcome.

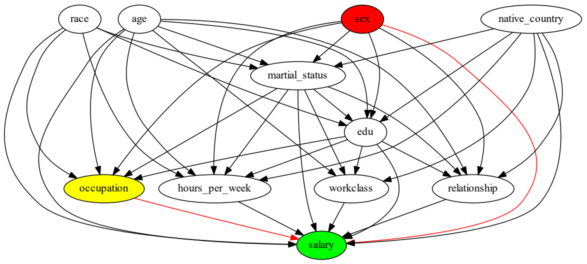


Figure 5: Causal graph for Adult data set.

	LR	SVM	KNN	NB	RF	NN
NST	0.078	0.105	0.289	0.129	0.293	0.102
UST	**0,030**	**0.097**	**0.232**	**0.083**	**0.213**	**0.095**

Table 2: Root Mean Square Error (RMSE) of NST and UST with different classifiers on the Adult data set. The smaller error is highlighted.

As no ground truth causal effects can be found from the Adult dataset, we treat the DAG in Figure 5 as the true causal graph and use the causal effects inferred from the DAG and the data set using do-calculus [18].

Unlike the synthetic data, the distribution of the discrimination score of the Adult data set approximates to a normal distribution with a single peak. for some classifiers (LR, SVM, NN), the RMSE values are small in general whereas for the other models, the RMSE values are bigger (see Table 2). However, for all models evaluated, the UST method outperforms NST with smaller RMSE.

Furthermore, from the discrimination score distribution shown in figure 6, we can see that the distributions of the discrimination scores estimated by UST are obviously closer to the distributions of the true discrimination score than NST in all experiments.

To demonstrate how UST reduces bias in individual discrimination detection, we select three individuals from the Adult data set and examine their discrimination scores estimated by UST and NST and the ground truth score. As shown in Table 3, discrimination scores estimated by NST deviate significantly from the true scores, and discrimination scores by UST are closer to the true scores.

To demonstrate the impact of collider bias on the estimation of discrimination score with a more realistic example, we use the Adult data set to build an LR classifier without using the collider variable (salary).

In this case, as expected, the NST method is effective, and as seen from Figure 7, the results of the NST are shown in Figure 7. As we can see, RMSE very small (0.034), in contrast to the RMSE when the collider is involved, which is 0.078, as shown in Table 2. This also

		NST		UST		Discrimination Score		
		$P(y\|sex = f\ or\ m, \mathbf{X} = \mathbf{x}_i)$		$P(y\|do(sex = f\ or\ m), do(\mathbf{X} = \mathbf{x}_i))$				
ID	sex	P_0	P_1	P_0	P_1	True	NST	UST
12335	f	0.682	0.627	0.561	0.606	0.071	-0.055	0.045
21479	m	0.538	0.622	0.506	0.617	0.103	0.084	0.111
30160	m	0.466	0.738	0.458	0.681	0.215	0.272	0.223

Table 3: Three examples showing fixed biases by UST.

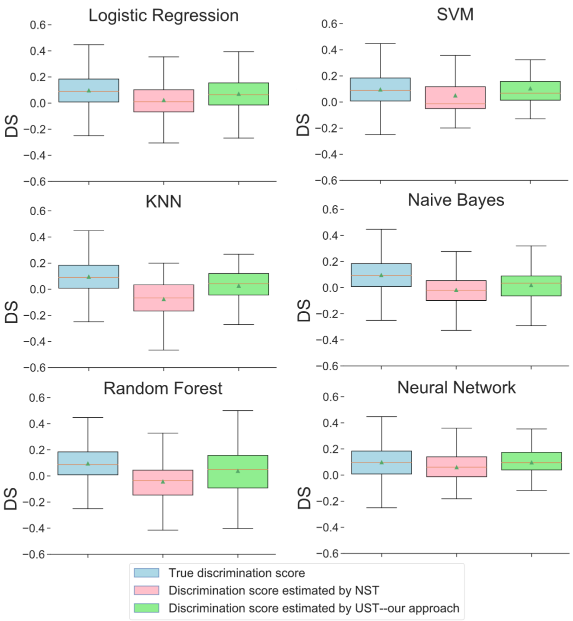


Figure 6: Comparison of the distributions of DS (discrimination scores) estimated by UST (unbiased situation test) and NST (naive situation test), w.r.t. different classifiers used, against the true discrimination score distributions. Here we use box plot because true discrimination score has Gaussian distribution.

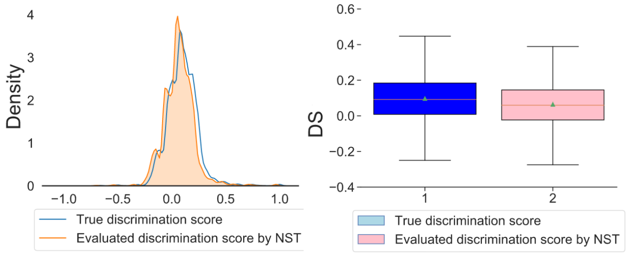


Figure 7: Comparison of the distributions of discrimination scores estimated by NST (naive situation test) on logistic regression classifier against the true discrimination score distributions. Here we use density plot and box plot. RMSE between true discrimination score and NST method is small.

verifies our proposition that collider will affect the estimation of the causal effect, and thus introduce bias in assessing the fairness of classifiers.

7 Related Work

7.1 Individual Fairness

Dwork et al. firstly used the concept of individual fairness [7] and focused on whether similar individuals are classified similarly. Individual-level fairness still is a challenging problem. Most traditional algorithms focus on group fairness [8, 9].

Some methods have been proposed to achieve individual fairness. Zemel et al. [23] utilized one intermediate transferred representation for original data to match their measurement. Luong et al. [17] adopted a variant of k-NN classification to discover and prevent unfairness. Lohia et al. [15] proposed one post-processing approach for increasing the fairness for individual and group fairness. These methods mainly come out from the data mining framework.

Recently some emerging methods have been introduced into the manifold with deep learning techniques. Madras et al. proposed a causal model [14] combined Judea Pearl's model [18] and deep latent variable model [5, 16, 19] with approximate inference. They defined a new SCM (Structural Causal Model) named Fair Causal VAE. They modeled the protected attribute as a confounder of the treatment and outcome. Speicher et al. [20] borrowed inequality indices from economics to give out one unified measurement for individual, and group unfairness decomposed the overall individual unfairness into within-group and between-group to tradeoff individual fairness.

7.2 Situation Test in Discrimination Detection

Luong et al. [17] proposed a K-NN method to find similar neighbors. The neighbors are all ranked by Manhattan distance. Then the difference can be computed by the proportion between a protected group and unprotected group. This method has obvious limitations. The discrimination score depends on the parameter K and distance function. The different parameter K and distance functions will produce completely different results. As a result, the fairness assessment may not be accurate.

To overcome the shortcoming of the K-NN method, Zhang et at.[24] proposed a causal Bayesian network based method. They defined a distance function just on the direct cause attribute of the outcome variable. They used the direct causal effect to represent the discrimination effect between protected attribute A and the outcome Y. However, they did not consider the impact of collider biases.

7.3 Unfairness Prevention

The methods for unfairness prevention can be divided into three categories based on which stage of classification they are applied to pre-processing, in-processing, and post-processing approaches.

Pre-processing methods modify the historical data with the aim that the model learned from the modified data is fairer. Given a data set X, Y, the work in [8] uses calibration to generate a new $\bar{X}$ from X so that the predictor trained on $\bar{X}, Y$ with no disparate impact. The work in [12] takes a different approach from calibration. It massages the data set by changing class labels and reweighs the data to remove unfairness.

In-processing methods add fairness constraint or modify the algorithms to achieve fairness. For example, in [22], the author proposed one mechanism for logistic regression and support vector machines with adding fairness constraints. A framework was proposed in [13] to adjust the predictive model to remove unfairness. In [4], a meta-algorithm was introduced as input a general class of fairness constraint for fairness guarantees.

Post-processing methods change unfair outcomes or predictions labels directly. In [13], the author relabelled the leaf nodes of the decision tree to reduce unfairness. Pranay K. Lohia [15] designed one bias detector for preventing algorithms from removing disparate impact for group fairness. [1] proposes a method to build a fair classifier based on the black-box classifier.

None of the methods consider collider bias.

8 Conclusions and Discussion

In this paper, We have discovered that in the situation test of the black box classifier, if the method of flipping protected attributes is used directly to predict the causal effect, it will inevitably be affected by collider variables so that the evaluation results will be biased. We have proposed relevant theoretical analysis and proofs, and based on this, we have developed a new unbiased situation test algorithm. We illustrated our approach by using synthetic and real-world data. Experimental results show that our algorithm can effectively eliminate the impact of collider variables on the causal effect of protected attributes.

Our method examines the fairness of existing classifiers from a different perspective and analyzes and resolves potential deviations in the commonly used situation test method. Due to the widespread use of the situation test method, solving this potential bias risk has practical value. Another advantage of our approach is that it can be used in the scenario when we want to assess a classifier but have no access to training data. In some practical application scenarios, the available data may be few, so our method has good application prospects.

The difficulty with our approach is that we need to use the statistical distribution information for collider variables. Such statistical information may be unavailable in some cases, and if inaccurate distribution information is used at the same time, it will affect the correctness of the method. In the future, we plan to use sampling to estimate the distribution of collider variables to solve this problem.

References

[1] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. 2018. A Reductions Approach to Fair Classification. In *International Conference on Machine Learning*. 60–69.

[2] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias risk assessments in criminal sentencing. *ProPublica* 23 (May 2016).

[3] Marc Bendick. 2007. Situation testing for employment discrimination in the United States of America. *Horizons stratégiques* (2007), 17–39.

[4] L Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K Vishnoi. 2019. Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 319–328.

[5] Irene Chen, Fredrik D Johansson, and David Sontag. 2018. Why is my classifier discriminatory?. In *Advances in Neural Information Processing Systems*. 3539–3550.

[6] Dheeru Dua and Casey Graff. 2017. *UCI Machine Learning Repository*. https://archive.ics.uci.edu/ml/datasets/Adult

[7] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. 214–226.

[8] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 259–268.

[9] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*. 3315–3323.

[10] Jenny Häggström. 2018. Data-driven confounder selection via Markov and Bayesian networks. *Biometrics* 74 (2018), 389–398.

[11] Markus Kalisch and Peter Bühlmann. 2012. Causal inference using graphical models with the R package pcalg. *Journal of Statistical Software* 47 (2012), 1–26.

[12] Faisal Kamiran and Toon Calders. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems* (2012), 1–33.

[13] Faisal Kamiran, Toon Calders, and Mykola Pechenizkiy. 2010. Discrimination aware decision tree learning. In *2010 IEEE International Conference on Data Mining*. 869–874.

[14] Jiuyong Li, Jixue Liu, Lin Liu, Thuc Duy Le, Saisai Ma, and Yizhao Han. 2017. Discrimination detection by causal effect estimation. In *2017 IEEE International Conference on Big Data (Big Data)*. 1087–1094.

[15] Pranay K Lohia, Karthikeyan Natesan Ramamurthy, Manish Bhide, Diptikalyan Saha, Kush R Varshney, and Ruchir Puri. 2019. Bias mitigation post-processing for individual and group fairness. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2847–2851.

[16] Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. 2017. Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems*. 6446–6456.

[17] Binh Thanh Luong, Salvatore Ruggieri, and Franco Turini. 2011. k-NN as an implementation of situation testing for discrimination discovery and prevention. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. 502–510.

[18] Judea Pearl. 2009. *Causality: models, reasoning and inference* (2nd. ed.). Cambridge university press.

[19] Uri Shalit, Fredrik D Johansson, and David Sontag. 2017. Estimating individual treatment effect: generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. 3076–3085.

[20] Till Speicher, Hoda Heidari, Nina Grgic-Hlaca, Krishna P Gummadi, Adish Singla, Adrian Weller, and Muhammad Bilal Zafar. 2018. A Unified Approach to Quantifying Algorithmic Unfairness: Measuring Individual and Group Unfairness via Inequality Indices. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2239–2248.

[21] Peter Spirtes, Clark N Glymour, Richard Scheines, David Heckerman, Christopher Meek, Gregory Cooper, and Thomas Richardson. 2000. *Causation, prediction, and search* (2nd. ed.). MIT press.

[22] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. 2017. Fairness Constraints: Mechanisms for Fair Classification. In *Proceedings of the 20th International Conference on Artificial Intelligence ans Statistics*, Vol. 54. 962–970.

[23] Richard Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning fair representations. In *International Conference on Machine Learning*. 325–233.

[24] Lu Zhang, Yongkai Wu, and Xintao Wu. 2016. Situation Testing-Based Discrimination Discovery: A Causal Inference Approach. In *International Joint Conferences on Artificial Intelligence(IJCAI)*, Vol. 16. 2718–2724.