

Discrimination detection by causal effect estimation

Jiuyong Li, Jixue Liu, Lin Liu, Thuc Duy Le, Saisai Ma, and Yizhao Han

University of South Australia

Adelaide, Australia

{jiuyong.li, jixue.liu, lin.liu, thuc.le, saisai.ma, yizhao.han}@unisa.edu.au

Abstract—With more and more decisions made by learnt algorithms from data, algorithmic discriminations have become a risk for civil rights. The detection of discrimination is a process of counterfactual reasoning. This paper proposes a general detection framework by combining a data mining method with a well established counterfactual reasoning method, potential outcome model. The potential outcome model supports operational definitions of global and local discriminations and discriminations by combined factors, while a data mining method makes the detection efficient. The proposed method, instantiated by association rule mining with potential outcome model based causal effect estimation, is evaluated with four real world data sets and is compared with a Bayesian network (BN) based detection method. It is able to detect not only global discriminations that are detected by the BN based method, but also local and combined discriminations that the BN based method cannot find. The proposed method is efficient, and scales well with the data set size and the number of attributes.

Index Terms—Discrimination detection; potential outcomes; association rules; causal effect;

I. INTRODUCTION

While big data technologies have transformed every aspect of our society, data driven technologies potentially cause social harms [1]. More and more decisions, such as those in health care, employment, housing, insurance, and education, have been made by computer models learnt from data to maximise a utility measure. The decisions made by these models may be discriminatory. We use an example to illustrate this. Assume that in a call centre, the queue for the service (e.g. for internet connection) is maintained by an algorithm learnt from historic data to prioritise the most profitable customers, and that one of the rules embedded in the algorithm is that the customers with phone numbers starting with a certain pattern (e.g. 833*) are the least profitable. When a customer with a phone number starting with 833 calls the centre for help, he/she will be predicted least profitable by the algorithm and his/her request will be pushed down in the queue so other more profitable callers will be served first. As a result, the customer may not receive a service if the centre receives a large number of calls on the day. The deprivation of the service from a less profitable customer by the algorithm is discriminatory.

Some data mining methods for detecting discriminations from data have been proposed. Most of them are correlation based, such as, extended lift [2], selection and contrast lift [3], olift [3], η -neutrality [4], balanced error rate [5], and a direct use of contingency table [6]. However, legally acceptable evidence of discrimination is based on counterfactual reasoning:

if the person had belonged to a different group, would the outcome have changed? Answering the counterfactual question is a process of causal inference. For example, when we assess whether females are discriminated in their payment, we will need to find out if being a female is a cause of a lower salary by unbiasedly estimating the causal effect of gender on salary.

Causal methods for discrimination detection head toward a right direction since they answer counterfactual questions. Currently emerging works using causal methods for discrimination detection [7, 8, 9, 10] are Bayesian network based methods. A causal Bayesian network is very good for visualising causal relationships, and helps define various types of discrimination cases.

A Bayesian network (BN) is structured as a directed acyclic graph where nodes denote variables and edges denote dependency relationships between nodes. A BN graphically represents the joint distribution of variables under the Markov assumption. When the causal sufficiency assumption and the faithfulness assumption [11] are made, an edge of a BN is potentially a causal relationship between two variables. A fair decision system represented by a BN should not contain an edge from a protected attributes (e.g. gender or age) to the outcome variable. In the case where such an edge exists, potential discrimination appears, and the degree of the discrimination is evaluated by estimating the causal effect of the edge [12, 13].

This BN based discrimination detection method has some limitations although it is a major step forward in discrimination detection.

Firstly, causal effect estimation using BNs has uncertainties because of equivalent classes of BNs. As the same joint distribution of a variable set can be represented by multiple BNs of the same equivalent class [11], the causal effect of the same pair of variables in different BNs is different. In addition, because of the high complexity of Bayesian network learning, many heuristics to improve efficiency for constructing BNs also result in uncertainties in learnt BNs.

Secondly, the definition of local discrimination using BN is not succinct and does not lead to efficient discovery. The location discrimination in [9] is defined on a sub data set of a context using a BN. In the discovery, constructing a BN is difficult, and constructing a number of Bayesian networks in context specific data sets is very difficult since a data set often has many possible contexts. Authors do not show how to find local discriminations in their work apart from presenting a conceptual definition.

Thirdly, it is very inefficient to find combined causes in a BN. Combined causes mean that two or more variables jointly cause an outcome. Naively, combined causes can be modelled directly by combining multiple variables to construct a new variable in BN construction and detecting the causal effect of the newly constructed variables on the outcome variable. However this will increase the number of variables exponentially, making constructing BNs impossible. Other solutions presented in [14] are still quite expensive.

This is why we propose a combination of the potential outcome model [15], a main model for causal effect estimation, and a data mining method, i.e. an association rule mining method in our instantiation, to discover discriminations in data. Our contributions are outlined as the following.

1) We propose a sound method for discrimination detection. It uses causal effect to measure discrimination. The method provides unified and succinct operational definitions for global and local discriminations by single or combined attributes. There are no other causality based methods detecting local discriminations or discriminations by combined attributes.

2) The method is efficient. The efficiency is from the use of an association rule discovery algorithm for candidate discrimination rules. The experimental results show that it is multiple orders of magnitude faster than a BN based detection method for high dimensional and large data sets.

II. PROBLEM DEFINITION

A. Notations

Given a data set \mathbf{R} containing all historical decision records of a system. Each record of the data set represents an individual and contains three types of attributes, protected, explanatory, and other attributes, in addition to the decision attribute D . The set of protected attributes $\mathbf{P}^c = \{P_1, P_2, \dots\}$, such as a Gender, Age, Race, Religion, and etc., are defined by the law and they cannot be used in decision making. The set of explanatory attributes $\mathbf{E}^c = \{E_1, E_2, \dots\}$ are supposed to be used for making decisions. The other attributes $\mathbf{O}^c = \{O_1, O_2, \dots\} = \mathbf{R} - \mathbf{P}^c - \mathbf{E}^c - \{D\}$ are the remaining attributes that are not protected and not used in decision making. The explanatory attributes and other attributes are disjoint. Users know which attributes have been used for making decisions. For example, home loan or insurance companies are required to explain to the regulatory authority the information used in decision making. For government organisations, policy transparency also requires that the factors for decision making are publicly known. Explanatory attributes may not be all used in building a decision system as in supervised learning, but they are claimed and legally allowed attributes used for making decisions. We use \mathbf{P} and \mathbf{O} to mean subsets of \mathbf{P}^c and \mathbf{O}^c respectively, including the empty set. \mathbf{PO} is a shorthand notation for $\{\mathbf{P} \cup \mathbf{O}\}$. The decision attribute D takes a value of 1 or 0 where 1 means the unfavourable decision.

For the simplicity of presentation, we assume that all attributes are binary. This assumption does not restrict the

generality of the proposed approach and the assumption affects only the matching methods [15] used in causal effect estimation of the potential outcome model [16].

B. Risk difference and bias in its estimation

Risk difference is a commonly used measure for discrimination detection. Risk difference in a population or a sub population is presented as the following: $\text{prob}(D = 1|P = 1) - \text{prob}(D = 1|P = 0)$.

If a group of people with $P = 1$ have a higher chance receiving an unfavourable decision than the other people in the population, this group is discriminated. British legislation for sex discrimination sets a threshold difference of 5% [17].

A major drawback of this criterion is the Simpson's paradox. When we observe a high risk difference in a group of people, the same high risk difference may not be observed in its subsets or supersets. For example, University of California, Berkeley was accused of discrimination against females since the admission rate of males was significantly higher than that of females in 1973. However, when examining the admission rates of all departments, the majority of departments showed a bias in favour of females [18]. Which statistics should readers believe?

We see that causal inference is a better way for detecting discriminations. Causal effect will give a unbiased estimation of the effect of a protected variable on the outcome. For example, to estimate whether the females are discriminated in their payment, a fair comparison should compare the payments of females and males if they have the same education level, the same type of positions, and the same experience in the same location, and so on. This is in contrast to the risk difference which is based on the whole population and may compare payments of people with different education levels and positions.

C. Definitions in the potential outcome framework

To detect discriminations in data, the fundamental question to answer is whether the decision will be changed if the value of a protected attribute changes. The average causal effect estimated in Rubin's potential outcome model [15] is a suitable measure to provide a quantitative answer to this question. We use a discrimination scenario to explain average causal effect.

Let P be a protected variable. Each individual has two potential outcomes corresponding to the two values of P (known as a treatment variable in the potential outcome model.). Let Y_i^1 be the potential outcome of individual i when $P = 1$ and Y_i^0 be the potential outcome of individual i when $P = 0$. Y_i^1 corresponds to our decision variable D_i . For example, let P be the gender. Y_i^1 is the outcome when the individual is male, and Y_i^0 is the outcome when the individual is female. As a person cannot be both male and female, we cannot get Y_i^1 and Y_i^0 at the same time.

Let us assume that we could observe both potential outcomes, Y_i^0 and Y_i^1 . The causal effect of P on the decision of individual i is quantified as $CE_i(P) = Y_i^1 - Y_i^0$. If $CE_i(P)$ is zero, this means that P does not cause a difference in the

decisions. If $CE_i(P)$ is larger than 0, this means that P causes a difference in the decisions and there is a potential discrimination since the same person receives different decisions just because of the difference in P .

When we aggregate the causal effects of all individuals in a target population, we obtain the average cause effect as the following. Let $E[\cdot]$ be the expectation of a random variable.

$$ACE(P) = E[Y^1|P = 1] - E[Y^0|P = 0]$$

The mean $E(\cdot)$ is over all individuals. However, if an individual is a female, only $Y^1|P = 1$ is observable, not $Y^0|P = 0$ as she cannot change to a male to be observed at the same time. In the same way, if an individual is a male, only $Y^0|P = 0$ is observable, not $Y^1|P = 1$. This leads to the situation where the mean over all the individuals is impossible. We resort some estimations for the non-observably values.

With the assumptions of $E(Y^1|P = 1) = E(Y^1|P = 0)$ and $E(Y^0|P = 1) = E(Y^0|P = 0)$ for a given strata \mathbf{e} of all explanatory variables \mathbf{E}^c , an unbiased estimation of the average causal effect of P on D in \mathbf{e} is the following [19].

$$ACE_e(P) = E(Y^1|P = 1, \mathbf{E}^c = \mathbf{e}) - E(Y^0|P = 0, \mathbf{E}^c = \mathbf{e})$$

The causal effect of the whole population is the weighted sum of the average causal effects of all strata.

$$ACE(P) = E[ACE_e(P)]$$

The above estimation is a standard and statistically sound solution for estimating causal effect by perfect stratification [19] or exact matching [16].

The above formulas use the decision values $D = 1$ and $D = 0$ to derive causal effect. An alternative and equivalent way is to use probabilities of decision values to derive causal effect, i.e. $\text{prob}(D = 1|P = 1, \mathbf{E}^c = \mathbf{e}) - \text{prob}(D = 1|P = 0, \mathbf{E}^c = \mathbf{e})$. For example, to estimate whether the females are discriminated in their payment, we estimate the causal effect of gender on salary as $ACE(\text{Gender}) = E(\text{prob}(\text{salary} = \text{low}|\text{female}, \mathbf{E}^c = \mathbf{e}) - \text{prob}(\text{salary} = \text{low}|\text{male}, \mathbf{E}^c = \mathbf{e}))$ where $\mathbf{E}^c = \mathbf{e}$ represents the same education level, the same profession, and the same experience in the same location, and so on. If causal effect $ACE(\text{Gender})$ is nonzero, the difference is caused by gender.

In a fair decision system, as long as the values for the explanatory attributes are the same, the decision should be the same regardless of the values of P . So, the expectation of $ACE(P)$ is zero.

Definition 1 (Global discrimination) *Attribute P is discriminatory if $|ACE(P)| > \alpha$ where α is a discrimination threshold.*

The global discrimination is defined on all strata of all explanatory variables. It is relatively easy to observe. In contrast, many other discriminations are local and hidden. For example, females in rural areas may be discriminated even though overall females are not discriminated.

Let us define the context based causal effect as the following.

$$\begin{aligned} ACE(P|O = o) &= E[ACE_e(P|O = o)], \text{ where} \\ ACE_e(P|O = o) &= E[Y^1|P = 1, O = o, \mathbf{E}^c = \mathbf{e}] - \\ &E[Y^0|P = 0, O = o, \mathbf{E}^c = \mathbf{e}] \end{aligned}$$

Definition 2 (Local discrimination) *Given a context $O = o$, an attribute P is discriminatory if $|ACE(P|O = o)| > \alpha$ where α is a discrimination threshold.*

In the above discussions, discriminations are defined on a single protected attribute. In some cases, the combined protected attributes cause discriminations, for example, old females are discriminated although either females or old people are not discriminated.

We define a combined protected attribute $\mathbf{P} = (P_1, P_2, \dots, P_l)$ to consist of l attributes. Given $\mathbf{p} = (p_1, p_2, \dots, p_l)$ where p_1, p_2, \dots, p_l are either 1 or 0, then the value of \mathbf{P} is defined as

$$\mathbf{P} = \begin{cases} 1 & \text{if } P_1 = p_1, P_2 = p_2, \dots, P_l = p_l \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

\mathbf{P} is also called a l -pattern.

Definition 3 (Discrimination of a combined attribute)

Let $\mathbf{P} = (P_1, P_2, \dots, P_l)$. The combined attribute \mathbf{P} is discriminatory if $|ACE(\mathbf{P}|O = o)| > \alpha$ where α is a discrimination threshold. When $O = \emptyset$, the discrimination is global and otherwise local.

Note that, the context in Definitions 2 and 3 can be a l -pattern $\mathbf{O} = (O_1, O_2, \dots, O_l) = 1$.

D. Causal effect adjustment

When there are more than one global discriminatory attributes, there might be an interaction between the attributes. For example, gender discrimination and race discrimination occur frequently together. The causal effect of gender on the decision may include some effect from race on the decision. It is good to know how the change of a discriminatory attribute P directly affects D .

In order to find direct causal effect of a single protected attribute (or a combined attribute) \mathbf{P} , we need to screen off the causal effect from other discriminatory attributes. The set of attributes for stratification in this case becomes $\mathbf{E}' = \mathbf{E}^c \cup \mathbf{P}'$ where \mathbf{P}' is the set of discriminatory attributes. The calculation of adjusted causal effect $ACE(P)$ uses attribute set $\mathbf{E}' \setminus \{P\}$ for stratification and this estimation excludes the effects of other discriminatory attributes. For the direct effect of a set \mathbf{P} of discriminatory attributes, the stratification set is $\mathbf{E}' \setminus \mathbf{P}$.

The above discussed method for causal effect estimation employs perfect stratification or exact matching of samples [16, 19]. Matching is to balance the distributions of covariates in the treatment group and control group in a data set

and to make them have similar distributions to reduce bias in causal effect estimation. A number of measures are commonly used for assessing the similarity of samples for matching. Examples are exact matching (samples having exactly same values for co-variables), Mahalanobis distance, and propensity score. There are also different techniques or procedures for matching samples based on the measures [16], and some typical examples are k : 1-nearest neighbourhood matching and subclassification (stratification).

III. ALGORITHM

The mathematical models in the previous section are not complex. However, computing them is still challenging as the possible combination of all sorts of attributes. In this section, we describe our framework of solutions. Our framework for discrimination detection from data consists of two steps. Firstly for a given data set, an association rule mining method is used to detect the signals of discriminations in it. Then the causal effects of the detected signals are estimated using the potential outcome model.

A. Candidate discriminatory rules

Definition 4 (Candidate discriminatory rules) $\mathbf{PO} = 1 \rightarrow D = 1$ is a candidate discriminatory rule if $\text{corr}(\mathbf{PO}, D) > \beta$ and $\text{prob}(\mathbf{PO} = 1, D = 1) > \gamma$ where β and γ are user defined parameters, $\text{corr}(\mathbf{PO}, D)$ stands for correlation between the combined attribute \mathbf{PO} and attribute D , and $\text{prob}(\mathbf{PO} = 1, D = 1)$ is called support of the association rule.

We use odds ratio [20] as a measure for $\text{corr}(\mathbf{PO}, D)$. \mathbf{PO} is called a pattern, and a pattern is frequent if its support is higher than γ . Note that a subset of \mathbf{PO} is also a pattern.

For the candidate discriminatory rule $\mathbf{PO} = 1 \rightarrow D = 1$, we test the following discriminatory cases.

- 1) When $\mathbf{O} = \emptyset$, \mathbf{P} is a global single (or combined) discriminatory attribute if $\text{ACE}(\mathbf{P}) > \alpha$.
- 2) When $\mathbf{O} \neq \emptyset$, \mathbf{P} is a local single (or combined) discriminatory attribute given $\mathbf{O} = 1$ when $\text{ACE}(\mathbf{P}|\mathbf{O} = 1) > \alpha$.

Definition 5 (Redundant candidate discriminatory rules)

Candidate discriminatory rule $\mathbf{P}_2\mathbf{Q}_2 = 1 \rightarrow D = 1$ is redundant if $\mathbf{P}_1\mathbf{Q}_1 = 1 \rightarrow D = 1$ is discriminatory, and $\mathbf{P}_1 \subseteq \mathbf{P}_2$ and $\mathbf{Q}_1 \subseteq \mathbf{Q}_2$.

Redundancy means that the testing of $\mathbf{P}_2\mathbf{Q}_2 = 1 \rightarrow D = 1$ is not necessary if $\mathbf{P}_1\mathbf{Q}_1 = 1 \rightarrow D = 1$ is true. This is true because of the nature of itemset supports. This property will help prune the search space in a level-wise search and evaluation process.

B. Forming strata

In Definition 2, exact matching [16] is used to unbiasedly estimate the average causal effect. The data set is firstly stratified according to values of the explanatory attributes. That

is, the data records are sorted by the values of explanatory attributes \mathbf{E}^c . The average causal effect is calculated in each stratum. The overall average causal effect is the weighted average over all strata.

C. Algorithm – DDCR

Our proposed algorithm, called Discrimination Discovery by Causal Rules (DDCR), is listed in Algorithm 1. It contains mainly two modules, global discrimination discovery and local discrimination discovery.

Lines 1 - 3 initiate variables, generate strata using explanatory variables by a quick sort algorithm in Line 3.

Lines 4 - 8 discover global discriminations for each protected attribute. It firstly tests the correlation between each protected attribute P and the decision attribute D . If the correlation is high enough, ACE is calculated. If the ACE is high enough, the attribute is a discriminatory attributes. \emptyset means that P 's discrimination does not have a context attribute.

Lines 9 - 18 discover combined discriminatory attributes and local discriminations. Firstly, frequent patterns are generated from itemised attribute sets \mathbf{P}' (the discovered global discriminatory attributes will be excluded) and \mathbf{O}^c . An itemset is like $\{P_1 = p_1, O_1 = o_1\}$ where P_1 and O_1 are attributes and p_1 and o_1 are respective values of the attributes. A pattern is an itemset. FP-growth [21] is used for the discovery.

The frequent pattern set will be tested for local discriminations including those combined attributes from 2 patterns to k_0 patterns level by level. The test is based on Definition 4. The candidates for redundant discriminatory attributes will be removed before the test.

Lines 19- 21 calculate the causal effect of each discriminatory attribute by removing the contribution from other discriminatory attributes. The discriminatory attributes are organised into the context. If each group has only one discriminatory attribute, this can be skipped.

Line 22 outputs the discovered discriminatory attributes.

D. Analysis of the algorithm

The algorithm is correct. It follows the well established potential outcome model to estimate causal effect as discussed in Section 2 without a heuristic or approximation. The search for combined discriminatory attributes and contexts of local discriminations is exhaustive given a minimum support constraint. It does not miss any signals of discriminations in a data set.

We analyse the complexity of the algorithm in three phases. In Phase I (Lines 1 - 3), the complexity is that of quick sort: $O(n \log n)$ where n is the number of records of a data set. In Phase II (Lines 4-8), the discovery of global discriminatory attributes involves the calculation of correlations and causal effects, and each calculation scans the data set once to count respective contingency table. The overall complexity is $O(n)$. In Phase III (lines 12-18), the most expensive part is finding the k_0 -frequent patterns. The time complexity is $O((|\mathbf{P}| + |\mathbf{O}|)^{k_0})$. The FP growth algorithm for

Algorithm 1 Discrimination Discovery by Causal Rules (DDCR)

Input: Data set \mathbf{R} with decision attribute D , protected attributes \mathbf{P}^c , explanatory attributes \mathbf{E}^c , and other attribute \mathbf{O}^c , discrimination threshold α , minimum odds ratio β , minimum support γ , maximum length of candidate rules k_0 .

Output: A set of discriminatory attributes (global and local, and single and combined) \mathbf{Z}

- 1: let discriminatory attribute set $\mathbf{Z} = \emptyset$
 - 2: let candidate attribute set $\mathbf{C} = \mathbf{P}^c$
 - 3: sort data set by values of \mathbf{E}^c to generate strata
 - 4: **for** each protected attribute $P \in \mathbf{C}$ **do**
 - 5: **if** $\text{OR}(P, D) \leq \beta$: next attribute
 - 6: **if** $\text{ACE}(P) > \alpha$:
 - 7: add (P, \emptyset) to \mathbf{Z} ; remove P from \mathbf{C}
 - 8: **end for**
 - 9: itemise \mathbf{P}^c and \mathbf{O}^c for pattern mining
 - 10: find frequent l -patterns using the minimum support γ which include at least one attribute in \mathbf{P} where $l \leq k_0$.
 - 11: let \mathbf{F} contain the discovered patterns sorted from the shortest to the longest
 - 12: **for** each pattern \mathbf{PO} in \mathbf{F} (the shortest first) **do**
 - 13: **if** \mathbf{PO} is redundant:
 - 14: remove \mathbf{PO} and move to the next pattern
 - 15: **if** $\text{OR}(\mathbf{PO}, D) \leq \beta$:
 - 16: remove \mathbf{PO} and move to the next pattern
 - 17: **if** $\text{ACE}(\mathbf{P}|\mathbf{O}) > \alpha$: add (\mathbf{P}, \mathbf{O}) to \mathbf{Z}
 - 18: **end for**
 - 19: **for** each group of discriminations organised by the same context **do**
 - 20: recalculate ACE when there are more than one discriminatory attribute as in Section II-D
 - 21: **end for**
 - 22: output \mathbf{Z}
-

TABLE I
DATA SETS USED IN EXPERIMENTS

Name	#Records	#attrs	Distributions
Adult	48842	14	23.9% & 76.1%
Census-income	299285	13	6.2% & 93.8%
Dutch census	60420	11	47.6% & 52.4%
Titanic	2201	6	32.3% & 67.7%

frequent pattern discovery scans the data set once. For each pattern, there is a need to scan the data set once more to work out correlation and the average causal effect. The overall complexity in this process is $O(n(|\mathbf{P}| + |\mathbf{O}|)^{k_0})$. This is also the overall complexity of the complete algorithm since the complexity of Phrases I + II is multiple orders of magnitude smaller than this.

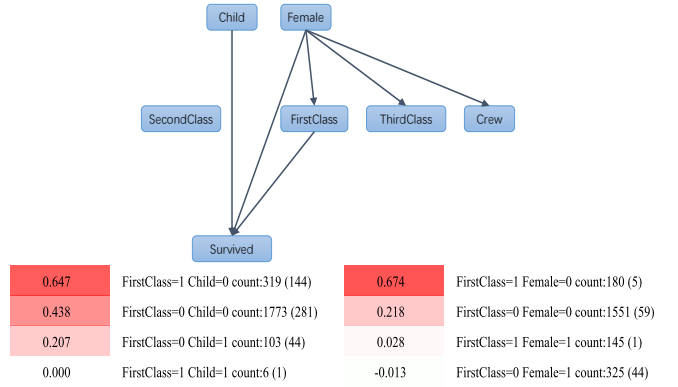


Fig. 1. Results of the causal BN based approach [9] on Titanic data set. Top: the BN learnt from the data set. Edges between the nodes in the same tier are omitted for clarity. Bottom: data strata for causal effect estimation. Values indicate risk differences in strata, and counts are represented as: strata size (the number of protected individuals in the strata). Left for attribute Female. Right for attribute Child.

IV. EXPERIMENTS

Discrimination detection is fundamentally a process of counterfactual reasoning. Causality based approaches are principled methods for this problem. In causal inference research, causal Bayesian network [12, 13] and potential outcome model [15, 22] have been widely used in various applications. Therefore, we compare the proposed method which uses the potential outcome model with a recent work based on causal Bayesian network [9].

A. Data sets and settings

To evaluate DDCR, the proposed discrimination detection method, we use three data sets used in previous discrimination detection research: the Adult, the Census Income, Dutch Census, and Titanic data sets. The Adult and the Census-Income (KDD) data sets contain the USA census data in 1994 (both available at <https://archive.ics.uci.edu/ml/datasets>, UCI Machine Learning Repository). The Dutch census contains the Dutch census data in 2001 [9]. The easily understandable data set, Titanic (<https://www.kaggle.com/c/titanic/data>), is used to illustrate our comparisons. A summary of the data sets is given in Table I. We leave interesting readers to read the attribute description following the link to our software package web page.

B. A case study on the Titanic data set

Discriminations are difficult to validate since we do not have labeled data sets as for classification. We use a data set that has known discriminations to show the power of our algorithm. We know that in the last few hours on the Titanic ship, females and children were preferably treated, and there existed discriminations against males.

We firstly show the results of the Bayes network based method [9]. The causal Bayesian network based method is a counterfactual based approach to discovering discriminations.

0.777	FirstClass=0 SecondClass=1 ThirdClass=0 Crew=0 Child=0 count:261 (93)
0.647	FirstClass=0 SecondClass=0 ThirdClass=0 Crew=1 Child=0 count:885 (23)
0.647	FirstClass=1 SecondClass=0 ThirdClass=0 Crew=0 Child=0 count:319 (144)
0.298	FirstClass=0 SecondClass=0 ThirdClass=1 Crew=0 Child=0 count:627 (165)
0.181	FirstClass=0 SecondClass=0 ThirdClass=1 Crew=0 Child=1 count:79 (31)
0.000	FirstClass=1 SecondClass=0 ThirdClass=0 Crew=0 Child=1 count:6 (1)
0.000	FirstClass=0 SecondClass=1 ThirdClass=0 Crew=0 Child=1 count:24 (13)
0.917	FirstClass=0 SecondClass=1 ThirdClass=0 Crew=0 Female=0 count:179 (11)
0.674	FirstClass=1 SecondClass=0 ThirdClass=0 Crew=0 Female=0 count:180 (5)
0.140	FirstClass=0 SecondClass=1 ThirdClass=0 Crew=0 Female=1 count:106 (13)
0.108	FirstClass=0 SecondClass=0 ThirdClass=1 Crew=0 Female=0 count:510 (48)
0.028	FirstClass=1 SecondClass=0 ThirdClass=0 Crew=0 Female=1 count:145 (1)
-0.009	FirstClass=0 SecondClass=0 ThirdClass=1 Crew=0 Female=1 count:196 (31)

Fig. 2. Results from DDCR on Titanic data set. Top: data strata for causal effect estimation for attribute Female. Bottom: data strata for causal effect estimation for attribute Child. Values indicate risk differences in strata, and counts are represented as: strata size (the number of protected individuals in the strata).

It takes two steps to identify discriminations. Firstly, it builds a Bayesian network. A candidate discriminatory attribute must be a parent node of the decision attribute. It then calculates direct causal effect between the candidate discriminatory attribute and the decision attribute. To find the direct causal effect, “path block” technique [13] has been employed to screen off causal effect of other attributes. The work in [9] shows that the set of all parents except the candidate discriminatory attribute are sufficient to form a block set for estimating direct causal effect. Then, the block set is used last for stratification, and risk differences in the strata are examined to determine discrimination.

The results of the Bayesian network approach [9] on the Titanic data set is shown in Figure 1. (1) Both the Female node and the Child node are candidates for discrimination since they are both parents of the survival node. Another parent node is First class node. (2) To test whether adults are discriminated, the data records are stratified by the First class and Child attributes. They form four strata and their risk differences are listed in Figure 1 (left). The first two strata show that the adults are discriminated. (3) To test whether males are discriminated, the data records are stratified by the First class and Female attributes. They form four strata and their risk differences are listed in Figure 1 (right). The first two strata show that males are discriminated. (4) Note that in the last two strata where Female is true, the risk differences are very close to zero and this is understandable since children and females have the same priority.

With DDCR, to test whether males are discriminated, the data records are stratified by cabin classes and crew status for causal effect estimation. The causal effect is estimated as 0.546. Similarly, the average causal effect of Child on survival is estimated as 0.287. Since both protected attributes are discriminatory, we rectify the causal effect of Female on Survival by adding the Child attribute into the stratification

attribute set. The rectified causal effect is 0.537. The rectified causal effect of Child on Survival is 0.272. To compare with Bayesian network based approach, we list the strata after the correction and their risk differences in Figure 2. Both methods give the same conclusions.

The causal effects estimated by DDCR are higher than the causal effects by the Bayesian network based approach. In this data set, we believe that our estimation is more accurate. The cabin classes and crew status are all possible information for situation that we can use for the detection and we have used them all. Note that the risk differences in the first, second and fourth strata in the upper table of Figure 2 have quite different causal effects. In the causal Bayesian network based approach, the causal effect is estimated in a single stratum (the second stratum in the bottom left table in Figure 1). When a stratum contains such heterogeneous subgroups, its average causal effect estimation likely contains a bias. This is a reason that we assert that our estimation is more accurate.

C. Compare the Bayesian network approach and DDCR

We continue comparing the DDCR method with the Bayesian network method [9] on the Adult, the Census income and the Dutch data sets.

The Bayesian networks are learned from the three data sets, where both protected attributes Gender and Race (or Marital status in Dutch data set) are candidates for discrimination since they have edges into the decision nodes. To test whether a candidate protected attribute is discriminatory, the set of other parent nodes of the decision node is used as a block set for calculate causal effect.

Both methods detect the same discriminatory attributes. The block sets for stratification in the BN approach are different from the explanatory attribute sets in DDCR. Even with the differences, the causal effects obtained from both methods are quite similar as shown in Figure 3. Gender in data sets Adult and Dutch is discriminatory, and other attributes are not discriminatory.

It is interesting to observe the close similarity of causal effect estimation of both methods although the stratification attributes are different. This makes DDCR practical since the explanatory attributes do not have to be precise. One exception is that any effect attribute of the decision attribute should not be included in the explanatory attribute set (and a block set). Such an inclusion will incur bias in causal effect estimation. We will have to rely on domain experts to sift effect attributes from causal attributes. Note that a Bayesian network could not separate causal and effect attributes either since edge orientation based on data is largely impossible. In our experiments, when we do not set tier for attributes in Bayesian network learning, most edges are undirected or bi-directed. This means that we do not know whether a node is a parent or a child of the decision node either.

D. Local discrimination and discrimination by combined attributes

DDCR can find local discriminations and discriminations by combined attributes. In Census data set, females are not



Fig. 3. Results from causal BN approach and DDCR. Left Two: results by the BN approach. Right two: results by DDCR. From the top to the bottom, Adult data set, Census income data set, and Dutch data set. Each block indicates a stratum, with a darker shade standing for a larger risk difference. The overall causal effect estimations of both methods are quite consistent.

discriminated globally since the average causal effect is 0.061. However, in private sector (context: $\text{work.Private}=1$), females are discriminated since the causal effect is 0.092, close to 0.1. Since we have only two protected attributes in each data set, we did not find discriminations of combined attributes. However, the following discovery shows the potential for such a finding. In the Dutch data set, $\text{Gender}=1$ and $\text{Marital status}=2$ have a significant higher causal effect than that of either Gender or Marital Status alone.

E. Efficiency

We compare the scalability of DDCR and the BN based approach with data set size and the number of attributes. We randomly sampled the Census Income data set into 50K, 100K, 150K, 200K and 250K for scalability study on data set size. We use the original Census Income data set, and take

each value as a binary attribute and obtain data set with 495 attributes. We randomly sampled 100K records, and 15, 20, 40, 60, 80 and 100 attributes including Gender attribute. Gender is the protected attribute for both methods. For DDCR, 10 randomly selected attributes are set as explanatory attributes. The comparisons were carried out using a desktop computer (Quad core CPU 3.4 GHz and 16 GB of memory).

DDCR is significantly faster than the BN based approach and is up to multiple orders of magnitude faster as shown in the left figure of Figure 4. This observation is consistent to their computational complexities.

BN based approach does not scale well with the number of attributes while DDCR does as shown in the middle figure of Figure 4. When the number of variables is 40, the BN based approach did not return results in two hours. The complexity of learning a Bayesian network is exponential to the number of attributes. Although some works reported building Bayesian networks with hundreds of variables, the networks are very sparse. DDCR scales well with the number of attributes.

DDCR scales well with the minimum support as shown in the right diagram of Figure 4.

V. CONCLUSION

Discrimination detection is crucial to advance civil rights in big data era. The detection of discriminations is a process of counterfactual reasoning. Bayesian network (BN) based methods have been proposed for the detection by counterfactual reasoning, but they are inefficient and not effective for local and combined discriminatory detection. This paper proposes a detection method by combining association rule mining with potential outcome model. The potential outcome model supports unified and succinct operational definitions for global and local discriminations and discrimination by combined attributes. The proposed method, DDCR, detects global discriminations as effectively as a BN based method and is also able to discover local and combined discriminations that a BN based method could not find. The method is very fast, and scales well with the data size and the number of attributes.

REFERENCES

- [1] Whitehouse, "Big data: seizing opportunities, preserving values," https://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf, 2014.
- [2] D. Pedreschi, S. Ruggieri, and F. Turini, "Discrimination-aware data mining," *ACM SIGKDD Intl. Conf. on Knowl. Disc. and Data Mining (KDD)*, 2008.
- [3] —, "Measuring discrimination in socially-sensitive decision records," *SIAM Intl. Conf. on Data Mining (SDM)*, 2009.
- [4] K. Fukuchi, J. Sakuma, and T. Kamishima, "Prediction with model-based neutrality," *Euro. Conf. on Machine Learning and Knowledge Discovery in Databases - Volume 8189*, pp. 499–514, 2013.

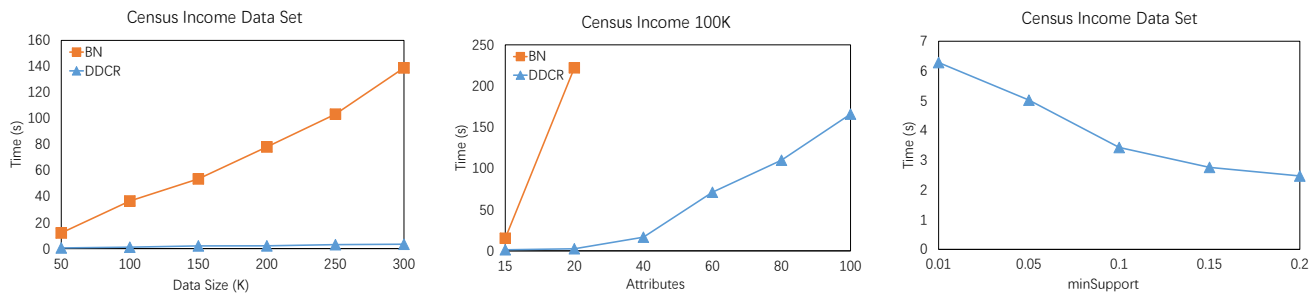


Fig. 4. Running time of DDCR with the data set size, the number of attributes and the minimum supports. The first two are compared with Bayesian network based approach.

- [5] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, “Certifying and removing disparate impact,” *ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining*, pp. 259–268, 2015.
- [6] S. Ruggieri, S. Hajian, F. Kamiran, and X. Zhang, “Anti-discrimination analysis using privacy attack strategies,” *Euro. Conf. Machine Learning and Knowledge Discovery in Databases ECML PKDD, Part II*, pp. 694–710, 2014.
- [7] K. Mancuhan and C. Clifton, “Combating discrimination using bayesian networks,” *Artificial Intelligence and Law*, vol. 22, no. 2, pp. 211–238, 2014.
- [8] L. Zhang, Y. Wu, and X. Wu, “Situation testing-based discrimination discovery: A causal inference approach,” *Intl. Joint Conf. on Artificial Intelligence*, p. 2718, 2016.
- [9] —, “On discrimination discovery using causal networks,” *Intl. Conf. on Social Computing, Behavioral-Cultural Modeling, Prediction and Behavior Representation in Modeling and Simulation*, 2016.
- [10] F. Bonchi, S. Hajian, B. Mishra, and D. Ramazzotti, “Exposing the probabilistic causal structure of discrimination,” <https://arxiv.org/abs/1510.00552>, 2015.
- [11] P. Spirtes, “Introduction to causal inference,” *Journal of Machine Learning Research*, vol. 11, no. May, pp. 1643–1662, 2010.
- [12] J. Pearl, *Causality Models, Reasoning, and Inference (2nd Edition)*. Cambridge University Press, 2009.
- [13] —, “An introduction to causal inference,” *Intl. Jnl. of Biostatistics*, vol. 6, no. 2, 2010.
- [14] S. Ma, J. Li, L. Liu, and T. D. Le, “Mining combined causes in large data sets,” *Knowl.-Based Syst*, vol. 92, pp. 104–111, 2016.
- [15] D. B. Rubin, “Causal inference using potential outcomes: Design, modeling, decisions,” *Journal of American Statistical Association*, vol. 100, no. 469, pp. 322–331, 2005.
- [16] E. A. Stuart, “Matching methods for causal inference: A review and a look forward,” *Statistical Science*, vol. 25, no. 1, pp. 1–21, 2010.
- [17] P. o. t. U. Kingdom, “Sex discrimination act 1975,” http://www.legislation.gov.uk/ukpga/1975/65/pdfs/ukpga_19750065_en.pdf, 1975.
- [18] P. J. Bickel, E. A. Hammel, and J. W. O’connell, “Sex bias in graduate admissions: data from berkeley.” *Science*, vol. 187, no. 4175, pp. 398–404, 1975.
- [19] S. L. Morgan and D. J. Harding, “Matching estimators of causal effects: Prospects and pitfalls in theory and practice,” *Sociological Methods & Research*, vol. 35, no. 1, pp. 3–60, 2006.
- [20] J. L. Fleiss, B. Levin, and M. C. Paik, *Statistical Methods for Rates and Proportions*, 3rd ed. Wiley, 2003.
- [21] J. Han, J. Pei, Y. Yin, and R. Mao, “Mining frequent patterns without candidate generation: A frequent-pattern tree approach,” *Data Mining and Knowledge Discovery*, vol. 8, pp. 53–87, 2004.
- [22] P. R. Rosenbaum and D. B. Rubin, “The central role of the propensity score in observational studies for causal effects,” *Biometrika*, vol. 70, no. 1, pp. 41–55, 1983.