On Discrimination Discovery Using Causal Networks

Lu Zhang(✉), Yongkai Wu, and Xintao Wu

University of Arkansas, Fayetteville, USA
{lz006,yw009,xintaowu}@uark.edu

Abstract. Discrimination discovery is an increasingly important task in the data mining field. The purpose of discrimination discovery is to unveil discriminatory practices on the protective attribute (e.g., gender) by analyzing the dataset of historical decision records. Different types of discrimination have been proposed in the literature. We aim to develop a framework that is able to deal with all types of discrimination. We make use of the causal networks, which effectively captures the existence of discrimination patterns and can provide quantitative evidence of discrimination in decision making. In this paper, we first propose a categorization for various discrimination. Then, we present our preliminary results on four types of discrimination, namely system-level direct discrimination, the system-level indirect discrimination, group-level discrimination, and individual level discrimination. We have conducted empirical assessments on real datasets. The results show great efficacy of our approach.

1 Introduction

Discrimination discovery has been an active research area recently [18,20]. Discrimination generally refers to an unjustified distinction of individuals based on their membership, or perceived membership, in a certain group, and often occurs when the group is treated less favorably than others. Laws and regulations disallow discrimination on several grounds, such as gender, age, marital status, sexual orientation, race, religion or belief, membership in a national minority, disability or illness, denoted as *protected attributes*. Various business models have been built around the collection and use of individual data including the above protected attributes to make important decisions like employment, credit, and insurance. Consumers have a right to learn why a decision was made against them and what information was used to make it, and whether he was fairly treated during the decision making process. Therefore, the historic data and the predictive algorithms must be carefully examined and monitored for potential discriminatory outcomes for disadvantaged groups.

Our society has endeavored to discover discrimination, however, we face several challenges. First, discrimination claims legally require plaintiffs to demonstrate a causal relationship between the challenged decision and a protected status characteristics. However, randomized experiments, which are gold-standard for causal relationship inferring in statistics, are not possible or not cost-effective

K.S. Xu et al. (Eds.): SBP-BRiMS 2016, LNCS 9708, pp. 83–93, 2016.
DOI: 10.1007/978-3-319-39931-7_9

in the context of discrimination analysis. In most cases, the causal relationship needs to be derived from the observational data not controlled experiments. Second, algorithmic decisions, which may not be directly based on protected attribute values, could still incur discrimination against the vulnerable classes of our society.

The state of the art of discrimination discovery [18,20] has developed different approaches for discovering discrimination. These approaches classify discrimination into different types including group discrimination, individual discrimination, direct and indirect discrimination. However, these work are mainly based on correlation or association-based measures which cannot be used to estimate the causal effect of the protected attributes on the decision. In addition, each of them targets one or two types of discrimination only. In real situations, several types of discrimination may present at the same time in a dataset. Thus, a single framework that is able to deal with all types of discrimination is a necessity.

We propose to investigate all types of discrimination in our research. We categorize various discrimination based on whether discrimination is across the whole system, occur in one subsystem, or happen to one individual, and whether discrimination is a direct effect or indirect effect on the decision. Then, we propose to develop a single unifying framework to capture and measure different discrimination types. We make use of the causal networks [21], which effectively captures the existence of discrimination patterns and can provide quantitative evidence of discrimination in decision making. Based on the causal networks, we present our preliminary results on system-level direct and indirect discrimination, group and individual-level discrimination. Empirical assessments for the system-level direct discrimination on two real datasets have been conducted. The results show great efficacy of our approach.

The rest of this paper is organized as follows. Section 2 presents the discrimination categorization. System-level direct and indirect discrimination is discussed in Sects. 3 and 4. Section 5 deals with group and individual-level discrimination. The experimental setup and results for system-level direct discrimination are shown in Sect. 6. Section 7 summarizes the related work. Finally, Sect. 8 concludes the paper.

2 Discrimination Categorization

We assume the historical dataset $\mathcal{D}$ contains a subset of explicitly specified protected-by-law attributes, some decision attributes, and other non-protected attributes. For ease of representation, we assume that there is only one protected attribute and one decision. We denote the protected attribute by C, associated with domain values of the protected group c^- (e.g., female) and the non-protected group c^+ (e.g., male); and denote the decision by E, associated with domain values of positive decision e^+ and negative decision e^-. Our formulation and analysis can be generalized to situations which involve multiple protected attributes and decisions.

Several types of discrimination have been proposed in the literature. In [18], discrimination is classified as group discrimination, individual discrimination,

direct and indirect discrimination. Accordingly, different types of discrimination discovery techniques have been developed, e.g., association rules for group discrimination discovery [16,17], situation testing for individual discrimination discovery [11], correlation analysis considering explanatory attributes for direct discrimination discovery [6,25], rule inference for indirect discrimination discovery [5,16], and fair classification for group/individual fairness [3].

All the above approaches are mainly based on correlation or association. In discrimination discovery, it is critical to derive causal relationship, and not merely association relationship. We need to determine what factors truly cause discrimination and not just which factors might predict discrimination. Besides, we need a unifying framework and a systematic approach for determining all types of discrimination rather than using different types of techniques for some specific types of discrimination. To this end, we first categorize various discrimination types with the following two dimensions:

- Discrimination Level. The decision making process can be modeled as a stochastic system where discrimination may happen. Discrimination can exist across the whole system, occur in one particular subsystem, or happen to one particular individual. We call them system-level discrimination, group-level discrimination, and individual-level discrimination, respectively.
- Discrimination Manner. Discrimination can be either the direct causal effect of C on E or indirect causal effect which passes the effect of C on E via some intermediate attributes. We call the former as direct discrimination and the latter indirect discrimination.

It is worth pointing out that a discrimination can combine two features mentioned above. As an example, there can be a direct discrimination at the system-level, thus forming a system-level direct discrimination.

For a quantitative measurement of discrimination, a general legal principle is to compare the proportion of positive decisions between the protected group and non-protected group [18]. The comparison can be measured by differences or rates of these proportions. In the proposed research, we will use *risk difference*, i.e., the difference in the the proportion of positive decisions between the protected group and non-protected group, as our discrimination measure. The results can be easily applied to other measures such as risk ratio, odds ratio, etc. In general, risk difference can be performed within a subpop-

Table 1. University admission: row 1 is the number of applicants and row 2 is the acceptance rate

(a) Case I				(b) Case II				(c) Case III			
Math		Biology		Math		Biology		Math		Biology	
Female	Male	Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
800	200	200	800	200	800	800	200	800	200	200	800
22 %	20 %	42 %	40 %	15 %	26 %	35 %	44 %	26 %	15 %	35 %	44 %

ulation under a partition using a subset of attributes. Formally, given a subpopulation $\mathbf{b}$ produced by a partition $\mathbf{B}$, risk difference can be denoted as $\Delta P|_{\mathbf{b}} = P(e^+|c^+, \mathbf{b}) - P(e^+|c^-, \mathbf{b})$. We say that C negatively affects E within subpopulation $\mathbf{b}$ if $\Delta P|_{\mathbf{b}} \geq \tau$, $(\tau > 0)$, where τ is a threshold for discrimination depending on the law. For instance, the 1975 British legislation for sex discrimination sets $\tau = 0.05$, namely a 5 % difference. In the following, we use an example to illustrate why it is imperative to develop this categorization framework and determine correctly the type of discrimination under investigation.

Illustrative Example. Suppose that in a university admission system, we have three attributes, the applicants' *gender*, *major applied*, and *admission decision*, and assume there are two majors, *math* and *biology*. They have different acceptance standards: the competition for *math* is more challenging than that for *biology*. Meanwhile, the choice of the major depends on the gender of an applicant, as males are more likely to apply for one major whereas females prefer the other. Table 1 shows three cases. In Case I, we see the overall admission rate is 36 % for males, but only 26 % for females. However, the claimed discrimination against the whole university may be groundless. This is because, when examining each major, no major is biased against females but to some extent in favor of females, i.e., $\Delta P|_{\{\text{math}\}} = -0.02$ and $\Delta P|_{\{\text{biology}\}} = -0.02$. In Case II, the overall admission rate of females is 31 %, which is slightly higher than that of males 30 %, showing that females and males have approximately equal chances to be accepted. However, there is clear discrimination against female, as in each major the admission rate of females is significantly lower than that of males ($\Delta P|_{\{\text{math}\}} = 0.11$ and $\Delta P|_{\{\text{biology}\}} = 0.09$). In Case III, the biases in the admission rates in the two majors are opposite, i.e., $\Delta P|_{\{\text{math}\}} = -0.11$ and $\Delta P|_{\{\text{biology}\}} = 0.09$. Hence there would be insufficient evidence to litigate the university for discrimination, since university-wide discrimination, probably because of a universal prejudice against females among admission officers, or a biased admission procedure commonly adopted by all majors, should be presented in each major of the university. As we can see, solely examining either the overall admission rates or the admission rates in any one major would lead to incorrect conclusions. On the other hand, if admission decision is made at the major level and a biased admission procedure could only be adopted by a particular major, the biology major could be litigated for discrimination against females.

The phenomenon shown in Case I and II is known as the Simpson's paradox [14], which indicates we need to consider other attributes correctly when determining and measuring the discrimination. The phenomenon shown in case III implies that system-level/group-level discrimination is a negative effect persisting in all subpopulations, given partition $\mathbf{B}$. This makes discrimination different from general causalities in that it is a persistent effect. In social and psychological sciences, three sources of discrimination are generally identified: prejudice, statistical thinking, and unintentionality [18]. All these factors can be considered as persistent across the system (for system-level discrimination) or the components within a subsystem (for group-level discrimination) and hardly change. Thus, discrimination should be considered as persistent and does not reverse or disappear under situations where the sources of discrimination are supposed to exist.

While direct discrimination is about the direct causal effect of C on E, indirect discrimination concerns about the indirect causal effects that may also be considered as discrimination. In this case, C does not necessarily have direct effect on E. Instead, it affects E via some apparently neutral attributes which are correlated with the protected attribute, hence eventually results in an unfair treatment of the protected group. Our proposed causal network based discrimination discovery framework attempts to capture and measure all types of discrimination shown in the above categorization. In this paper, we present our preliminary results for system-level direct discrimination and system-level indirect discrimination.

3 System-Level Direct Discrimination

System-level direct discrimination deals with the direct causal effect of C on E across the whole system. The direct causal effect of C on E is captured by the direct arc from C to E, i.e., $C \rightarrow E$ in the causal network. Thus, not all causal paths but only the *direct* arc may represent discrimination. As discussed above, discrimination cannot be inferred directly from the presence of the direct arc due to the intrinsic differences between discrimination and general causalities. In addition to the presence of the direct arc, we need to measure the exact causal effect carried by the arc under a correct partition $\mathbf{B}$. In order the do this, we need to suppress all other influences, some of which are spurious, some of which, although causal, can be explained by other attributes and hence are not regarded as discrimination. In other words, partition $\mathbf{B}$ must suppress influences by all other attributes. Otherwise, it cannot generate a correct and meaningful partition.

We employ the "path blocking" technique [15] to suppress all other influences. A path can be blocked by conditioning on a set of nodes not containing the two end-nodes. Upon blocked, the effect originally transmitted through the path is suppressed in each subpopulation under the partition defined by the set of nodes. If all paths other than arc $C \rightarrow E$ are blocked, all undesired influences are suppressed. We refer to the set of nodes using which we can measure the exact causal effect carried by $C \rightarrow E$ as the *block set*. As defined in [15], a node set $\mathbf{S}$ not containing C or E blocks a path p between nodes C and E if either (1) p contains at least one noncollider X in set $\mathbf{S}$, or (2) p contains at least one collider X, and X and all its descendants are outside set $\mathbf{S}$. A block set should block all paths from C to E. In addition, as we cannot measure the exact causal effect of $C \rightarrow E$ if we have the knowledge about the consequences caused by E, no E's descendant should be contained in the block set.

We consider the system-level direct discrimination a persistent effect across the system (e.g., university-wide discrimination should cause bias in each major of the university). Thus, given a block set $\mathbf{B}$, the discriminatory effect presents if $\Delta P|_{\mathbf{b}} \geq \tau$ holds for each subpopulation $\mathbf{b}$. If there are multiple block sets in a causal network, we observe that inconsistent conclusions can be drawn according to different block sets. If discrimination does exist, the discriminatory

effect must present under each correct partition. Thus, a discrimination claim is not convincing if inconsistent conclusions drawn under different partitions. Therefore, to make a discrimination claim, we need to examine all block sets and they must reach a consistent conclusion.

Based on the above analysis, we propose our discrimination criterion. Formally, we use $\mathbf{B}$ to denote a block set as well as its defined partition, and use $\mathbf{b}$ to denote each subpopulation.

Definition 1. *Discrimination is considered to present if inequality* $\Delta P|_{\mathbf{b}} > \tau$ *holds for each instance* $\mathbf{b}$ *associated with each block set* $\mathbf{B}$.

In real situations, $\Delta P|_{\mathbf{b}}$s may vary from one subpopulation to another due to randomness in the decision making process and sampling. The $\Delta P|_{\mathbf{b}}$ values of a few instances $\mathbf{b}$ could be less than τ or even negative although the majority of $\Delta P|_{\mathbf{b}}$ values are significantly greater than τ. To better captures discrimination under the context of randomness and sampling, we propose a relaxed (τ, α)-discrimination criterion, which examines whether the likelihood requirement $P(\Delta P|_{\mathbf{B}} \geq \tau) \geq \alpha$ holds. In addition, we also propose an efficient way to test the requirement in Definition 1: instead of examining each block set $\mathbf{B}$, examining one node set $\mathbf{Q}$, which is the set of E's all parents except C, i.e., $\mathbf{Q} = \mathrm{Par}(E) \backslash \{C\}$, is sufficient for guaranteeing the requirement. Please refer to our technical report [23] for details.

4 System-Level Indirect Discrimination

While direct discrimination is about the direct causal effect through $C \rightarrow E$, indirect discrimination concerns about the indirect causal effects that are transmitted through intermediate attributes along the causal paths from C to E other than the direct arc $C \rightarrow E$. These intermediate attributes are correlated with C, hence the indirect effect eventually results in an unfair treatment of the protected group. A well-known example of indirect discrimination is redlining, where the residential Zip code of the individual is used for making decisions such as granting a loan. Although the Zip code is apparently a neutral attribute, it correlates with race due to the racial makeups of certain areas. Thus, the use of the Zip code can indirectly lead to racial discrimination.

Not all indirect causal effects of C on E should be considered as indirect discrimination. From a legal perspective, the absence of indirect discrimination can be proved if the defendant can provide an objective and reasonable justification on the using of the attributes correlated with the protected attribute. Consider a loan application dataset which contains three attributes, gender C, loan status E, and income X. The causal structure $C \rightarrow X \rightarrow E$ shows being female is the actual cause of the low income and is the indirect cause of loan denial through low income. The use of attribute X can be legally justified because of an actual legitimate causal relationship of X and E, i.e., a loan is denied if the applicant has low income. The high correlation between income X and gender C may be due to the fact that the women in the dataset tend to be underpaid.

In this case, the causal effect of gender on loan denial should not be considered as discrimination.

We refer to the attributes on the causal paths whose usage cannot be legally justified as the *redlining* attributes. Formally, the causal path $C \rightarrow \cdots \rightarrow X \rightarrow \cdots \rightarrow E$, denoted as p, corresponds to an indirect discrimination if X is a redlining attribute. We propose to identify the redlining attributes by examining the relationship represented by each arc along each causal path from C to E. If any relationship cannot be legally justified, the node that emanates the arc representing an unjustified relationship is identified as the redlining attribute. We propose to measure the indirect causal effect through the paths that each contains at least one redlining attribute. Similarly, indirect discrimination can be claimed if persistent negative effects are measured.

To measure the indirect causal effect through a set of paths $\mathbf{p}$, we propose a simple three-step approach. First, we measure the direct causal effect through $C \rightarrow E$ by blocking all paths from C to E other than arc $C \rightarrow E$. Second, we measure the combined effect by blocking all paths from C to E other than $\mathbf{p}$ and arc $C \rightarrow E$. At last, the indirect causal effect through p can be identified by the difference of the above two measurements.

5 Group and Individual-Level Discrimination

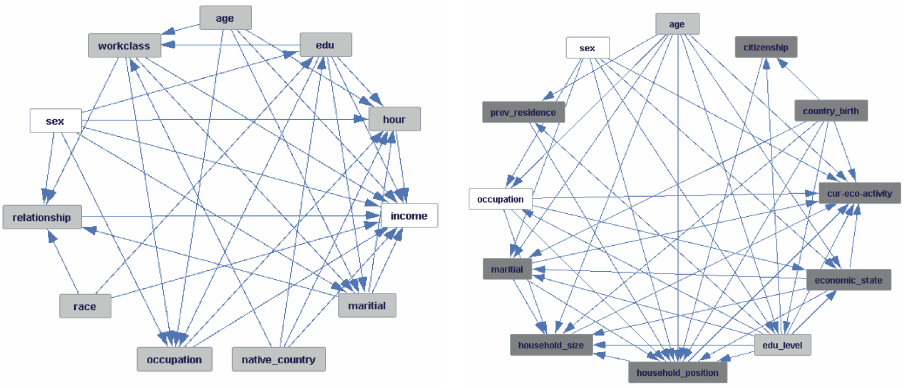
Group-level discrimination occurs in a particular subsystem other than across the whole system. The group G can be specified by analysts to denote a subsystem. It is determined by a subset of profiling attributes. For example, when we determine whether there exists group-level discrimination in a particular major (e.g., CS) in university admission, G contains all applicants in CS. When adapting discrimination discovery techniques for system-level discrimination to group-level discrimination, we should note that the determination of block set $\mathbf{B}$ needs to be adjusted based on the given group G to form a partition within the given group. For instance, when focusing on group-level discrimination in CS major, $\mathbf{B}$ may contain test scores. Then, for each test score $\mathbf{b}$, group-level discrimination can be claimed after we examine $\Delta P|_{\mathbf{b}}$ across all test scores among CS applicants.

Individual-level discrimination requires to identify discrimination for a specific individual, i.e., an entire record in the dataset. It can be considered as a special case of group-level discrimination, in which the values of all profiling attributes are given. To deal with individual-level discrimination, we propose to find two neighborhood groups that contain similar individuals from the protected group and the non-protected group. The individual is considered as discriminated if significant difference is observed between the decisions from the two groups. We propose to use the causal networks as the guideline of finding the neighborhood group. The causal structure of the system and the causal effect of each attribute on the decision can be used to facilitate the similarity measurement. Please refer to our technical report [24] for the details of our work on individual-level discrimination.

6 Experiments

We present our preliminary results for system-level direct discrimination discovery using two real data sets: the Adult dataset [10] and the Dutch Census of 2001 [13], which are widely used in discrimination discovery literature. The causal networks are constructed and presented by utilizing an open-source software TETRAD [4] which is a platform for causal modeling. We employ the original PC algorithm [21] and the significance level $\alpha = 0.01$ for network learning. The threshold τ is set as 0.05.

The Adult dataset consists of 48842 tuples with 11 attributes. Each tuple corresponds to an individual and describes the individual's personal information such as `age`, `eduation`, `sex`, `occupation`, `income`, etc. Since the computational complexity of the PC algorithm is an exponential function of the number of attributes and their domain sizes, for computational feasibility we binarize each attribute's domain values into two classes to reduce the domain sizes. For numerical attribute such as `age` or `income`, the domain values are binarized into low and high classes based on the median. For categorical attribute such as `eduation` or `occupation`, we select the domain value with the largest number of tuples as one class, and other domain values are combined as another class.



(a) Adult (b) Dutch Census

Fig. 1. Causal networks

We treat `sex` (female and male) as the protective attribute and `income` (low_income and high_income) as the decision. The causal network is shown in Fig. 1a. We observe an arc pointing from `sex` to `income`, indicating that the two attributes are causally related and further examination is required for discovering discrimination. We find all the block sets $\mathbf{B}$. Some subpopulations contain zero tuple from the dataset. On ignoring these subpopulations, the value of $\Delta P|_{\mathbf{b}}$ ranges from -0.614 to 0.524 across all the other subpopulations. Based on our criterion, we consider there is no discrimination against females in the Adult dataset.

Another dataset Dutch Census consists of 60421 tuples each of which is described by 12 attributes. Similarly, we binarize the domain values of attribute `age` due to its large domain size. We treat `sex` (female and male) as the protective attribute and `occupation` (occupation_w_low_income, occupation_w_high_income) as the decision. The causal network is shown in Fig. 1b. An arc from `sex` to `occupation` is observed in the network. We find all the block sets $\mathbf{B}$. The value of $\Delta P|_{\mathbf{b}}$ ranges from 0.062 to 0.435 across all the subpopulations, implying that females are discriminated in obtaining occupations with high income. Therefore, discrimination against females is detected in the Dutch dataset.

7 Related Work

A number of data mining techniques have been proposed to discover and measure discrimination in the literature. Pedreschi et al. [16,17] proposed to extract from the dataset classification rules, each of which consists of the protective attribute, the decision, and a set of context attributes. If the presence of the protective attribute increases the confidence of a classification rule, this classification rule is regarded as a discriminatory decision pattern in the data set. Then, discrimination can be unveiled by searching all discriminatory decision patterns. Based on that, the authors in [12] further proposed to use the Bayesian network to compute the confidence of the classification rules for detecting discrimination. Bonchi et al. in [1] proposed a random walk method based on the Suppes-Bayes causal network. Differently, conditional discrimination, where part of discrimination may be explained by other legally grounded attributes, was studied in [25]. In [22], the authors proposed the use of loglinear modeling to capture and measure discrimination and developed a method for discrimination prevention by modifying significant coefficients from the fitted loglinear model.

For individual discrimination, Luong et al. in [11] exploited the idea of situation testing. For each member of protected group with a negative decision outcome, testers with similar characteristics are searched for in a dataset. When there are significantly different decision outcomes between the testers of the protected group and the testers of the unprotected group, the negative decision can be considered as discrimination. For indirect discrimination, the authors in [5,16] studied the data mining task of discovering the attributes values that can act as a proxy to the protected groups and lead to discriminatory decisions indirectly. The authors in [19] adopted an approach based on rule inference to deal with the indirect discovery. The authors in [3] addressed the problem of fair classification that achieves both group fairness, i.e., the proportion of members in a protected group receiving positive classification is identical to the proportion in the population as a whole, and individual fairness, i.e., similar individuals should be treated similarly.

Another issue related to anti-discrimination is discrimination prevention, which aims to build non-discriminatory predictive models when the historical data contains discrimination [2,7–9]. Proposed methods focus on either modifying the historic data to remove discrimination, or tweaking the predictive model

to make it discrimination free. In all the methods, discrimination needs to be identified and measured first before it can be removed. Our work complements discrimination prevention in that we provide a formal criterion and measure for discrimination, which advances theoretical understanding related to both discrimination discovery and prevention.

8 Conclusions and Future Work

We categorize different discrimination types based on discrimination level and discrimination manner. We investigated the problem of discrimination discovery for system-level direct discrimination and indirect discrimination. We establish a discrimination models based on the causal networks. In the future work, we plan the extend the results to other types of discrimination. Using our discrimination criteria, we will also study the problem of discrimination prevention, which aim to remove discrimination by modifying the based data before conducing predictive analysis.

Acknowledgment. This work was supported in part by U.S. National Institute of Health (1R01GM103309).

References

1. Bonchi, F., Hajian, S., Mishra, B., Ramazzotti, D.: Exposing the probabilistic causal structure of discrimination. arXiv preprint (2015). arXiv:1510.00552
2. Calders, T., Verwer, S.: Three naive bayes approaches for discrimination-free classification. Data Min. Knowl. Discov. **21**(2), 277–292 (2010)
3. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness through awareness. In: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, pp. 214–226. ACM (2012)
4. Glymour, C., et al.: The TETRAD project (2004). http://www.phil.cmu.edu/tetrad
5. Hajian, S., Domingo-Ferrer, J.: A methodology for direct and indirect discrimination prevention in data mining. TKDE **25**(7), 1445–1459 (2013)
6. Hajian, S., Domingo-Ferrer, J., Monreale, A., Pedreschi, D., Giannotti, F.: Discrimination-and privacy-aware patterns. Data Min. Knowl. Discov. **29**(6), 1–50 (2014)
7. Kamiran, F., Calders, T.: Classifying without discriminating. In: International Conference on Computer, Control and Communication, pp. 1–6 (2009)
8. Kamiran, F., Calders, T.: Data preprocessing techniques for classification without discrimination. Knowl. Inf. Syst. **33**(1), 1–33 (2012)
9. Kamishima, T., Akaho, S., Sakuma, J.: Fairness-aware learning through regularization approach. In: ICDMW, pp. 643–650. IEEE (2011)
10. Lichman, M.: UCI machine learning repository (2013). http://archive.ics.uci.edu/ml
11. Luong, B.T., Ruggieri, S., Turini, F.: K-nn as an implementation of situation testing for discrimination discovery and prevention. In: KDD, pp. 502–510. ACM (2011)

12. Mancuhan, K., Clifton, C.: Combating discrimination using bayesian networks. Artif. Intell. Law **22**(2), 211–238 (2014)
13. Statistics Netherlands. Volkstelling (2001). https://sites.google.com/site/faisalkamiran/
14. Pearl, J.: Comment: understanding simpsons paradox. Am. Stat. **68**(1), 8–13 (2014)
15. Pearl, J., et al.: Causal inference in statistics: an overview. Stat. Surv. **3**, 96–146 (2009)
16. Pedreschi, D., Ruggieri, S., Turini, F.: Measuring discrimination in socially-sensitive decision records. In: SIAM SDM (2009)
17. Pedreshi, D., Ruggieri, S., Turini, F.: Discrimination-aware data mining. In: KDD, pp. 560–568. ACM (2008)
18. Romei, A., Ruggieri, S.: A multidisciplinary survey on discrimination analysis. Knowl. Eng. Rev. **29**(05), 582–638 (2014)
19. Ruggieri, S., Hajian, S., Kamiran, F., Zhang, X.: Anti-discrimination analysis using privacy attack strategies. In: Calders, T., Esposito, F., Hüllermeier, E., Meo, R. (eds.) ECML PKDD 2014, Part II. LNCS, vol. 8725, pp. 694–710. Springer, Heidelberg (2014)
20. Ruggieri, S., Pedreschi, D., Turini, F.: Data mining for discrimination discovery. ACM TKDD **4**(2), 9 (2010)
21. Spirtes, P., Glymour, C.N., Scheines, R.: Causation, Prediction, and Search, vol. 81. MIT press, Cambridge (2000)
22. Wu, Y., Wu, X.: Using loglinear model for discrimination discovery and prevention. Technical report, DPL-2015-002, University of Arkansas (2015)
23. Zhang, L., Wu, Y., Wu, X.: Causal Bayeisan network-based discrimination discovery and removal. Technical report, DPL-2016-001, University of Arkansas (2016)
24. Zhang, L., Yongkai, W., Xintao, W., Discovery, situation testing-based discrimination : a causal inference approach. Technical report, DPL-2016-002, University of Arkansas (2016)
25. Zliobaite, I., Kamiran, F., Calders, T.: Handling conditional discrimination. In: ICDM 2011, pp. 992–1001. IEEE (2011)