2016 IEEE International Conference on Data Science and Advanced Analytics

Using Loglinear Model for Discrimination Discovery and Prevention

Yongkai Wu, Xintao Wu
University of Arkansas
Email: {yw009,xintaowu}@uark.edu

***Abstract*—Discrimination discovery and prevention has received intensive attention recently. 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. However, existing discrimination discovery and prevention approaches are often limited to examining the relationship between one decision attribute and one protected attribute and do not sufficiently incorporate the effects due to other non-protected attributes. In this paper we develop a single unifying framework that aims to capture and measure discriminations between multiple decision attributes and protected attributes in addition to a set of non-protected attributes. Our approach is based on loglinear modeling. The coefficient values of the fitted loglinear model provide quantitative evidence of discrimination in decision making. The conditional independence graph derived from the fitted graphical loglinear model can be effectively used to capture the existence of discrimination patterns based on Markov properties. We further develop an algorithm to remove discrimination. The idea is modifying those significant coefficients from the fitted loglinear model and using the modified model to generate new data. Our empirical evaluation results show effectiveness of our proposed approach.**

***Keywords*-discrimination discovery; discrimination prevention; loglinear modeling;**

I. Introduction

Discrimination discovery and prevention has been an active research area recently [1]–[4]. Discrimination discovery is the data mining problem of unveiling discriminatory practices by analyzing a dataset of historical decision records and discrimination prevention aims to remove discrimination by modifying the biased data before conducting predictive analysis. 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 (e.g., the Fair Credit Reporting Act or Equal Credit Opportunity Act) 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.

The state of the art of discrimination discovery [3], [4] has developed different approaches for discovering different types of discrimination including group discrimination, individual discrimination, direct and indirect discrimination, and conditional discrimination. Proposed methods for discrimination prevention [2], [5]–[7] are either based on data preprocessing or algorithm tweaking. Preventing discrimination when training a classifier consists of balancing two contrasting objectives: maximizing accuracy of the extracted classification model and minimizing the number of predictions that are discriminatory. However, those methods are often limited to examining the relationship between one decision attribute and one protected attribute without considering other attributes included in the historical dataset. Hence it is imperative to develop approaches that can capture, measure and remove discriminations between multiple decision attributes and protected attributes given a set of non-protected attributes.

A. Summary of Our Contribution

We develop a single unifying framework that aims to capture and measure discriminations based on the use of loglinear modeling. Loglinear modeling is the statistical method to model categorical variables and their multi-way relationships through a set of estimation and modeling strategies. The fitted loglinear model represents the intrinsic interaction effects via statistical dependencies. For low dimensional data we simply build a loglinear model, examine its coefficient values, use a derived metric to effectively quantify the magnitude of the discriminations. For high dimensional data we limit our model fitting to graphical loglinear models to improve interpretability and reduce the computational cost of fitting the model. We build from the fitted graphical loglinear model a graph representing the conditional independence structures of attributes in the historic dataset. We show how to determine the existence of discrimination patterns and interpret them based on Markov properties. To our best knowledge, this is the first work to capture and quantify discriminations between multiple protected attributes and decision attributes given a set of non-protected attributes.

We solve the problem of discrimination prevention by modifying those significant coefficients from the fitted loglinear model and using the modified model to generate new data. We design strategies of effectively changing the coefficient values of the fitted model to meet different dis-

978-1-5090-5206-6/16 $31.00 © 2016 IEEE
DOI 10.1109/DSAA.2016.18

crimination requirements. The proposed approach is shown to effectively remove discrimination while preserving the utility.

II. Related Work

A. Anti-discrimination Learning

Anti-discrimination learning either focused on using data mining to discover and measure discrimination [3] or dealt with preventing discrimination when building data mining models [1], [2]. Discrimination can be classified as group discrimination, individual discrimination, direct and indirect discrimination, and conditional discrimination [4].

Group discrimination refers to discrimination against a subgroup described by a combination of multiple protected and non-protected attributes. It has been studied in [8], [9] where classification rules are extracted from a dataset of historical records and then are ranked according to some measures of discrimination. Rules with high ranks indicate possible discrimination against groups of protected-by-law. Individual discrimination requires to measure the amount of discrimination for a specific individual, i.e., an entire record in the dataset. The authors [10] exploited the idea of situation testing to discover individual discrimination. For each member of the protected group with a negative decision outcome, testers with similar characteristics are searched for in a dataset of historical decision records. 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. In [11], the authors proposed to use the causal networks as the guideline of finding the similar records. The causal structure of the underlying data and the causal effect of each attribute on the decision are used to facilitate the similarity measurement. Conditional discrimination, i.e., part of discrimination which may be explained by other legally grounded attributes, was studied in [5], [12]. The task was to evaluate to which extent the discrimination apparent for a group is explainable on a legal ground. Indirect discrimination discovery refers to 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 [1], [9], [13].

Proposed methods for discrimination prevention are either based on data preprocessing or algorithm tweaking. Data preprocessing methods [2], [5] modify the historic data to remove discriminatory patterns according to some discrimination measure before learning a prediction model. In [5], the authors analyzed the issue of conditional non-discrimination in classifier design and developed local techniques, local massaging and preferential sampling, for handling conditional discrimination, i.e., removing the bad discrimination when one of the attributes is considered to be explanatory for the discrimination. Proposed methods for discrimination prevention using model tweaking include the tweaking of decision tree [6], naive Bayes classifier [7], and logistic regression [14]. All the methods require some tweak of predictive models. For example, in [6], the authors developed a strategy for relabeling the leaf nodes of a decision tree to make it discrimination free. Preventing discrimination when training a classifier consists of balancing two contrasting objectives: maximizing accuracy of the extracted classification model and minimizing the number of predictions that are discriminatory.

The authors in [15] proposed the use of Bayesian networks to compute the confidence of the classification rules for discrimination discovery. In [16], the authors proposed the use of the causal networks to categorize various types of discrimination. The authors in [17] proposed the use of Suppes-Bayes causal network and developed several random-walk-based methods over the causal structure to detect different types of discrimination. However, the Suppes-Bayes causal network learned from a data set with h categorical attributes and s samples has m nodes where each node corresponds to a Bernoulli variable of the type $< attribute = value >$. The time complexity of the network construction algorithm is $\mathcal{O}(sm)$ and the space required is $\mathcal{O}(m^2)$, which is impractical with the large number of attribute-value pairs.

B. Loglinear Modeling

Loglinear modeling [18] is a discrete multivariate statistical technique that is designed specifically for analyzing categorical data and its derived contingency table. It is used to measure the strength of interactions among categorical attributes without conceptually distinguishing between a dependent variable and independent variables. In the data mining area, Loglinear modeling has been applied to data compression [19], multi-item associations [20], and data cube exploration [21].

For a data set with d categorical attributes $A_1, \ldots, A_d$, we use $n_{i_1 \ldots i_d}$ to denote the number of records in the cell $i_1 \ldots i_d$ where i_r denotes the r-th domain value of attribute A_r. We define the log of anticipated value $m_{i_1 \ldots i_n}$ as a linear additive function of the coefficients, γ terms, which capture contributions from various higher level group-bys.

For instance, in a 4-dimensional table with dimensions A,B,C,D, we use (i, j, k, l, n_{ijkl}) to denote the cell in a 4-D space, where $i = 1, \cdots, I$,$j = 1, \cdots, J$,$k = 1, \cdots, K$,$l = 1, \cdots, L$. Equation 1 shows the saturated loglinear model which contains the 4-factor effects, all the possible 3-factor effects, and so on up to the 1-factor effects and the mean γ. For example, γ_i^A is 1-factor effect, γ_{ij}^{AB} is 2-factor effect which shows the dependency within the distributions of the associated attributes A, B.

$$\begin{aligned} \log m_{ijkl} \quad = \quad & \gamma + \gamma_i^A + \gamma_j^B + \gamma_k^C + \gamma_l^D \\ & + \gamma_{ij}^{AB} + \gamma_{ik}^{AC} + \gamma_{il}^{AD} + \gamma_{jk}^{BC} + \gamma_{jl}^{BD} + \gamma_{kl}^{CD} \\ & + \gamma_{ijk}^{ABC} + \gamma_{ijl}^{ABD} + \gamma_{ikl}^{ACD} + \gamma_{jkl}^{BCD} \\ & + \gamma_{ijkl}^{ABCD} \end{aligned} \tag{1}$$

Equation 2 shows the linear constraints among coefficients, where a dot "." means that the parameter has been summed over the index (e.g., $\gamma_{i.}^{AB} = \sum_{j=1}^{J} \gamma_{ij}^{AB}$). In short, the constraints specify that the loglinear parameters sum to 0 over all indices.

$$\begin{aligned} \gamma_{.}^{A} = \gamma_{.}^{B} = \gamma_{.}^{C} = \gamma_{.}^{D} = 0 \\ \gamma_{i.}^{AB} = \gamma_{.j}^{AB} = \gamma_{i.}^{AC} = \gamma_{.k}^{AC} = \cdots = \gamma_{.l}^{CD} = 0 \\ \cdots \\ \gamma_{ijk.}^{ABCD} = \gamma_{ij.l}^{ABCD} = \gamma_{i.kl}^{ABCD} = \gamma_{.jkl}^{ABCD} = 0 \end{aligned} \tag{2}$$

The coefficient corresponding to a k-way effect is obtained by subtracting from the average value all the coefficients from $(k-1)$-way effects, $(k-2)$-way effects, and so on up to the 1-way effects and the grand mean. Equation 3 shows how to compute the coefficients in a 4-dimensional table.

$$\begin{aligned} \gamma &= \frac{n_{....}}{IJKL} \\ \gamma_i^A &= \frac{n_{i...}}{JKL} - \gamma \\ &\cdots \\ \gamma_{ij}^{AB} &= \frac{n_{ij..}}{KL} - \gamma_i^A - \gamma_j^B - \gamma \\ \gamma_{ijk}^{ABC} &= \frac{n_{ijk.}}{L} - \gamma_{ij}^{AB} - \gamma_{ik}^{AC} - \gamma_{jk}^{BC} - \gamma_i^A - \gamma_j^B - \gamma_k^C - \gamma \\ &\cdots \end{aligned} \tag{3}$$

where $n_{....}$ is the grand sum and $n_{i...}$ is the sum over all values along i-th member of dimension A. We can see γ_i^A denotes how much the average of the values along i-th member of dimension A differs from the overall average.

It is obvious that a large number of models can be used to fit a given data set. For an d-dimensional loglinear model, there are a total 2^{2^d} possible models (determined by which parameters of the saturated model are set to zero). Various strategies of model selection have been investigated [22].

Using the fitted Loglinear model, we can obtain not only the single-factor effects but also two-factor or higher order-factor effects, which captures multi-way interactions. Contingency tables are used in categorical data analysis to analyze expected patterns produced by various combinations of certain variable levels. The parameters of the loglinear model provide the interactions between variables. The γ_i^A parameter is expressed as a deviation from the grand-sum parameter γ; it reflects the extent to which membership in the i-th category of A is different from the average across all categories of A. A 2-way interaction between A and B indicates that a partial association exists between them. γ_{ij}^{AB} term in the model can be expressed as a deviation from the γ_i^A term. A 3-way interaction γ_{ijk}^{ABC} means that the A-B relation is not constant but changes for each category of C. The parameter of γ_{ijk}^{ABC} is derived as the difference between the average value of the A-B relation and the value exhibited at the k-th level of C. It tells how the A-B relation is enhanced, depressed, or changed in direction when compared with the average A-B relation.

It is a common practice for researchers and practitioners to restrict model selection to a subclass of loglinear models, e.g., hierarchical models or graphical models. A loglinear model is *hierarchical* if the model includes all lower-order terms composed from variables contained in a higher order term. Hierarchical models are nested models in which when an high-factor interaction is present, all interactions of lower order between the variables of that interaction are also present. For example, if a 3-way interaction (γ^{ABC}) is present, the hierarchical model must also include all 2-way interaction ($\gamma^{AB}, \gamma^{AC}, \gamma^{BC}$) as well as the single variable ($\gamma^A, \gamma^B, \gamma^C$) and the grand mean ($\gamma$).

In graphical loglinear models, the vertices correspond to variables and an edge between a pair of variables captures the conditional dependence. For instance, we add the edge A-B when γ^{AB} is contained in the fitted loglinear model. A probability model is graphical if, for every clique in its conditional independence graph, all possible dependencies implied by the clique are present. We can interpret conditional independence relationships based on the well known Markov properties.

- Pairwise Markov Property: A and B have no edge between them iff A and B are conditionally independent given all other variables.
- Global Markov Property: Two sets of variables $\mathbf{U}$ and $\mathbf{V}$ are separated by a third set of variables $\mathbf{W}$, if and only if $\mathbf{U} \perp \mathbf{V}|\mathbf{W}$.
- Local Markov Property: A is conditionally independent of its non-neighbors in the graph $\mathbf{G}$, given its neighbors $bd(A)$, i.e., $A \perp \mathbf{G} \backslash \{A \cup bd(A)\} | bd(A)$.

III. Discrimination Detection and Interpretation

Our approach is to use the loglinear model to fit the data and examine its coefficients to interpret the interaction effects among decision variable, sensitive attributes, and other explanatory attributes. Loglinear modeling can simultaneously test relationships between multiple decision variables and multiple explanatory variables. In the paper, we use bold capital letters (e.g., $\mathbf{X}$) to denote a subset of variables and use the regular calligraphic letters (e.g., X) to denote a variable of the subset. In the context of anti-discrimination learning, we assume the historical dataset contains a subset of protected-by-law attributes $\mathbf{X}$, a subset of decision attributes $\mathbf{Y}$, and other non-protected attributes $\mathbf{Z}$. Obviously, $X \in \mathbf{X}$ denotes a protected attribute and $Y \in \mathbf{Y}$ denotes an attribute recording one historical decision. We assume that protected attributes and decision attributes are explicitly specified as input of the discrimination analysis.

A. Three Dimension Case

We assume in this subsection there is one protected attribute X, one decision attribute Y, and one non-protected attribute Z in the historical dataset. The dataset can be represented as one 3-way contingency table. We use $(i, j, k, \{n_{ijk}\})$ to denote cell in a 3-D space, where n_{ijk} denotes the number of records in cell (i, j, k). The counts $\{n_{ijk}\}$ follow a multinomial distribution $M(n, \{\pi_{ijk}\})$ where π_{ijk} denotes the probability of a record in cell (i, j, k). For simplicity, we write $\pi_{ijk} = P(X = i, Y = j, Z = k)$, which represents the joint distribution of X, Y, Z. Let $m_{ijk} = E(n_{ijk})$ denote the mean where $n_{ijk} = n\pi_{ijk}$ for the multinomial distribution.

1) Detecting Discrimination Structure: To determine whether there is a discrimination, we first fit the model and then examine the coefficients of the fitted model. The fitted loglinear model represents the intrinsic interaction effects via statistical dependencies. In our discrimination analysis, our focus is on the relationship between the protected attribute X and the decision attribute Y.

Table I shows all possible hierarchical models for a 3-D dataset. A hierarchical model includes all lower-order terms composed from variables contained in a higher-order term and can be abbreviated by giving the terms of higher order. For example, the symbol (XY, XZ) denotes the model containing $\gamma^{XY}, \gamma^{XZ}, \gamma^{X}, \gamma^{Y}, \gamma^{Z}$ and γ. In Table I, No 1 shows the main effects model, No 2-4 show three models which contain only one 2-factor interaction, No 5-7 show three models which contain two 2-factor interactions, No 8 shows the all 2-factor interaction model, and No 9 shows the saturated model.

We can easily see loglinear models shown in No 1, 3, 4, and 7 do not contain γ^{XY}, which indicates the dataset does not contain discrimination. The loglinear model shown as No 1 indicates all three variables are mutually independent. The loglinear model shown in No 3 indicates decision attribute Y is jointly independent of X and Z although X and Z are dependent.

We can see γ^{XY} term is included in the loglinear models shown in No 2, 5, 6, 8 and 9. All these models indicate the existence of discrimination in the underlying data set. The loglinear model (XY, Z), shown in No 2 and the loglinear model (XY, XZ) shown in No 5 clearly indicate the discrimination due to sensitive attribute X and the third variable Z does not help explain any discrimination. The difference between No 2 and No 5 is the existence of conditional dependence between X and Z in No 5. The loglinear model (XY, YZ) shown in No 6 indicates both X and Z have their own association with Y, however X and Z are conditionally independent given Y.

The loglinear model (XY, XZ, YZ) shown in No 8 contains all 2-factor interactions and the loglinear model (XYZ) shown in No 9 contains the 3-factor interaction in addition to all 2-factor interactions. A 2-way interaction between X and Y in a model in which X and Y are involved in no higher order interactions simply means that X and Y share a relation that is constant across levels of all other variables in the table. A 3-way interaction involving X, Y and Z means that the relation between any two of these variables changes across the levels of the third.

It is worth pointing out that the loglinear model shown in No 4 indicates the decision attribute Y is dependent on the non-protected attribute Z but independent with the sensitive attribute X when Z is considered. In the loglinear model shown in No 7, the decision attribute Y is conditionally independent with the protected attribute X given the non-protected attribute Z. Conditional independence of X and Y given Z (in short $X \perp Y|Z$) holds if for all i, j, k, $P(X = i, Y = j|Z = k) = P(X = i|Z = k)P(Y = j|Z = k)$. However, for both models No 4 and 7, there may exist strong marginal association between X and Y when Z is ignored. This corresponds to the well known Simpson's paradox phenomenon [23].

2) Quantifying Magnitude of Discrimination: To quantify the discrimination's magnitude, we use the coefficient values captured in the fitted loglinear model. The coefficient values indicate the strength of the association between variables, and therefore can be used to quantify the magnitude of discrimination. Recall the values of γ^{XY} show the 2-factor interaction effects between the protected attribute X (e.g., sex) and the decision attribute Y (e.g., admission) for each combination of two attributes X and Y (e.g., female and decline). Similarly, the values of γ^{XYZ} indicates that more information about the X-Y relation is provided when levels of Z are distinguished. If γ_{ijk}^{XYZ} is positive, it means that the X-Y relation in the (i, j)-th cell of the X-Y subtable in the k-th level of Z is more positive than the average X-Y relation over all levels of Z; if it is negative, the X-Y relation in that cell is more negative than the average. In [24], the authors showed that full information about the pattern of the XYZ interaction cannot be given by an inspection of the highest order parameter alone.

To quantify the interaction between X and Y after incorporating the third variable Z, denoted as $I^{XY|Z}$, we need to combine the influence of lower order coefficient γ^{XY} and the influence of higher order coefficient γ^{XYZ}. To do this, we add the appropriate 3-way interaction parameter to the XY parameter to determine the XY relation for each level of Z:

$$I_{ij|k}^{XY|Z} = \gamma_{ij}^{XY} + \gamma_{ijk}^{XYZ} = n_{ijk} - \frac{n_{i.k}}{J} - \frac{n_{.jk}}{I} + \frac{n_{..k}}{IJ} \quad (4)$$

By comparing the values $I^{XY|Z}$ across different combinations of X and Y, we can quantify the magnitude of difference and conclude whether there exist significant discriminations. It is worth pointing out that our loglinear model based discrimination analysis covers the general case where the protected attribute has multiple domain values

Table I
LOGLINEAR MODELS FOR THREE-DIMENSIONAL TABLE. DIS: DISCRIMINATION

No	Loglinear Model	Generators	Dis
1	$logm_{ijk} = \gamma + \gamma_i^X + \gamma_j^Y + \gamma_k^Z$	(X, Y, Z)	N
2	$logm_{ijk} = \gamma + \gamma_i^X + \gamma_j^Y + \gamma_k^Z + \gamma_{ij}^{XY}$	(XY, Z)	Y
3	$logm_{ijk} = \gamma + \gamma_i^X + \gamma_j^Y + \gamma_k^Z + \gamma_{ik}^{XZ}$	(XZ, Y)	N
4	$logm_{ijk} = \gamma + \gamma_i^X + \gamma_j^Y + \gamma_k^Z + \gamma_{jk}^{YZ}$	(X, YZ)	N
5	$logm_{ijk} = \gamma + \gamma_i^X + \gamma_j^Y + \gamma_k^Z + \gamma_{ij}^{XY} + \gamma_{ik}^{XZ}$	(XY, XZ)	Y
6	$logm_{ijk} = \gamma + \gamma_i^X + \gamma_j^Y + \gamma_k^Z + \gamma_{ij}^{XY} + \gamma_{jk}^{YZ}$	(XY, YZ)	Y
7	$logm_{ijk} = \gamma + \gamma_i^X + \gamma_j^Y + \gamma_k^Z + \gamma_{ik}^{XZ} + \gamma_{jk}^{YZ}$	(XZ, YZ)	N
8	$logm_{ijk} = \gamma + \gamma_i^X + \gamma_j^Y + \gamma_k^Z + \gamma_{ij}^{XY} + \gamma_{ik}^{XZ} + \gamma_{jk}^{YZ}$	(XY, XZ, YZ)	Y
9	$logm_{ijk} = \gamma + \gamma_i^X + \gamma_j^Y + \gamma_k^Z + \gamma_{ij}^{XY} + \gamma_{ik}^{XZ} + \gamma_{jk}^{YZ} + \gamma_{ijk}^{XYZ}$	(XYZ)	Y

(e.g., race has domain values of black, white, and asian) and the decision attribute also has multiple domain values (e.g., decision has domain values accept, reject, and waiting-list). In previous work of discrimination analysis, researchers only consider the binary case for both the decision attribute and the protected attribute. There are two widely used metrics to measure discrimination: risk difference (RD) is defined as $P(Y = +|X = p) - P(Y = +|X = n)$, and risk ratio ($RR$) is defined as $P(Y = +|X = p)/P(Y = +|X = n)$ where $+$ denotes favorable decision and p denotes the protected group. When the difference is smaller than some threshold θ, discrimination does not occur. The value of threshold used for discrimination depends on the law. For instance, the 1975 British legislation for sex discrimination sets $\theta = 0.05$, namely a 5% difference. The U.S. legislation for employment discrimination sets the risk ratio threshold as 1.25 (known as the four-fifths rule). In statistics, Odds Ratio (OR), $\frac{P(Y=+|X=p)(1-P(Y=+|X=n))}{P(Y=+|X=n)(1-P(Y=+|X=p))}$, can also be used to quantify the discrimination.

Result 1: For the $I \times J \times K$ table with 3 dimensions X, Y and Z, given a fixed domain value k of Z, for any two values of the protected attribute X, i and i', and any two values of the decision attribute Y, j and j', we have

$$logOR = I_{ij|k}^{XY|Z} + I_{i'j'|k}^{XY|Z} - I_{i'j|k}^{XY|Z} - I_{ij'|k}^{XY|Z}. \quad (5)$$

Specially, when the decision attribute only has two domain values, we have

$$logOR = 2I_{ij|k}^{XY|Z} - 2I_{i'j|k}^{XY|Z}. \quad (6)$$

When both the decision attribute and the protected attribute have two domain values, we have

$$logOR = 4I_{ij|k}^{XY|Z}. \quad (7)$$

We skip all proofs due to space limits.

B. High Dimensional Data

One advantage of loglinear modeling is to be able to capture the high-order interactions for any number of protected attributes and decision attributes in one fitted model. Our goal is to derive a fitted model such that it is complex enough to provide a good fit to the data and is simple to capture all significant effects without overfitting the data. For high dimensional data, as the number of possible interaction terms increases, the loglinear model fitting can be computationally prohibitive and the interpretation of multi-factor interactions can be tedious.

1) Detecting Discrimination Structure: When there are multiple protected attributes $\mathbf{X}$ and decision attributes $\mathbf{Y}$ in addition to non-protected attributes $\mathbf{Z}$, discrimination analysis is becoming challenging. Our idea is to use the graphical model and interpret the conditional independence relationships between subsets of X and Y given subsets of Z based one graphical structure.

We restrict model selection to hierarchical models in the screening process. Specifically, we adopt the classic fitting approach that consists of fitting the model having only single-factor terms, then the model having only single-factor and 2-factor terms, then the model having only 3-factor and lower order terms, and so forth. We then choose one from the above all k-way interaction models which best fit the data. We apply the strategy proposed by Goodman et al [25], i.e., screening out those insignificant γ terms from the fitted model by using the χ^2 and the degree of freedom to compare the models. In essence, the χ^2 test statistic reflects how well a variable improves the model's goodness-of-fit beyond the fit produced by a baseline model without this variable.

We then find the smallest graphical model containing the fitted hierarchical model. This is a common practice to interpret loglinear models in terms of conditional independence. Since the fitted model is a submodel of that graphical model, all the conditional independence structure of the graphical model also holds for the fitted model. We are interested in the relationship between the protected attributes $\mathbf{X}$ and the decision attributes $\mathbf{Y}$ conditioning on non-protected attributes $\mathbf{Z}$.

Result 2: From the conditional independence graph constructed from the graphical loglinear model, we have:

1) there is no discrimination between the protected attribute X and the decision attribute Y if there is no edge between them.

2) for $\mathbf{U} \subseteq \mathbf{X}$, $\mathbf{V} \subseteq \mathbf{Y}$, and $\mathbf{W} \subseteq \mathbf{Z}$, if $\mathbf{U}$ and $\mathbf{V}$ are separated by $\mathbf{W}$, there is no discrimination between $\mathbf{U}$ and $\mathbf{V}$.
3) there is no discrimination between the decision attribute Y and sensitive attributes $\mathbf{X}$ according to the local Markov property if its neighbors $bd(Y) \subseteq \mathbf{Z}$.

It is worth pointing out that the graphical loglinear modeling can be easily used to detect the occurrence of Simpson's paradox like phenomena. Simpson's paradox happens when $P(\mathbf{Y}|\mathbf{X}, \mathbf{Z}) \neq P(\mathbf{Y}|\mathbf{X})$. The non-occurrence of Simpson's paradox like phenomena is related to collapsibility. We say that the conditional relationship $\mathbf{Y}|\mathbf{X}$ is collapsible over $\mathbf{Z}$ if $P(\mathbf{Y}|\mathbf{X}, \mathbf{Z}) = P(\mathbf{Y}|\mathbf{X})$. When the model is a graphical model, the global Markov property tells us the collapsibility can happen iff $\mathbf{Z}$ separates the graph into disjoint subgraph containing $\mathbf{X}$ and $\mathbf{Y}$. In other words, this can happen iff every path from a variable in $\mathbf{X}$ to a variable in $\mathbf{Y}$ goes through at least one variable in $\mathbf{Z}$.

The conditional independence graph can also be used to interpret marginal independence under some circumstances. Marginal independence occurs if there is no chain in the graph that connects two groups of variables. In general association in marginal tables differs from association structures found in the full table. For example, X and Y can be conditionally independent given $Z = k$, even if variables X and Y are marginally dependent. Marginal dependence means that the association is considered in the marginal table obtained from collapsing over the categories of the other variables, i.e., the other variables are ignored.

2) Quantifying Magnitude of Discrimination: In many cases, the best fitted model contains higher order interactions involving several variables. Simply put, a higher order interaction means that the nature of the association between two variables depends on the values (levels) of one or more other variables. Recall in the 3-D context, we are interested in the X-Y relationship and its changes over levels of Z. Our definition and procedure shown in Equation 4 readily generalizes to the d-th order interactions. The coefficients of an d-way interaction are made up of the contrast between the average value of an $(d-1)$-way interaction over all levels of the remaining d-th variable and the value of that $(d-1)$-way interaction exhibited at a particular level of the d-th variable. For example, in a 4-way interaction involving A, B, C, and D, as shown in Equation 1,

When we investigate a 2-way interaction, $\gamma_{ij|kl}^{AB|CD}$, across combinations of levels of C and D (i.e., A is the protected attribute, B is the decision attribute, and both C and D are non-protected attributes), our procedure involves adding the appropriate 3-way and 4-way parameter terms to the γ_{ij}^{AB} parameter:

$$\begin{aligned} I_{ij|kl}^{AB|CD} &= \gamma_{ij}^{AB} + \gamma_{ijk}^{ABC} + \gamma_{ijl}^{ABD} + \gamma_{ijkl}^{ABCD} \\ &= n_{ijkl} - \frac{n_{i.kl}}{J} - \frac{n_{.jkl}}{I} + \frac{n_{..kl}}{IJ} \end{aligned} \tag{8}$$

When we aim to determine the pattern of change for the 3-way interaction, γ_{ijk}^{ABC}, across levels of D (e.g., both A and C are protected attributes, B is the decision attribute, and D is the non-protected attribute), we can add the 4-way term, γ_{ijkl}^{ABCD}, to the 3-way term, γ_{ijk}^{ABC}. The relationship between the interaction and odds ratio shown in Result 1 can also be readily extended to high dimensional cases, i.e., replacing Z with $\mathbf{Z}$ and k with $\mathbf{k}$ where $\mathbf{Z}$ denotes the subset of non-protected attributes and $\mathbf{k}$ denotes the corresponding indices of $\mathbf{Z}$.

IV. Discrimination Prevention

Previous works of discrimination prevention tried to address this challenging problem by introducing a reverse discrimination in the training data [26] or pushing constraints into the trained classifiers [7], [8]. These works only consider the association or correlation between one protected attribute X and the decision attribute Y and do not take into account any effects due to non-protected attributes. Our discrimination discovery based on loglinear modeling can deal with any number of protected and decision attributes. The conditional independence graph derived from the fitted graphical loglinear model can effectively capture the existence or non-existence of discrimination patterns based on Result 2. Moreover, the coefficient values of the fitted loglinear model provide quantitative evidence of discrimination in decision making.

One naive approach is to remove all the coefficients that contain both a sensitive attribute and a decision attribute (e.g., γ^{XYZ}, γ^{XY}) from the fitted model. The newly generated dataset is guaranteed to have no discrimination between protected attributes and decision attributes. This strategy achieves the acceptance probabilities to be equal across different groups.

In practice, rather than removing those significant coefficients completely, we can modify the coefficient values such that the odds ratio values for all $X \in \mathbf{X}$ and $Y \in \mathbf{Y}$ are below a given threshold θ. Recall in Result 1, the odds ratio is determined by interaction parameters $I_{ij|\mathbf{k}}^{XY|\mathbf{Z}}$ that can be further calculated using coefficients (γ terms) of the fitted loglinear model. Hence our algorithm is to modify those interaction parameters $I_{ij|\mathbf{k}}^{XY|\mathbf{Z}}$ such that the odds ratio of every possible combination of X and Y is below the given threshold θ (Lines 4-9 of Algorithm 1). This procedure is equivalent to modifying the corresponding coefficients (γ terms) of the fitted loglinear model, as shown in Equation 4. The number of records in the corresponding cells are then modified as shown in Line 11 and a new table is returned. In our empirical evaluation, we will examine how the discrimination (in terms of risk difference and risk ratio) and the utility loss (in terms of the χ^2) vary with different θ values.

Algorithm 1 Removing Discrimination

Input: contingency table $T = (i_1, \cdots, i_d, n_{i_1,\cdots,i_d})$; attributes $A_1, \cdots, A_d$ as well as their assignments $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$; threshold θ

Output: new table $\hat{T}$

Build loglinear model from contingency table T;
for each pair of $X \in \mathbf{X}$ and $Y \in \mathbf{Y}$ **do**
  Calculate $I^{XY|\mathbf{Z}}_{ij|\mathbf{k}}$ over all $i \in [1, I]$ and $j \in [1, J]$;
  Calculate $\delta = max_{i,i' \in [1,I] j,j' \in [1,J]} \{I^{XY|\mathbf{Z}}_{ij|\mathbf{k}} + I^{XY|\mathbf{Z}}_{i'j'|\mathbf{k}} - I^{XY|\mathbf{Z}}_{i'j|\mathbf{k}} - I^{XY|\mathbf{Z}}_{ij'|\mathbf{k}}\}$
  if $\delta > \theta$ **then**
    $\hat{I}^{XY|\mathbf{Z}}_{ij\mathbf{k}} \leftarrow \frac{\theta}{\delta} I^{XY|\mathbf{Z}}_{ij\mathbf{k}}$ for all $i \in [1, I]$ $j \in [1, J]$;
  else
    $\hat{I}^{XY|\mathbf{Z}}_{ij\mathbf{k}} \leftarrow I^{XY|\mathbf{Z}}_{ij\mathbf{k}}$;
  end if
  for each cell $n_{i_1 \cdots i_d}$ **do**
    $\hat{n}_{i_1 \cdots i_d} \leftarrow e^{\hat{I}^{XY|\mathbf{Z}}_{ij\mathbf{k}} - I^{XY|\mathbf{Z}}_{ij\mathbf{k}}} \times n_{i_1 \cdots i_d}$;
  end for
end for
return new table $\hat{T} = (i_1, \cdots, i_d, \hat{n}_{i_1 \cdots i_d})$;

V. Evaluation

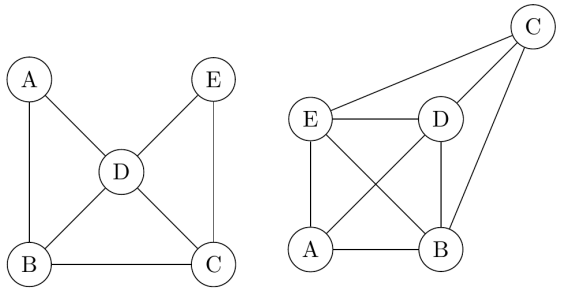
A. Dataset

We use two real datasets, *Adult* and *Dutch Census*, from the UCI Repository of Machine Learning Databases. These two datasets are typically used in discrimination analysis literature. The *Adult* dataset consists of 30169 tuples (after removing those tuples with missing values) with 14 attributes. The predictive task is to classify individuals into high and low income classes. It is well known that a number of attributes in the *Adult* dataset are weakly related with gender such as workclass, education, occupation, race, capital loss, native country. In our experiments, we select five attributes, gender, workclass, education, native country and income, denoted as A-E correspondingly. Their domain sizes are 2, 8, 16, 41, and 2 accordingly. The second dataset, *Dutch Census*, consists of 60420 tuples with 12 attributes. In our experiment, we select five attributes, sex, household position, household size, education level and income denoted as A-E correspondingly and their domain sizes are 2, 8, 6, 6, and 2 accordingly. The decision attribute is income with two domain values of high income and low income. For loglinear model fitting, we use the R package called *MASS* [1]. We limit the numbers of attributes for both datasets as five because the *MASS* package cannot deal with more than 10 attributes. In our future work, we plan to examine the recent development of loglinear modeling for high dimensional data (e.g., [27]) for our discrimination detection and prevention.

[1] http://cran.r-project.org/web/packages/MASS/index.html

B. Detecting Discrimination

1) Adult Dataset: For the *Adult* dataset, our fitted loglinear model is $(AB, AD, BC, BD, CD, CE, DE)$. Its χ^2 and G^2 are 7563 and 6886 respectively with the degree of freedom of 19840, which is much better than other models. For example, χ^2 and G^2 of the independence model (containing only 1-factor coefficients) are 53139 and 36962 respectively with the degree of freedom of 20927. Due to space limitations, we skip discussions of coefficients of the fitted loglinear models.



(a) Adult (b) Dutch

Figure 1. Conditional independence graph

Figure 1(a) shows the conditional independence graph from the fitted graphical loglinear model for the *Adult* dataset. Recall the pairwise property in Result 2, we know there is no discrimination between the protected attribute X and the decision attribute Y if there is no edge between them. We can see from Figure 1(a) that there is no discrimination between Gender (A) and income (E) because of the absence of the edge A-E. However, when native country (D) is considered as the protected attribute, there is significant discrimination between native country and income, as indicated by the presence of edge D-E in the conditional independence graph. Note that the conditional independence graph represents explicitly the relationship among all attributes. This would be very useful to answer "what if" questions when protected attributes and decision attributes are not a-priori specified.

We emphasize our loglinear modeling based discrimination analysis approach can straightforwardly deal with multiple protected attributes and decision attributes. For example, when both gender (A) and workclass (B) are considered as protected attributes and income (E) is considered as decision attribute in the *Adult* dataset, based on the global Markov property shown in Result 2, we conclude there is no discrimination between gender, workclass and income because AB and E are separated by non-protected attributes, education (C) and native country (D), as shown in Figure 1(a). Similarly, we can reach the conclusion that the *Adult* dataset does not have discrimination based on the local Markov property of Result 2 because the neighbors of

income (E) do not contain any protected attributes such as gender and workclass.

2) Dutch Dataset: For the *Dutch* dataset, our fitted hierarchical loglinear model is (ABC, ABD, ABE, ACD, ACE, ADE, BCD, BCE, BDE, CDE). Its χ^2 and G^2 are 467 and 383 respectively with the degree of freedom of 620. On the contrary, the independence model has $\chi^2 = 215032$, $G^2 = 2145$ and the degree of freedom 1132 whereas the pairwise loglinear model has $\chi^2 = 5620$, $G^2 = 2145$ and the degree of freedom 1002.

Figure 1(b) shows the conditional independence graph for the *Dutch* dataset. We can see that there is discrimination between sex (A) and income (E) because of the presence of the edge A-E. Actually for the *Dutch* dataset, the neighbors of income (E) contain all other attributes (A-D), thus based on the local Markov property discrimination always exists no matter how we choose protected attributes. This can also be derived based on the global Markov property because no protected attributes can be separated from the from the decision attribute E by any subsets of non-protected attributes.

C. Removing Discrimination

Our discrimination removal algorithm 1 modifies coefficients in the fitted loglinear model that contribute to discrimination and uses the new modified model to generate synthetic data. We examine how the utility loss and discrimination change with varying θ values (from 0 to 1 incremented by 0.1). For each θ value, we generate a new data set. To measure data utility loss, we follow the common practice and use χ^2 test statistic, i.e., the sum of squared residuals comparing the contingency table of the original data with the contingency table of the generated synthetic data cell-by-cell. The χ^2 statistic in our context captures both the error due to loglinear model fitting and the error due to coefficient modification in our discrimination removal algorithm. To examine how the generated synthetic data affects the performance of predictive models, we also measure utility loss in terms of the loss of classification accuracy. We measure the discriminations by reporting the percentage of subtables that violate the law requirements based on the traditional risk difference and risk ratio. Specifically, the threshold of risk difference is 0.05 and the threshold of risk ratio is 1.25.

1) Utility Loss vs. Discrimination: For the *Adult* data, we treat D as the protected attribute and E as the decision attribute. Figure 2 (left column) shows how the utility loss (χ^2) and the discrimination (in terms of risk difference and risk ratio) change with varying θ values for the *Adult* data. We observe that the larger the θ (which corresponds less distortion), the better utility of the generated data, the more discrimination the generated data has. From the results, we see the algorithm removes all discrimination (in terms of both risk difference and risk ratio) and achieves good utility preservation by setting θ with 0.4.

For the *Dutch* data, we treat A as the protected attribute and E as the decision attribute. Figure 2 (right column) shows how χ^2, risk difference and risk ratio change with varying θ values for the *Dutch* data. We observe the similar pattern as that for *Adult*, i.e., less distortion preserves more utility while incurring more discrimination. One difference is that the generated data with $\theta = 0.4$ indicates the existence of discrimination in terms of risk ratio, as shown in Figure 2(f), whereas the same generated data indicates no discrimination in terms of risk difference.

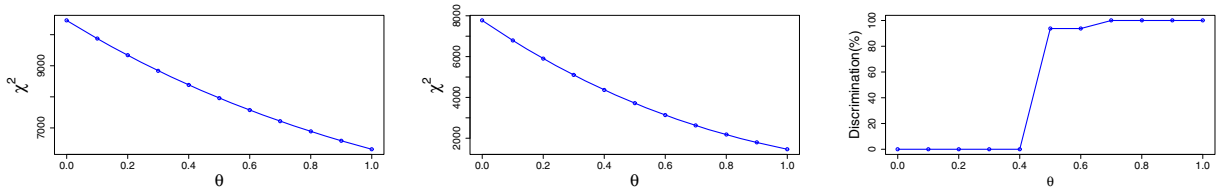
2) Classification Accuracy vs. Threshold θ: The synthetic data, which is guaranteed discrimination-free, can be used for any analysis. In this section, we conduct experiments to measure loss of classification accuracy. We use two standard classifiers: decision tree (J48) and naive Bayes classifier (NB). Figures 3 shows how classification accuracy changes with varying θ values for *Adult* (left column) and *Dutch* (right column) data respectively. The Y-axis of these plots represents classification accuracy and each point is for a specific value of θ which is varied from 0 to 1. In each figure, we also report the baseline, i.e., the classification accuracy on the original data. All results reported are obtained using 5-fold cross-validation.

It is observed that for the *Adult* data the classification accuracy values of both decision tree and naive Bayes classifier increase as the value of θ is increased from 0 to 0.3. This is mainly due to the distortion decrease in the discrimination removal process. The classification accuracy values keep almost unchanged when the value of θ is greater than 0.3. The accuracy gap from the baseline is mainly due to the fitting error of the loglinear model.

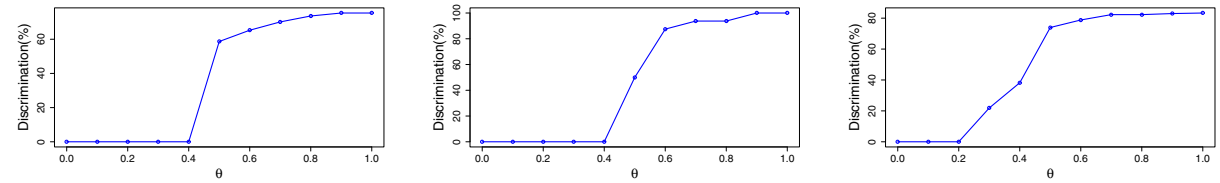
For the *Dutch* data, we can observe from Figure 3(b) that the classification accuracy of J48 increases significantly when θ is increased from 0.8 to 1.0. However, the classification accuracy of naive Bayes classifier is not significantly improved for naive Bayes classifier, as shown in Figure 3(d).

3) Removing Discrimination for Multiple Protected Attributes: Our loglinear model based discrimination removal algorithm covers the general case where there are multiple protected attributes and decision attributes in addition to a set of non-protected attributes. A protected attribute may have multiple domain values (e.g., race has domain values of black, white, and asian) and a decision attribute may also have multiple domain values (e.g., decision has domain values accept, reject, and waiting-list). As the fitted loglinear model can inherently capture and quantify discriminations between multiple protected attributes and decision attributes, our discrimination removal algorithm can also effectively remove those multi-attribute discriminations.

Table II shows the results of applying our discrimination removal algorithm over *Dutch* data where both A and C are considered as protected attributes and E as the decision one.

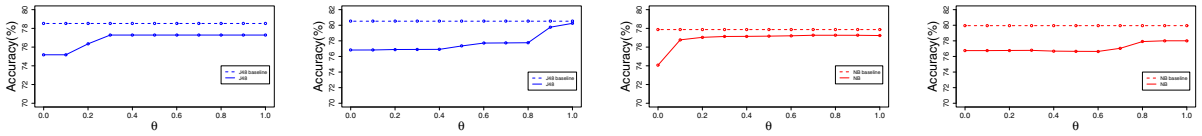


(a) Utility loss (χ^2) - Adult (b) Utility loss (χ^2) - Dutch (c) Risk difference (RD) - Adult



(d) Risk difference (RD) - Dutch (e) Risk ratio (RR) - Adult (f) Risk ratio (RR) - Dutch

Figure 2. Utility loss and discrimination vs. varying θ



(a) Decision tree - Adult (b) Decision tree - Dutch (c) Naive Bayes - Adult (d) Naive Bayes - Dutch

Figure 3. Classification accuracy vs. varying θ

Table II
UTILITY LOSS AND DISCRIMINATION VS. VARYING θ FOR TWO PROTECTED ATTRIBUTES AC AND DECISION ATTRIBUTE E OF DUTCH DATA

θ	Protected attribute A			Protected attribute C			Protected attributes AC		
	χ^2	$RD(\%)$	RR (%)	χ^2	$RD(\%)$	$RR(\%)$	χ^2	$RD(\%)$	$RR(\%)$
0.0	7769	0.0	0.0	7789	0.0	0.0	20214	0.0	0.0
0.1	6795	0.0	0.0	7076	0.0	0.0	17859	0.0	0.0
0.2	5906	0.0	0.0	6431	0.0	0.0	15713	0.0	54.1
0.3	5098	0.0	21.8	5852	0.0	30.5	13763	75.0	85.4
0.4	4370	0.0	38.1	5324	0.0	50.0	11999	93.7	87.5
0.5	3717	58.6	73.9	4843	47.2	63.8	10412	95.8	89.5
0.6	3135	65.2	78.8	4405	66.6	69.4	8997	95.8	89.5
0.7	2622	70.1	82.2	4003	73.6	76.3	7730	95.8	97.9
0.8	2175	73.6	82.2	3640	79.1	79.1	6611	95.8	97.9
0.9	1791	75.3	82.9	3313	81.9	81.9	5637	95.8	100.0
1.0	1459	75.3	83.3	3020	86.1	84.7	4811	97.9	100.0

As shown in the right-most column block, the utility loss (χ^2) decreases as θ increases whereas risk difference and risk ratio increase accordingly. We also include the results of applying our discrimination removal algorithm with only one single protected attribute, e.g., for A in the second block, and for C shown in the third block, respectively. Comparing these three blocks, we see that for a fixed θ value, the utility loss (χ^2) is larger in the case of multiple protected attributes than in the cases of a single protected attribute.

D. Discussion

Our evaluation here focuses on the use of two discrimination metrics, risk difference and risk ratio, which have been adopted by legislations. It is worth pointing out that odds ratio (OR) is more appropriately used for quantifying discrimination from the statistical viewpoint. As shown in Section III-A2, our θ threshold is closely related to odds ratio and actually θ is the logarithm value of odds ratio. Thus we would argue the use of θ as an indicative metric for discrimination.

We would point out that it is challenging to determine the optimal θ threshold to maximize the utility while minimizing discrimination risk. We argue it is more appropriate to examine the utility preservation while keeping an upper bound on the risk measure. We also argue it is better to use the utility loss metric (χ^2) than the classification accuracy (tied with one particular classifier) to quantify the utility loss.

VI. Conclusions and Future Work

We have developed a single unifying framework that captures and measures discriminations based on loglinear modeling. The derived conditional independence graph represents the conditional independence structures of attributes in the historic dataset. Based on Markov properties, we have shown how to determine the existence of discrimination patterns and how to interpret them. We derived the interaction metric to quantify the discrimination (i.e., the protected attributes' effects on the decision attribute) and showed its relationship with odds ratio. We developed a method of solving the problem of discrimination prevention by modifying those significant coefficients from the fitted loglinear model and using the modified model to generate the new data. We designed strategies of effectively changing the coefficient values of the fitted loglinear model to meet different discrimination requirements.

In our future work, we will study how to incorporate background knowledge in model fitting. We can express domain knowledge as some constraints when building graphical loglinear models. For example, we can enforce no edge between gender and race in the conditional independence graph since they are biologically independent. We will also examine the recent development of loglinear modeling for high dimensional data (e.g., [27]) and extend our discrimination detection and prevention to the high dimensional data scenario. We will develop discrimination analysis techniques based on logistic regression, which is appropriate for analyzing a mixed set of nominal/ordinal and interval variables.

Acknowledgement

This work was supported in part by U.S. National Science Foundation (1646654) and U.S. National Institute of Health (1R01GM103309).

References

[1] S. Hajian and J. Domingo-Ferrer, "A methodology for direct and indirect discrimination prevention in data mining," *TKDE*, vol. 25, no. 7, pp. 1445–1459, 2013.

[2] F. Kamiran and T. Calders, "Data preprocessing techniques for classification without discrimination," *KAIS*, vol. 33, no. 1, pp. 1–33, 2012.

[3] S. Ruggieri, D. Pedreschi, and F. Turini, "Data mining for discrimination discovery," *TKDD*, vol. 4, no. 2, p. 9, 2010.

[4] A. Romei and S. Ruggieri, "A multidisciplinary survey on discrimination analysis," *The Knowledge Engineering Review*, vol. 29, no. 05, pp. 582–638, 2014.

[5] I. Zliobaite, F. Kamiran, and T. Calders, "Handling conditional discrimination," *ICDM*, IEEE, 2011, pp. 992–1001.

[6] F. Kamiran, T. Calders, and M. Pechenizkiy, "Discrimination aware decision tree learning," *ICDM*, IEEE, 2010, pp. 869–874.

[7] T. Calders and S. Verwer, "Three naive bayes approaches for discrimination-free classification," *DMKD*, vol. 21, no. 2, pp. 277–292, 2010.

[8] D. Pedreshi, S. Ruggieri, and F. Turini, "Discrimination-aware data mining," *KDD*, ACM, 2008, pp. 560–568.

[9] D. Pedreschi, S. Ruggieri, and F. Turini, "Measuring discrimination in socially-sensitive decision records," *SDM*, 2009.

[10] B. T. Luong, S. Ruggieri, and F. Turini, "K-nn as an implementation of situation testing for discrimination discovery and prevention," *KDD*, ACM, 2011, pp. 502–510.

[11] L. Zhang, Y. Wu, and X. Wu. "Situation Testing-based discrimination discovery: a causal inference approach," *IJCAI*, 2016.

[12] S. Hajian, J. Domingo-Ferrer, A. Monreale, D. Pedreschi, and F. Giannotti, "Discrimination-and privacy-aware patterns," *Data Mining and Knowledge Discovery*, pp. 1–50, 2014.

[13] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," *KDD*. ACM, 2015, pp. 259–268.

[14] T. Kamishima, S. Akaho, and J. Sakuma, "Fairness-aware learning through regularization approach," in *ICDMW*. IEEE, 2011, pp. 643–650.

[15] K. Mancuhan and C. Clifton, "Combating discrimination using bayesian networks," *Artificial intelligence and law*, vol. 22, no. 2, pp. 211–238, 2014.

[16] Lu Zhang, Yongkai Wu, and Xintao Wu. "On discrimination discovery using causal networks," *SBP-BRimS*, 2016.

[17] F. Bonchi, S. Hajian, B. Mishra, and D. Ramazzotti. "Exposing the probabilistic causal structure of discrimination," *CoRR*, abs/1510.00552, 2015.

[18] A. Agresti, *Categorical data analysis*, Wiley Series in Probability and Statistics. Wiley-Interscience, 2002.

[19] D. Barbará and X. Wu, "Loglinear based quasi cubes," *JIIS*, vol. 16, no. 3, pp. 255–276, 2001.

[20] X. Wu, D. Barbar, and Y. Ye, "Screening and interpreting multi-item associations based on log-linear modeling." *KDD*. ACM, 2003, pp. 276–285.

[21] S. Sarawagi, R. Agrawal, and N. Megiddo, "Discovery-driven exploration of olap data cubes," in *EDBT*, 1998, pp. 168–182.

[22] Y. M. Bishop, S. E. Fienberg, and P. W. Holland, *Discrete Multivariate Analysis: Theory and Practice*. The MIT Press, 1975.

[23] C. R. Blyth, "On simpson's paradox and the sure-thing principle," *Journal of the American Statistical Association*, vol. 67, no. 338, pp. 364–366, 1972.

[24] G. C. Elliott, "Interpreting higher order interactions in log-linear analysis." *Psychological Bulletin*, vol. 103, no. 1, p. 121, 1988.

[25] L. A. Goodman, "The analysis of multidimensional contingency tables: Stepwise procedures and direct estimation methods for building models for multiple classifications," *Technometrics*, vol. 13, no. 1, pp. 33–61, 1971.

[26] F. Kamiran and T. Calders, "Classification with no discrimination by preferential sampling," in *BeneLearn Conference on Machine Learning*, 2010.

[27] F. Petitjean, L. Allison, and G. I. Webb, "A statistically efficient and scalable method for log-linear analysis of high-dimensional data," *ICDM*, 2014, pp. 480–489.