A Methodology for Direct and Indirect Discrimination Prevention in Data Mining

Sara Hajian and Josep Domingo-Ferrer, *Fellow*, *IEEE*

Abstract—Data mining is an increasingly important technology for extracting useful knowledge hidden in large collections of data. There are, however, negative social perceptions about data mining, among which potential privacy invasion and potential discrimination. The latter consists of unfairly treating people on the basis of their belonging to a specific group. Automated data collection and data mining techniques such as classification rule mining have paved the way to making automated decisions, like loan granting/denial, insurance premium computation, etc. If the training data sets are biased in what regards discriminatory (sensitive) attributes like gender, race, religion, etc., discriminatory decisions may ensue. For this reason, antidiscrimination techniques including discrimination discovery and prevention have been introduced in data mining. Discrimination can be either direct or indirect. Direct discrimination occurs when decisions are made based on sensitive attributes. Indirect discrimination occurs when decisions are made based on nonsensitive attributes which are strongly correlated with biased sensitive ones. In this paper, we tackle discrimination prevention in data mining and propose new techniques applicable for direct or indirect discrimination prevention individually or both at the same time. We discuss how to clean training data sets and outsourced data sets in such a way that direct and/or indirect discriminatory decision rules are converted to legitimate (nondiscriminatory) classification rules. We also propose new metrics to evaluate the utility of the proposed approaches and we compare these approaches. The experimental evaluations demonstrate that the proposed techniques are effective at removing direct and/or indirect discrimination biases in the original data set while preserving data quality.

Index Terms—Antidiscrimination, data mining, direct and indirect discrimination prevention, rule protection, rule generalization, privacy

1 Introduction

In sociology, discrimination is the prejudicial treatment of an individual based on their membership in a certain group or category. It involves denying to members of one group opportunities that are available to other groups. There is a list of antidiscrimination acts, which are laws designed to prevent discrimination on the basis of a number of attributes (e.g., race, religion, gender, nationality, disability, marital status, and age) in various settings (e.g., employment and training, access to public services, credit and insurance, etc.). For example, the European Union implements the principle of equal treatment between men and women in the access to and supply of goods and services in [3] or in matters of employment and occupation in [4]. Although there are some laws against discrimination, all of them are reactive, not proactive. Technology can add proactivity to legislation by contributing discrimination discovery and prevention techniques.

Services in the information society allow for automatic and routine collection of large amounts of data. Those data are often used to train association/classification rules in view of making automated decisions, like loan granting/denial, insurance premium computation, personnel selection, etc. At first sight, automating decisions may give a sense of fairness: classification rules do not guide themselves by personal preferences. However, at a closer look, one realizes that classification rules are actually learned by the system (e.g., loan granting) from the training data. If the training data are inherently biased for or against a particular community (e.g., foreigners), the learned model may show a discriminatory prejudiced behavior. In other words, the system may infer that just being foreign is a legitimate reason for loan denial. Discovering such potential biases and eliminating them from the training data without harming their decision-making utility is therefore highly desirable. One must prevent data mining from becoming itself a source of discrimination, due to data mining tasks generating discriminatory models from biased data sets as part of the automated decision making. In [12], it is demonstrated that data mining can be both a source of discrimination and a means for discovering discrimination.

Discrimination can be either direct or indirect (also called systematic). Direct discrimination consists of rules or procedures that explicitly mention minority or disadvantaged groups based on sensitive discriminatory attributes related to group membership. Indirect discrimination consists of rules or procedures that, while not explicitly mentioning discriminatory attributes, intentionally or unintentionally could generate discriminatory decisions. Redlining by financial institutions (refusing to grant mortgages or insurances in urban areas they consider as deteriorating) is an archetypal example of indirect discrimination, although certainly not the only one. With a slight abuse of

• *The authors are with Department of Computer Engineering and Maths, UNESCO Chair in Data Privacy, Av. Països Catalans 26, E-43007 Tarragona, Catalonia. E-mail: {sara.hajian, josep.domingo}@urv.cat.*

Manuscript received 20 Aug. 2011; revised 20 Jan. 2012; accepted 17 Mar. 2012; published online 22 Mar. 2012.
Recommended for acceptance by E. Ferrari.
For information on obtaining reprints of this article, please send e-mail to: tkde@computer.org, and reference IEEECS Log Number TKDE-2011-08-0503.
Digital Object Identifier no. 10.1109/TKDE.2012.72.

language for the sake of compactness, in this paper indirect discrimination will also be referred to as *redlining* and rules causing indirect discrimination will be called *redlining rules* [12]. Indirect discrimination could happen because of the availability of some background knowledge (rules), for example, that a certain zip code corresponds to a deteriorating area or an area with mostly black population. The background knowledge might be accessible from publicly available data (e.g., census data) or might be obtained from the original data set itself because of the existence of nondiscriminatory attributes that are highly correlated with the sensitive ones in the original data set.

1.1 Related Work

Despite the wide deployment of information systems based on data mining technology in decision making, the issue of antidiscrimination in data mining did not receive much attention until 2008 [12]. Some proposals are oriented to the *discovery and measure* of discrimination. Others deal with the *prevention* of discrimination.

The discovery of discriminatory decisions was first proposed by Pedreschi et al. [12], [15]. The approach is based on mining classification rules (the inductive part) and reasoning on them (the deductive part) on the basis of quantitative measures of discrimination that formalize legal definitions of discrimination. For instance, the US Equal Pay Act [18] states that: "a selection rate for any race, sex, or ethnic group which is less than four-fifths of the rate for the group with the highest rate will generally be regarded as evidence of adverse impact." This approach has been extended to encompass statistical significance of the extracted patterns of discrimination in [13] and to reason about affirmative action and favoritism [14]. Moreover it has been implemented as an Oracle-based tool in [16]. Current discrimination discovery methods consider each rule individually for measuring discrimination without considering other rules or the relation between them. However, in this paper we also take into account the relation between rules for discrimination discovery, based on the existence or nonexistence of discriminatory attributes.

Discrimination prevention, the other major antidiscrimination aim in data mining, consists of inducing patterns that do not lead to discriminatory decisions even if the original training data sets are biased. Three approaches are conceivable:

- *Preprocessing*. Transform the source data in such a way that the discriminatory biases contained in the original data are removed so that no unfair decision rule can be mined from the transformed data and apply any of the standard data mining algorithms. The preprocessing approaches of data transformation and hierarchy-based generalization can be adapted from the privacy preservation literature. Along this line, [7], [8] perform a controlled distortion of the training data from which a classifier is learned by making minimally intrusive modifications leading to an unbiased data set. The preprocessing approach is useful for applications in which a data set should be published and/or in which data mining needs to be performed also by *external parties* (and not just by the data holder).
- *In-processing*. Change the data mining algorithms in such a way that the resulting models do not contain unfair decision rules. For example, an alternative approach to cleaning the discrimination from the original data set is proposed in [2] whereby the nondiscriminatory constraint is embedded into a decision tree learner by changing its splitting criterion and pruning strategy through a novel leaf relabeling approach. However, it is obvious that in-processing discrimination prevention methods must rely on new special-purpose data mining algorithms; standard data mining algorithms cannot be used.
- *Postprocessing*. Modify the resulting data mining models, instead of cleaning the original data set or changing the data mining algorithms. For example, in [13], a confidence-altering approach is proposed for classification rules inferred by the CPAR algorithm. The postprocessing approach does not allow the data set to be published: only the modified data mining models can be published (knowledge publishing), hence data mining can be performed by the data holder only.

One might think of a straightforward preprocessing approach consisting of just removing the discriminatory attributes from the data set. Although this would solve the direct discrimination problem, it would cause much information loss and in general it would not solve indirect discrimination. As stated in [12] there may be other attributes (e.g., Zip) that are highly correlated with the sensitive ones (e.g., Race) and allow inferring discriminatory rules. Hence, there are two important challenges regarding discrimination prevention: one challenge is to consider both direct and indirect discrimination instead of only direct discrimination; the other challenge is to find a good tradeoff between discrimination removal and the quality of the resulting training data sets and data mining models.

Although some methods have already been proposed for each of the above-mentioned approaches (preprocessing, in-processing, postprocessing), discrimination prevention stays a largely unexplored research avenue. In this paper, we concentrate on discrimination prevention based on preprocessing, because the preprocessing approach seems the most flexible one: it does not require changing the standard data mining algorithms, unlike the in-processing approach, and it allows data publishing (rather than just knowledge publishing), unlike the postprocessing approach.

1.2 Contribution and Plan of This Paper

Discrimination prevention methods based on preprocessing published so far [7], [8] present some limitations, which we next highlight:

- They attempt to detect discrimination in the original data only for one discriminatory item and based on a single measure. This approach cannot guarantee that the transformed data set is really discrimination free, because it is known that discriminatory behaviors can often be hidden behind several discriminatory items, and even behind combinations of them.
- They only consider direct discrimination.

- They do not include any measure to evaluate how much discrimination has been removed and how much information loss has been incurred.

In this paper, we propose preprocessing methods which overcome the above limitations. Our new data transformation methods (i.e., rule protection and rule generalization (RG)) are based on measures for both direct and indirect discrimination and can deal with several discriminatory items. Also, we provide utility measures. Hence, our approach to discrimination prevention is broader than in previous work.

In our earlier work [5], we introduced the initial idea of using rule protection and rule generalization for direct discrimination prevention, but we gave no experimental results. In [6], we introduced the use of rule protection in a different way for indirect discrimination prevention and we gave some preliminary experimental results. In this paper, we present a *unified approach to direct and indirect discrimination prevention*, with finalized algorithms and all possible data transformation methods based on rule protection and/or rule generalization that could be applied for direct or indirect discrimination prevention. We specify the different features of each method. Since methods in our earlier papers [5], [6] could only deal with either direct or indirect discrimination, the methods described in this paper are new.

As part of this effort, we have developed metrics that specify which records should be changed, how many records should be changed, and how those records should be changed during data transformation. In addition, we propose new utility measures to evaluate the different proposed discrimination prevention methods in terms of data quality and discrimination removal for both direct and indirect discrimination. Based on the proposed measures, we present extensive experimental results for two well-known data sets and compare the different possible methods for direct or indirect discrimination prevention to find out which methods could be more successful in terms of low information loss and high discrimination removal.

The rest of this paper is organized as follows. Section 2 introduces some basic definitions and concepts that are used throughout the paper. Section 3 describes our proposal for direct and indirect discrimination prevention. Section 4 shows the tests we have performed to assess the validity and quality of our proposal and compare different methods. Finally, Section 5 summarizes conclusions and identifies future research topics in the field of discrimination prevention.

2 Background

In this section, we briefly review the background knowledge required in the remainder of this paper. First, we recall some basic definitions related to data mining [17]. After that, we elaborate on measuring and discovering discrimination.

2.1 Basic Definitions

- A *data set* is a collection of data objects (records) and their attributes. Let $\mathcal{DB}$ be the original data set.
- An *item* is an attribute along with its value, e.g., $\text{Race} = \text{black}$.
- An *item set*, i.e., X, is a collection of one or more items, e.g., $\{\text{Foreign worker} = \text{Yes}, \text{City} = \text{NYC}\}$.
- A *classification rule* is an expression $X \rightarrow C$, where C is a class item (a yes/no decision), and X is an item set containing no class item, e.g., $\{\text{Foreign worker} = \text{Yes}, \text{City} = \text{NYC}\} \rightarrow \text{Hire} = \text{no}$. X is called the premise of the rule.
- The *support* of an item set, $supp(X)$, is the fraction of records that contain the item set X. We say that a rule $X \rightarrow C$ is *completely supported* by a record if both X and C appear in the record.
- The *confidence* of a classification rule, $conf(X \rightarrow C)$, measures how often the class item C appears in records that contain X. Hence, if $supp(X) > 0$ then

$$conf(X \rightarrow C) = \frac{supp(X, C)}{supp(X)}. \quad (1)$$

  Support and confidence range over $[0, 1]$.
- A *frequent classification rule* is a classification rule with support and confidence greater than respective specified lower bounds. Support is a measure of statistical significance, whereas confidence is a measure of the strength of the rule. Let $\mathcal{FR}$ be the database of frequent classification rules extracted from $\mathcal{DB}$.
- The *negated item set, i.e.,* $\neg X$ is an item set with the same attributes as X, but the attributes in $\neg X$ take any value except those taken by attributes in X. In this paper, we use the $\neg$ notation for item sets with binary or nonbinary categorical attributes. For a binary attribute, e.g., $\{\text{Foreign worker} = \text{Yes/No}\}$, if X is {Foreign worker = Yes}, then $\neg X$ is {Foreign worker=No}. If X is binary, it can be converted to $\neg X$ and vice versa, that is, the negation works in both senses. In the previous example, we can select the records in $\mathcal{DB}$ such that the value of the Foreign worker attribute is "Yes" and change that attribute's value to "No," and conversely. However, for a nonbinary categorical attribute, e.g., $\{\text{Race} = \text{Black/White/Indian}\}$, if X is $\{\text{Race} = \text{Black}\}$, then $\neg X$ is $\{\text{Race} = \text{White}\}$ or $\{\text{Race} = \text{Indian}\}$. In this case, $\neg X$ can be converted to X without ambiguity, but the conversion of X into $\neg X$ is not uniquely defined, which we denote by $\neg X \Rightarrow X$. In the previous example, we can select the records in $\mathcal{DB}$ such that the Race attribute is "White" or "Indian" and change that attribute's value to "Black"; but if we want to negate $\{\text{Race} = \text{Black}\}$, we do not know whether to change it to $\{\text{Race} = \text{White}\}$ or $\{\text{Race} = \text{Indian}\}$. In this paper, we use only nonambiguous negations.

2.2 Potentially Discriminatory and Nondiscriminatory Classification Rules

Let DI_s be the set of predetermined discriminatory items in DB (e.g., $\text{DI}_\text{s} = \{\text{Foreign worker} = \text{Yes}, \text{Race} = \text{Black}, \text{Gender} = \text{Female}\}$). Frequent classification rules in $\mathcal{FR}$ fall into one of the following two classes:

1. A classification rule $X \rightarrow C$ is *potentially discriminatory* (PD) when $X = A, B$ with $A \subseteq DI_s$ a nonempty discriminatory item set and B a nondiscriminatory item set. For example, $\{\text{Foreign worker} = \text{Yes}, \text{City} = \text{NYC}\} \rightarrow \text{Hire} = \text{No}$.

2. A classification rule $X \rightarrow C$ is *potentially nondiscriminatory* (PND) when $X = D, B$ is a nondiscriminatory item set. For example, $\{\text{Zip} = 10451, \text{City} = \text{NYC}\} \rightarrow \text{Hire} = \text{No}$, or $\{\text{Experience} = \text{Low}, \text{City} = \text{NYC}\} \rightarrow \text{Hire} = \text{No}$

The word "potentially" means that a PD rule could probably lead to discriminatory decisions. Therefore, some measures are needed to quantify the direct discrimination potential. Also, a PND rule could lead to discriminatory decisions in combination with some background knowledge; e.g., if the premise of the PND rule contains the zip code as an attribute and one knows that zip code 10451 is mostly inhabited by foreign people. Hence, measures are needed to quantify the indirect discrimination potential as well.

2.3 Direct Discrimination Measure

Pedreschi et al. [12], [13] translated the qualitative statements in existing laws, regulations, and legal cases into quantitative formal counterparts over classification rules and they introduced a family of measures of the degree of discrimination of a PD rule. One of these measures is the *extended lift* ($elift$).

Definition 1. *Let $A, B \rightarrow C$ be a classification rule such that $\text{conf}(B \rightarrow C) > 0$. The extended lift of the rule is*

$$\text{elift}(A, B \rightarrow C) = \frac{\text{conf}(A, B \rightarrow C)}{\text{conf}(B \rightarrow C)}. \quad (2)$$

The idea here is to evaluate the discrimination of a rule as the gain of confidence due to the presence of the discriminatory items (i.e., A) in the premise of the rule. Whether the rule is to be considered discriminatory can be assessed by thresholding $elift$ as follows.

Definition 2. *Let $\alpha \in R$ be a fixed threshold[1] and let A be a discriminatory item set. A PD classification rule $c = A, B \rightarrow C$ is α-protective w.r.t.* elift *if* $\text{elift}(c) < \alpha$. *Otherwise, c is α-discriminatory.*

The purpose of direct discrimination discovery is to identify α-discriminatory rules. In fact, α-discriminatory rules indicate biased rules that are directly inferred from discriminatory items (e.g., $\text{Foreign worker} = \text{Yes}$). We call these rules direct α-discriminatory rules.

In addition to $elift$, two other measures $slift$ and $olift$ were proposed by Pedreschi et al. in [13]. The reason is that different measures of discriminating power of the mined decision rules can be defined, according to the various antidiscrimination regulations in different countries. Yet the protection methods are similar no matter the measure adopted (see discussion in the Appendix, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TKDE.2012.72).

2.4 Indirect Discrimination Measure

The purpose of indirect discrimination discovery is to identify redlining rules. In fact, redlining rules indicate biased rules that are indirectly inferred from nondiscriminatory items (e.g., Zip = 10451) because of their correlation with discriminatory ones. To determine the redlining rules, Pedreschi et al. in [12] stated the theorem below which gives a lower bound for α-discrimination of PD classification rules, given information available in PND rules (γ, δ), and information available from background rules (β_1, β_2). They assume that background knowledge takes the form of classification rules relating a nondiscriminatory item set D to a discriminatory item set A within the context B.

Theorem 1. *Let $r : D, B \rightarrow C$ be a PND classification rule, and let*

$$\gamma = conf(r : D, B \rightarrow C)\delta = conf(B \rightarrow C) > 0.$$

Let A be a discriminatory item set, and let β_1, β_2 such that

$$\text{conf}(r_{b1} : A, B \rightarrow D) \geq \beta_1$$

$$\text{conf}(r_{b2} : D, B \rightarrow A) \geq \beta_2 > 0.$$

Call

$$f(x) = \frac{\beta_1}{\beta_2}(\beta_2 + x - 1)$$

$$elb(x, y) = \begin{cases} f(x)/y & \text{if} f(x) > 0 \\ 0 & otherwise. \end{cases}$$

It holds that, for $\alpha \geq 0$, if $elb(\gamma, \delta) \geq \alpha$, the PD classification rule $r' : A, B \rightarrow C$ is α-discriminatory.

Based on the above theorem, the following formal definitions of redlining and nonredlining rules are presented:

Definition 3. *A PND classification rule $r : D, B \rightarrow C$ is a* redlining rule *if it could yield an α-discriminatory rule $r' : A, B \rightarrow C$ in combination with currently available background knowledge rules of the form $r_{b1} : A, B \rightarrow D$ and $r_{b2} : D, B \rightarrow A$, where A is a discriminatory item set. For example,* $\{\text{Zip} = 10451, \text{City} = \text{NYC}\} \rightarrow \text{Hire} = \text{No}$.

Definition 4. *A PND classification rule $r : D, B \rightarrow C$ is a* nonredlining or legitimate rule *if it cannot yield any α-discriminatory rule $r' : A, B \rightarrow C$ in combination with currently available background knowledge rules of the form $r_{b1} : A, B \rightarrow D$ and $r_{b2} : D, B \rightarrow A$, where A is a discriminatory item set. For example,* $\{\text{Experience} = \text{Low}, \text{City} = NYC\} \rightarrow \text{Hire} = \text{No}$.

We call α-discriminatory rules that ensue from redlining rules *indirect α-discriminatory rules*.

3 A Proposal for Direct and Indirect Discrimination Prevention

In this section, we present our approach, including the data transformation methods that can be used for direct and/or indirect discrimination prevention. For each method, its algorithm and its computational cost are specified.

3.1 The Approach

Our approach for direct and indirect discrimination prevention can be described in terms of two phases:

- **Discrimination measurement**. Direct and indirect discrimination discovery includes identifying

1. Note that α is a fixed threshold stating an acceptable level of discrimination according to laws and regulations. For example, the four-fifths rule of US Federal Legislation sets $\alpha = 1.25$.

α-discriminatory rules and redlining rules. To this end, first, based on predetermined discriminatory items in $\mathcal{DB}$, frequent classification rules in $\mathcal{FR}$ are divided in two groups: PD and PND rules. Second, direct discrimination is measured by identifying α-discriminatory rules among the PD rules using a direct discrimination measure ($elift$) and a discriminatory threshold (α). Third, indirect discrimination is measured by identifying redlining rules among the PND rules combined with background knowledge, using an indirect discriminatory measure (elb), and a discriminatory threshold (α). Let $\mathcal{MR}$ be the database of direct α-discriminatory rules obtained with the above process. In addition, let $\mathcal{RR}$ be the database of redlining rules and their respective indirect α-discriminatory rules obtained with the above process.

- **Data transformation**. Transform the original data $\mathcal{DB}$ in such a way to remove direct and/or indirect discriminatory biases, with minimum impact on the data and on legitimate decision rules, so that no unfair decision rule can be mined from the transformed data. In the following sections, we present the data transformation methods that can be used for this purpose.

As mentioned before, background knowledge might be obtained from the original data set itself because of the existence of nondiscriminatory attributes that are highly correlated with the sensitive ones in the original data set. Let BK be a database of background rules that is defined as

$$\mathcal{BK} = \{r_{b2} : D, B \rightarrow A | A \text{ discriminatory item set and } supp\,(D, B \rightarrow A) \geq ms\}.$$

In fact, $\mathcal{BK}$ is the set of classification rules $D, B \rightarrow A$ with a given minimum support ms that shows the correlation between the discriminatory item set A and the nondiscriminatory item set D with context B. Although rules of the form $r_{b1} : A, B \rightarrow D$ (in Theorem 1) are not included in $\mathcal{BK}$, $conf(r_{b1} : A, B \rightarrow D)$ could be obtained as $supp(r_{b2} : D, B \rightarrow A)/supp(B \rightarrow A)$.

3.2 Data Transformation for Direct Discrimination

The proposed solution to prevent direct discrimination is based on the fact that the data set of decision rules would be free of direct discrimination if it only contained PD rules that are α-protective or are instances of at least one nonredlining PND rule. Therefore, a suitable data transformation with minimum information loss should be applied in such a way that each α-discriminatory rule either becomes α-protective or an instance of a nonredlining PND rule. We call the first procedure *direct rule protection* (DRP) and the second one *rule generalization*.

3.2.1 Direct Rule Protection

In order to convert each α-discriminatory rule into an α-protective rule, based on the direct discriminatory measure (i.e., Definition 2), we should enforce the following inequality for each α-discriminatory rule $r' : A, B \rightarrow C$ in $\mathcal{MR}$, where A is a discriminatory item set:

$$elift(r') < \alpha. \tag{3}$$

By using the statement of the $elift$ Definition, Inequality (3) can be rewritten as

$$\frac{conf(r' : A, B \rightarrow C)}{conf(B \rightarrow C)} < \alpha. \tag{4}$$

Let us rewrite Inequality (4) in the following way:

$$conf(r' : A, B \rightarrow C) < \alpha \cdot conf(B \rightarrow C). \tag{5}$$

So, it is clear that Inequality (3) can be satisfied by decreasing the confidence of the α-discriminatory rule $r' : A, B \rightarrow C$ to a value less than the right-hand side of Inequality (5), without affecting the confidence of its base rule $B \rightarrow C$. A possible solution for decreasing

$$conf(r' : A, B \rightarrow C) = \frac{supp(A, B, C)}{supp(A, B)}, \tag{6}$$

is to perturb the discriminatory item set from $\neg A$ to A in the subset $\mathcal{DB}_c$ of all records of the original data set which completely support the rule $\neg A, B \rightarrow \neg C$ and have minimum impact on other rules; doing so increases the denominator of Expression (6) while keeping the numerator and $conf(B \rightarrow C)$ unaltered.

There is also another way to provide direct rule protection. Let us rewrite Inequality (4) in the following different way:

$$conf(B \rightarrow C) > \frac{conf(r' : A, B \rightarrow C)}{\alpha}. \tag{7}$$

It is clear that Inequality (3) can be satisfied by increasing the confidence of the base rule ($B \rightarrow C$) of the α-discriminatory rule $r' : A, B \rightarrow C$ to a value higher than the right-hand side of Inequality (7), without affecting the value of $conf(r' : A, B \rightarrow C)$. A possible solution for increasing Expression

$$conf(B \rightarrow C) = \frac{supp(B, C)}{supp(B)}, \tag{8}$$

is to perturb the class item from $\neg C$ to C in the subset $\mathcal{DB}_c$ of all records of the original data set which completely support the rule $\neg A, B \rightarrow \neg C$ and have minimum impact on other rules; doing so increases the numerator of Expression (8) while keeping the denominator and $conf(r' : A, B \rightarrow C)$ unaltered.

Therefore, there are two methods that could be applied for direct rule protection. One method (Method 1) changes the discriminatory item set in some records (e.g., gender changed from male to female in the records with granted credits) and the other method (Method 2) changes the class item in some records (e.g., from grant credit to deny credit in the records with male gender). Similar data transformation methods could be applied to obtain direct rule protection with respect to other measures (i.e., $slift$ and $olift$); see details in the Appendix, available in the online supplemental material.

3.2.2 Rule Generalization

Rule generalization is another data transformation method for direct discrimination prevention. It is based on the fact

that if each α-discriminatory rule $r' : A, B \rightarrow C$ in the database of decision rules was an instance of at least one nonredlining (legitimate) PND rule $r : D, B \rightarrow C$, the data set would be free of direct discrimination.

In rule generalization, we consider the relation between rules instead of discrimination measures. The following example illustrates this principle. Assume that a complainant claims discrimination against foreign workers among applicants for a job position. A classification rule $\{\text{Foreign worker} = \text{Yes}, \text{City} = \text{NYC}\} \rightarrow \text{Hire} = \text{No}$ with high $elift$ supports the complainant's claim. However, the decision maker could argue that this rule is an instance of a more general rule $\{\text{Experience} = \text{Low}, \text{City} = \text{NYC}\} \rightarrow \text{Hire} = \text{No}$. In other words, foreign workers are rejected because of their low experience, not just because they are foreign. The general rule rejecting low-experienced applicants is a legitimate one, because experience can be considered a genuine/legitimate requirement for some jobs. To formalize this dependency among rules (i.e., r' is an instance of r), Pedreschi et al. in [14] say that a PD classification rule $r' : A, B \rightarrow C$ is an instance of a PND rule $r : D, B \rightarrow C$ if rule r holds with the same or higher confidence, namely $conf(r : D, B \rightarrow C) \geq conf(r' : A, B \rightarrow C)$, and a case (record) satisfying discriminatory item set A in context B satisfies legitimate item set D as well, namely $conf(A, B \rightarrow D) = 1$. The two conditions can be relaxed in the following definition.

Definition 5. *Let $p \in [0, 1]$. A classification rule $r' : A, B \rightarrow C$ is a p-instance of $r : D, B \rightarrow C$ if both conditions below are true:*

- ***Condition 1:*** $\text{conf}(r) \geq p \cdot \text{conf}(r')$
- ***Condition 2:*** $\text{conf}(r'' : A, B \rightarrow D) \geq p$.

Then, if r' is a p-instance of r (where p is 1 or a value near 1), r' is free of direct discrimination. Based on this concept, we propose a data transformation method (i.e., rule generalization) to transform each α-discriminatory r' in $\mathcal{MR}$ into a p-instance of a legitimate rule. An important issue to perform rule generalization is to find a suitable PND rule ($r : D, B \rightarrow C$) or, equivalently, to find a suitable D (e.g., Experience = Low). Although choosing nonredlining rules, as done in this paper, is a way to obtain legitimate PND rules, sometimes it is not enough and a semantic hierarchy is needed to find the most suitable legitimate item set.

At any rate, rule generalization can be attempted for α-discriminatory rules r' for which there is at least one nonredlining PND rule r satisfying at least one of the two conditions of Definition 5. If any of the two conditions does not hold, the original data should be transformed for it to hold. Let us assume that Condition 2 is satisfied but Condition 1 is not. Based on the definition of p-instance, to satisfy the first condition of Definition 5, we should enforce for each α-discriminatory rule $r' : A, B \rightarrow C$ in $\mathcal{MR}$ the following inequality, with respect to its PND rule $r : D, B \rightarrow C$:

$$conf(r : D, B \rightarrow C) \geq p \cdot conf(r' : A, B \rightarrow C). \quad (9)$$

Let us rewrite Inequality (9) in the following way:

$$conf(r' : A, B \rightarrow C) \leq \frac{conf(r : D, B \rightarrow C)}{p}. \quad (10)$$

So, it is clear that Inequality (9) can be satisfied by decreasing the confidence of the α-discriminatory rule ($r' : A, B \rightarrow C$) to values less than the right-hand side of Inequality (10), without affecting the confidence of rule $r : D, B \rightarrow C$ or the satisfaction of Condition 2 of Definition 5. The confidence of r' was previously specified in Expression (6). A possible solution to decrease this confidence is to perturb the class item from C to $\neg C$ in the subset $\mathcal{DB}_c$ of all records in the original data set which completely support the rule $A, B, \neg D \rightarrow C$ and have minimum impact on other rules; doing so decreases the numerator of Expression (6) while keeping its denominator, $conf(r : D, B \rightarrow C)$ and also $conf(r'' : A, B \rightarrow D)$ (Condition 2 for rule generalization) unaltered.

Let us see what happens if Condition 1 of Definition 5 is satisfied but Condition 2 is not. In this case, based on the definition of p-instance, to satisfy Condition 2 we should enforce the following inequality for each α-discriminatory rule $r' : A, B \rightarrow C$ in $\mathcal{MR}$ with respect to its PND rule $r : D, B \rightarrow C$:

$$conf(r'' : A, B \rightarrow D) \geq p. \quad (11)$$

Inequality (11) must be satisfied by increasing the confidence of rule $r'' : A, B \rightarrow D$ to a value higher than p, without affecting the satisfaction of Condition 1. However, any effort at increasing the confidence of r'' impacts on the confidence of the r or r' rules and might threaten the satisfaction of Condition 1 of Definition 5; indeed, in order to increase the confidence of r'' we must either decrease $supp(A, B)$ (which increases $conf(r')$) or change $\neg D$ to D for those records satisfying A and B (which decreases $conf(r)$). Hence, rule generalization can only be applied if Condition 2 is satisfied without any data transformation.

To recap, we see that rule generalization can be achieved provided that Condition 2 is satisfied, because Condition 1 can be reached by changing the class item in some records (e.g., from "Hire no" to "Hire yes" in the records of foreign and high-experienced people in NYC city).

3.2.3 Direct Rule Protection and Rule Generalization

Since rule generalization might not be applicable for all α-discriminatory rules in $\mathcal{MR}$, rule generalization cannot be used alone for direct discrimination prevention and must be combined with direct rule protection. When applying both rule generalization and direct rule protection, α-discriminatory rules are divided into two groups:

- α-discriminatory rules r' for which there is at least one nonredlining PND rule r satisfying Condition 2 of Definition 5. For these rules, rule generalization is performed unless direct rule protection requires less data transformation (in which case direct rule protection is used).
- α-discriminatory rules such that there is no such PND rule. For these rules, direct rule protection is performed.

We use the following algorithm (step numbers below refer to pseudocode Algorithm 5 in the Appendix, available in the online supplemental material) to select the most appropriate discrimination prevention approach for each

α-discriminatory rule. First, for each α-discriminatory rule in $\mathcal{MR}$ of type $r' : A, B \rightarrow C$, a collection D_{pn} of nonredlining PND rules of type $r : D, B \rightarrow C$ is found (Step 2). Then, the conditions of Definition 5 are checked for each rule in D_{pn}, for $p \geq 0.8$ (Steps 4-18). Three cases arise depending on whether Conditions 1 and 2 hold:

- **Case 1: There is at least one rule $r \in D_{pn}$ such that both Conditions 1 and 2 of Definition 5 hold.** In this case, r' is a p-instance of r for $p \geq 0.8$ and no transformation is required (Steps 19-20).
- **Case 2: There is no rule in D_{pn} satisfying both Conditions 1 and 2 of Definition 5, but there is at least one rule satisfying Condition 2.** In this case (Step 23), the PND rule r_b in D_{pn} should be selected (Step 24) which requires the minimum data transformation to fulfill Condition 1. A smaller difference between the values of the two sides of Condition 1 for each r in D_{pn} indicates a smaller required data transformation. In this case, the α-discriminatory rule is transformed by rule generalization (Step 25).
- **Case 3: No rule in D_{pn} satisfies Condition 2 of Definition 5.** In this case (Step 21), rule generalization is not possible and direct rule protection should be performed (Step 22).

For the α-discriminatory rules to which rule generalization can be applied, it is possible that rule protection can be achieved with a smaller data transformation. For these rules the algorithm *should select the approach with minimum transformation* (Steps 31-36). The algorithm, pseudocoded in the Appendix, available in the online supplemental material, yields as output a database $\mathcal{TR}$ with all $r' \in \mathcal{MR}$, their respective rule r_b, and their respective discrimination prevention approaches ($TR_{r'}$).

3.3 Data Transformation for Indirect Discrimination

The proposed solution to prevent indirect discrimination is based on the fact that the data set of decision rules would be free of indirect discrimination if it contained no redlining rules. To achieve this, a suitable data transformation with minimum information loss should be applied in such a way that redlining rules are converted to nonredlining rules. We call this procedure *indirect rule protection* (IRP).

3.3.1 Indirect Rule Protection

In order to turn a redlining rule into an nonredlining rule, based on the indirect discriminatory measure (i.e., elb in Theorem 1), we should enforce the following inequality for each redlining rule $r : D, B \rightarrow C$ in $\mathcal{RR}$:

$$elb(\gamma, \delta) < \alpha. \tag{12}$$

By using the definitions stated when introducing elb in Theorem 1,[2] Inequality (12) can be rewritten as

$$\frac{\frac{conf(r_{b1})}{conf(r_{b2})}(conf(r_{b2}) + conf(r : D, B \rightarrow C) - 1)}{conf(B \rightarrow C)} < \alpha. \tag{13}$$

2. It is worth noting that β_1 and β_2 are lower bounds for $conf(r_{b1})$ and $conf(r_{b2})$, respectively, so it is correct if we replace β_1 and β_2 with $conf(r_{b1})$ and $conf(r_{b2})$ in the elb formulation.

Note that the discriminatory item set (i.e., A) is not removed from the original database $\mathcal{DB}$ and the rules $r_{b1} : A, B \rightarrow D$ and $r_{b2} : D, B \rightarrow A$ are obtained from $\mathcal{DB}$, so that their confidences might change as a result of data transformation for indirect discrimination prevention. Let us rewrite Inequality (13) in the following way:

$$conf(r_{b1} : A, B \rightarrow D) < \frac{\alpha \cdot conf(B \rightarrow C) \cdot conf(r_{b2})}{conf(r_{b2}) + conf(r : D, B \rightarrow C) - 1}. \tag{14}$$

Clearly, in this case Inequality (12) can be satisfied by decreasing the confidence of rule $r_{b1} : A, B \rightarrow D$ to values less than the right-hand side of Inequality (14) without affecting either the confidence of the redlining rule or the confidence of the $B \rightarrow C$ and r_{b2} rules. Since the values of both inequality sides are dependent, a transformation is required that decreases the left-hand side of the inequality without any impact on the right-hand side. A possible solution for decreasing

$$conf(A, B \rightarrow D) = \frac{supp(A, B, D)}{supp(A, B)}, \tag{15}$$

in Inequality (14) to the target value is to perturb the discriminatory item set from $\neg A$ to A in the subset $\mathcal{DB}_c$ of all records of the original data set which completely support the rule $\neg A, B, \neg D \rightarrow \neg C$ and have minimum impact on other rules; this increases the denominator of Expression (15) while keeping the numerator and $conf(B \rightarrow C)$, $conf(r_{b2} : D, B \rightarrow A)$, and $conf(r : D, B \rightarrow C)$ unaltered.

There is another way to provide indirect rule protection. Let us rewrite Inequality (13) as Inequality (16), where the confidences of r_{b1} and r_{b2} rules are not constant

$$conf(B \rightarrow C) > \frac{\frac{conf(r_{b1})}{conf(r_{b2})}(conf(r_{b2}) + conf(r : D, B \rightarrow C) - 1)}{\alpha}. \tag{16}$$

Clearly, in this case Inequality (12) can be satisfied by increasing the confidence of the base rule ($B \rightarrow C$) of the redlining rule $r : D, B \rightarrow C$ to values greater than the right-hand side of Inequality (16) without affecting either the confidence of the redlining rule or the confidence of the r_{b1} and r_{b2} rules. A possible solution for increasing Expression (8) in Inequality (16) to the target value is to perturb the class item from $\neg C$ to C in the subset $\mathcal{DB}_c$ of all records of the original data set which completely support the rule $\neg A, B, \neg D \rightarrow \neg C$ and have minimum impact on other rules; this increases the numerator of Expression (8) while keeping the denominator and $conf(r_{b1} : A, B \rightarrow D)$, $conf(r_{b2} : D, B \rightarrow A)$, and $conf(r : D, B \rightarrow C)$ unaltered.

Hence, like in direct rule protection, there are also two methods that could be applied for indirect rule protection. One method (Method 1) changes the discriminatory item set in some records (e.g., from nonforeign worker to foreign worker in the records of hired people in NYC city with Zip $\neq 10451$) and the other method (Method 2) changes the class item in some records (e.g., from "Hire yes" to "Hire no" in the records of nonforeign worker of people in NYC city with Zip $\neq 10451$).

TABLE 1
Direct and Indirect Rule Protection Methods

	Method 1	Method 2
Direct Rule Protection	$\neg A, B \rightarrow \neg C \Rightarrow A, B \rightarrow \neg C$	$\neg A, B \rightarrow \neg C \Rightarrow \neg A, B \rightarrow C$
Indirect Rule Protection	$\neg A, B, \neg D \rightarrow \neg C \Rightarrow A, B, \neg D \rightarrow \neg C$	$\neg A, B, \neg D \rightarrow \neg C \Rightarrow \neg A, B, \neg D \rightarrow C$

3.4 Data Transformation for Both Direct and Indirect Discrimination

We deal here with the key problem of transforming data with minimum information loss to prevent *at the same time* both direct and indirect discrimination. We will give a preprocessing solution to *simultaneous direct and indirect discrimination prevention*. First, we explain when direct and indirect discrimination could simultaneously occur. This depends on whether the original data set ($\mathcal{DB}$) contains discriminatory item sets or not. Two cases arise:

- Discriminatory item sets (i.e., A) did not exist in the original database $\mathcal{DB}$ or have previously been removed from it due to privacy constraints or for preventing discrimination. However, if background knowledge from publicly available data (e.g., census data) is available, indirect discrimination remains possible. In fact, in this case, only PND rules are extracted from $\mathcal{DB}$ so only indirect discrimination could happen.
- At least one discriminatory item set (i.e., A) is not removed from the original database ($\mathcal{DB}$). So it is clear that PD rules could be extracted from $\mathcal{DB}$ and direct discrimination could happen. However, in addition to direct discrimination, indirect discrimination might occur because of background knowledge obtained from $\mathcal{DB}$ itself due to the existence of nondiscriminatory items that are highly correlated with the sensitive (discriminatory) ones. Hence, in this case both direct and indirect discrimination could happen.

To provide both direct rule protection (DRP) and indirect rule protection (IRP) at the same time, an important point is the relation between the data transformation methods. Any data transformation to eliminate direct α-discriminatory rules should not produce new redlining rules or prevent the existing ones from being removed. Also any data transformation to eliminate redlining rules should not produce new direct α-discriminatory rules or prevent the existing ones from being removed.

For subsequent use in this section, we summarize in Table 1 the methods for DRP and IRP described in Sections 3.2.1 and 3.3.1 above. We can see in Table 1 that DRP and IRP operate the same kind of data transformation: in both cases Method 1 changes the discriminatory item set, whereas Method 2 changes the class item. Therefore, *in principle* any data transformation for DRP (resp. IRP) not only does not need to have a negative impact on IRP (resp. DRP), but both kinds of protection could even be beneficial to each other.

However, there is a difference between DRP and IRP: the set of records chosen for transformation. As shown in Table 1, in IRP the chosen records should not satisfy the D item set (chosen records are those with $\neg A, B, \neg D \rightarrow \neg C$), whereas DRP does not care about D at all (chosen records are those with $\neg A, B \rightarrow \neg C$). The following interactions between direct and indirect rule protection become apparent.

Lemma 1. *Method 1 for DRP cannot be used if simultaneous DRP and IRP are desired.*

Proof. Method 1 for DRP might undo the protection provided by Method 1 for IRP, as we next justify. Method 1 for DRP decreases $conf(A, B \rightarrow C)$ until the direct rule protection requirement (Inequality (5)) is met and Method 1 for IRP needs to decrease $conf(A, B \rightarrow D)$ until the indirect rule protection requirement is met (Inequality (14)). Assume that decreasing $conf(A, B \rightarrow C)$ to meet the direct rule protection requirement is achieved by changing y (how y is obtained will be discussed in Section 3.6) number of records with $\neg A, B, \neg C$ to records with $A, B, \neg C$ (as done by Method 1 for DRP). This actually could increase $conf(A, B \rightarrow D)$ if z among the changed records, with $z \leq y$, turn out to satisfy D. This increase can undo the protection provided by Method 1 for IRP (i.e., $conf(A, B \rightarrow D) < IRP_{req1}$, where

$$IRP_{req1} = \frac{\alpha \cdot conf(B \rightarrow C) \cdot conf(r_{b2})}{conf(r_{b2}) + conf(r : D, B \rightarrow C) - 1}$$

if the new value

$$conf(A, B \rightarrow D) = \frac{supp(A, B, D) + z}{supp(A, B) + y}$$

is greater than or equal to IRP_{req1}, which happens if $z \geq IRP_{req1} \cdot (supp(A, B) + Y) - supp(A, B, D)$. □

Lemma 2. *Method 2 for IRP is beneficial for Method 2 for DRP. On the other hand, Method 2 for DRP is at worst neutral for Method 2 for IRP.*

Proof. Method 2 for DRP and Method 2 for IRP are both aimed at increasing $conf(B \rightarrow C)$. In fact, Method 2 for IRP changes a subset of the records changed by Method 2 for DRP. This proves that Method 2 for IRP is beneficial for Method 2 for DRP. On the other hand, let us check that, in the worst case, Method 2 for DRP is neutral for Method 2 for IRP: such a worst case is the one in which all changed records satisfy D, which could result in increasing *both* sides of Inequality (16) by an equal amount (due to increasing $conf(B \rightarrow C)$ and $conf(D, B \rightarrow C)$); even in this case, there is no change in whatever protection is achieved by Method 2 for IRP. □

Thus, we conclude that Method 2 for DRP and Method 2 for IRP are the only methods among those described that can be applied to achieve simultaneous direct and indirect discrimination prevention. In addition, in the cases where either only direct or only indirect discrimination exist, there is no interference between the described methods: Method 1

for DRP, Method 2 for DRP, and Rule Generalization can be used to prevent direct discrimination; Method 1 for IRP and Method 2 for IRP can be used to prevent indirect discrimination. In what follows, we propose algorithms based on the described methods that cover direct and/or indirect discrimination prevention.

3.5 The Algorithms

We describe in this section our algorithms based on the direct and indirect discrimination prevention methods proposed in Sections 3.2, 3.3, and 3.4. There are some assumptions common to all algorithms in this section. First, we assume the class attribute in the original data set $\mathcal{DB}$ to be binary (e.g., denying or granting credit). Second, we consider classification rules with negative decision (e.g., denying credit) to be in $\mathcal{FR}$. Third, we assume the discriminatory item sets (i.e., A) and the nondiscriminatory item sets (i.e., D) to be binary or nonbinary categorical.

3.5.1 Direct Discrimination Prevention Algorithms

We start with direct rule protection. Algorithm 1 details Method 1 for DRP. For each direct α-discriminatory rule r' in $\mathcal{MR}$ (Step 3), after finding the subset $\mathcal{DB}_c$ (Step 5), records in $\mathcal{DB}_c$ should be changed until the direct rule protection requirement (Step 10) is met for each respective rule (Steps 10-14).

Algorithm 1. DIRECT RULE PROTECTION (METHOD 1)

1: Inputs: $\mathcal{DB}$, $\mathcal{FR}$, $\mathcal{MR}$, α, DI_s
2: Output: $\mathcal{DB}'$ (transformed data set)
3: **for** each $r' : A, B \rightarrow C \in \mathcal{MR}$ **do**
4: $\quad \mathcal{FR} \leftarrow \mathcal{FR} - \{r'\}$
5: $\quad \mathcal{DB}_c \leftarrow$ All records completely supporting $\neg A$, $B \rightarrow \neg C$
6: $\quad$ **for** each $db_c \in \mathcal{DB}_c$ **do**
7: $\quad\quad$ Compute $impact(db_c) = |\{r_a \in \mathcal{FR} | db_c$ supports the premise of $r_a\}|$
8: $\quad$ **end for**
9: $\quad$ Sort $\mathcal{DB}_c$ by ascending impact
10: $\quad$ **while** $conf(r') \geq \alpha \cdot conf(B \rightarrow C)$ **do**
11: $\quad\quad$ Select first record in $\mathcal{DB}_c$
12: $\quad\quad$ Modify discriminatory item set of db_c from $\neg A$ to A in $\mathcal{DB}$
13: $\quad\quad$ Recompute $conf(r')$
14: $\quad$ **end while**
15: **end for**
16: Output: $\mathcal{DB}' = \mathcal{DB}$

Among the records of $\mathcal{DB}_c$, one should change those with lowest impact on the other (α-protective or nonredlining) rules. Hence, for each record $db_c \in \mathcal{DB}_c$, the number of rules whose premise is supported by db_c is taken as the impact of db_c (Step 7), that is $impact(db_c)$; the rationale is that changing db_c impacts on the confidence of those rules. Then, the records db_c with minimum $impact(db_c)$ are selected for change (Step 9), with the aim of scoring well in terms of the utility measures proposed in the next section. We call this procedure (Steps 6-9) *impact minimization* and we reuse it in the pseudocodes of the rest of algorithms specified in this paper.

Algorithm 2 details Method 2 for DRP. The parts of Algorithm 2 to find subset $\mathcal{DB}_c$ and perform *impact minimization* (Step 4) are the same as in Algorithm 1. However, the transformation requirement that should be met for each α-discriminatory rule in $\mathcal{MR}$ (Step 5) and the kind of data transformation are different (Steps 5-9).

Algorithm 2. DIRECT RULE PROTECTION (METHOD 2)

1: Inputs: $\mathcal{DB}$, $\mathcal{FR}$, $\mathcal{MR}$, α, DI_s
2: Output: $\mathcal{DB}'$ (transformed data set)
3: **for** each $r' : A, B \rightarrow C \in \mathcal{MR}$ **do**
4: $\quad$ Steps 4-9 Algorithm 1
5: $\quad$ **while** $conf(B \rightarrow C) \leq \frac{conf(r')}{\alpha}$ **do**
6: $\quad\quad$ Select first record in $\mathcal{DB}_c$
7: $\quad\quad$ Modify the class item of db_c from $\neg C$ to C in $\mathcal{DB}$
8: $\quad\quad$ Recompute $conf(B \rightarrow C)$
9: $\quad$ **end while**
10: **end for**
11: Output: $\mathcal{DB}' = \mathcal{DB}$

As mentioned in Section 3.2.3, rule generalization cannot be applied alone for solving direct discrimination prevention, but it can be used in combination with Method 1 or Method 2 for DRP. In this case, after specifying the discrimination prevention method (i.e., direct rule protection or rule generalization) to be applied for each α-discriminatory rule based on the algorithm in Section 3.2.3, Algorithm 3 should be run to combine rule generalization and one of the two direct rule protection methods.

Algorithm 3. DIRECT RULE PROTECTION AND RULE GENERALIZATION

1: Inputs: $\mathcal{DB}$, $\mathcal{FR}$, $\mathcal{TR}$, $p \geq 0.8$, α, DI_s
2: Output: $\mathcal{DB}'$ (transformed data set)
3: **for** each $r' : A, B \rightarrow C \in \mathcal{TR}$ **do**
4: $\quad \mathcal{FR} \leftarrow \mathcal{FR} - \{r'\}$
5: $\quad$ **if** $TR_{r'} = RG$ **then**
6: $\quad\quad$ // Rule Generalization
7: $\quad\quad \mathcal{DB}_c \leftarrow$ All records completely supporting $A,\ B, \neg D \rightarrow C$
8: $\quad\quad$ Steps 6-9 Algorithm 1
9: $\quad\quad$ **while** $conf(r') > \frac{conf(r_b : D, B \rightarrow C)}{p}$ **do**
10: $\quad\quad\quad$ Select first record in $\mathcal{DB}_c$
11: $\quad\quad\quad$ Modify class item of db_c from C to $\neg C$ in $\mathcal{DB}$
12: $\quad\quad\quad$ Recompute $conf(r')$
13: $\quad\quad$ **end while**
14: $\quad$ **end if**
15: $\quad$ **if** $TR_{r'} = DRP$ **then**
16: $\quad\quad$ // Direct Rule Protection
17: $\quad\quad$ Steps 5-14 Algorithm 1 or Steps 4-9 Algorithm 2
18: $\quad$ **end if**
19: **end for**
20: Output: $\mathcal{DB}' = \mathcal{DB}$

Algorithm 3 takes as input $\mathcal{TR}$, which is the output of the algorithm in Section 3.2.3, containing all $r' \in \mathcal{MR}$ and their respective $TR_{r'}$ and r_b. For each α-discriminatory rule r' in $\mathcal{TR}$, if $TR_{r'}$ shows that rule generalization should be performed (Step 5), after determining the records that should be changed for *impact minimization* (Steps 7-8), these records should be changed until the rule generalization requirement is met (Steps 9-13). Also, if $TR_{r'}$ shows that

direct rule protection should be performed (Step 15), based on either Method 1 or Method 2, the relevant sections of Algorithms 1 or 2 are called, respectively (Step 17).

3.5.2 Indirect Discrimination Prevention Algorithms

A detailed algorithm implementing Method 2 for IRP is provided in [6], from which an algorithm implementing Method 1 for IRP can be easily derived. For the sake of brevity and due to similarity with the previous algorithms, we do not recall those two algorithms for IRP here.

3.5.3 Direct and Indirect Discrimination Prevention Algorithms

Algorithm 4 details our proposed data transformation method for simultaneous direct and indirect discrimination prevention. The algorithm starts with redlining rules. From each redlining rule ($r : X \rightarrow C$), more than one indirect α-discriminatory rule ($r' : A, B \rightarrow C$) might be generated because of two reasons: 1) existence of different ways to group the items in X into a context item set B and a nondiscriminatory item set D correlated to some discriminatory item set A; and 2) existence of more than one item in DI_s. Hence, as shown in Algorithm 4 (Step 5), given a redlining rule r, proper data transformation should be conducted for all indirect α-discriminatory rules $r' : (A \subseteq DI_s), (B \subseteq X) \rightarrow C$ ensuing from r.

Algorithm 4. DIRECT AND INDIRECT DISCRIMINATION PREVENTION

1: Inputs: $\mathcal{DB}$, $\mathcal{FR}$, $\mathcal{RR}$, $\mathcal{MR}$, α, DI_s
2: Output: $\mathcal{DB}'$ (transformed data set)
3: **for** each $r : X \rightarrow C \in \mathcal{RR}$, where $D, B \subseteq X$ **do**
4: $\gamma = conf(r)$
5: **for** each $r' : (A \subseteq DI_s), (B \subseteq X) \rightarrow C \in \mathcal{RR}$ **do**
6: $\beta_2 = conf(r_{b2} : X \rightarrow A)$
7: $\Delta_1 = supp(r_{b2} : X \rightarrow A)$
8: $\delta = conf(B \rightarrow C)$
9: $\Delta_2 = supp(B \rightarrow A)$
10: $\beta_1 = \frac{\Delta_1}{\Delta_2}$ $//conf(r_{b1} : A, B \rightarrow D)$
11: Find $\mathcal{DB}_c$: all records in $\mathcal{DB}$ that completely support $\neg A, B, \neg D \rightarrow \neg C$
12: Steps 6-9 Algorithm 1
13: **if** $r' \in \mathcal{MR}$ **then**
14: **while** $(\delta \leq \frac{\beta_1(\beta_2+\gamma-1)}{\beta_2 \cdot \alpha})$ and $(\delta \leq \frac{conf(r')}{\alpha})$ **do**
15: Select first record db_c in $\mathcal{DB}_c$
16: Modify the class item of db_c from $\neg C$ to C in $\mathcal{DB}$
17: Recompute $\delta = conf(B \rightarrow C)$
18: **end while**
19: **else**
20: **while** $\delta \leq \frac{\beta_1(\beta_2+\gamma-1)}{\beta_2 \cdot \alpha}$ **do**
21: Steps 15-17 Algorithm 4
22: **end while**
23: **end if**
24: **end for**
25: **end for**
26: **for** each $r' : (A, B \rightarrow C) \in \mathcal{MR} \setminus \mathcal{RR}$ **do**
27: $\delta = conf(B \rightarrow C)$
28: Find $\mathcal{DB}_c$: all records in $\mathcal{DB}$ that completely support $\neg A, B \rightarrow \neg C$
29: Step 12
30: **while** $(\delta \leq \frac{conf(r')}{\alpha})$ **do**
31: Steps 15-17 Algorithm 4
32: **end while**
33: **end for**
34: Output: $\mathcal{DB}' = \mathcal{DB}$

If some rules can be extracted from $\mathcal{DB}$ as both direct and indirect α-discriminatory rules, it means that there is overlap between $\mathcal{MR}$ and $\mathcal{RR}$; in such case, data transformation is performed until both the direct and the indirect rule protection requirements are satisfied (Steps 13-18). This is possible because, as we showed in Section 3.4, the same data transformation method (Method 2 consisting of changing the class item) can provide both DRP and IRP. However, if there is no overlap between $\mathcal{MR}$ and $\mathcal{RR}$, the data transformation is performed according to Method 2 for IRP, until the indirect discrimination prevention requirement is satisfied (Steps 19-23) for each indirect α-discriminatory rule ensuing from each redlining rule in $\mathcal{RR}$; this can be done without any negative impact on direct discrimination prevention, as justified in Section 3.4. Then, for each direct α-discriminatory rule $r' \in \mathcal{MR} \setminus \mathcal{RR}$ (that is only directly extracted from $\mathcal{DB}$), data transformation for satisfying the direct discrimination prevention requirement is performed (Steps 26-33), based on Method 2 for DRP; this can be done without any negative impact on indirect discrimination prevention, as justified in Section 3.4. Performing rule protection or generalization for each rule in $\mathcal{MR}$ by each of Algorithms 1-4 has no adverse effect on protection for other rules (i.e., rule protection at Step $i + x$ to make r' protective cannot turn into discriminatory a rule r made protective at Step i) because of the two following reasons: the kind of data transformation for each rule is the same (change the discriminatory item set or the class item of records) and there are no two α-discriminatory rules r and r' in $\mathcal{MR}$ such that $r = r'$.

3.6 Computational Cost

The computational cost of Algorithm 1 can be broken down as follows:

- Let m be the number of records in $\mathcal{DB}$. The cost of finding subset $\mathcal{DB}_c$ (Step 5) is $O(m)$.
- Let k be the number of rules in $\mathcal{FR}$ and h the number of records in subset $\mathcal{DB}_c$. The cost of computing $impact(db_c)$ for all records in $\mathcal{DB}_c$ (Steps 6-8) is $O(hk)$.
- The cost of sorting $\mathcal{DB}_c$ by ascending impact (Step 9) is $O(h \log h)$. Then, the cost of the *impact minimization* procedure (Steps 6-9) in all algorithms is $O(hk + h \log h)$.
- During each iteration of the inner loop (Step 10), the number of records supporting the premise of rule $r' : A, B \rightarrow C$ is increased by one. After d iterations, the confidence of $r' : A, B \rightarrow C$ will be $conf(r' : A, B \rightarrow C)^{(d)} = \frac{N_{ABC}}{N_{AB}+d}$, where N_{ABC} is the number of records supporting rule r' and N_{AB} is the number of records supporting the premise of rule r'. If we let $DRP_{req1} = \alpha \cdot conf(B \rightarrow C)$, the inner loop (Step 10) is iterated until $conf(r' : A, B \rightarrow C)^{(d)} < DRP_{req1}$ or

equivalently $\frac{N_{ABC}}{N_{AB}+d} < DRP_{req1}$. This inequality can be rewritten as

$$d > \left(\frac{N_{ABC}}{DRP_{req1}} - N_{BC}\right).$$

From this last inequality we can derive that $d = \lceil \frac{N_{ABC}}{DRP_{req1}} - N_{BC} \rceil$. Hence, iterations in the inner loop (Step 10) will stop as soon as the first integer value greater than (or equal) $\frac{N_{ABC}}{DRP_{req1}} - N_{BC}$ is reached. Then, the cost spent on the inner loop to satisfy the direct rule protection requirement (Steps 10-14) will be

$$O\left(m * \left\lceil \frac{N_{ABC}}{DRP_{req1}} - N_{BC} \right\rceil\right).$$

Therefore, assuming n is the number of α-discriminatory rules in $\mathcal{MR}$ (Step 3), the total computational time of Algorithm 1 is bounded by $O(n*\{m + hk + h\log h + dm\})$, where $d = \lceil \frac{N_{ABC}}{DRP_{req1}} - N_{BC} \rceil$.

The *impact minimization* procedure substantially increases the complexity. Without computing the impact, the time complexity of Algorithm 1 decreases to $O(n*\{m + dm\})$. In addition, it is clear that the execution time of Algorithm 1 increases linearly with the number m of original data records as well as the number k of frequent classification rules and the number n of α-discriminatory rules.

The computational cost of the other algorithms can be computed similarly, with some small differences. In summary, the total computational time of Algorithm 2 is also bounded by $O(n*\{m + hk + h\log h + dm\})$, where $d = \lceil (N_B * DRP_{req2}) - N_{BC} \rceil$, N_{BC} is the number of records supporting rule $B \rightarrow C$, N_B is the number of records supporting item set B and $DRP_{req2} = \frac{conf(r')}{\alpha}$. The computational cost of Algorithm 3 is the same as the last ones with the difference that $d = \lceil N_{ABC} - (RG_{req} * N_{AB}) \rceil$, where $RG_{req} = \frac{conf(r_b)}{p}$, or $d = \lceil (N_B * DRP_{req2}) - N_{BC} \rceil$, depending on whether rule generalization or direct rule protection is performed.

Finally, assuming f is the number of indirect α-discriminatory rules in $\mathcal{RR}$ and n is the number of direct α-discriminatory rules in $\mathcal{MR}$ that no longer exist in $\mathcal{RR}$, the total computational time of Algorithm 4 is bounded by

$$O((f+n) * \{m + hk + h\log h + dm\}),$$

where $d = \lceil (N_B * max_{req}) - N_{BC} \rceil$ and

$$max_{req} = \max\left(\frac{\beta_1(\beta_2 + \gamma - 1)}{\beta_2 \cdot \alpha}, \frac{conf(r')}{\alpha}\right).$$

4 Experiments

This section presents the experimental evaluation of the proposed direct and/or indirect discrimination prevention approaches and algorithms. To obtain $\mathcal{FR}$ and $\mathcal{BK}$ we used the Apriori algorithm [1], which is a common algorithm to extract frequent rules. All algorithms and utility measures were implemented using the C# programming language. The tests were performed on an 2.27 GHz Intel Core i3 machine, equipped with 4 GB of RAM, and running under Windows 7 Professional.

First, we describe the data sets used in our experiments. Then, we introduce the new utility measures we propose to evaluate direct and indirect discrimination prevention methods in terms of their success at discrimination removal and impact on data quality. Finally, we present the evaluation results of the different methods and also the comparison between them.

4.1 Data Sets

Adult data set: We used the Adult data set [10], also known as Census Income, in our experiments. This data set consists of 48,842 records, split into a "train" part with 32,561 records and a "test" part with 16,281 records. The data set has 14 attributes (without class attribute). We used the "train" part in our experiments. The prediction task associated with the Adult data set is to determine whether a person makes more than 50K$ a year based on census and demographic information about people. The data set contains both categorical and numerical attributes.

For our experiments with the Adult data set, we set $DI_s = \{\text{Sex} = \text{Female}, \text{Age} = \text{Young}\}$. Although the Age attribute in the Adult data set is numerical, we converted it to categorical by partitioning its domain into two fixed intervals: Age ≤ 30 was renamed as Young and Age > 30 was renamed as old.

German credit data set: we also used the German Credit data set [11]. This data set consists of 1,000 records and 20 attributes (without class attribute) of bank account holders. This is a well-known real-life data set, containing both numerical and categorical attributes. It has been frequently used in the antidiscrimination literature [12], [7]. The class attribute in the German Credit data set takes values representing good or bad classification of the bank account holders. For our experiments with this data set, we set $DI_s = \{\text{Foreign worker} = \text{Yes}, \text{Personal Status} = \text{Female and not Single}, \text{Age} = \text{Old}\}$; (cut-off for Age = Old: 50 years old).

4.2 Utility Measures

Our proposed techniques should be evaluated based on two aspects. On the one hand, we need to measure the success of the method in removing all evidence of direct and/or indirect discrimination from the original data set; on the other hand, we need to measure the impact of the method in terms of information loss (i.e., data quality loss). To measure discrimination removal, four metrics were used:

- **Direct discrimination prevention degree (DDPD)**. This measure quantifies the percentage of α-discriminatory rules that are no longer α-discriminatory in the transformed data set. We define DDPD as

$$DDPD = \frac{|\mathcal{MR}| - |\mathcal{MR}'|}{|\mathcal{MR}|},$$

where $\mathcal{MR}$ is the database of α-discriminatory rules extracted from $\mathcal{DB}$ and $\mathcal{MR}'$ is the database of α-discriminatory rules extracted from the transformed data set $\mathcal{DB}'$. Note that $|\cdot|$ is the cardinality operator.
- **Direct discrimination protection preservation (DDPP)**. This measure quantifies the percentage of

TABLE 2
Adult Data Set: Utility Measures for Minimum Support 2 Percent and Confidence 10 Percent for All the Methods

Methods	α	p	No. Redlining Rules	No. Indirect α-Disc. Rules	No. Direct α-Disc. Rules	Discrimination Removal				Data Quality	
						Direct		Indirect			
						DDPD	DDPP	IDPD	IDPP	MC	GC
Removing. Disc. Attributes	n.a.	n.a	n.a.	n.a	n.a.	n.a.	n.a.	n.a.	n.a.	66.08	0
DRP (Method 1)	1.2	n.a	n.a.	n.a	274	100	100	n.a.	n.a.	4.16	4.13
DRP (Method 2)	1.2	n.a.	n.a.	n.a.	274	100	100	n.a.	n.a.	0	0
DRP (Method 1) + RG	1.2	0.9	n.a.	n.a.	274	100	100	n.a.	n.a.	4.1	4.1
DRP (Method 2) + RG	1.2	0.9	n.a.	n.a.	274	91.58	100	n.a.	n.a.	0	0
IRP (Method 1)	1.1	n.a.	21	30	n.a.	n.a.	n.a.	100	100	0.54	0.38
IRP (Method 2)	1.1	n.a.	21	30	n.a.	n.a.	n.a.	100	100	0	0
DRP(Method 2) + IRP(Method 2)	1.1	n.a.	21	30	280	100	100	100	100	0	0
No of Freq. Class. Rules: 5,092						No. of Back. Know. Rules: 2089					

Value "n.a." denotes that the respective measure is not applicable.

the α-protective rules in the original data set that remain α-protective in the transformed data set. It is defined as

$$DDPP = \frac{|\mathcal{PR} \bigcap \mathcal{PR}'|}{|\mathcal{PR}|},$$

where $\mathcal{PR}$ is the database of α-protective rules extracted from the original data set $\mathcal{DB}$ and $\mathcal{PR}'$ is the database of α-protective rules extracted from the transformed data set $\mathcal{DB}'$.

- **Indirect discrimination prevention degree (IDPD).** This measure quantifies the percentage of redlining rules that are no longer redlining in the transformed data set. It is defined like DDPD but substituting $\mathcal{MR}$ and $\mathcal{MR}'$ with the database of redlining rules extracted from $\mathcal{DB}$ and $\mathcal{DB}'$, respectively.
- **Indirect discrimination protection preservation (IDPP).** This measure quantifies the percentage of nonredlining rules in the original data set that remain nonredlining in the transformed data set. It is defined like DDPP but substituting $\mathcal{PR}$ and $\mathcal{PR}'$ with the database of nonredlining extracted from $\mathcal{DB}$ and $\mathcal{DB}'$, respectively.

Since the above measures are used to evaluate the success of the proposed method in direct and indirect discrimination prevention, ideally their value should be 100 percent. To measure data quality, we use two metrics proposed in the literature as information loss measures in the context of rule hiding for privacy-preserving data mining (PPDM) [19].

- **Misses cost** (MC). This measure quantifies the percentage of rules among those extractable from the original data set that cannot be extracted from the transformed data set (side effect of the transformation process).
- **Ghost cost** (GC). This measure quantifies the percentage of the rules among those extractable from the transformed data set that were not extractable from the original data set (side effect of the transformation process).

MC and GC should ideally be 0 percent. However, MC and GC may not be 0 percent as a side effect of the transformation process.

4.3 Evaluation of the Methods

We implemented the algorithms for all proposed methods for direct and/or indirect discrimination prevention, and we evaluated them in terms of the proposed utility measures. We report the performance results in this section.

Tables 2 and 3 show the utility scores obtained by our methods on the Adult data set and the German Credit data set, respectively. Within each table, the first row relates to the simple approach of deleting discriminatory attributes, the next four rows relate to direct discrimination prevention methods, the next two ones relate to indirect discrimination prevention methods and the last one relates to the combination of direct and indirect discrimination.

Table 2 shows the results for minimum support 2 percent and minimum confidence 10 percent. Table 3 shows the results for minimum support 5 percent and minimum confidence 10 percent. In Tables 2 and 3, the results of direct

TABLE 3
German Credit Data Set: Utility Measures for Minimum Support 5 Percent and Confidence 10 Percent for all Methods

Methods	α	p	No. Redlining Rules	No. Indirect α-Disc. Rules	No. Direct α-Disc. Rules	Discrimination Removal				Data Quality	
						Direct		Indirect			
						DDPD	DDPP	IDPD	IDPP	MC	GC
Removing. Disc. Attributes	n.a.	n.a	n.a.	n.a	n.a.	n.a.	n.a.	n.a.	n.a.	64.35	0
DRP (Method 1)	1.2	n.a	n.a.	n.a	991	100	100	n.a.	n.a.	15.44	13.52
DRP (Method 2)	1.2	n.a.	n.a.	n.a.	991	100	100	n.a.	n.a.	0	4.06
DRP (Method 1) + RG	1.2	0.9	n.a.	n.a.	991	100	100	n.a.	n.a.	13.34	12.01
DRP (Method 2) + RG	1.2	0.9	n.a.	n.a.	991	100	100	n.a.	n.a.	0.01	4.06
IRP (Method 1)	1	n.a.	37	42	n.a.	n.a.	n.a.	100	100	1.62	1.47
IRP (Method 2)	1	n.a.	37	42	n.a.	n.a.	n.a.	100	100	0	0.96
DRP(Method 2) + IRP(Method 2)	1	n.a.	37	42	499	99.97	100	100	100	0	2.07
No of Freq. Class. Rules: 32,340						No. of Back. Know. Rules: 22,763					

Value "n.a." denotes that the respective measure is not applicable.

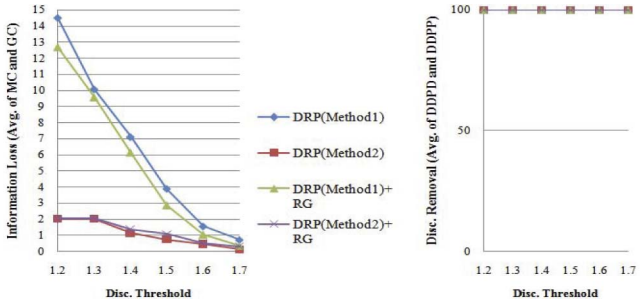

Fig. 1. Information loss (left) and discrimination removal degree (right) for direct discrimination prevention methods for $\alpha \in [1.2, 1.7]$.

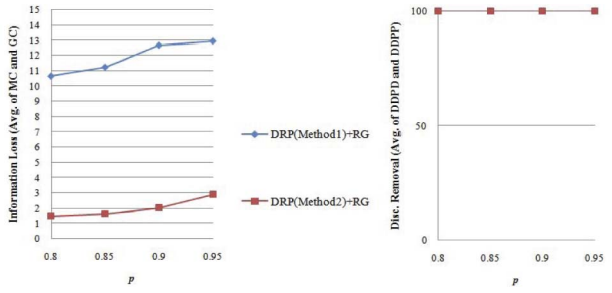

Fig. 2. Information loss (left) and discrimination removal (right) degree for direct discrimination prevention methods for $p \in [0.8, 0.95]$.

discrimination prevention methods are reported for discriminatory threshold $\alpha = 1.2$ and, in the cases where direct rule protection is applied in combination with rule generalization, we used $p = 0.9$, and $DI_s = \{\text{Sex} = \text{Female}, \text{Age} = \text{Young}\}$ in the Adult data set, and $DI_s = \{\text{Foreign worker} = \text{Yes}, \text{Personal Status} = \text{Female and not Single}, \text{Age} = \text{Old}\}$ in the German Credit data set. In addition, in Table 2, the results of the indirect discrimination prevention methods and both direct and indirect discrimination prevention are reported for discriminatory threshold $\alpha = 1.1$ and $DI_s = \{\text{Sex} = \text{Female}, \text{Age} = \text{Young}\}$; in Table 3, these results are reported for $\alpha = 1$ and $DI_s = \{\text{Foreign worker} = \text{Yes}\}$).

We selected the discriminatory threshold values and DI_s for each data set in such a way that the number of redlining rules and α-discriminatory rules extracted from $\mathcal{DB}$ could be suitable to test all our methods. In addition to the scores of utility measures, the number of redlining rules, the number of indirect α-discriminatory rules, and the number of direct α-discriminatory rules are also reported in Tables 2 and 3. These tables also show the number of frequent classification rules found, as well as the number of background knowledge rules related to this experiment.

As shown in Tables 2 and 3, we get very good results for all methods in terms of discrimination removal: DDPD, DDPP, IDPD, IDPP are near 100 percent for both data sets. In terms of data quality, the best results for direct discrimination prevention are obtained with Method 2 for DRP or Method 2 for DRP combined with Rule Generalization. The best results for indirect discrimination prevention are obtained with Method 2 for IRP. This shows that lower information loss is obtained with the methods changing the class item (i.e., Method 2) than with those changing the discriminatory item set (i.e., Method 1). As mentioned above, in direct discrimination prevention, rule generalization cannot be applied alone and must be applied in combination with direct rule protection; however, direct rule protection can be applied alone. The results in the last row of the above tables (i.e., Method 2 for DRP + Method 2 for IRP) based on Algorithm 4 for the case of simultaneous direct and indirect discrimination demonstrate that the proposed solution achieves a high degree of simultaneous direct and indirect discrimination removal with very little information loss.

For all methods, Tables 2 and 3 show that we obtained lower information loss in terms of MC and GC in the Adult data set than in the German Credit data set. In terms of discrimination removal, results on both data sets were almost the same. In addition, the highest value of information loss is obtained by the simple approach of removing discriminatory attributes (first row of each table): as it could be expected, entirely suppressing the discriminatory attributes is much more information damaging than modifying the values of these attributes in a few records.

After the above general results and comparison between methods, we now present more specific results on each method for different parameters α and p. Fig. 1 shows on the left the degree of information loss (as average of MC and GC) and on the right the degree of discrimination removal (as average of DDPD and DDPP) of direct discrimination prevention methods for the German Credit data set when the value of the discriminatory threshold α varies from 1.2 to 1.7, p is 0.9, the minimum support is 5 percent and the minimum confidence is 10 percent. The number of direct α-discriminatory rules extracted from the data set is 991 for $\alpha = 1.2$, 415 for $\alpha = 1.3$, 207 for $\alpha = 1.4$, 120 for $\alpha = 1.5$, 63 for $\alpha = 1.6$, and 30 for $\alpha = 1.7$, respectively. As shown in Fig. 1, the degree of discrimination removal provided by all methods for different values of α is also 100 percent. However, the degree of information loss decreases substantially as α increases; the reason is that, as α increases, the number of α-discriminatory rules to be dealt with decreases. In addition, as shown in Fig. 1, the lowest information loss for most values of α is obtained by Method 2 for DRP.

In addition, to demonstrate the impact of varying p on the utility measures in the methods using Rule Generalization, Fig. 2(left) shows the degree of information loss and Fig. 2(right) shows the degree of discrimination removal for different values of $p(0.8, 0.85, 0.9, 0.95)$ and $\alpha = 1.2$ for the German Credit data set. Although the values of DDPD and DDPP achieved for different values of p remain almost the same, increasing the value of p leads to an increase of MC and GC because, to cope with the rule generalization requirements, more data records must be changed.

Tables 4 and 5 show the utility measures obtained by running Algorithm 4 to achieve simultaneous direct and indirect discrimination prevention (i.e., Method 2 for DRP + Method 2 for IRP) on the Adult and German credit data sets, respectively. In Table 4, the results are reported for different values of $\alpha \in [1, 1.5]$; in Table 5 different values of $\alpha \in [1, 1.4]$ are considered. We selected these α intervals in such a way that, with respect to the predetermined discriminatory items in this experiment for the Adult data set (i.e., $DI_s = \{\text{Sex} = \text{Female}, \text{Age} = \text{Young}\}$) and the German Credit data set (i.e., $DI_s = \{\text{Foreign worker} = \text{Yes}\}$), both direct α-discriminatory and redlining rules could be extracted. The reason is

TABLE 4
Adult Data Set: Utility Measures for Minimum Support 2 Percent and Confidence 10 Percent for Direct and Indirect Rule Protection; Columns Show the Results for Different Values of α

α	No. of redlining rules	No. of Indirect α-Disc. Rules	No. of Direct α-Disc. rules	Discrimination Removal				Data Quality	
				Direct		Indirect			
				DDPD	DDPP	IDPD	IDPP	MC	GC
α=1	43	71	804	89.45	100	95.35	100	0	0.03
α=1.1	21	30	280	100	100	100	100	0	0
α=1.2	9	14	140	100	100	100	100	0	0
α=1.3	0	0	67	100	100	n.a.	100	0	0.01
α=1.4	0	0	32	100	100	n.a.	100	0	0
α=1.5	0	0	7	100	100	n.a.	100	0	0
No of Freq. Class. Rules: 5,092				No. of Back. Know. Rules: 2,089					

Value "n.a." denotes that the respective measure is not applicable.

TABLE 5
German Credit Data Set: Utility Measures for Minimum Support 5 Percent and Confidence 10 Percent for Direct and Indirect Rule Protection; Columns Show the Results for Different Values of α

α	No. of redlining rules	No. of Indirect α-Disc. Rules	No. of Direct α-Disc. rules	Discrimination Removal				Data Quality	
				Direct		Indirect			
				DDPD	DDPP	IDPD	IDPP	MC	GC
α=1	37	42	499	99.97	100	100	100	0	2.07
α=1.1	0	0	312	100	100	n.a.	100	0	2.07
α=1.2	0	0	26	100	100	n.a.	100	0	1.01
α=1.3	0	0	14	100	100	n.a.	100	0	1.01
α=1.4	0	0	9	100	100	n.a.	100	0	0.69
No of Freq. Class. Rules: 32,340				No. of Back. Know. Rules: 22,763					

Value "n.a." denotes that the respective measure is not applicable.

that we need to detect some cases with both direct and indirect discrimination to be able to test our method. Moreover, we restricted the lower bound to limit the number of direct α-discriminatory and redlining rules. In addition to utility measures, the number of redlining rules, the number of indirect α-discriminatory rules, and the number of direct α-discriminatory rules are also reported for different values of α.

The values of both direct discrimination removal measures (i.e., DDPD and DDPP) and indirect discrimination removal measures (i.e., IDPD and IDPP) shown in Tables 4 and 5 demonstrate that the proposed solution achieves a high degree of both direct and indirect discrimination prevention for different values of the discriminatory threshold. The important point is that, by applying the proposed method, we get good results for both direct and indirect discrimination prevention at the same time. In addition, the values of MC and GC demonstrate that the proposed solution incurs low information loss.

Tables 4 and 5 show that we obtained lower information loss in terms of the GC measure in the Adult data set than in the German Credit data set. Another remark on these tables is that, although no redlining rules are detected in the Adult data set for $\alpha \geq 1.3$ and in the German Credit data set for $\alpha \geq 1.1$, the IDPP measure is computed and reported to show that in the cases where only direct discrimination exists, the elimination of direct discrimination by Algorithm 4 does not have a negative impact on indirect discrimination (i.e., nonredlining rules do not become redlining rules).

Fig. 3 illustrates the effect of the *impact minimization* procedure, described in Section 3.5.1, on execution times and information loss of Method 1 for DRP, respectively. As shown in this figure (right) *impact minimization* has a noticeable effect on information loss (decreasing MC and GC). However, as discussed in Section 3.6 and shown in Fig. 3(left), *impact minimization* substantially increases the execution time of the algorithm. For other methods, the same happens. Fig. 3(left) also shows that, by increasing α, the number of α-discriminatory rules and hence the execution time are decreased. Additional experiments are presented in the Appendix, available in the online supplemental material, to show the effect of varying the minimum support and the minimum confidence on the proposed techniques.

5 Conclusions and Future Work

Along with privacy, discrimination is a very important issue when considering the legal and ethical aspects of data mining. It is more than obvious that most people do not want to be discriminated because of their gender, religion, nationality, age, and so on, especially when those attributes are used for making decisions about them like giving them a job, loan, insurance, etc.

The purpose of this paper was to develop a new preprocessing discrimination prevention methodology including different data transformation methods that can prevent

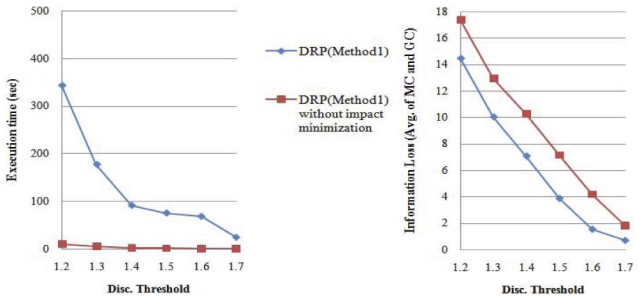

Fig. 3. Execution times (left) and Information loss degree (right) of Method 1 for DRP for $\alpha \in [1.2, 1.7]$ with and without impact minimization.

direct discrimination, indirect discrimination or both of them at the same time. To attain this objective, the first step is to measure discrimination and identify categories and groups of individuals that have been directly and/or indirectly discriminated in the decision-making processes; the second step is to transform data in the proper way to remove all those discriminatory biases. Finally, discrimination-free data models can be produced from the transformed data set without seriously damaging data quality. The experimental results reported demonstrate that the proposed techniques are quite successful in both goals of removing discrimination and preserving data quality.

The perception of discrimination, just like the perception of privacy, strongly depends on the legal and cultural conventions of a society. Although we argued that discrimination measures based on $elift$ and elb are reasonable, as future work we intend to explore measures of discrimination different from the ones considered in this paper. This will require us to further study the legal literature on discrimination in several countries and, if substantially different discrimination definitions and/or measures were to be found, new data transformation methods would need to be designed.

Last but not least, we want to explore the relationship between discrimination prevention and privacy preservation in data mining. It would be extremely interesting to find synergies between rule hiding for privacy-preserving data mining and rule hiding for discrimination removal. Just as we were able to show that indirect discrimination removal can help direct discrimination removal, it remains to be seen whether privacy protection can help antidiscrimination or viceversa. The connection with current privacy models, like differential privacy, is also an intriguing research avenue.

Disclaimer and Acknowledgments

The authors are with the UNESCO Chair in Data Privacy, but this paper does not commit UNESCO. Thanks go to Antoni Martínez-Ballesté and three anonymous referees for their help. Partial support is acknowledged from Spanish projects TSI2007-65406-C03-01, TIN2011-27076-C03-01 and CONSOLIDER CSD2007-00004, Catalan project 2009SGR1135, and European FP7 project "DwB." The second author is an ICREA Acadèmia Researcher.

References

[1] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules in Large Databases," *Proc. 20th Int'l Conf. Very Large Data Bases,* pp. 487-499, 1994.

[2] T. Calders and S. Verwer, "Three Naive Bayes Approaches for Discrimination-Free Classification," *Data Mining and Knowledge Discovery,* vol. 21, no. 2, pp. 277-292, 2010.

[3] European Commission, "EU Directive 2004/113/EC on Anti-Discrimination," http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2004:373:0037:0043:EN:PDF, 2004.

[4] European Commission, "EU Directive 2006/54/EC on Anti-Discrimination," http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2006:204:0023:0036:en:PDF, 2006.

[5] S. Hajian, J. Domingo-Ferrer, and A. Martínez-Ballesté, "Discrimination Prevention in Data Mining for Intrusion and Crime Detection," *Proc. IEEE Symp. Computational Intelligence in Cyber Security (CICS '11),* pp. 47-54, 2011.

[6] S. Hajian, J. Domingo-Ferrer, and A. Martínez-Ballesté, "Rule Protection for Indirect Discrimination Prevention in Data Mining," *Proc. Eighth Int'l Conf. Modeling Decisions for Artificial Intelligence (MDAI '11),* pp. 211-222, 2011.

[7] F. Kamiran and T. Calders, "Classification without Discrimination," *Proc. IEEE Second Int'l Conf. Computer, Control and Comm. (IC4 '09),* 2009.

[8] F. Kamiran and T. Calders, "Classification with no Discrimination by Preferential Sampling," *Proc. 19th Machine Learning Conf. Belgium and The Netherlands,* 2010.

[9] F. Kamiran, T. Calders, and M. Pechenizkiy, "Discrimination Aware Decision Tree Learning," *Proc. IEEE Int'l Conf. Data Mining (ICDM '10),* pp. 869-874, 2010.

[10] R. Kohavi and B. Becker, "UCI Repository of Machine Learning Databases," http://archive.ics.uci.edu/ml/datasets/Adult, 1996.

[11] D.J. Newman, S. Hettich, C.L. Blake, and C.J. Merz, "UCI Repository of Machine Learning Databases," http://archive.ics.uci.edu/ml, 1998.

[12] D. Pedreschi, S. Ruggieri, and F. Turini, "Discrimination-Aware Data Mining," *Proc. 14th ACM Int'l Conf. Knowledge Discovery and Data Mining (KDD '08),* pp. 560-568, 2008.

[13] D. Pedreschi, S. Ruggieri, and F. Turini, "Measuring Discrimination in Socially-Sensitive Decision Records," *Proc. Ninth SIAM Data Mining Conf. (SDM '09),* pp. 581-592, 2009.

[14] D. Pedreschi, S. Ruggieri, and F. Turini, "Integrating Induction and Deduction for Finding Evidence of Discrimination," *Proc. 12th ACM Int'l Conf. Artificial Intelligence and Law (ICAIL '09),* pp. 157-166, 2009.

[15] S. Ruggieri, D. Pedreschi, and F. Turini, "Data Mining for Discrimination Discovery," *ACM Trans. Knowledge Discovery from Data,* vol. 4, no. 2, article 9, 2010.

[16] S. Ruggieri, D. Pedreschi, and F. Turini, "DCUBE: Discrimination Discovery in Databases," *Proc. ACM Int'l Conf. Management of Data (SIGMOD '10),* pp. 1127-1130, 2010.

[17] P.N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining.* Addison-Wesley, 2006.

[18] United States Congress, *US Equal Pay Act,* http://archive.eeoc.gov/epa/anniversary/epa-40.html, 1963.

[19] V. Verykios and A. Gkoulalas-Divanis, "A Survey of Association Rule Hiding Methods for Privacy," *Privacy-Preserving Data Mining: Models and Algorithms,* C.C. Aggarwal and P.S. Yu, eds., Springer, 2008.

Sara Hajian is working toward the PhD degree in data security and privacy at the Universitat Rovira i Virgili, and the MSc degree in computer science from Iran University of Science and Technology in 2008.

Josep Domingo-Ferrer received the MSc and PhD degrees in computer science from the Universitat Autnoma de Barcelona in 1988 and 1991, respectively, and the MSc degree in mathematics. He is a distinguished professor at the Universitat Rovira i Virgili, where he holds the UNESCO chair in Data Privacy. More information can be found at http://crises-deim.urv.cat/jdomingo. He is a fellow of the IEEE.