Rule Protection for Indirect Discrimination Prevention in Data Mining

Sara Hajian, Josep Domingo-Ferrer, and Antoni Martínez-Ballesté

Universitat Rovira i Virgili
Department of Computer Engineering and Mathematics
UNESCO Chair in Data Privacy
{sara.hajian,josep.domingo,antoni.martinez}@urv.cat

Abstract. Services in the information society allow automatically and routinely collecting large amounts of data. Those data are often used to train classification rules in view of making automated decisions, like loan granting/denial, insurance premium computation, etc. If the training datasets are biased in what regards sensitive attributes like gender, race, religion, etc., discriminatory decisions may ensue. Direct discrimination occurs when decisions are made based on biased sensitive attributes. Indirect discrimination occurs when decisions are made based on non-sensitive attributes which are strongly correlated with biased sensitive attributes. This paper discusses how to clean training datasets and outsourced datasets in such a way that legitimate classification rules can still be extracted but indirectly discriminating rules cannot.

Keywords: Anti-discrimination, Indirect discrimination, Discrimination prevention, Data mining, Privacy.

1 Introduction

Automated data collection in the information society facilitates automating decision making as well. Superficially, 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 trained on the collected data. If those training data are biased, the learned model will be biased. For example, if the data are used to train classification rules for loan granting and most of the Brazilians in the training dataset were denied their loans, the leaned rules will also show biased behavior toward Brazilian and it is a discriminatory reason for loan denial. Unfairly treating people on the basis of their belonging to a specific group (race, ideology, gender, etc.) is known as discrimination and is legally punished in many democratic countries.

1.1 Discrimination-Aware Data Mining

The literature in law and social sciences distinguishes direct and indirect discrimination (the latter is also called systematic). Direct discrimination consists

V. Torra et al. (Eds.): MDAI 2011, LNAI 6820, pp. 211–222, 2011.

of rules or procedures that explicitly impose "disproportionate burdens" on minority or disadvantaged groups (*i.e.* discriminatory rules) based on sensitive attributes related to group membership (*i.e.* discriminatory attributes). Indirect discrimination consists of rules or procedures that, while not explicitly mentioning discriminatory attributes, impose the same disproportionate burdens, intentionally or unintentionally. This effect and its exploitation is often referred to as *redlining* and indirectly discriminating rules can be called *redlining rules* [1]. The term "redlining" was invented in the late 1960s by community activists in Chicago [2]. The authors of [1] also support this claim: even after removing the discriminatory attributes from the dataset, discrimination persists because there may be other attributes that are highly correlated with the sensitive (discriminatory) ones or there may be background knowledge from publicly available data (*e.g.* census data) allowing inference of the discriminatory knowledge (rules).

The existing literature on anti-discrimination in computer science mainly elaborates on data mining models and related techniques. Some proposals are oriented to the *discovery* and *measure* of discrimination [1,3,4,7]. Others deal with the *prevention* of discrimination. Although some methods have been proposed, discrimination prevention stays a largely unexplored research avenue. Clearly, a straightforward way to handle discrimination prevention would consist of removing discriminatory attributes from the dataset. However in terms of indirect discrimination, as stated in [1,2] there may be other attributes that are highly correlated with the sensitive ones or there may be background knowledge from publicly available data that allow for the inference of discrimination rules. Hence, one might decide to remove also those highly correlated attributes as well. Although this would solve the discrimination problem, in this process much useful information would be lost. Hence, one challenge regarding discrimination prevention is considering indirect discrimination other than direct discrimination and another challenge is to find an optimal trade-off between anti-discrimination and usefulness of the training data.

1.2 Contribution and Paper Organization

The main contributions of this paper are as follows: (1) a new preprocessing method for indirect discrimination prevention based on data transformation that can consider several discriminatory attributes and their combinations; (2) some measures for evaluating the proposed method in terms of its success in discrimination prevention and its impact on data quality. Although some methods have recently been proposed for discrimination prevention [2,5,6,10], such works only consider direct discrimination. Their approaches cannot guarantee that the transformed dataset is really discrimination-free, because it is known that discriminatory behaviors can be hidden behind non-discriminatory items. To the best of our knowledge this is the first work that proposes a discrimination prevention method for indirect discrimination.

In this paper, Section 2 elaborates on the discovery of indirect discrimination. Section 3 presents our proposed method. Evaluation measures and experimental evaluation are presented in Section 4. Conclusions are drawn in Section 5.

2 Discovering Discrimination

In this section, we present some background concepts that are used throughout the paper. Moreover, we formalize the finding of indirect discrimination.

2.1 Background

A *dataset* is a collection of records and their attributes. Let $\mathcal{DB}$ be the original dataset. An *item* is an attribute along with its value, *e.g.* `Race=black`. An *itemset* is a collection of one or more items. A *classification rule* is an expression $X \to C$, where X is an itemset, containing no class items, and C is a class item, *e.g.* `Class=bad`.

The *support* of an itemset, $supp(X)$, is the fraction of records that contain the itemset X. We say that a rule $X \to C$ *is completely supported* if both X and C appear in the record. The *confidence* of a classification rule, $conf(X \to C)$, measures how often the class item C appears in records that contain X. A *frequent classification rule* is a classification rule with a support or confidence greater than a specified lower bound. Let $\mathcal{FR}$ be the database of frequent classification rules extracted from $\mathcal{DB}$.

With the assumption that discriminatory items in $\mathcal{DB}$ are predetermined (*e.g.* `Race=black`), rules fall into one of the following two classes with respect to discriminatory and non-discriminatory items in $\mathcal{DB}$: (i) a classification rule is *potentially discriminatory* (PD) when $X = A, B$ with A a non-empty discriminatory itemset and B a non-discriminatory itemset (*e.g.* {`Race=black, City=NYC`}→ `Class=bad`); (ii) a classification rule is *potentially non-discriminatory* (PND) when $X = D, B$ is a non-discriminatory itemset (*e.g.* {`Zip=10451, City=NYC`} → `Class=bad`). Let assume that the notation $X(D, B)$ means $X = D, B$. Let $\mathcal{PR}$ a database of frequent classification rules with PD and PND classification rules. The word "potentially" means that a PD rule could probably lead to discriminatory decisions, so some measures are needed to quantify the discrimination potential (direct discrimination). Also, a PND rule could lead to discriminatory decisions if combined with some background knowledge (indirect discrimination); *e.g.*, if the premise of the PND rule contains the `Zip=10451` itemset, relying on additional background knowledge one knows that zip 10451 is mostly inhabited by black people.

Pedreschi *et al.*[1,4] introduced a family of measures of the degree of discrimination of a PD rule. One of these measures is *extended lift* measure ($elift$):

$$elift(A, B \to C) = \frac{conf(A, B \to C)}{conf(B \to C)}$$

Whether the rule is to be considered discriminatory can be assessed by using a threshold: Let $\alpha \in R$ be a fixed threshold[1] and let A be a discriminatory itemset. A PD classification rule $c : A, B \to C$ is *α-protective* w.r.t. $elift$ if $elift(c) < \alpha$. Otherwise, c is *α-discriminatory.*

[1] Note that α is a fixed threshold stating an acceptable level of discrimination according to laws and regulations.

2.2 Indirect Discrimination Formalization

In terms of indirect discrimination, the purpose of discrimination discovery is identifying PND rules that are to a certain extent equivalent to α-discriminatory rules or, in other words, identifying redlining rules. To determine the redlining rules, Pedreschi *et al.* in [1] 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 non-discriminatory itemset D to a discriminatory itemset A within the context B.

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

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

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

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

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

Call

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

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

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, we propose the following formal definitions of redlining and non-redlining rules.

Definition 1. *A PND classification rule $r : X(D, B) \rightarrow C$ 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 itemset.*

Definition 2. *A PND classification rule $r : X(D, B) \rightarrow C$ is a* non-redlining 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 itemset.*

Note that the correlation between the discriminatory itemset A and the non-discriminatory itemset D with context B indicated by the background rules r_{b1} and r_{b2} holds with confidences at least β_1 and β_2, respectively; however, it is not a completely certain correlation. Let $\mathcal{RR}$ be the database of redlining rules extracted from database $\mathcal{DB}$.

3 A Proposal for Indirect Discrimination Prevention

In this section we present a new indirect discrimination prevention method. The method transforms the source data by removing indirect discriminatory biases so that no unfair decision rule can be indirectly mined from the transformed data. The proposed solution is based on the fact that the dataset of decision rules would be free of indirect discrimination if it contained no redlining rule.

For discrimination prevention using preprocessing, we should transform data by removing all evidence of discrimination in the form of α-discriminatory rules and redlining rules. In [10] and [11] we concentrated on direct discrimination and considered α-discriminatory rules. In this paper, we focus on indirect discrimination and consider redlining rules. For these rules, a suitable data transformation with minimum information loss should be applied in such a way that those redlining rules are converted to non-redlining rules.

As mentioned above, based on the definition of the indirect discriminatory measure (*i.e. elb*), to convert redlining rules into non-redlining rules, we should enforce the following inequality for each redlining rule $r : D, B \rightarrow C$ in $\mathcal{RR}$:

$$elb(\gamma, \delta) < \alpha \tag{1}$$

By using the definitions in the statement of Theorem 1, Inequality (1) can be rewritten as

$$\frac{\frac{conf(r_{b1})}{conf(r_{b2})}(conf(r_{b2}) + conf(r : D, B \rightarrow C) - 1)}{conf(B \rightarrow C)} < \alpha \tag{2}$$

To enforce the above inequality, there can be two situations:

- **Case 1:** Assume that discriminatory items (*i.e.* A) are removed from the original database ($\mathcal{DB}$), and the r_{b1} and r_{b2} rules are obtained from publicly available data so that their confidences are constant. Let us rewrite Inequality (2) in the following way

$$conf(r : D, B \rightarrow C) < \frac{\alpha \cdot conf(B \rightarrow C) \cdot conf(r_{b2})}{conf(r_{b1})} - (conf(r_{b2}) + 1) \tag{3}$$

  It is clear that Inequality (2) can be satisfied by decreasing the confidence of redlining rule ($r : D, B \rightarrow C$) to values less than the right-hand side of Inequality (3).
- **Case 2:** Assume that discriminatory items (*i.e.* A) are not removed from the original database ($\mathcal{DB}$), and the rules r_{b1} and r_{b2} might be obtained from $\mathcal{DB}$ so that their confidences might change by data transformation. This could be more useful to detect the non-discriminatory items that are highly correlated with the discriminatory ones and thereby discover the possibly discriminatory rules that could inferred from them. Let us rewrite Inequality (2) as Inequality (4), where the confidences of r_{b1} and r_{b2} rules are not constant.

$$conf(B \to C) > \frac{\frac{conf(r_{b1})}{conf(r_{b2})}(conf(r_{b2}) + conf(r : D, B \to C) - 1)}{\alpha} \tag{4}$$

Clearly, in this case Inequality (2) can be satisfied by increasing the confidence of the base rule ($B \to C$) of the redlining rule ($r : D, B \to C$) to values greater than the right-hand side of Inequality (4) without affecting either the confidence of the redlining rule or the confidence of the r_{b1} and r_{b2} rules.

The detailed process of our preprocessing discrimination prevention method for indirect discrimination is described by means of the following phases:

- *Phase 1.* Use Pedreschi's measure on each PND rule to discover the patterns of indirect discrimination emerged from the available data and also the background knowledge. It consists of the following steps: (i) extract frequent classification rules from $\mathcal{DB}$ using Apriori [9]; (ii) divide the rules into PD and PND, with respect to the predetermined discriminatory items in the dataset; (iii) for each PND rule, compute *elb* to determine the collection of redlining rules. Let $\mathcal{RR}$ be a database of redlining rules and their respective α-discriminatory rules ensuing from those rules through combination with background knowledge rules.
- *Phase 2.* Transform the original data to convert each redlining rule to a non-redlining rule without seriously affecting the data or other rules. Algorithms 1 and 2 show the steps of this phase.
- *Phase 3.* Evaluate the transformed dataset with the discrimination prevention and information loss measures of Section 4.1 below, to check whether they are free of discrimination and useful enough.

The second phase will be explained in detail in the following subsection.

3.1 Data Transformation Method

The data transformation method should increase or decrease some rule confidences as proposed in the previous section with minimum impact on data quality. In terms of the measures defined in Section 4.1 below, we should maximize the discrimination prevention measures and minimize the information loss measures. It is worth mentioning that data transformation methods were previously used for knowledge hiding [8] in privacy-preserving data mining (PPDM). Here we propose a data transformation method for hiding discriminatory and redlining rules.

Algorithms 1 and 2 detail our proposed data transformation method for each of the aforementioned cases. Without loss of generality, we assume that the class attribute C is binary (any non-binary class attribute can be expressed as the Cartesian product of binary class attributes).

1. **No discriminatory attributes in the dataset.** For each redlining rule in this case, Inequality (3) should be enforced. Note that $conf(r_{b2} : D, B \to A)$

and $conf(r_{b1} : A, B \rightarrow D)$ are constant. The values of both sides of Inequality (3) are not independent; hence, 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

$$conf(r : D, B \rightarrow C) = \frac{supp(D, B, C)}{supp(D, B)} \tag{5}$$

In inequality (3) to the target value is to perturb item D from $\neg D$ to D in the subset $\mathcal{DB}_c$ of all records of the original dataset which completely support the rule $\neg D, B \rightarrow \neg C$ and have minimum impact on other rules to increase the denominator of Expression (5) while keeping the numerator and $conf(B \rightarrow C)$ fixed.

2. **Discriminatory attributes in the dataset.** For each redlining rule in this case, Inequality (4) should be enforced. Note that in this case $conf(r_{b2} : D, B \rightarrow A)$ and $conf(r_{b1} : A, B \rightarrow D)$ might not be constant. So it is clear that the values of both inequality sides are dependent; hence, a transformation is required that increases the left-hand side of the inequality without any impact on the right-hand side. A possible solution for increasing

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

in Inequality (4) to the target value is to perturb item C from $\neg C$ to C in the subset $\mathcal{DB}_c$ of all records of the original dataset 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 (6) while keeping the denominator and $conf(r_{b1} : A, B \rightarrow D)$, $conf(r_{b2} : D, B \rightarrow A)$, and $conf(r : D, B \rightarrow C)$ fixed.

In Algorithms 1 and 2, records in $\mathcal{DB}_c$ should be changed until the transformation requirement is met for each redlining rule. Among the records of $\mathcal{DB}_c$, one should change those with lowest impact on the other (non-redlining) 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, 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, with the aim of scoring well in terms of the four utility measures proposed in the next section.

Background Information. In order to implement the proposed data transformation method for indirect discrimination prevention, we simulate the availability of a large set of background rules under the assumption that the dataset contains the discriminatory items. Let BK_s be a database of background rules be defined as

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

In fact, $\mathcal{BK}$ is the set of classification rules $X \rightarrow A$ with a given minimum support ms and A a discriminatory itemset. Although rules of the form r_{b1} :

Algorithm 1.

```
Inputs: $\mathcal{DB}$, $\mathcal{FR}$, $\mathcal{RR}$, $\alpha$, $DI_s$
Output: $\mathcal{DB}'$: the transformed dataset
for each $r : X(D, B) \rightarrow C \in \mathcal{RR}$ do
  $\gamma = conf(r)$
  for each $r' : (A \subseteq DI_s), (B \subseteq X) \rightarrow C$ do
    $\beta_2 = conf(r_{b2} : X \rightarrow A)$
    $\Delta_1 = supp(r_{b2} : X \rightarrow A)$
    $\delta = conf(B \rightarrow C)$
    $\Delta_2 = Supp(B \rightarrow A)$
    $\beta_1 = \frac{\Delta_1}{\Delta_2}$        $//conf(r_{b1} : A, B \rightarrow D)$
    Find $\mathcal{DB}_c$: all records in $\mathcal{DB}$ that completely support $\neg D, B \rightarrow \neg C$
    for each $db_c \in \mathcal{DB}_c$ do
      Compute $impact(db_c) = |\{r_a \in \mathcal{FR} | db_c$ supports the premise of $r_a\}|$
    end for
    Sort $\mathcal{DB}_c$ by ascending impact
    while $\gamma \geq \frac{\alpha \cdot \delta \cdot \beta_2}{\beta_1} - (\beta_2 + 1)$ do
      Select first record $db_c$ in $\mathcal{DB}_c$
      Modify $D$ item of $db_c$ from $\neg D$ to $D$ in $\mathcal{DB}$
      Recompute $\gamma = conf(r : X \rightarrow C)$
    end while
  end for
end for
Output: $\mathcal{DB}' = \mathcal{DB}$
```

$A, B \rightarrow D$ are not included in $\mathcal{BK}$, $conf(r_{b1} : A, B \rightarrow D)$ could be obtained as $supp(r_{b2} : D, B \rightarrow A)/supp(B \rightarrow A)$.

From each redlining rule $(r : X(D, B) \rightarrow C)$ in combination with background knowledge, more than one α-discriminatory rule $r' : A, B \rightarrow C$ might be generated because of two reasons: 1) existence of different sub-itemsets $D, B \subseteq X$ such that X can be written as D, B and 2) existence of more than one item in the set of predetermined discriminatory items (DI_s). Hence, given a redlining rule (r), proper data transformation should be conducted for all α-discriminatory rules $r' : (A \subseteq DI_s), (B \subseteq X) \rightarrow C$ ensuing from r.

4 Experimental Evaluation

This section presents an experimental evaluation of our solution for indirect discrimination prevention. First, we present the utility measures that we propose to evaluate our solution. Finally, we report the experimental results.

4.1 Utility Measures

Two aspects are relevant to evaluate the performance of our indirect discrimination prevention method, namely the success of the method in removing all evidence of indirect discrimination from the original dataset (degree of discrimination prevention) and the impact of the method on data quality (degree of

Algorithm 2.

```
Inputs: $\mathcal{DB}$, $\mathcal{FR}$, $\mathcal{RR}$, $\alpha$, $DI_s$
Output: $\mathcal{DB}'$: the transformed dataset
for each $r : X(D, B) \rightarrow C \in \mathcal{RR}$ do
  $\gamma = conf(r)$
  for each $r' : (A \subseteq DI_s), (B \subseteq X) \rightarrow C$ do
    $\beta_2 = conf(r_{b2} : X \rightarrow A)$
    $\Delta_1 = supp(r_{b2} : X \rightarrow A)$
    $\delta = conf(B \rightarrow C)$
    $\Delta_2 = Supp(B \rightarrow A)$
    $\beta_1 = \frac{\Delta_1}{\Delta_2}$        $//conf(r_{b1} : A, B \rightarrow D)$
    Find $\mathcal{DB}_c$: all records in $\mathcal{DB}$ that completely support $\neg A, B, \neg D \rightarrow \neg C$
    for each $db_c \in \mathcal{DB}_c$ do
      Compute $impact(db_c) = |\{r_a \in \mathcal{FR} | db_c$ supports the premise of $r_a\}|$
    end for
    Sort $\mathcal{DB}_c$ by ascending impact
    while $\delta \leq \frac{\beta_1(\beta_2+\gamma-1)}{\beta_2 \cdot \alpha}$ do
      Select first record $db_c$ in $\mathcal{DB}_c$
      Modify $C$ item of $db_c$ from $\neg C$ to $C$ in $\mathcal{DB}$
      Recompute $\delta = conf(B \rightarrow C)$
    end while
  end for
end for
Output: $\mathcal{DB}' = \mathcal{DB}$
```

information loss). A discrimination prevention method should provide a good trade-off between both aspects above. We propose the following measures for evaluating our solution:

- *Discrimination Prevention Degree* (DPD). This measure quantifies the percentage of redlining rules that are no longer redlining in the transformed dataset. It is defined as

$$DPD = \frac{|\mathcal{RR}| - |\mathcal{RR}'|}{|\mathcal{RR}|}$$

  where $\mathcal{RR}$ is the database of redlining rules extracted from $\mathcal{DB}$, $\mathcal{RR}'$ is the database of redlining rules extracted from the transformed dataset $\mathcal{DB}'$, and $|\cdot|$ is the cardinality operator.
- *Discrimination Protection Preservation* (DPP). This measure quantifies the percentage of the non-redlining rules in the original dataset that remain non-redlining in the transformed dataset. It is defined as

$$DPP = \frac{|\mathcal{NR} \bigcap \mathcal{NR}'|}{|\mathcal{NR}|}$$

  where $\mathcal{NR}$ is the database of non-redlining rules extracted from the original dataset $\mathcal{DB}$, and $\mathcal{NR}'$ is the database of non-redlining rules extracted from the transformed dataset $\mathcal{DB}'$.

- *Misses Cost* (MC). This measure quantifies the percentage of rules among those extractable from the original dataset that cannot be extracted from the transformed dataset (side-effect of the transformation process). It is defined as

$$MC = \frac{|\mathcal{FR}| - |\mathcal{FR} \bigcap \mathcal{FR}'|}{|\mathcal{FR}|}$$

  where $\mathcal{FR}'$ is the database of frequent classification rules extracted from the transformed dataset $\mathcal{DB}'$.
- *Ghost Cost* (GC). This measure quantifies the percentage of the rules among those extractable from the transformed dataset that could not be extracted from the original dataset (side-effect of the transformation process). It is defined as

$$GC = \frac{|\mathcal{FR}'| - |\mathcal{FR} \bigcap \mathcal{FR}'|}{|\mathcal{FR}'|}$$

  where $\mathcal{FR}'$ is the database of frequent classification rules extracted from the transformed dataset $\mathcal{DB}'$.

The DPD and DPP measures are used to evaluate the success of the proposed method in indirect discrimination prevention; ideally they should be 100%. The MC and GC measures are used for evaluating the degree of information loss (impact on data quality); ideally they should be 0% (MC and GC may not be 0% as a side-effect of the transformation process). MC and GC were previously proposed and used as information loss measures for knowledge hiding in PPDM [8].

4.2 Results

We use the German Credit Dataset [12] in our experiments, since it is a well-known and frequently used dataset in the context of anti-discrimination. In this dataset, we consider the following set of predetermined discriminatory items (DI_s): female and not single as *personal status*, unemployed or unskilled non resident as *job*, the attributes marking the individual as *foreign worker* and *old-aged*.

In this section, we present the experimental evaluation of the proposed method. For the first phase we have used Apriori [9]. The algorithms and the utility measures corresponding to the second and third phases of the proposed solution, respectively, were implemented using Microsoft Visual Studio 2008 with C# programming language. The tests were performed on an 2.27 GHz Intel® Core$^{\text{TM}}$i3 machine, equipped with 4 GB of RAM, and running under Windows 7 Professional.

In order to evaluate our proposed solution we need to simulate the background knowledge rules. Hence, we assume that the original dataset $\mathcal{DB}$ contains discriminatory attributes and implement Algorithm 2. The values of utility measures for minimum support 0.5% and minimum confidence 10% are presented in Table 1. In this experiment, the number of frequent classification rules extracted

Table 1. Utility measures for minimum support 0.5% and minimum confidence 10%

	N. Redlining Rules	**N. α-Disc. Rules**	**MC**	**GC**	**DPD**	**DPP**	**Execution time (sec)**
α= 0.6	1	2	0	0.21	100	100	11
α= 0.5	2	5	0.34	0.49	100	100	27
α= 0.4	3	7	0.52	0.47	100	99.95	49
α= 0.3	11	28	1.62	1.97	90.90	99.81	125

from $\mathcal{DB}$ is 7690 and the number of background knowledge rules is 7416. As shown, the results are reported for different values of $\alpha \in [0.3, 0.6]$. We selected the upper bound (0.6) because, with respect to our predetermined discriminatory items, redlining rules could be extracted from $\mathcal{DB}$. We restrict the lower bound to limit the number of redlining rules extracted form $\mathcal{DB}$. Other than utility measures, the number of redlining rules and the number of α-discriminatory rules that could be generated from those redlining rules are also reported for different values of α.

As shown in Table 1, the values of DDP and DPD demonstrate that the proposed solution achieves a high degree of indirect discrimination prevention in different cases (*i.e.* different values of α). In addition, the values of MC and GC demonstrate that the proposed solution incurs little information loss, especially when α is not too small. By decreasing the value of α, the number of redlining rules is increased, which causes more data transformation to be done, thereby increasing MC and GC. As presented in Table 1, the execution time of the algorithm increases linearly with the number of redlining rules and α-discriminatory rules.

5 Conclusions

To the best of our knowledge, we have presented the first method for preventing indirect discrimination in data mining due to biased training datasets. Our contribution in this paper concentrates on producing training data which are free or nearly free from indirect discrimination while preserving their usefulness to data mining algorithms. In order to prevent indirect discrimination in a dataset, a first step consists in discovering whether there exists indirect discrimination. If any discrimination is found, the dataset is modified until discrimination is brought below a certain threshold or is entirely eliminated. In the future, we want to present a unified discrimination prevention approach based on the discrimination hiding idea that encompasses both direct and indirect discrimination.

Disclaimer and Acknowledgments. The authors are with the UNESCO Chair in Data Privacy, but they are solely responsible for the views expressed in this paper, which do not necessarily reflect the position of UNESCO nor commit that organization. This work was partly supported by the Spanish Government through projects TSI2007-65406-C03-01 “E-AEGIS”, CONSOLIDER INGENIO

2010 CSD2007-00004 “ARES” and TSI-020100-2009-720 “everification”, by the Government of Catalonia under grant 2009 SGR 01135, and by the European Commission under FP7 project “DwB”. The second author is partly supported as an ICREA Acadèmia Researcher by the Government of Catalonia.

References

1. Pedreschi, D., Ruggieri, S., Turini, F.: Discrimination-aware data mining. In: Proc. of the 14th ACM International Conference on Knowledge Discovery and Data Mining (KDD 2008), pp. 560–568. ACM, New York (2008)
2. Kamiran, F., Calders, T.: Classification without discrimination. In: Proc. of the 2nd IEEE International Conference on Computer, Control and Communication (IC4 2009). IEEE, Los Alamitos (2009)
3. Ruggieri, S., Pedreschi, D., Turini, F.: Data mining for discrimination discovery. ACM Transactions on Knowledge Discovery from Data 4(2) Article 9 (2010)
4. Pedreschi, D., Ruggieri, S., Turini, F.: Measuring discrimination in socially-sensitive decision records. In: Proc. of the 9th SIAM Data Mining Conference (SDM 2009), pp. 581–592. SIAM, Philadelphia (2009)
5. Kamiran, F., Calders, T.: Classification with No Discrimination by Preferential Sampling. In: Proc. of the 19th Machine Learning Conference of Belgium and, The Netherlands (2010)
6. Calders, T., Verwer, S.: Three naive Bayes approaches for discrimination-free classification. Data Mining and Knowledge Discovery 21(2), 277–292 (2010)
7. Pedreschi, D., Ruggieri, S., Turini, F.: Integrating induction and deduction for finding evidence of discrimination. In: Proc. of the 12th ACM International Conference on Artificial Intelligence and Law (ICAIL 2009), pp. 157–166. ACM, New York (2009)
8. Verykios, V., Gkoulalas-Divanis, A.: A survey of association rule hiding methods for privacy. In: Aggarwal, C.C., Yu, P.S. (eds.) Privacy- Preserving Data Mining: Models and Algorithms. Springer, Heidelberg (2008)
9. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules in large databases. In: Proc. of the 20th International Conference on Very Large Data Bases, pp. 487–499. VLDB (1994)
10. Hajian, S., Domingo-Ferrer, J., Martínez-Ballesté, A.: Discrimination prevention in data mining for intrustion and crime detection. In: Proc. of the IEEE Symposium on Computational Intelligence in Cyber Security (CICS 2011), pp. 47–54. IEEE, Los Alamitos (2011)
11. Hajian, S., Domingo-Ferrer, J., Martínez-Ballesté, A.: Rule generalization and protection for discrimination prevention in data mining (submitted)
12. Newman, D.J., Hettich, S., Blake, C.L., Merz, C.J.: UCI repository of machine learning databases (1998), http://archive.ics.uci.edu/ml