Fair Pattern Discovery

Sara Hajian
Universitat Rovira i Virgili
Department of Computer Engineering and Maths
Tarragona, Catalonia
sara.hajian@urv.cat

Anna Monreale
University of Pisa
Department of Computer Science
Pisa, Italy
annam@di.unipi.it

Dino Pedreschi
University of Pisa
Department of Computer Science
Pisa, Italy
pedre@di.unipi.it

Josep Domingo-Ferrer
Universitat Rovira i Virgili
Department of Computer Engineering and Maths
Tarragona, Catalonia
josep.domingo@urv.cat

Fosca Giannotti
ISTI-CNR
Pisa, Italy
fosca.giannotti@isti.cnr.it

ABSTRACT

Data mining is gaining societal momentum due to the ever increasing availability of large amounts of human data, easily collected by a variety of sensing technologies. We are assisting to unprecedented opportunities of understanding human and society behavior that unfortunately is darkened by several risks for human rights: one of this is the *unfair* discrimination based on the extracted patterns and profiles. Consider the case when a set of patterns extracted from the personal data of a population of individual persons is released for subsequent use in a decision making process, such as, *e.g.*, granting or denying credit. Decision rules based on such patterns may lead to unfair discrimination, depending on what is represented in the training cases. In this context, we address the discrimination risks resulting from publishing frequent patterns. We present a set of pattern sanitization methods, one for each discrimination measure used in the legal literature, for fair (discrimination-protected) publishing of frequent pattern mining results. Our proposed pattern sanitization methods yield discrimination-protected patterns, while introducing reasonable (controlled) pattern distortion. Finally, the effectiveness of our proposals is assessed by extensive experiments.

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Data Mining

General Terms

Algorithms, Legal Aspects

SAC'14 March 24-28, 2014, Gyeongju, Korea.

Keywords

Data mining, Frequent pattern discovery, Anti-discrimination

1. INTRODUCTION

Data mining plays a key role in extracting useful knowledge hidden in the so-called *big data*, especially data describing human and social activities. Unfortunately, the new opportunities to extract knowledge and understand human and social complex phenomena increase hand in hand with the risks of violation of fundamental human rights, such as privacy and non-discrimination. *Discrimination* refers to unfair or unequal treatment of people based on membership to a category, group or minority, without regard to individual merit. Human rights laws not only have concern about data protection [6] but also prohibit discrimination [2, 7] against protected groups on the grounds of race, color, religion, nationality, sex, marital status, age and pregnancy; and in a number of settings, like credit and insurance, personnel selection and wages, and access to public services. Clearly, preserving the great benefits of data mining within a discrimination-aware technical ecosystem would lead to a wider social acceptance of a multitude of new services and applications based on the knowledge discovery process. The issue of anti-discrimination has been recently considered from a data mining perspective [16]. Some proposals are oriented to the discovery and measurement of discrimination using data mining, while others deal with preventing data mining from becoming itself a source of discrimination, due to automated decision making based on discriminatory models extracted from inherently biased datasets.

Consider the case in which a set of patterns extracted (mined) from the personal data of a population of individual persons is released for subsequent use in a decision making process, such as, *e.g.*, granting or denying credit. Decision rules based on such patterns may lead to unfair discrimination, depending on what is represented in the training cases. The following example illustrates this point. Assume a credit institution, *e.g.*, a bank, wants to release among its employees the rules to grant/deny credit, for the purpose of supporting future decision making. Assume that such rules have been mined from decision records accumulated during past years in a certain city, such as those illustrated in Table

1. This would allow releasing a rule such that *Sex = female, Credit_history=no-taken → Credit_approved = no*. Clearly, using such a rule for credit scoring is discriminatory against new women applicants.

Table 1: A data table of personal decision records

Sex	Job	Credit_ history	...	Credit_ approved
Male	Writer	No-taken	...	Yes
Female	Lawyer	Paid-duly	...	No
Male	Veterinary	Paid-delay	...	Yes
...	...	...	...	...

This simple example shows that protecting against discrimination is needed when disclosing a set of patterns. This line of reasoning also permeates the comprehensive reform of the data protection law proposed in 2012 by the European Commission, currently under approval by the European Parliament, which introduces measures based on profiling and discrimination within a broader concept of privacy and personal data[1]. The contributions of this paper, towards the above stated aim, are summarized as follows. First, we define a natural scenario of pattern mining from personal data records containing potentially discriminatory attributes and decision attributes, and characterize the problem statement of publishing a collection of patterns which is discrimination-free. Second, we propose new pattern sanitization methods, one for each discrimination measure used in the legal literature, for discrimination prevention when publishing frequent patterns. Third, we theoretically and empirically show that the proposed algorithm is effective at protecting against discrimination threats while introducing reasonable (controlled) pattern distortion. In [10], the authors introduced the initial idea of achieving anti-discrimination in frequent pattern discovery and providing a sanitization approach suitable for one of the legal measures of discrimination.

In this paper, we present a complete framework for obtaining fair frequent patterns by considering all measures of discrimination used in the legal literature. In other words, we propose new algorithms which are parametric to different measures of discrimination. Moreover, we evaluate the impact of our fairness transformation on the classification task. In particular, we compare the accuracy of a rule-based classifier from the original and sanitized patterns.

The rest of this article is organized as follows. Section 2 reviews related work in detail. Section 3 introduces the basic concepts used throughout the paper. Section 4 introduces the notion of discrimination-protected (fair) frequent patterns. Section 5 describes our methods and algorithms to make the collection of patterns fair. Section 6 reports on the evaluation of our sanitization methods. Finally, Section 7 concludes the paper and identifies future research topics in this context.

2. RELATED WORK

According to current legislation, discrimination occurs when a group is treated "less favorably" [2] than others, or when "a higher proportion of people not in the group is able to comply" [7] with a qualifying criterion. As mentioned above, the issue of anti-discrimination has recently been considered from a data mining perspective [16], under the name of discrimination-aware data analysis. A substantial part of the existing literature on anti-discrimination in data mining is oriented to *discovering* and *measuring* discrimination. Other contributions deal with *preventing* discrimination. Summaries of contributions in discrimination-aware data analysis are collected in [4]. Pedreschi *et al.* [16, 17, 18] have introduced data mining approaches for discrimination discovery. These approaches have followed the legal principle of *under-representation* to unveil contexts of possible discrimination against *protected-by-law* groups (*e.g.*, women). This is done by extracting classification rules from a dataset of historical decision records (inductive part); then, rules are ranked according to some *legally-grounded* measures of discrimination (deductive part).

Beyond discrimination discovery, preventing knowledge-based decision support systems from making discriminatory decisions is a more challenging issue. Discrimination prevention approaches can be classified according to the phase of the data mining process in which they operate: pre-processing methods [11, 9, 15, 5] transform data in such a way that the discriminatory biases contained in the original data are removed so that no unfair decision models can be mined from the transformed data; in-processing methods [12, 3, 13] modify data mining algorithms in such a way that the resulting models do not contain unfair decisions; finally, post-processing pays attention to modifying the resulting data mining models, instead of cleaning the original dataset or changing the data mining algorithms. In this paper, we focus on the post-processing approach in frequent pattern mining. The reason is that frequent patterns are very basic structures, that can be used as basis to generate more complex mining models such as association rules, classification or clustering models.

Note that the actual discovery and prevention of discriminatory situations and practices may be an extremely difficult task [18]. A huge number of possible contexts may, or may not, be the theater for discrimination. Consider the case of gender discrimination in credit approval. Although an analyst may observe that no discrimination occurs in general, *i.e.*, when considering the whole set of available decision records, it may turn out that it is extremely difficult for new women applicants to obtain credit. Hence, many small or large such contexts may exist that conceal discrimination, and therefore all possible specific situations should be considered as candidates, consisting of all possible combinations of attributes and attribute values. Moreover, one might think of a straightforward pre-processing approach consisting of just removing the discriminatory attributes from the dataset. Although this would solve the discrimination problem, it would cause much information loss. And as stated in [16] there may be other attributes that are highly correlated with the sensitive ones and allow inferring discriminatory rules. Hence, another challenge regarding discrimination prevention is to find a good trade-off between discrimination removal and the quality of the resulting data mining models.

3. BASIC DEFINITIONS

Let $\mathcal{I} = \{i_1, \ldots, i_n\}$ be a set of items, where each *item* i_j has the form *attribute=value* (*e.g.*, *Sex=female*). An

[1] `http://ec.europa.eu/justice/newsroom/data-protection/news/120125_en.htm`

itemset $X \subseteq \mathcal{I}$ is a collection of one or more items, *e.g.* {*Sex=female, Credit_history=no-taken*}. A *database* is a collection of data objects (records) and their attributes; more formally, a (transaction) database $\mathcal{D} = \{r_1, \ldots, r_m\}$ is a set of data records or transactions where each $r_i \subseteq \mathcal{I}$. Civil rights laws [2, 7] explicitly identify the groups to be protected against discrimination, such as minorities and disadvantaged people, *e.g.*, women. In our context, these groups can be represented as items, *e.g.*, *Sex=female*, which we call potentially discriminatory (PD) items; a collection of PD items can be represented as an itemset, *e.g.*, {*Sex=female, Foreign_worker=yes*}, which we call PD itemset or protected-by-law (or protected for short) groups, denoted by DI_b. The attributes in a database $\mathcal{D}$ can be classified in several non-disjoint categories. *PD attributes* are those that can take PD items as values; for instance, *Race* and *Gender* where DI_b:{*Sex=female, Race=black*}. A *decision (class) attribute* is one taking as values *yes* or *no* to report the outcome of a decision made on an individual; an example is the attribute *credit_approved*, which can be *yes* or *no*. The *support* of an itemset X in a database $\mathcal{D}$ is the number of records that contain X, *i.e.* $supp_{\mathcal{D}}(X) = |\{r_i \in \mathcal{D} | X \subseteq r_i\}|$, where $|\ .\ |$ is the cardinality operator. Given a support threshold σ, an itemset X is called *σ-frequent* in a database $\mathcal{D}$ if $supp_{\mathcal{D}}(X) \geq \sigma$. A σ-frequent itemset is also called *σ-frequent pattern*. The collection of all σ-frequent patterns in $\mathcal{D}$ is denoted by $\mathcal{F}(\mathcal{D}, \sigma)$. The frequent pattern mining problem is formulated as follows: given a database $\mathcal{D}$ and a support threshold σ, find all σ-frequent patterns, *i.e.* the collection $\mathcal{F}(\mathcal{D}, \sigma)$. Several algorithms have been proposed for finding $\mathcal{F}(\mathcal{D}, \sigma)$. In this paper we use the Apriori algorithm [1], which is a very common choice. From patterns it is possible to derive classification rules. A *classification rule* is an expression $X \rightarrow C$, where C is a class item and X is an itemset containing no class item, *e.g. Sex=female, Cedit_history=no-taken* $\rightarrow$ *Credit_approved=no*. The itemset X is called the premise of the rule. The *confidence* of a classification rule, $conf_{\mathcal{D}}(X \rightarrow C)$, measures how often the class item C appears in records that contain X. Hence, if $supp_{\mathcal{D}}(X) > 0$ then

$$conf_{\mathcal{D}}(X \rightarrow C) = \frac{supp_{\mathcal{D}}(X, C)}{supp_{\mathcal{D}}(X)} \tag{1}$$

Confidence ranges over $[0, 1]$. We omit the subscripts in $supp_{\mathcal{D}}(\cdot)$ and $conf_{\mathcal{D}}(\cdot)$ when there is no ambiguity. Also, the notation readily extends to negated itemsets $\neg X$. A *frequent classification rule* is a classification rule with support and confidence greater than respective specified lower bounds.

4. FAIR FREQUENT PATTERN SET

Given DI_b and starting from a dataset $\mathcal{D}$ of historical decision records, the approach is to extract frequent classification rules of the form $A, B \rightarrow C$, called PD rules, to unveil contexts B of possible discrimination, where the non-empty protected group $A \subseteq DI_b$ suffers from over-representation with respect to the *negative* decision C (C is a class item reporting a negative decision, such as credit denial, application rejection, job firing, and so on). In other words, A is under-represented w.r.t. the corresponding positive decision $\neg C$. As an example, rule *Sex=female, Credit_history=no-taken* $\rightarrow$ *Credit_approved=no* is a PD rule about denying credit (the decision C) to women (the protected group A) among those who are new applicants (the context B), where DI_b:{*Sex=female*}. Starting from the above definition of PD classification rule, we define when a frequent pattern is PD.

DEFINITION 1. *Given protected groups DI_b, a frequent pattern $p \in \mathcal{F}(\mathcal{D}, \sigma)$ is said to be a PD if: (1) p contains a class item C denying some benefit, i.e., $C \subset p$, and (2) $\exists p' \subset p$ s.t. $p' \subseteq DI_b$.*

In other words, a frequent pattern $p : \{A, B, C\}$ is a PD pattern if a PD classification rule $A, B \rightarrow C$ can be derived from it. As an example, pattern {*Sex=female, Credit_history=no-taken, Credit_approved=no*} is a PD pattern, where DI_b : {*Sex=female*}. Then, the degree of under-representation should be measured over each PD rule by one of the *legally-grounded* measures introduced in Pedreschi *et al.* [17].

DEFINITION 2. *Let $A, B \rightarrow C$ be a PD classification rule extracted from $\mathcal{D}$ with $conf(\neg A, B \rightarrow C) > 0$. The selection lift[2] (slift) of the rule is*

$$slift(A, B \rightarrow C) = \frac{conf(A, B \rightarrow C)}{conf(\neg A, B \rightarrow C)} \tag{2}$$

In fact, *slift* is the ratio of the proportions of benefit denial, *e.g.*, credit denial, between the protected and unprotected groups, *e.g.* women and men resp., in the given context, *e.g.* new applicants.

DEFINITION 3. *Let $A, B \rightarrow C$ be a PD classification rule extracted from $\mathcal{D}$ with $conf(B \rightarrow C) > 0$. The* extended lift [3] (elift) *of the rule is*

$$elift(A, B \rightarrow C) = \frac{conf(A, B \rightarrow C)}{conf(B \rightarrow C)} \tag{3}$$

In fact, *elift* is the ratio of the proportions of benefit denial, *e.g.* credit denial, between the protected groups and all people who were not granted the benefit in the given context, *e.g.* women versus all men and women who were denied credit, in the given context, *e.g.* those who are new applicants. Although the measures introduced so far are defined in terms of ratios, measures based on the difference of confidences have been considered on the legal side as well.

DEFINITION 4. *Let $A, B \rightarrow C$ be a PD classification rule extracted from $\mathcal{D}$. The* difference measures *are defined as*

$$slift_d(A, B \rightarrow C) = conf(A, B \rightarrow C) - conf(\neg A, B \rightarrow C) \tag{4}$$

$$elift_d(A, B \rightarrow C) = conf(A, B \rightarrow C) - conf(B \rightarrow C) \tag{5}$$

Difference-based measures range over $[-1, 1]$. Lastly, the following measures are also defined in terms of ratios and known as chance measures.

DEFINITION 5. *Let $A, B \rightarrow C$ be a PD classification rule extracted from $\mathcal{D}$. The* chance measures *are defined as*

$$slift_c(A, B \rightarrow C) = \frac{1 - conf(A, B \rightarrow C)}{1 - conf(\neg A, B \rightarrow C)} \tag{6}$$

$$elift_c(A, B \rightarrow C) = \frac{1 - conf(A, B \rightarrow C)}{1 - conf(B \rightarrow C)} \tag{7}$$

[2]Discrimination on the basis of an attribute happens if a person with an attribute is treated less favorably than a person without the attribute.

[3]Discrimination occurs when a higher proportion of people not in the group is able to comply.

For $slift$ and $elift$, the values of interest (potentially indicating discrimination) are those greater than 1; for $slift_d$ and $elift_d$, they are those greater than 0; and for $slift_c$ and $elift_c$, they are those less than 1. On the legal side, different measures are adopted worldwide. For example, UK law mentions mostly $slift_d$. The EU court of justice has made more emphasis in $slift$, and US laws courts mainly refer to $slift_c$. Whether the rule has to be considered discriminatory or not can be assessed by thresholding one of the above measures as follows.

DEFINITION 6. *Let f be one of the measures in Definitions 3 or 4. Given protected groups DI_b and $\alpha \in R$, a fixed threshold*[4]*, a PD classification rule $r : A, B \rightarrow C$, where C denies some benefit and $A \subseteq DI_b$, is α-protective w.r.t. f if $f(r) < \alpha$. Otherwise, c is α-discriminatory.*

EXAMPLE 1. *Let $f = slift$, $\alpha = 1.25$ and DI_b :{Sex=female}. Assume that, in the data set of Table 1, the total number of new women applicants and the number of new women applicants who are denied credit are 34 and 20, respectively, and the total number of new men applicants and the number of new men applicants who are denied credit are 47 and 19, respectively. The PD classification rule r* :Sex=female, Credit_history=no-taken → Credit_approved=no *extracted from Table 1 is 1.25-discriminatory, because $slift(r) = \frac{20/34}{19/47} =$* 1.45. □

Based on Definitions 1 and 6, we introduce the notions of α-protective and α-discriminatory patterns.

DEFINITION 7. *Let f be one of the measures in Definitions 3-4. Given protected groups DI_b and $\alpha \in R$ a fixed threshold, a PD pattern $p : \{A, B, C\}$, where C denies some benefit and $A \subseteq DI_b$, is α-protective w.r.t. f if the classification rule $r : A, B \rightarrow C$ is α-protective. Otherwise, p is α-discriminatory.*

EXAMPLE 2. *Continuing Example 1, a PD pattern p* :{Sex = female, Credit_history=no-taken, Credit_approved = no} *extracted from Table 1, is 1.25-discriminatory because rule r is 1.25-discriminatory, where r is* Sex = female, Credit_history=no-taken → Credit_approved = no. □

Based on Definition 7, we introduce the notion of discrimination-protected (fair) pattern set.

DEFINITION 8 (α-PROTECTIVE PATTERN SET). *Given a collection of frequent patterns $\mathcal{F}(\mathcal{D}, \sigma)$, discrimination measure f, a discrimination threshold α, and protected groups DI_b, $\mathcal{F}(\mathcal{D}, \sigma)$ is α-protective w.r.t. DI_b and f if $\nexists p \in \mathcal{F}(\mathcal{D}, \sigma)$ s.t. p is an α-discriminatory pattern.*

5. ACHIEVING A FAIR FREQUENT PATTERN SET

In order to generate a discrimination-protected (*i.e.* an α-protective) version of $\mathcal{F}(\mathcal{D}, \sigma)$, we propose an approach including two steps. First, detecting α-discriminatory patterns in $\mathcal{F}(\mathcal{D}, \sigma)$ w.r.t. discriminatory measure f, DI_b and α as discussed in Section 4. We propose Algorithm 1 for detecting α-discriminatory patterns in $\mathcal{F}(\mathcal{D}, \sigma)$. The algorithm starts by obtaining the subset $\mathcal{D}_{PD}$ which contains the PD patterns in $\mathcal{F}(\mathcal{D}, \sigma)$ found according to C and DI_b (Line 4). For each pattern $p : \{A, B, C\}$ in $\mathcal{D}_{PD}$, where $A \subseteq DI_b$, the value of f (one of the measures in Definitions 3-4) regarding its PD rule $r : X \rightarrow C$, where $X = A, B$, is computed to determine the subset $\mathcal{D}_D$ which contains the α-discriminatory patterns in $\mathcal{F}(\mathcal{D}, \sigma)$ (Lines 5-13). After obtaining $\mathcal{D}_D$, the second step of our approach is sanitization for each pattern in $\mathcal{D}_D$, in order to make it α-protective. In the sequel, we study and propose a pattern sanitization solution for each possible measure of discrimination f.

Algorithm 1 DETECTING α-DISCRIMINATORY PATTERNS

```
1: Inputs: Database D, FP := F(D,σ), DI_b, discrimination measure f, α, C =class item with a negative decision value
2: Output: D_D: α-discriminatory patterns in FP
3: Function DetDiscPatt(FP, D, DI_b, f, α, C)
4: D_PD ← All patterns ⟨p : A, B, C, supp(p)⟩ ∈ FP with p∩C ≠ ∅ and p ∩ DI_b ≠ ∅
5: for all p ∈ D_PD do
6:    X = p\C
7:    r = X → C
8:    Compute f(r) using FP and D where f is one of the measures from Definitions 3-4
9:    if f(r) ≥ α then
10:      Add p in D_D
11:   end if
12: end for
13: return D_D
14: End Function
```

5.1 Anti-discrimination pattern sanitization for slift and its variants

According to Definition 7, to make an α-discriminatory pattern $p : \{A, B, C\}$ α-protective where $f = slift$, we should enforce the following inequality

$$slift(A, B \rightarrow C) < \alpha \tag{8}$$

where $A \subseteq DI_b$ and C denies some benefit. By using the definitions of confidence and $slift$ (Expressions (1) and (2), resp.), Inequality (8) can be rewritten as

$$\frac{\frac{supp(A,B,C)}{supp(A,B)}}{\frac{supp(\neg A,B,C)}{supp(\neg A,B)}} < \alpha. \tag{9}$$

Then, it is clear that Inequality (8) can be satisfied by decreasing the left-hand side of Inequality (9) to a value less than the discriminatory threshold α, which can be done in the following way:

- *Anti-discrimination pattern sanitization where $f = slift$.* Increase the support of the pattern $\{A, B\}$ and all its subsets by a specific value Δ_{slift} to satisfy Inequality (9). This increment decreases the numerator of equation $\frac{\frac{supp(A,B,C)}{supp(A,B)}}{\frac{supp(\neg A,B,C)}{supp(\neg A,B)}}$ while keeping the denominator unaltered.

Modifying the support of the subsets of respective patterns accordingly is needed to avoid contradictions (maintain compatibility) among the released patterns. In fact, *anti-discrimination pattern sanitization* makes pattern p α-protective

[4] α states an acceptable level of discrimination according to laws and regulations. For example, the U.S. Equal Pay Act [19] 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 amounts to using $slift$ with $\alpha = 1.25$.

by decreasing the proportion of people in the protected group and given context who were not granted the benefit (*e.g.* decreasing the proportion of new women applicants who were denied credit). Let us compute the value Δ_{slift} to be used in anti-discrimination pattern sanitization where $f = slift$. The support of the pattern $\{A, B\}$ should be increased to satisfy Inequality (9):

$$slift(A, B \rightarrow C) = \frac{\frac{supp(A,B,C)}{supp(A,B)+\Delta_{slift}}}{\frac{supp(\neg A,B,C)}{supp(\neg A,B)}} < \alpha.$$

The above equality can be rewritten as

$$\Delta_{slift} > \frac{supp(A, B, C) \times supp(\neg A, B)}{supp(\neg A, B, C) \times \alpha} - supp(A, B). \quad (10)$$

Hence, taking Δ_{slift} equal to the ceiling of the right-hand side of Equation (10) suffices to make $p : \{A, B, C\}$ α-protective w.r.t. $f = slift$. Considering the definitions of $slift_d$ and $slift_c$ (Expressions (4) and (6), resp.), a similar method can make pattern p α-protective w.r.t. $f = slift_d$ and $f = slift_c$. The value of Δ_{slift_d} and Δ_{slift_c} can be computed in the same way as we compute Δ_{slift}.

Example 3. *Continuing Examples 1 and 2, pattern* p : {Sex = female, Credit_history=no-taken, Credit_approved = no} *can be made* 1.25*-protective by increasing the support of pattern* {Sex=female, Credit_history=no-taken} *and all its subsets by* $\Delta_{slift} = 6$*, which is the value resulting from Inequality (10).* □

5.2 Anti-discrimination pattern sanitization for elift and its variants

According to Definition 7, to make an α-discriminatory pattern $p : \{A, B, C\}$ α-protective where $f = elift$, we should enforce the following inequality

$$elift(A, B \rightarrow C) < \alpha \quad (11)$$

where $A \subseteq DI_b$ and C denies some benefit. By using the definitions of confidence and $elift$ (Expressions (1) and (3), resp.), Inequality (11) can be rewritten as

$$\frac{\frac{supp(A,B,C)}{supp(A,B)}}{\frac{supp(B,C)}{supp(B)}} < \alpha. \quad (12)$$

Then, it is clear that Inequality (11) can be satisfied by decreasing the left-hand side of Inequality (12) to a value less than the discriminatory threshold α. A similar anti-discrimination pattern sanitization proposed for $f = slift$ cannot make pattern p α-protective w.r.t. $f = elift$ because increasing the support of pattern $\{A, B\}$ and all its subsets by a specific value can decrease the numerator of equation $\frac{\frac{supp(A,B,C)}{supp(A,B)}}{\frac{supp(B,C)}{supp(B)}}$ and decrease the denominator of it as well. Then, making pattern $p : \{A, B, C\}$ α-protective w.r.t. $f = elift$ is possible using an alternative pattern sanitization method:

- *Anti-discrimination pattern sanitization where* $f = elift$. Increase the support of the pattern $\{B, C\}$ and all its subsets by a specific value Δ_{elift} to satisfy Inequality (12). This increment increases the denominator of equation $\frac{\frac{supp(A,B,C)}{supp(A,B)}}{\frac{supp(B,C)}{supp(B)}}$ while keeping the numerator unaltered.

In fact, the above method makes pattern p α-protective w.r.t. $elift$ by increasing the proportion of people in the given context who were not granted the benefit (*e.g.* increasing the proportion of new applicants who were denied credit while the proportion of new women applicants who were denied credit is unaltered). Let us compute the value Δ_{elift} to be used in anti-discrimination pattern sanitization where $f = elift$. The support of the pattern $\{B, C\}$ should be increased to satisfy Inequality (12):

$$elift(A, B \rightarrow C) = \frac{\frac{supp(A,B,C)}{supp(A,B)}}{\frac{supp(B,C)+\Delta_{elift}}{supp(B)+\Delta_{elift}}} < \alpha.$$

Since the value of α is higher than 1 and $\frac{supp(A,B,C)}{supp(A,B)} \leq \alpha$, from the above equality we obtain

$$\Delta_{elift} > \frac{\alpha \times supp(A, B) \times supp(B, C) - supp(A, B, C) \times supp(B)}{supp(A, B, C) - \alpha \times supp(A, B)}. \quad (13)$$

Hence, taking Δ_{elift} equal to the ceiling of the right-hand side of Equation (13) suffices to make $p : \{A, B, C\}$ α-protective w.r.t. $f = elift$. Considering the definitions of $elift_d$ and $elift_c$ (Expressions (4) and (6), resp.), a similar method can make pattern p α-protective w.r.t. $f = elift_d$ and $f = elift_c$. The values of Δ_{elift_d} and Δ_{elift_c} can be computed in the same way as Δ_{elift}.

5.3 Discrimination analysis

An essential property of a valid anti-discrimination pattern sanitization method is not to produce new discrimination as a result of the transformations it performs. The following theorem shows that all the methods described above satisfy this property.

Theorem 1. *Anti-discrimination pattern sanitization methods for making* $\mathcal{F}(\mathcal{D}, \sigma)$ α*-protective w.r.t.* f *do not generate new discrimination as a result of their transformations, where* f *is one of the measures from Definition 2-4.*

Proof. It is enough to show that anti-discrimination pattern sanitization methods to make each α-discriminatory pattern in $\mathcal{F}(\mathcal{D}, \sigma)$ α-protective w.r.t. f cannot make α-protective patterns in $\mathcal{F}(\mathcal{D}, \sigma)$ α-discriminatory. Consider two PD patterns $p_1 : \{A, B, C\}$ and $p_2 : \{A', B', C\}$, where $A, A' \subseteq DI_b$ and $p_1 \neq p_2$. The following possible relations between p_1 and p_2 are conceivable:

- $A = A'$ and $B \neq B'$, special case: $B' \subset B$
- $A \neq A'$ and $B = B'$, special case: $A' \subset A$
- $A \neq A'$ and $B \neq B'$, special case: $A' \subset A$ and $B' \subset B$

In all the above special cases (*i.e.* $p_2 \subset p_1$), making p_1 α-protective w.r.t. f involves increasing $supp(A', B')$ by Δ_{slift} or Δ_{slift_d} where $f = slift$ or $f = slift_d$, resp., and involves increasing $supp(B', C)$ and $supp(B')$ by Δ_{elift}, Δ_{elift_d} where $f = elift$ or $f = elift_d$, respectively. This cannot make p_2 less α-protective w.r.t. f; actually, it can even make p_2 more protective because increasing $supp(A', B')$ can decrease $slift(A', B' \rightarrow C)$ and $slift_d(A', B' \rightarrow C)$ and increasing $supp(B', C)$ and $supp(B')$ can decrease $elift$ $(A', B' \rightarrow C)$ and $elift_d(A', B' \rightarrow C)$. On the other hand, making p_2 α-protective w.r.t. f cannot make p_1 less or more protective since there is no overlap between the modified

patterns to make p_2 α-protective and the patterns whose changing support can change $f(A, B \rightarrow C)$. Otherwise (no special cases), making p_1 (resp. p_2) α-protective w.r.t. f cannot make p_2 (resp. p_1) less or more protective since there is no overlap between the modified patterns to make p_1 (resp. p_2) α-protective w.r.t. f and the patterns whose changing support can change $f(A', B' \rightarrow C)$ (resp. $f(A, B \rightarrow C)$). Hence, the theorem holds. □

5.4 Pattern sanitization algorithm

From the analysis in the previous section, by using the proposed anti-discrimination pattern sanitization methods an α-protective version of $\mathcal{F}(\mathcal{D}, \sigma)$ w.r.t. f can be obtained. We propose Algorithm 2 for doing so.

There are two assumptions in this algorithm: first, the class attribute is binary; second, protected groups DI_b correspond to nominal attributes. Given an original pattern set $\mathcal{F}(\mathcal{D}, \sigma)$, denoted by $\mathcal{FP}$ for short, a discriminatory threshold α, a discrimination measure f, protected groups DI_b and a class item C which denies some benefit, Algorithm 2 starts by obtaining $\mathcal{D}_D$, which contains α-discriminatory patterns in $\mathcal{FP}$ (Line 3). It uses Algorithm 1 to do this.

Algorithm 2 ANTI-DISCRIMINATION PATTERN SANITIZATION

1: Inputs: Database $\mathcal{D}$, $\mathcal{FP} := \mathcal{F}(D, \sigma)$, $\mathcal{D}_D$, DI_b, discrimination measure f, α, C = class item with negative decision value
2: Output: $\mathcal{FP}^*$: α-protective version of $\mathcal{FP}$
3: $\mathcal{D}_D \leftarrow$ **Function** ANTIDISCPATTSANIT($\mathcal{FP}$, $\mathcal{D}$, $\mathcal{D}_D$, DI_b, f, α, C) // Algorithm 1
4: **for all** $p \in \mathcal{D}_D$ **do**
5: Compute $impact(p) = |\{p' : A', B', C\} \in \mathcal{D}_D \text{ s.t. } p' \subset p|$
6: **end for**
7: Sort $\mathcal{D}_D$ by descending $impact$
8: **for all** $p : \{A, B, C\} \in \mathcal{D}_D$ **do**
9: $X = p \backslash C$
10: Compute Δ_f for pattern p w.r.t. the value of f using $\mathcal{D}$, $\mathcal{FP}$ and α
11: **if** $\Delta_f \geq 1$ **then**
12: **if** $f = slift$ or $f = slift_d$ **then**
13: $p_t = X$
14: **else if** $f = elift$ or $elift_d$ **then**
15: $Y = p \cap DI_b$
16: $p_t = p \backslash Y$
17: **end if**
18: **end if**
19: $\mathcal{D}_s = \{p_s \in \mathcal{FP} | p_s \subseteq p_t\}$
20: **for all** $\langle p_s, supp(p_s)\rangle \in \mathcal{D}_s$ **do**
21: $supp(p_s) = supp(p_s) + \Delta_f$
22: **end for**
23: **end for**
24: **return** $\mathcal{FP}^* = \mathcal{FP}$
25: **End Function**

As we showed in Theorem 1, given two α-discriminatory patterns $p_1 : \{A, B, C\}$ and $p_2 : \{A', B', C\}$, where $p_2 \subset p_1$, making p_1 α-protective w.r.t. f can make also p_2 less discriminatory or even α-protective, depending on the value of α and the support of patterns. This justifies why, among the patterns in $\mathcal{D}_D$, Algorithm 2 transforms first those with maximum impact on making other patterns α-protective w.r.t. f. For each pattern $p \in \mathcal{D}_D$, the number of patterns in $\mathcal{D}_D$ which are subsets of p is taken as the impact of p (Lines 4-6), that is $impact(p)$. Then, the patterns in $\mathcal{D}_D$ will be made α-protective w.r.t. f by descending order of $impact$ (Line 7). Thus, the patterns with maximum $impact(p)$ will be made α-protective first, with the aim of minimizing the pattern distortion. Finally, the algorithm performs anti-discrimination pattern sanitization to make each pattern p in $\mathcal{D}_D$ α-protective using anti-discrimination pattern sanitization methods w.r.t. f (Lines 8-23).

6. EXPERIMENTAL ANALYSIS

This section presents the experimental evaluation of the approach we proposed in this paper. First, we describe the utility measures and then the empirical results. In the sequel, $\mathcal{FP}$ denotes the set of frequent patterns extracted from database $\mathcal{D}$ by the Apriori algorithm [1]; and $\mathcal{FP}^*$ denotes the α-protective version of $\mathcal{FP}$ obtained by Algorithm 2. We used the Adult and German credit datasets from the UCI Repository of Machine Learning Databases [8]. These are well-known real-life datasets, containing both numerical and categorical attributes, frequently used in anti-discrimination research. The Adult dataset is also known as Census Income consists of 48,842 records, split into a "train" part with 32,561 records and a "test" part with 16,281 records. The dataset has 14 attributes (without the class attribute). We used the "train" part to obtain $\mathcal{FP}$. The prediction task associated to the Adult dataset is to determine whether a person makes more than 50K$ a year based on census and demographic information about people. For our experiments with the Adult dataset, we set DI_b : {*Sex = female, Age = young*} (cut-off for Age = young: 30 years old). The German credit dataset consists of 1,000 records and 20 attributes (without class attribute) of bank account holders. The class attribute in the German credit dataset takes values representing good or bad classification of the bank account holders. For our experiments with this dataset, we set DI_b : {*Age = old, Foreign worker = yes*} (cut-off for Age = old: 50 years old).

6.1 Utility measures

To assess the information loss incurred to achieve fairness in frequent pattern discovery, we use the following measures.

- **Patterns with changed support**. Fraction of original frequent patterns in $\mathcal{FP}$ which have their support changed in any transformed pattern set $\mathcal{FP}^*$:

$$\frac{|\{\langle I, supp(I)\rangle \in \mathcal{FP} : supp_{\mathcal{FP}}(I) \neq supp_{\mathcal{FP}^*}(I)\}|}{|\mathcal{FP}|}$$

- **Pattern distortion error**. Average distortion w.r.t. the original support of frequent patterns:

$$\frac{1}{|\mathcal{FP}|} \cdot \sum_{I \in \mathcal{FP}} \left(\frac{supp_{\mathcal{FP}^*}(I) - supp_{\mathcal{FP}}(I)}{supp_{\mathcal{FP}}(I)}\right)$$

The purpose of these measures is assessing the distortion introduced when making $\mathcal{FP}$ discrimination-free. Moreover, we measure the impact of our pattern sanitization method for making $\mathcal{FP}$ α-protective on the accuracy of a classifier using the CMAR (*i.e.* classification based on multiple association rules) approach [14]. Below, we describe the process of our evaluation:

1. The original data are first divided into training and testing sets, $\mathcal{D}_{train}$ and $\mathcal{D}_{test}$.
2. The original frequent patterns $\mathcal{FP}$ are extracted from the training set $\mathcal{D}_{train}$ by the Apriori algorithm [1].

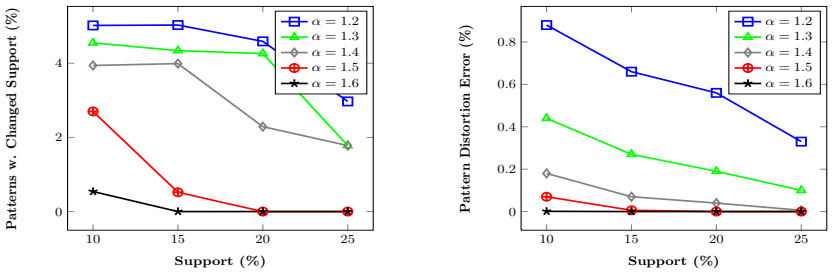

Figure 1: Pattern distortion scores to make the Adult dataset α-protective

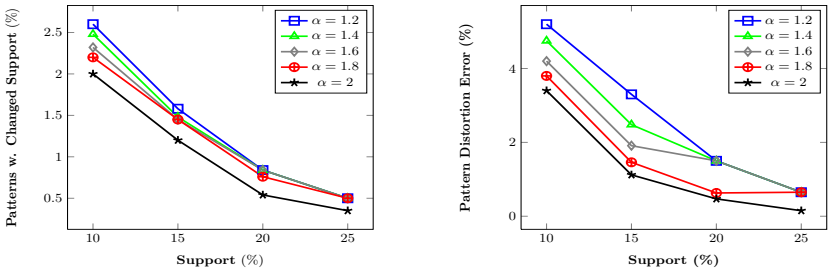

Figure 2: Pattern distortion scores to make the German dataset α-protective

3. Given $\mathcal{FP}$, $\mathcal{FP}^*$ is generated by Algorithm 2.

4. Patterns in $\mathcal{FP}^*$ which contain the class item are selected as a candidate patterns for classification. They are donated by $\mathcal{FP}_s^*$.

5. To classify each new object (record) in $\mathcal{D}_{test}$, the subset of patterns matching the new record in $\mathcal{FP}_s^*$ is found.

6. If all the patterns matching the new object have the same class item, then that class value is assigned to the new record. If the patterns are not consistent in terms of class items, the patterns are divided into groups according to class item values (*e.g.* denying credit and accepting credit). Then, the effects of the groups should be compared to yield the strongest group. A strongest group is composed of a set of patterns highly positively correlated and that have good support. To determine this, for each pattern $p : \{X, C\}$, the value of $\max \chi^2$ is computed as follows:

$$\max \chi^2 = (\min\{supp(X), supp(C)\} - \frac{supp(X) \times supp(C)}{|\mathcal{D}_{train}|})^2 \times |\mathcal{D}_{train}| \times e$$

where

$$e = \frac{1}{supp(X) \times supp(C)} + \frac{1}{supp(X) \times (|\mathcal{D}_{train}| - supp(C))} + \frac{1}{(|\mathcal{D}_{train}| - supp(X)) \times supp(C)} + \frac{1}{(|\mathcal{D}_{train}| - supp(X))(|\mathcal{D}_{train}| - supp(C))}$$

Thus, for each group of patterns, the *weighted* χ^2 measure of the group is computed[5] as $\sum \frac{\chi^2 \times \chi^2}{max\chi^2}$. The class item of the group with maximum *weighted* χ^2 is assigned to the new record.

7. After obtaining the predicted class item of each record in $\mathcal{D}_{test}$, the accuracy of the classifier can be simply computed w.r.t. observed and predicted class items (*i.e. contingency table).*

6.2 Pattern distortion

Fig. 1 and Fig. 2 show pattern distortion scores observed after making $\mathcal{FP}$ α-protective in Adult and German credit, respectively. We take $f = slift$ and we show the results for varying values of α and support σ. It can be seen that distortion scores increase with smaller α and smaller σ, because the number of α-discriminatory patterns increases. Comparing the two datasets, pattern distortion scores in German credit are higher than those in Adult, even taking the same values of α and σ. This is mainly due to the difference in the number of α-discriminatory patterns detected in the two datasets. We performed the same experiments for discrimination measures other than *slift* and we observed a similar behavior. Different measures of discrimination lead to more or less number of α-discriminatory patterns and consequently more or less pattern distortion scores. Summarizing, we observe that our pattern sanitization method yields discrimination-protected patterns by introducing reasonable pattern distortion.

6.3 Classification accuracy

Tables 2 show the accuracy of classifiers obtained from $\mathcal{FP}$ and $\mathcal{FP}^*$ in the Adult and German credit datasets for $f = slift$, $\sigma = 10\%$ and different values of α. We do not observe a significant difference between the accuracy of the

[5]For computing χ^2 see `http://www.csc.liv.ac.uk/~frans/Notes/chiTesting.html`

classifier obtained from an α-protective version of the original pattern set and the accuracy of the classifier obtained from the original pattern set. In addition, the accuracy of the classifier decreases with smaller α. When comparing the two datasets, we observe less accuracy for the German credit dataset; this is consistent with the higher distortion observed above for this dataset. Note that the low values of accuracy in Tables 2 are related to worst-case scenario (minimum value of α).

Table 2: Accuracy of classifiers: Adult dataset (left) and German credit dataset (right)

α	$\mathcal{FP}$	$\mathcal{FP}^*$
1.2	0.744	0.691
1.6	0.744	0.739
1.8	0.744	0.740

α	$\mathcal{FP}$	$\mathcal{FP}^*$
1.2	0.7	0.645
1.6	0.7	0.658
1.8	0.7	0.674

7. CONCLUSION AND FUTURE RESEARCH

In this paper, we have investigated the problem of discrimination-aware frequent pattern publishing. In particular, we have proposed methods for sanitizing patterns mined from a transaction database in such a way that discriminatory inferences cannot be inferred on the released patterns. For each measure of discrimination used in the legal literature, we have proposed a solution for obtaining a discrimination-free collection of patterns, thus achieving a robust (and formal) notion of fairness in the resulting pattern collection. Further, we have presented empirical results on the utility of the protected data. Specifically, we evaluate the distortion introduced by our methods and its effects on classification. It turns out that the utility loss caused by anti-discrimination is marginal. This result supports the practical deployment of our methods. Genuine occupational requirement refers to detecting that part of the discrimination may be explainable by other attributes, *e.g.*, denying credit to women may be explainable by the fact that most of them have low salary or delay in returning previous credits. We plan to extend our algorithms to also take into account the legal concept of genuine occupational requirement for making an original pattern set protected only against unexplainable discrimination.

8. ACKNOWLEDGMENTS

This work has been supported by the European FET OPEN projects "LIFT" and "MODAP", by the European FP7 infrastructure project "DwB", by Spain's TIN2011-27076-C03-01 and CSD2007-00004 "ARES", and by Catalonia's 2009 SGR 1135. The fourth author is partially funded by ICREA. The first and fourth authors are with the UNESCO Chair in Data Privacy, but the views in this paper do not commit UNESCO.

9. REFERENCES

[1] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In *VLDB*, pp. 487-499, 1994.

[2] Australian Legislation. (a) Equal Opportunity Act – Victoria State, (b) Anti-Discrimination Act – Queensland State, 2008. `http://www.austlii.edu.au`

[3] T. Calders and S. Verwer. Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21(2):277-292, 2010.

[4] B. Custers, T. Calders, B. Schermer and T. Z. Zarsky (eds.). *Discrimination and Privacy in the Information Society - Data Mining and Profiling in Large Databases.* Studies in Applied Philosophy, Epistemology and Rational Ethics 3. Springer, 2013.

[5] C. Dwork, M. Hardt, T. Pitassi, O. Reingold and R. S. Zemel. Fairness through awareness. In *ITCS 2012*, pp. 214-226. ACM, 2012.

[6] European Union Legislation. Directive 95/46/EC, 1995.

[7] European Union Legislation, (a) Race Equality Directive,2000/43/EC, 2000; (b) Employment Equality Directive, 2000/78/EC, 2000; (c) Equal Treatment of Persons, European Parliament legislative resolution, P6_TA(2009)0211, 2009.

[8] A. Frank and A. Asuncion. UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science, 2010. `http://archive.ics.uci.edu/ml/datasets`

[9] S. Hajian and J. Domingo-Ferrer. A methodology for direct and indirect discrimination prevention in data mining. *IEEE TKDE*, 25(7): 1445-1459, 2013.

[10] S. Hajian, A. Monreale, D. Pedreschi, J. Domingo-Ferrer and F. Giannotti. Injecting discrimination and privacy awareness into pattern discovery. In *IEEE ICDM Workshops*, pp. 360-369, 2012.

[11] F. Kamiran and T. Calders. Data preprocessing techniques for classification without discrimination. *KAIS*, 33(1): 1-33, 2011.

[12] F. Kamiran, T. Calders and M. Pechenizkiy. Discrimination aware decision tree learning. In *ICDM*, pp. 869-874. IEEE, 2010.

[13] T. Kamishima, S. Akaho, H. Asoh, J. Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *ECML/PKDD*, LNCS 7524, pp. 35-50, 2012.

[14] W. Li, J. Han and J. Pei. CMAR: accurate and efficient classification based on multiple class-association rules. In *ICDM*, pp. 369-376, 2001.

[15] B. L. Loung, S. Ruggieri and F. Turini. k-NN as an implementation of situation testing for discrimination discovery and prevention. In *KDD*, pp. 502-510, 2011.

[16] D. Pedreschi, S. Ruggieri and F. Turini. Discrimination-aware data mining. In *KDD*, pp. 560-568, 2008.

[17] D. Pedreschi, S. Ruggieri and F. Turini. Measuring discrimination in socially-sensitive decision records. In *SDM 2009*, pp. 581-592. SIAM, 2009.

[18] S. Ruggieri, D. Pedreschi and F. Turini. Data mining for discrimination discovery. *ACM TKDD*, 4(2), Article 9, 2010.

[19] United States Congress, *US Equal Pay Act*, 1963.