Discrimination-aware Data Mining

Dino Pedreschi Salvatore Ruggieri Franco Turini

Dipartimento di Informatica, Università di Pisa
L.go B. Pontecorvo 3, 56127 Pisa, Italy
{pedre,ruggieri,turini}@di.unipi.it

ABSTRACT

In the context of civil rights law, discrimination refers to unfair or unequal treatment of people based on membership to a category or a minority, without regard to individual merit. Rules extracted from databases by data mining techniques, such as classification or association rules, when used for decision tasks such as benefit or credit approval, can be discriminatory in the above sense. In this paper, the notion of discriminatory classification rules is introduced and studied. Providing a guarantee of non-discrimination is shown to be a non trivial task. A naïve approach, like taking away all discriminatory attributes, is shown to be not enough when other background knowledge is available. Our approach leads to a precise formulation of the redlining problem along with a formal result relating discriminatory rules with apparently safe ones by means of background knowledge. An empirical assessment of the results on the German credit dataset is also provided.

Categories and Subject Descriptors

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

General Terms

Algorithms, Economics, Legal Aspects

Keywords

Discrimination, Classification Rules

1. INTRODUCTION

The word *discrimination* originates from the Latin *discriminare*, which means to "distinguish between". In social sense, however, discrimination refers specifically to an action based on prejudice resulting in unfair treatment of people on the basis of their membership to a category, without regard to individual merit. As an example, U.S. federal laws [17] prohibit discrimination on the basis of race, color, religion, nationality, sex, marital status, age and pregnancy in a number of settings, including: credit/insurance scoring (Equal Credit Opportunity Act); sale, rental, and financing of housing (Fair Housing Act); personnel selection and wage (Intentional Employment Act, Equal Pay Act, Pregnancy Discrimination Act).

Concerning the research side, the issue of discrimination in credit, mortgage, insurance, labor market, education and other human activities has attracted much interest of researchers in economics and human sciences since late '50s, when a theory on the economics of discrimination was proposed [3]. The literature has given evidence of unfair treatment in racial profiling and redlining [14], mortgage discrimination [9], personnel selection discrimination [6, 7], and wages discrimination [8].

In data mining and machine learning, classification models are constructed on the basis of historical data exactly with the purpose of discrimination in the original Latin sense: i.e. distinguishing between elements of different classes, in order to unveil the reasons of class membership, or to predict it for unclassified samples. In either cases, classification models can be adopted as a support to decision making, clearly also in socially sensitive tasks such as the access of applicants to benefits, to public services, to credit. Now the question that naturally arises is the following. While classification models used for decision support can potentially guarantee less arbitrary decisions, can they be discriminating in the social, negative sense? The answer is clearly yes: it is evident that relying on mined models for decision making does not put ourselves on the safe side. Rather dangerously, learning from historical data may mean to discover traditional prejudices that are endemic in reality, and to assign to such practices the status of general rules, maybe unconsciously, as these rules can be deeply hidden within a classifier.

Surprisingly, despite the risk of discrimination poses clear ethical and legal obstacles to the practical application of data mining in socially sensitive decision making, to the best of our knowledge, there is no prior work on the issue. In this paper, we tackle the problem of discrimination in data mining in a rule-based setting, by introducing the notion of *discriminatory classification rules*, as a criterion to identify the potential risks of discrimination.

2. CONTROLLING DISCRIMINATION

The first natural approach to formally tackle the problem is to specify a set of selected attribute values (or, at an extreme, an attribute as a whole) as *potentially discriminatory*: examples include female gender, ethnic minority, low-level job, specific age range. However, this simple ap-

KDD'08, August 24–27, 2008, Las Vegas, Nevada, USA.

proach is flawed, in that discrimination may be the result of several joint characteristics that are not discriminatory in isolation. For instance, black cats crossing your path are typically discriminated as signs of bad luck, but no superstition is independently associated to being a cat, being black or crossing a path. In other words, the condition that describes a (minority) population that may be the object of discrimination should be stated as a conjunction of attributes values: pregnant women, minority ethnicity in disadvantaged neighborhoods, senior people in weak economic conditions, and so on. Coherently, we qualify as potentially discriminatory (PD) some selected itemsets, not necessarily single items nor whole attributes. Two consequences of this approach should be considered. First, single PD items or attributes are just a particular case in this more general setting. Second, PD itemsets are closed under intersection: the conjunction of two PD itemsets is a PD itemset as well, coherently with the intuition that the intersection of two disadvantaged minorities is a, possibly empty, smaller (even more disadvantaged) minority as well. In our approach, we assume that the analyst interested in studying discrimination compiles a list of PD itemsets with reference to attribute and attribute values that are present either in the data, or in his/her background knowledge, or in both.

Discrimination has been identified in law and social study literature as either *direct* or *indirect* (sometimes called systematic). Direct discrimination consists of rules or procedures that explicitly impose "disproportionate burdens" on minority or disadvantaged groups. Indirect discrimination consists of rules or procedures that, while not explicitly mentioning discriminatory attributes, intentionally or not impose the same disproportionate burdens.

Direct discrimination is modelled through *potentially discriminatory rules*, which are classification rules $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$ that contain potentially discriminatory itemsets $\mathbf{A}$ in their premises. We show in Sect. 4.1 that there is always a unique split of the premise into a PD part and a non PD part. A PD rule does not necessarily provide evidence of discriminatory actions. In order to measure the "disproportionate burdens" that a rule imposes, the notion of *α-protection* is introduced as a measure of the discrimination power of a PD classification rule. The idea is to define such a measure as the relative gain in confidence of the rule due to the presence of the discriminatory itemsets. The α parameter is the key for tuning the desired level of protection against discrimination. PD classification rules are extracted (see Fig. 1 left) from a dataset containing discriminatory itemsets. This is the case, for instance, when:

- *internal auditors* or *regulation authorities* want to identify discriminatory rules to the purpose of discovering malpractices that emerge from the historical transactions; they collect the dataset of past transactions and enrich it with potentially discriminatory itemsets in order to extract discriminatory PD rules;

- *data miners* want to extract models from a dataset that contains potentially discriminatory attributes that are essential for the purpose of classification, such as in the case of gender, age, and job type. Using these attributes for building classifiers is perfectly legal: it is their use in discriminatory decisions that may be illegal! Thus, data miners must remove from the set of extracted PD rules the discriminatory ones;

- *consumer advisor councils* or *regulation authorities* want to identify certain expected discriminatory PD rules to the purpose of checking that the results of specific positive discrimination policies – or affirmative actions, that tend to favor some disadvantaged categories – actually emerge from the historical transactions.

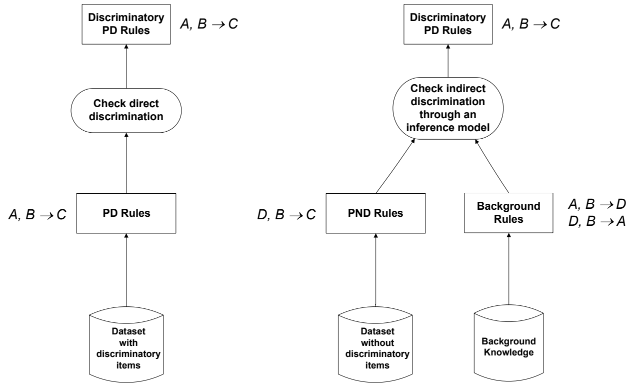
Discriminatory PD Rules
A, B → C
Check direct discrimination
A, B → C
PD Rules
Dataset with discriminatory items
Discriminatory PD Rules
A, B → C
Check indirect discrimination through an inference model
D, B → C
PND Rules
Background Rules
A, B → D
D, B → A
Dataset without discriminatory items
Background Knowledge

Figure 1: Modelling the process of direct (left) and indirect (right) discrimination control.

Concerning indirect discrimination, we consider rules $\mathbf{D}, \mathbf{B} \rightarrow \mathbf{C}$ that are potentially non-discriminatory (PND), i.e., that do not contain PD itemsets. They are extracted (see Fig. 1 right) from a dataset which may or may NOT contain PD itemsets. While apparently safe, PND rules may lead to discrimination as well. As an example, assume that the PND rule "rarely give credit to persons from neighborhood 10451 from NYC" is extracted. This may be or may be not a redlining rule. In order to unveil its nature, we have to rely on additional *background knowledge*. If we know that in NYC people from neighborhood 10451 are in majority black race, then using the rule above is like using the rule "rarely give credit to black-race persons from neighborhood 10451 of NYC", which is definitely discriminatory. This use case resembles the situation described in privacy-preserving data mining [2, 15], where an anonymized dataset coupled with external knowledge might allow for the inference of the identity of individuals. In our framework, we assume that background knowledge takes the form of association rules relating a non-discriminatory itemset $\mathbf{D}$ to a discriminatory itemset $\mathbf{A}$ within the context $\mathbf{B}$. Examples of background knowledge include the one originating from publicly available data (e.g., census data), from privately owned data (e.g., market surveys) or from experts or common sense (e.g., expert rules about customer behavior). Again, internal auditors, regulation authorities, consumer advisory councils, and data miners are interested for their own purposes in checking indirect discrimination by identifying PND rules that are to a certain extent equivalent to discriminatory PD rules. In order to model such a situation, we consider an *inference model*, i.e., a strategy that an analyst, provided with background knowledge, can pursue in order to unveil discriminatory PD rules starting from PND ones.

As an example of the overall processes shown in Fig. 1, consider the rules:

```
a. city=NYC                b. race=black, city=NYC
   ==> class=bad              ==> class=bad
   -- conf:(0.25)             -- conf:(0.75)
```

Rule (a) can be translated into the statement "people who live in NYC are assigned the bad credit class" 25% of times. Rule (b) concentrates on "black people from NYC". In this case, the additional (discriminatory) item in the premise increases the confidence of the rule up to 75%! α-protection is intended to detect rules where such an increase is lower than a fixed threshold α.

In direct discrimination, rules such as (a) and (b) above are extracted from the dataset and then α-protection can be easily checked (see Fig. 1 left). For instance, if the threshold for acceptable α-protection has been fixed to 3, rule (b) is classified as discriminatory. Tackling indirect discrimination is more challenging. Continuing the example, consider the classification rule:

```
c. neighborhood=10451, city=NYC
   ==> class=bad
   -- conf:(0.95)
```

extracted from a dataset where potentially discriminatory itemsets, such as race=black, are NOT present (see Fig. 1 right). Taken in isolation, rule (c) cannot be considered discriminatory or not. Assume now to know that people from neighborhood 10451 are in majority black, i.e., the following association rule holds:

```
d. neighborhood=10451, city=NYC
   ==> race=black
   -- conf:(0.80)
```

Despite rule (c) contains no discriminatory item, it leads to the (discriminatory) decision of denying credit to a minority sub-group (black people) which has been "redlined" by their ZIP code. In other words, the PD rule:

```
e. race=black, neighborhood=10451, city=NYC
   ==> class=bad
```

can be inferred from (c) and (d), together with a lower bound of 94% for its confidence. Such a lower bound shows a disproportionate burden (94% / 25%, i.e., 3.7 times) over black people living in neighborhood 10451. We will show a formal theorem that allows us to derive the lower bound for α-protection of (e) starting from PND rules (a) and (c) and a lower bound on the confidence of the background rule (d). Clearly, the proposed inference model provides *sufficient* conditions for checking indirect discrimination. If the inferred lower bound is not as high as to conclude non α-protection, we cannot state that an analyst has no other means to derive the same conclusion, e.g., by using another inference model or additional background knowledge.

The German credit case study

Throughout the paper, we illustrate the notions introduced by analysing the public domain German credit dataset [11], consisting of 1000 transactions representing the good/bad credit class of bank account holders. The dataset include nominal (or discretized) attributes on *personal properties*: checking account status, duration, savings status, property magnitude, type of housing; on *past/current credits and requested credit*: credit history, credit request purpose, credit request amount, installment commitment, existing credits, other parties, other payment plan; on *employment status*: job type, employment since, number of dependents, own telephone; and on *personal attributes*: personal status and gender, age, resident since, foreign worker.

3. BASIC DEFINITIONS

3.1 Association and Classification Rules

We recall the notions of itemsets, association rules and classification rules from standard definitions [1, 10, 18]. Let $\mathcal{R}$ be a relation with attributes $a_1, \ldots, a_n$. A class attribute is a fixed attribute c of the relation. An a-item is an expression $a = v$, where a is an attribute and $v \in dom(a)$, the domain of a. We assume that $dom(a)$ is finite for every attribute a. A c-item is called a class item. An item is any a-item. Let I be the set of all items. A transaction is a subset of I, with exactly one a-item for every attribute a. A database of transactions, denoted by $\mathcal{D}$, is a set of transactions. An itemset $\mathbf{X}$ is a subset of I. We denote by 2^I the set of all itemsets. As usual in the literature, we write $\mathbf{X}, \mathbf{Y}$ for $\mathbf{X} \cup \mathbf{Y}$. For a transaction T, we say that T verifies $\mathbf{X}$ if $\mathbf{X} \subseteq T$. The support of an itemset $\mathbf{X}$ w.r.t. a non-empty transaction database $\mathcal{D}$ is the ratio of transactions in $\mathcal{D}$ verifying $\mathbf{X}$: $supp_{\mathcal{D}}(\mathbf{X}) = |\{ T \in \mathcal{D} \mid \mathbf{X} \subseteq T \}|/|\mathcal{D}|$, where $|\ |$ is the cardinality operator. An association rule is an expression $\mathbf{X} \rightarrow \mathbf{Y}$, where $\mathbf{X}$ and $\mathbf{Y}$ are itemsets. $\mathbf{X}$ is called the *premise* (or the *body*) and $\mathbf{Y}$ is called the *consequence* (or the *head*) of the association rule. We say that $\mathbf{X} \rightarrow \mathbf{Y}$ is a *classification rule* if $\mathbf{Y}$ is a class item and $\mathbf{X}$ contains no class item. We refer the reader to [10, 18] for a discussion of the integration of classification and association rule mining. The support of $\mathbf{X} \rightarrow \mathbf{Y}$ w.r.t. $\mathcal{D}$ is defined as: $supp_{\mathcal{D}}(\mathbf{X} \rightarrow \mathbf{Y}) = supp_{\mathcal{D}}(\mathbf{X}, \mathbf{Y})$. The confidence of $\mathbf{X} \rightarrow \mathbf{Y}$, defined when $supp_{\mathcal{D}}(\mathbf{X}) > 0$, is:

$$conf_{\mathcal{D}}(\mathbf{X} \rightarrow \mathbf{Y}) = supp_{\mathcal{D}}(\mathbf{X}, \mathbf{Y})/supp_{\mathcal{D}}(\mathbf{X}).$$

Support and confidence range over $[0, 1]$. We omit the subscripts in $supp_{\mathcal{D}}()$ and $conf_{\mathcal{D}}()$ when clear from the context. Since the seminal paper by Agrawal and Srikant [1], a number of well explored algorithms [5] have been introduced in order to extract *frequent* itemsets, i.e. itemsets with a specified minimum support, and valid association rules, i.e. rules with a specified minimum confidence.

3.2 Extended Lift

We introduce a key concept for our purposes.

Definition 3.1. *[Extended lift] Let $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$ be an association rule such that $conf(\mathbf{B} \rightarrow \mathbf{C}) > 0$. We define the extended lift of the rule with respect to $\mathbf{B}$ as:*

$$\frac{conf(\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C})}{conf(\mathbf{B} \rightarrow \mathbf{C})}.$$

We call $\mathbf{B}$ the context, and $\mathbf{B} \rightarrow \mathbf{C}$ the base-rule.

Intuitively, the extended lift expresses the relative variation of confidence due to the addition of the extra itemset $\mathbf{A}$ in the premise of the base rule $\mathbf{B} \rightarrow \mathbf{C}$. In general, the extended lift ranges over $[0, \infty[$. However, if association rules with a minimum support $ms > 0$ are considered, it ranges over $[0, 1/ms]$. Similarly, if association rules with base-rules with a minimum confidence $mc > 0$ are considered, it ranges over $[0, 1/mc]$. The extended lift can be traced back to the well-known measure of lift [16], defined as:

$$lift_{\mathcal{B}}(\mathbf{A} \rightarrow \mathbf{C}) = conf_{\mathcal{B}}(\mathbf{A} \rightarrow \mathbf{C})/supp_{\mathcal{B}}(\mathbf{C}),$$

when $\mathcal{B} = \{T \in \mathcal{D} \mid \mathbf{B} \subseteq T\}$. When $\mathbf{B}$ is empty, the extended lift reduces to the standard lift.

4. MEASURING DISCRIMINATION

4.1 Discriminatory Itemsets and Rules

Our starting point consists of flagging at syntax level those itemsets which might potentially lead to discrimination in the sense explained in the introduction. A set of itemsets $\mathcal{I} \subseteq 2^I$ is downward closed if when $\mathbf{A}_1 \in \mathcal{I}$ and $\mathbf{A}_2 \in \mathcal{I}$ then $\mathbf{A}_1, \mathbf{A}_2 \in \mathcal{I}$.

Definition 4.1. *[PD/PND itemset] A set of potentially discriminatory (PD) itemsets $\mathcal{I}_d$ is any downward closed set. Itemsets in $2^{\mathcal{I}} \setminus \mathcal{I}_d$ are called potentially non-discriminatory (PND).*

Any itemset $\mathbf{X}$ can be uniquely split into a PD part $\mathbf{A}$ and a PND part $\mathbf{B} = \mathbf{X} \setminus \mathbf{A}$ by setting $\mathbf{A}$ to the largest subset of $\mathbf{X}$ that belongs to $\mathcal{I}_d$[1]. A simple way of defining PD itemsets is to take those that are built from a pre-defined set of items, i.e., to reduce to the case where the granularity of discrimination is at the level of items.

Example 4.2. *For the German credit dataset, we fix $\mathcal{I}_d = 2^{I_d}$, where I_d is the set of the following (discriminatory) items:* `personal_status=female div/sep/mar` *(female and not single),* `age=(52.6-inf)` *(senior people),* `job=unemp/unskilled non res` *(unskilled or unemployed non-resident), and* `foreign_worker=yes` *(foreign workers). Notice that the PD part of an itemset $\mathbf{X}$ is now easily identifiable as $\mathbf{X} \cap I_d$, and the PND part as $\mathbf{X} \setminus I_d$.*

It is worth noting that discriminatory items do not necessarily coincide with sensitive attributes with respect to pure privacy protection. For instance, gender is generally considered a non-sensitive attribute, whereas it can be discriminatory in many decision contexts. Moreover, note that we use the adjective *potentially* both for PD and PND itemsets. As we will discuss later on, also PND may unveil (indirect) discrimination. The notion of potential (non-)discrimination is now extended to classification rules.

Definition 4.3. *[PD/PND classification rule] A classification rule $\mathbf{X} \rightarrow \mathbf{C}$ is potentially discriminatory (PD) if $\mathbf{X} = \mathbf{A}, \mathbf{B}$ with $\mathbf{A}$ non-empty PD itemset and $\mathbf{B}$ PND itemset. It is potentially non-discriminatory (PND) if $\mathbf{X}$ is a PND itemset.*

It is worth noting that PD rules can be either extracted from a dataset that contain PD itemsets or inferred as shown in Fig. 1 right. PND rules can be extracted from a dataset which may or may not contain PD itemsets.

Example 4.4. *Consider Ex. 4.2, and the rules:*

```
a. personal_status=female div/sep/mar
   savings_status=no known savings
   ==> class=bad

b.  savings_status=no known savings
   ==> class=bad
```

`(a)` *is a PD rule since its premise contains an item belonging to I_d. On the contrary,* `(b)` *is a PND rule. Notice that* `(b)` *is the base rule of* `(a)` *if we consider as context the PND part of its premise.*

4.2 α-protection

We start concentrating on PD classification rules as the potential source of discrimination. In order to capture the idea of when a PD rule may lead to discrimination, we introduce the key concept of α-protective classification rules.

Definition 4.5. *[α-protection] Let $c = \mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$ be a PD classification rule, where $\mathbf{A}$ is a PD and $\mathbf{B}$ is a PND itemset, and let:*

$$\gamma = conf(\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}) \qquad \delta = conf(\mathbf{B} \rightarrow \mathbf{C}) > 0.$$

For a given threshold $\alpha \geq 0$, we say that c is α-protective if $elift(\gamma, \delta) < \alpha$, where:

$$elift(\gamma, \delta) = \gamma/\delta.$$

c is called α-discriminatory if $elift(\gamma, \delta) \geq \alpha$.

Intuitively, the definition assumes that the extended lift of c w.r.t. $\mathbf{B}$ is a measure of the degree of discrimination of $\mathbf{A}$ in the context $\mathbf{B}$. α-protection states that the added (potentially discriminatory) information $\mathbf{A}$ increases the confidence of concluding an assertion $\mathbf{C}$ under the base hypothesis $\mathbf{B}$ only by an acceptable factor, bounded by α.

Example 4.6. *Consider again Ex. 4.2. Fix $\alpha = 3$ and consider the classification rules:*

```
a. personal_status=female div/sep/mar
   savings_status=no known savings
   ==> class=bad
   -- supp:(0.013) conf:(0.27) elift:(1.52)

b.  age=(52.6-inf)
   personal_status=female div/sep/mar
   purpose=used car
   ==> class=bad
   -- supp:(0.003) conf:(1) elift:(6.06)
```

Rule `(a)` *can be translated as follows: if we know nothing about the savings of a person asking for credit, then assign bad credit class (or bad credit class has been assigned in past) to non-single women 52% more than the average. The support of the rule is 1.3%, its confidence 27%, and its extended lift 1.52. Hence, the rule is α-protective. Also, the confidence of the base rule:*

```
savings_status=no known savings ==> class=bad
```

is $0.27/1.52 = 17.8\%$. *Rule* `(b)` *states that senior non-single women that want to buy a used car are assigned the bad credit class with a probability more than 6 times higher than the average one for those that ask credit for the same purpose. The support of the rule is 0.3%, its confidence 100%, and its extended lift 6.06. Hence the rule is α-discriminatory. Finally, the confidence of the base rule*

```
purpose=used car ==> class=bad
```

is $1/6.06 = 16.5\%$.

4.3 Strong α-protection

When the class is binary, the concept of α-protection must be strengthened, as highlighted by the next example.

Example 4.7. *The following PD classification rule is extracted from the German credit dataset with minimum support of 1%:*

[1] Notice that $\mathbf{A}$ is univocally defined. If there were two maximal $\mathbf{A}_1 \neq \mathbf{A}_2$ subsets belonging to $\mathcal{I}_d$, then $\mathbf{A}_1, \mathbf{A}_2$ would belong to $\mathcal{I}_d$ as well since $\mathcal{I}_d$ is downward closed. But then $\mathbf{A}_1$ or $\mathbf{A}_2$ would not be maximal.

```
ExtractCR()
  $\mathcal{C}$ = { class items }
  $\mathcal{PD}_{group} = \mathcal{PND}_{group} = \emptyset$ for $group \geq 0$
  ForEach $k$ s.t. there exists $k$-frequent itemsets
    $\mathcal{F}_k$ = { $k$-frequent itemsets }
    ForEach $\mathbf{Y} \in \mathcal{F}_k$ with $\mathbf{Y} \cap \mathcal{C} \neq \emptyset$
      $\mathbf{C} = \mathbf{Y} \cap \mathcal{C}$
      $\mathbf{X} = \mathbf{Y} \setminus \mathbf{C}$
      $s = supp(\mathbf{Y})$
      $s' = supp(\mathbf{X})$        // found in $\mathcal{F}_{k-1}$
      $conf = s/s'$
      $\mathbf{A}$ = largest subset of $\mathbf{X}$ in $\mathcal{I}_d$
      $group = |\mathbf{X} \setminus \mathbf{A}|$
      If $|\mathbf{X}| = 0$
        add $\mathbf{X} \rightarrow \mathbf{C}$ to $\mathcal{PND}_{group}$ with confidence $conf$
      Else
        add $\mathbf{X} \rightarrow \mathbf{C}$ to $\mathcal{PD}_{group}$ with confidence $conf$
      EndIf
    EndForEach
  EndForEach
```

```
CheckAlphaPDCR($\alpha$)
  ForEach $group$ s.t. $\mathcal{PD}_{group} \neq \emptyset$
    ForEach $\mathbf{X} \rightarrow \mathbf{C} \in \mathcal{PD}_{group}$
      $\mathbf{A}$ = largest subset of $\mathbf{X}$ in $\mathcal{I}_d$
      $\mathbf{B} = \mathbf{X} \setminus \mathbf{A}$
      $\gamma = conf(\mathbf{X} \rightarrow \mathbf{C})$
      $\delta = conf(\mathbf{B} \rightarrow \mathbf{C})$    // found in $\mathcal{PND}_{group}$
      If $elift(\gamma, \delta) \geq \alpha$        // resp., $glift(\gamma, \delta) \geq \alpha$
        output $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$
      EndIf
    EndForEach
  EndForEach
```

Figure 2: Extraction of PD and PND classification rules (left) and direct checking of α-discrimination (right).

```
a-good. personal_status=female div/sep/mar
        purpose=used car
        checking_status=no checking
        ==> class=good
        -- supp:(0.011) conf:(0.846)
        -- conf_base:(0.963) elift:(0.88)
```

Rule a-good *has an extended lift of* 0.88*. Intuitively, this means that* good *credit class is assigned to non-single women* less *than the average of people that want to buy an used car and have no checking status. As a consequence, one can deduce that the* bad *credit class is assigned* more *than the average of people in the same context, i.e. the rule:*

```
a-bad.  personal_status=female div/sep/mar
        purpose=used car
        checking_status=no checking
        ==> class=bad
        -- supp:(0.002) conf:(0.154)
        -- conf_base:(0.037) elift:(4.15)
```

It is worth noting that the confidence of rule a-bad in the example is equal to 1 minus the confidence of a-good, and the same holds for the confidence of base rules. This property holds in general for binary classes. For a binary attribute a with $dom(a) = \{v_1, v_2\}$, we write $\neg(a = v_1)$ for $a = v_2$ and $\neg(a = v_2)$ for $a = v_1$.

Lemma 4.8. *Assume that the class attribute is binary. Let $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$ be a classification rule, and let:*

$$\gamma = conf(\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}) \qquad \delta = conf(\mathbf{B} \rightarrow \mathbf{C}) < 1,$$

We have that $conf(\mathbf{B} \rightarrow \neg\mathbf{C}) > 0$ and:

$$\frac{conf(\mathbf{A}, \mathbf{B} \rightarrow \neg\mathbf{C})}{conf(\mathbf{B} \rightarrow \neg\mathbf{C})} = \frac{1-\gamma}{1-\delta}.$$

As an immediate consequence, the extraction or the inference of an α-protective rule $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$ allows the calculation of the extended lift of the dual rule $\mathbf{A}, \mathbf{B} \rightarrow \neg\mathbf{C}$, which could be α-discriminatory. We strengthen the notion of α-protection to take into account such an implication.

Definition 4.9. *[Strong α-protection] Let $c = \mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$ be a PD classification rule, where $\mathbf{A}$ is a PD and $\mathbf{B}$ is a PND itemset, and let:*

$$\gamma = conf(\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}) \qquad \delta = conf(\mathbf{B} \rightarrow \mathbf{C}) > 0.$$

For a given threshold $\alpha \geq 1$, we say that c is strongly α-protective if $glift(\gamma, \delta) < \alpha$, where:

$$glift(\gamma, \delta) = \begin{cases} \gamma/\delta & \text{if } \gamma \geq \delta \\ (1-\gamma)/(1-\delta) & \text{otherwise} \end{cases}$$

If $glift(\gamma, \delta) \geq \alpha$, we say that c is strongly α-discriminatory.

The $glift()$ function ranges over $[1, \infty[$. If classification rules with a minimum support $ms > 0$ are considered, it ranges over $[1, 1/ms]$. Moreover, for $1 > \delta > 0$:

$$glift(\gamma, \delta) = max\{elift(\gamma, \delta), elift(1-\gamma, 1-\delta)\}.$$

5. DIRECT DISCRIMINATION

Let us consider the case of direct discrimination, as modelled in Fig. 1 left and with α-protection as the underlying measure of discrimination. Given a set of PD classification rules $\mathcal{A}$ and a threshold α, the problem of checking (strong) α-protection consists of finding the largest subset of $\mathcal{A}$ containing only (strong) α-protective rules. This problem is solvable by directly checking the inequality of Def. 4.5 (resp., Def. 4.9), provided that the elements of the inequality are available. We define a checking algorithm that starts from the set of frequent itemsets, namely itemsets with a given minimum support. This is the output of any of the several frequent itemset extraction algorithms available at the FIMI repository [5]. The algorithm is reported in Fig. 2. On the left hand side of the figure, the extraction of PD and PND classification rules is reported. It requires a single scan of frequent itemsets ordered by the itemset size k. For k-frequent itemsets that include a class item, a single classification rule is produced in output. The confidence of the rule can be computed by looking only at itemsets of length $k-1$. The rules in output are distinguished between PD and PND rules, based on the presence of discriminatory items in

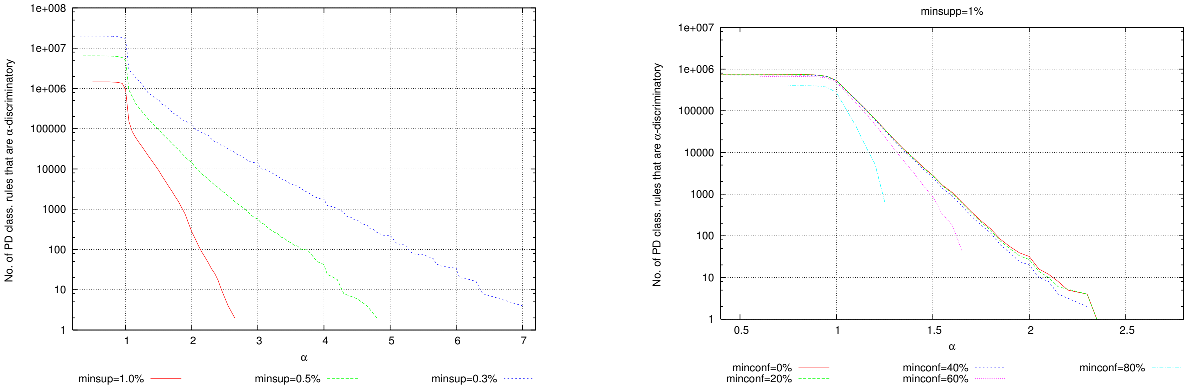

Figure 3: Left: distributions of α-discriminatory PD classification rules. Right: contribution of setting minimum confidence for base rules.

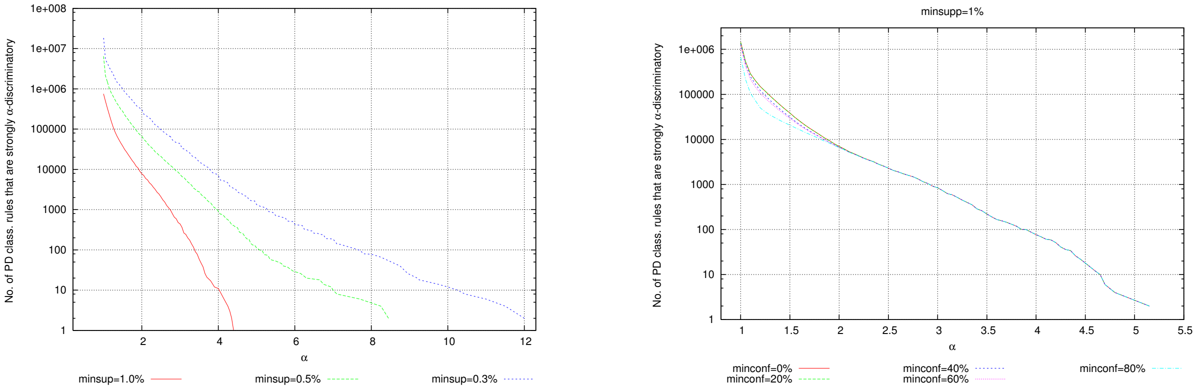

Figure 4: Left: distributions of strongly α-discriminatory PD classification rules. Right: contribution of setting minimum confidence for base rules.

their premises. Moreover, the rules are grouped on the basis of the size *group* of the PND part of the premise. The output is a collection of PD rules $\mathcal{PD}_{group}$ and a collection of PND rules $\mathcal{PND}_{group}$. On the right hand side of Fig. 2, the extended lift of a classification rule $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C} \in \mathcal{PD}_{group}$ is computed from its confidence and the confidence of the base rule $\mathbf{B} \rightarrow \mathbf{C} \in \mathcal{PND}_{group}$.

The German credit case study

The left-hand side of Fig. 3 (resp., Fig. 4) shows the distribution of α-discriminatory PD rules (resp., strong α-discriminatory PD rules) for minimum support of 1%, 0.5% and 0.3%. The figures highlight how lower support values increase the number and the proportion of PD rules and the maximum α. Notice that, for a same minimum support, α reaches higher values in Fig. 4 than in Fig. 3, since strong α-discrimination of a rule implicitly takes into account the complementary class rule, which may have a support lower than the minimum (see e.g., `(a-bad)` in Ex. 4.7). We report two sample PD rules with decreasing support and increasing extended lift.

```
a1. personal_status=female div/sep/mar
    employment=1<=X<4
    property_magnitude=real estate
    job=skilled
    ==> class=bad
    -- supp:(0.011) conf:(0.48) elift:(2.39)
```

```
a2. age=(52.6-inf)
    employment=1<=X<4
    savings_status=>=1000
    ==> class=bad
    -- supp:(0.002) conf:(1) elift:(9)
```

Rule `a1` states that among the people employed since one to four years, having a real estate property and with skilled job, the status of being woman and not single leads to having assigned the bad credit class 2.39 times more than the average. The rule has confidence 48%, which means that the base rule has confidence 0.48/2.39 = 20%. Rule `a2` reaches a lift of 9 when compared to the base rule:

```
employment=1<=X<4
savings_status=>=1000
==> class=bad
-- supp:(0.002) conf:(0.11)
```

People with large savings are usually given good credit. However, only 2 cases out of 18 (i.e., 11%) are assigned `class=bad`. Both of them are senior people!

In addition to minimum support, a widely adopted parameter for controlling rule generation is minimum confidence. The right-hand side of Fig. 3 shows how the confidence threshold of the base rule affects the distribution of α-discriminatory PD rules. Lower confidence thresholds lead to fewer number of discriminatory rules and lower maximum extended lift values. This is consistent with the observation that the extended lift ranges over $[0, 1/mc]$, where mc is the minimum confidence threshold of base rules.

This is not the case for strong α-protection, where acting on minimum confidence of the base rule does not turn out to be an effective control mechanism, as shown in Fig. 4 right.

6. INDIRECT DISCRIMINATION

Let us consider the case of indirect discrimination, as modelled in Fig. 1 right. The next example highlights a PND rule which leads to discrimination, and the background knowledge that allows for unveiling this.

Example 6.1. *Consider again the German credit dataset, but assume now that discriminatory items have been removed from it. Also, consider the following itemset:*

```
B = credit_history=critical/other existing credit
    residence_since=(2.8-inf)
    savings_status=<100
    checking_status=nochecking
```

The following PND classification rules can be extracted:

```
dbc. age=(-inf-30.2], B        bc. B
     ==> class=bad                 ==> class=bad
     -- conf:(0.167)               -- conf:(0.027)
```

Rule (dbc) *states that young people in the context* **B** *of people with critical credit history, residence since 2.8 years at least, with savings at most for 100 units, and with no checkings, are assigned the bad credit scoring with a confidence of 16.7%. Rule* (bc) *is obtained from* (dbc) *by discarding the item* `age=(-inf-30.2]` *in the premise, and it has a confidence of 2.7%. As discussed in Sect. 2, without any further information, we cannot say whether rule* (dbc) *is discriminatory or not. Assume now to know (by some background knowledge) that in the context* **B** *above, the set of persons satisfying* `age=(-inf-30.2]` *is somewhat related to the set of persons satisfying the discriminatory item* `personal_status=female div/sep/mar`*. If the two sets were exactly the same, we could replace* `age=(-inf-30.2]` *in rule* (dbc) *with the discriminatory item. This would lead us to the PD classification rule:*

```
abc. personal_status=female div/sep/mar, B
     ==> class=bad
```

with $glift(0.167, 0.027) = 6.19$*, which is considerably high.*

In case the two sets of persons coincide only to some extent, we can still obtain some lower bound for the $glift()$ *of* (abc)*. In particular, assume that young people in the context* **B***, contrarily to the average case, are almost all non-single women:*

```
dba. age=(-inf-30.2], B
     ==> personal_status=female div/sep/mar
     -- conf:(0.95)
```

Is this enough to conclude that non-single women in the context are discriminated? We cannot say that: for instance, if non-single women in the context are at 99% older than 30.2 *years, the remaining 1% is involved in the decisions fired by rule* (dbc)*, hence women in the context are not discriminated by these decisions. As a consequence, we need further information about the proportion of non-single women that are younger than* 30.2 *years. Assume to know that such a proportion is at least 70%, i.e. :*

```
abd. personal_status=female div/sep/mar, B
     ==> age=(-inf-30.2]
     -- conf:(0.7)
```

By means of the forthcoming Thm. 6.2, we can state that a lower bound for the $glift()$ *value of* (abc) *is* 3.19*. As a consequence, the rule* (abc) *is at least* 3.19*-discriminatory, i.e., non-single women in the context are imposed by* (abc) *a burden of at least* 3.19 *times than the average of people in the context. Since the German credit dataset contains the discriminatory items, we can calculate the actual* $glift()$ *value for* (abc)*, which turns out to be* 3.37*.*

We formalize the intuitions of this example in the next result, which derives 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). The non-trivial proof of the theorem (see [12]) relies on the inclusion-exclusion principle for boolean formulas over items, and is omitted for lack of space.

Theorem 6.2. *Let* $\mathbf{D}, \mathbf{B} \rightarrow \mathbf{C}$ *be a PND classification rule, and let:*

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

Let $\mathbf{A}$ *be a PD itemset and let* β_1, β_2 *such that:*

$$\begin{aligned} conf(\mathbf{A}, \mathbf{B} \rightarrow \mathbf{D}) &\geq \beta_1 \\ conf(\mathbf{D}, \mathbf{B} \rightarrow \mathbf{A}) &\geq \beta_2 > 0. \end{aligned}$$

Called:

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

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

$$glb(x, y) = \begin{cases} f(x)/y & \text{if } f(x) \geq y \\ f(1-x)/(1-y) & \text{elseif } f(1-x) > 1-y \\ 1 & \text{otherwise} \end{cases}$$

we have:

(i) $1 - f(1 - \gamma) \geq conf(\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}) \geq f(\gamma),.$

(ii) *for* $\alpha \geq 0$*, if* $elb(\gamma, \delta) \geq \alpha$*, the PD classification rule* $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$ *is* α*-discriminatory,*

(iii) *for* $\alpha \geq 1$*, if* $glb(\gamma, \delta) \geq \alpha$*, the PD classification rule* $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$ *is strongly* α*-discriminatory.* □

It is worth noting that β_1 and β_2 are lower bounds for the confidence values of $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{D}$ and $\mathbf{D}, \mathbf{B} \rightarrow \mathbf{A}$ respectively. This amounts to stating that the correlation between $\mathbf{A}$ and $\mathbf{D}$ in context $\mathbf{B}$ within the dataset must be known only with some approximation as background knowledge. Moreover, as β_1 and β_2 tend to 1, the lower and upper bounds in *(i)* tend to γ. Also, $f(\gamma)$ is monotonic w.r.t both β_1 and β_2, but an increase of β_1 leads to a proportional improvement of the precision of lower and upper bounds, while an increase of β_2 leads to a more than proportional improvement.

Example 6.3. *Reconsider Ex. 6.1. We have* $\gamma = 0.167, \delta = 0.027, \beta_1 = 0.7$*, and* $\beta_2 = 0.95$*. The lower bound for the* $glift()$ *value of rule* `abc` *is computed as follows. Called:*

$$f(x) = \frac{0.7}{0.95}(0.95 + x - 1),$$

we have $f(0.167) = 0.086 > 0.027$*, and then* $glb(0.833, 0.973)$ $= f(0.167)/0.027 = 3.19$*.*

Recalling the redlining example, an application of Thm. 6.2 allows us to conclude that black people (`race=black`) are discriminated in a context (`city=NYC`) because almost all people living in a certain neighborhood (`neighborhood=10451`) are black (this is β_2) and almost all black people live in that neighborhood (this is β_1). In general, this is not the case, since black people live in many different neighborhoods. Moreover, in the redlining example we had to provide, as background knowledge, only the approximation β_2. However, notice that the conclusion of the example is slightly different from the one above, stating that black people who live in a certain neighborhood (`race=black, neighborhood=10451`) are discriminated w.r.t. people in the context (`city=NYC`). Such an inference can be modelled as an instance of Thm. 6.2.

Example 6.4. *Rules* (a) *and* (c) *from Sect. 2*

```
a. city=NYC              c. neighborhood=10451, city=NYC
   ==> class=bad            ==> class=bad
   -- conf:(0.25)           -- conf:(0.95)
```

are instances respectively of $\mathbf{B} \rightarrow \mathbf{C}$ *and* $\mathbf{D}, \mathbf{B} \rightarrow \mathbf{C}$ *in Thm. 6.2, with* $\mathbf{B}$ = `city=NYC`, $\mathbf{D}$ = `neighborhood=10451` *and* $\mathbf{C}$ = `class=bad`. *Hence,* $\gamma = 0.95$ *and* $\delta = 0.25$.

What should be a set of PD itemsets for reasoning about redlining? Certainly, `neighborhood=10451` *alone cannot be considered discriminatory. However, the pair* $\mathbf{A}$ = `race=black, neighborhood=10451` *might denote a possible discrimination against black people in a specific neighborhood. In general, all conjunctions of items of minority races and neighborhoods is a source of potential discrimination. This set of itemsets is downward closed, albeit not in the form of* 2^J *for a set of items* J*. As background knowledge, we can now refer to census data, reporting distribution of population over the territory. So, we can easily gather statistics such as rule* (d) *from Sect. 2, which can be rewritten as:*

```
d. neighborhood=10451, city=NYC
   ==> race=black, neighborhood=10451
   -- conf:(0.8)
```

This is an instance of $\mathbf{D}, \mathbf{B} \rightarrow \mathbf{A}$ *in Thm. 6.2. The other expected background rule is* $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{D}$*, which readily has confidence 100%, i.e.* $\beta_1 = 1$*, since* $\mathbf{A}$ *contains* $\mathbf{D}$*. So, we have not to take it into account in this redlining example, which therefore represents a simpler inference problem than the one considered in Thm. 6.2. By the conclusion of the theorem, we obtain lower bounds for the confidence and the extended lift of* $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$*, i.e., rule* (e) *from Sect. 2:*

```
e. race=black, neighborhood=10451, city=NYC
   ==> class=bad
```

Confidence of (e) *is at least* $1/0.8(0.8 + 0.95 - 1) = 0.9375$*, and then its extended lift (w.r.t. the context* `city=NYC`*) is at least* $0.9375/0.25 = 3.75$*. Summarizing, the classification rule* (e) *is at least 3.75-discriminatory or, in intuitive words,* (c) *is a redlining rule imposing a "disproportionate burden" (of 3.75 times than the average of NYC people) over black-race people living in neighborhood 10451.*

Given a set of PND classification rules $\mathcal{PND}$ and a set of background rules $\mathcal{BR}$, we define the *absolute recall* at α as the number of α-discriminatory PD rules that are inferrable by Thm. 6.2. In order to test the proposed inference model,

```
      CheckAlphaPNDCR($\alpha$)
      ForEach $g$ s.t. $\mathcal{PND}_g \neq \emptyset$
        ForEach $\mathbf{X} \rightarrow \mathbf{C} \in \mathcal{PND}_g$
          $\gamma = conf(\mathbf{X} \rightarrow \mathbf{C})$
          $generateContexts = true$
          ForEach $\mathbf{X} \rightarrow \mathbf{A} \in \mathcal{BR}_g$ order by
                              $conf(\mathbf{X} \rightarrow \mathbf{A})$ descending
            $\beta_2 = conf(\mathbf{X} \rightarrow \mathbf{A})$
            $s = supp(\mathbf{X} \rightarrow \mathbf{A})$
(i)         If $\beta_2 > 1 - \gamma$ or $\beta_2 > \gamma$
              If $generateContexts$
                $generateContexts = false$
                $\mathcal{V} = \emptyset$
                ForEach $\mathbf{B} \subseteq \mathbf{X}$
                  // found in $\mathcal{PND}_{g'}$ with $g' = |\mathbf{B}| \leq g$
                  $\delta = conf(\mathbf{B} \rightarrow \mathbf{C})$
(iii)             If $\beta_2(1 - \alpha\delta) \geq 1 - \gamma$
                        or $\beta_2(1 - \alpha(1 - \delta)) \geq \gamma$
                    $\mathcal{V} = \mathcal{V} \cup \{(\mathbf{B}, \delta)\}$
                  EndIf
                EndForEach
              EndIf
              ForEach $(\mathbf{B}, \delta) \in \mathcal{V}$
(iii)           If $\beta_2(1 - \alpha\delta) \geq 1 - \gamma$
                      or $\beta_2(1 - \alpha(1 - \delta)) \geq \gamma$
                  // found in $\mathcal{BR}_{g'}$ with $g' = |\mathbf{B}| \leq g$
                  $\beta_1 = s/supp(\mathbf{B} \rightarrow \mathbf{A})$
                  If $glb(\gamma, \delta) \geq \alpha$
                    output $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$
                  EndIf
                Else
                  $\mathcal{V} = \mathcal{V} \setminus \{(\mathbf{B}, \delta)\}$
                EndIf
              EndForEach
            EndIf
          EndForEach
        EndForEach
      EndForEach
```

Figure 5: Algorithm for checking indirect strong α-discrimination. Here $\mathcal{BR}_g$ is $\{\mathbf{X} \rightarrow \mathbf{A} \in \mathcal{BR} \mid |\mathbf{X}| = g\}$.

we simulate the availability of a large set of background rules under the hypothesis that the dataset contains the discriminatory items, e.g., as in the German credit case. We define:

$$\mathcal{BR} = \{\mathbf{X} \rightarrow \mathbf{A} \mid \mathbf{X} \text{ PND}, \mathbf{A} \text{ PD}, supp(\mathbf{X} \rightarrow \mathbf{A}) \geq ms\ \}$$

as the set of association rules $\mathbf{X} \rightarrow \mathbf{A}$ with a given minimum support. While rules of the form $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{D}$ seem not to be included in the background rule set, we observe that $conf(\mathbf{A}, \mathbf{B} \rightarrow \mathbf{D})$ can be obtained as $supp(\mathbf{D}, \mathbf{B} \rightarrow \mathbf{A})/supp(\mathbf{B} \rightarrow \mathbf{A})$, where both rules in the ratio are of the required form. Notice that the set $\mathcal{BR}$ contains the most precise background rules that an analyst could use, in the sense that the values for β_1 and β_2 in Thm. 6.2 do coincide with the confidence values they limit. Next, for each candidate rule $\mathbf{X} \rightarrow \mathbf{C}$ in $\mathcal{PND}$, we have to enumerate all sub-itemsets $\mathbf{D}, \mathbf{B} \subseteq \mathbf{X}$ (which are $2^{|\mathbf{X}|}$) such that $\mathbf{X}$ can be written as $\mathbf{D}, \mathbf{B}$. What we will be looking for to speed up the enumeration and checking process is some necessary conditions on the inequalities to be checked that restrict the search space. Let us start considering necessary conditions for $elb(\gamma, \delta) \geq \alpha$. If $\alpha = 0$ the expression is always true, so

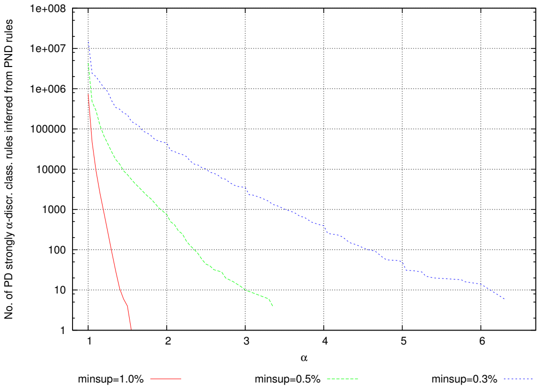

Figure 6: Distribution of absolute recall.

we concentrate on the case $\alpha > 0$. By definition of $elb()$, $elb(\gamma, \delta) \geq \alpha > 0$ happens only if $f(\gamma) > 0$ and $f(\gamma)/\delta \geq \alpha$, which can respectively be rewritten as:

$$\textit{(i)}\ \beta_2 > 1 - \gamma \qquad \textit{(ii)}\ \beta_1(\beta_2 + \gamma - 1) \geq \alpha\delta\beta_2.$$

Therefore, *(i)* is a necessary condition for $elb(\gamma, \delta) \geq \alpha$. From *(ii)* and $\beta_1 \leq 1$, we can conclude $elb(\gamma, \delta) \geq \alpha$ only if $\beta_2 + \gamma - 1 \geq \alpha\delta\beta_2$, i.e.:

$$\textit{(iii)}\ \beta_2(1 - \alpha\delta) \geq 1 - \gamma.$$

Therefore, *(iii)* is a necessary condition for $elb(\gamma, \delta) \geq \alpha$ as well. The selectivity of conditions *(i,iii)* lies in the fact that checking *(iii)* involves no lookup at rules $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{D}$; and checking *(i)* involves no lookup at rules $\mathbf{B} \rightarrow \mathbf{C}$. Moreover, condition *(iii)* is monotonic w.r.t β_2, hence if we scan association rules $\mathbf{X} \rightarrow \mathbf{A}$ ordered by descending confidence, we can stop checking it as soon as it is false. Finally, we observe that similar necessary conditions can be derived for $glb(\gamma, \delta) \geq \alpha$. The generate&test algorithm that incorporates the necessary conditions is shown in Fig. 5.

The German credit case study

With reference to the presented test framework, Fig. 6 plots the distribution of the absolute recall of the proposed inference model by varying α and minimum support. Even for high values of α, the number of indirectly discriminatory rules is considerably high. We report below the execution times of the **CheckAlphaPNDCR()** procedure (on a PC with Xeon 2.4Ghz and 2Gb main memory) for rules in $\mathcal{PND}$ and $\mathcal{BR}$ having minimum support of 1% and without/with the optimizations discussed earlier.

	without checks	**with checks**	**ratio**
$\alpha = 2.0$	10m21s	3m12s	31.0%
$\alpha = 1.8$	10m21s	3m15s	31.4%
$\alpha = 1.6$	10m21s	3m23s	32.7%
$\alpha = 1.4$	10m21s	3m49s	36.9%

The table shows a gain in the execution time up to 69%.

7. RELATED WORK AND CONCLUSIONS

To the best of our knowledge, this paper is the first to address the discrimination problem in data mining models. Nevertheless, discrimination has been recognized as an issue in the tutorial [4, Slide 19] where the danger of building classifiers capable of racial discrimination in home loans has been put forward. Technically, we measured discrimination through a generalization of lift to cope with *contexts*, specified as non-discriminatory itemsets. In this sense, there is a relation with the work of [13], where the notion of *conditional association rules* has been studied. A conditional rule $\mathbf{A} \Leftrightarrow \mathbf{C}/\mathbf{B}$ denotes a context $\mathbf{B}$ in which itemsets $\mathbf{A}$ and $\mathbf{C}$ are equivalent, namely where $conf(\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}) = 1$ and $conf(\neg\mathbf{A}, \mathbf{B} \rightarrow \neg\mathbf{C}) = 1$.

Summarizing, we have investigated how discrimination may be hidden in data mining models. Our study considered classification rules, which occur in a variety of approaches including decision trees, rule-based classifiers, and association rule-based classifiers. As the contributions of the paper, we have modelled both direct and indirect discrimination, introduced (strong) α-protection as a measure of the discriminatory power of a rule, and, as far as indirect discrimination is concerned, devised an inference model as a formal result that is able to infer discriminatory rules from apparently safe ones and some background knowledge.

8. REFERENCES

[1] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In *Proc. of VLDB 1994*, pages 487–499. Morgan Kaufmann, 1994.

[2] R. Agrawal and R. Srikant. Privacy-preserving data mining. In *Proc. of SIGMOD 2000*, pages 439–450. ACM, 2000.

[3] G. S. Becker. *The Economics of Discrimination.* University of Chicago Press, 1957.

[4] C. Clifton. Privacy preserving data mining: How do we mine data when we aren't allowed to see it? Tutorial at *KDD 2003.* `http://www.cs.purdue.edu/homes/clifton`.

[5] B. Goethals. Frequent Itemset Mining Implementations Repository, `http://fimi.cs.helsinki.fi`.

[6] H. Holzer, S. Raphael, and M. Stoll. Black job applicants and the hiring officer's race. *Industrial and Labor Relations Review*, 57(2):267–287, 2004.

[7] D.H. Kaye and M. Aickin, editors. *Statistical Methods in Discrimination Litigation.* Marcel Dekker, Inc., 1992.

[8] P. Kuhn. Sex discrimination in labor markets: The role of statistical evidence. *The American Economic Review*, 77:567–583, 1987.

[9] M. LaCour-Little. Discrimination in mortgage lending: A critical review of the literature. *J. of Real Estate Literature*, 7:15–50, 1999.

[10] B. Liu, W. Hsu, and Y. Ma. Integrating classification and association rule mining. In *Proc. of KDD 1998*, pages 80–86. AAAI Press, 1998.

[11] D.J. Newman, S. Hettich, C.L. Blake, and C.J. Merz. UCI repository of machine learning databases, 1998. `http://archive.ics.uci.edu/ml`.

[12] D. Pedreschi, S. Ruggieri, and F. Turini. Discrimination-aware data mining. Tech. Rep. 07-19, Dip. Inf., Univ. of Pisa, 2007. `http://compass2.di.unipi.it/TR`.

[13] J. Rauch and M. Simunek. Mining for association rules by 4ft-Miner. In *Proc. of INAP 2001*, pages 285–295. Prolog Association of Japan, 2001. `http://lispminer.vse.cz`.

[14] G. D. Squires. Racial profiling, insurance style: Insurance redlining and the uneven development of metropolitan areas. *J. of Urban Affairs*, 25(4):391–410, 2003.

[15] L. Sweeney. *Computational Disclosure Control: A Primer on Data Privacy Protection.* PhD thesis, MIT, 2001.

[16] P.-N. Tan, V. Kumar, and J. Srivastava. Selecting the right objective measure for association analysis. *Inf. Syst.*, 29(4):293–313, 2004.

[17] U.S. Federal Legislation. `http://www.usdoj.gov`.

[18] X. Yin and J. Han. CPAR: Classification based on Predictive Association Rules. In *Proc. of SIAM DM 2003*, SIAM, 2003.