Data Mining for Discrimination Discovery

SALVATORE RUGGIERI, DINO PEDRESCHI, and FRANCO TURINI
Università di Pisa

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. Discrimination in credit, mortgage, insurance, labor market, and education has been investigated by researchers in economics and human sciences. With the advent of automatic decision support systems, such as credit scoring systems, the ease of data collection opens several challenges to data analysts for the fight against discrimination. In this article, we introduce the problem of discovering discrimination through data mining in a dataset of historical decision records, taken by humans or by automatic systems. We formalize the processes of direct and indirect discrimination discovery by modelling protected-by-law groups and contexts where discrimination occurs in a classification rule based syntax. Basically, classification rules extracted from the dataset allow for unveiling contexts of unlawful discrimination, where the degree of burden over protected-by-law groups is formalized by an extension of the lift measure of a classification rule. In direct discrimination, the extracted rules can be directly mined in search of discriminatory contexts. In indirect discrimination, the mining process needs some background knowledge as a further input, for example, census data, that combined with the extracted rules might allow for unveiling contexts of discriminatory decisions. A strategy adopted for combining extracted classification rules with background knowledge is called an inference model. In this article, we propose two inference models and provide automatic procedures for their implementation. An empirical assessment of our results is provided on the German credit dataset and on the PKDD Discovery Challenge 1999 financial dataset.

Categories and Subject Descriptors: H.2.8 [**Database Management**]: Database Applications—*Data mining*

General Terms: Algorithms, Economics, Legal Aspects

Additional Key Words and Phrases: Discrimination, classification rules

ACM Reference Format:

Ruggieri, S., Pedreschi, D., and Turini, F. 2010. Data mining for discrimination discovery. ACM Trans. Knowl. Discov. Data. 4, 2, Article 9 (May 2010), 40 pages.
DOI = 10.1145/1754428.1754432 http://doi.acm.org/10.1145/1754428.1754432

A preliminary version of this article appeared in *Proceedings of the ACM SIGKDD International Conference on Knowldge Discovery and Data Mining* [Pedreschi et al. 2008].
Corresponding author's address: S. Ruggieri, Dipartimento di Informatica, Università di Pisa, L.go B. Pontecorvo 3, 56127, Pisa, Italy; email: ruggieri@di.unipi.it.

DOI 10.1145/1754428.1754432 http://doi.acm.org/10.1145/1754428.1754432

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 [U.S. Federal Legislation 2009] 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 wages (Intentional Employment Discrimination Act, Equal Pay Act, Pregnancy Discrimination Act). Other U.S. federal laws exist on discrimination in public programs or activities, such as public accommodations, education, health care, academic programs, student services, nursing homes, adoptions, senior citizens centers, hospitals, transportation. Several authorities (regulation boards, consumer advisory councils, commissions) are settled to monitor discrimination compliances in U.S., European Union and many other countries.

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 [Becker 1957]. The literature in those research fields has given evidence of unfair treatment in racial profiling and redlining [Squires 2003], mortgage discrimination [LaCour-Little 1999], personnel selection discrimination [Holzer et al. 2004; Kaye and Aickin 1992], and wages discrimination [Kuhn 1987].

The importance of data collection and data analysis for the fight against discrimination is emphasized in legal studies promoted by the European Commission [Makkonen 2007]. The possibility of accessing to historical data concerning decisions made in socially-sensitive tasks is the starting point for discovering discrimination. However, if available decision records accessible for inspection increase, the data available to decision makers for drawing their decisions increase at a much higher pace, together with ever more intelligent decision support systems, capable of assisting the decision process, and sometimes to automate the process entirely. As a result, the actual discovery of discriminatory situations and practices, hidden in the decision records under analysis, may reveal an extremely difficult task. The reason for this difficulty is twofold.

First, personal data in decision records are highly dimensional, that is, characterized by many multivalued variables: as a consequence, a huge number of possible contexts may, or may not, be the theater for discrimination. To see this point, consider the case of gender discrimination in credit approval: although an analyst may observe that no discrimination occurs in general, that is, when considering the whole available decision records, it may turn out that it is extremely difficult for aged women to obtain car loans. Many small or large niches may exist that conceal discrimination, and therefore all possible specific situations should be considered as candidates, consisting of all possible combinations

of variables and variable values: personal data, demographics, social, economic and cultural indicators, etc. Clearly, the anti-discrimination analyst is faced with a huge range of possibilities, which make her work hard: albeit the task of checking some known suspicious situations can be conducted using available statistical methods, the task of discovering niches of discrimination in the data is unsupported.

The second source of complexity is indirect discrimination: often, the feature that may be object of discrimination, for example, the race or ethnicity, is not directly recorded in the data. Nevertheless, racial discrimination may be equally hidden in the data, for instance in the case where a *redlining* practice is adopted: people living in a certain neighborhood get frequently credit denial, and by demographic data we can learn that most of people living in that neighborhood belong to the same ethnic minority. Once again, the anti-discrimination analyst is faced with a large space of possibly discriminatory situations: how can she highlight all interesting discriminatory situations that emerge from the data, both directly and in combination with further background knowledge in her possession (e.g., census data)?

The goal of our research is precisely to address the problem of discovering discrimination in historical decision records by means of data mining techniques. Generally, data mining is perceived as an enemy of fair treatment and as a possible source of discrimination, and certainly this may be the case, as we discuss in the following. Nonetheless, we will show that data mining can also be fruitfully put at work as a powerful aid to the anti-discrimination analyst, capable of automatically discovering the patterns of discrimination that emerge from the available data with stronger evidence.

Traditionally, classification models are constructed on the basis of historical data exactly with the purpose of discrimination in the original Latin sense: that is, 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. For instance, a large body of literature [Baesens et al. 2003; Hand 2001; Hand and Henley 1997; Thomas 2000; Viaene et al. 2001; Vojtek and Kočenda 2006] refers to classification models as the basis of scoring systems to predict the reliability of a mortgage/credit card debtor or the risk of taking up an insurance. Furthermore, data mining screening systems have recently been proposed for personnel selection [Chien and Chen 2008]. 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. For instance, if it is a current malpractice to deny pregnant women access to certain job positions, there is a high chance of finding a strong association in the historical data between pregnancy and access

denial, and therefore we run the risk of learning discriminatory decisions. This use of classification and prediction models may therefore exacerbate the risks of discrimination in socially sensitive decision making. However, as we show in this article, data mining also provides a powerful tool for discovering discrimination, both in the records of decisions taken by human decision makers, and in the recommendations provided by classification models, or any combinations thereof.

In this article, we tackle the problem of discovering discrimination within a rule-based setting, by introducing the notion of *discriminatory classification rules*, as a criterion to identify and analyse the potential risk of discrimination. By mining all discriminatory classification rules from a dataset of historical decision records, we offer a sound and practical method to discover niches of direct and indirect discrimination hidden in the data, as well as a criterion to measure discrimination in any such contexts. This extends our KDD 2008 paper [Pedreschi et al. 2008] in many respects: besides providing a detailed account of the theoretical aspects, under a conservative extension of the syntax of frequent itemsets, it offers a new perspective on the problem of discrimination discovery; an extended framework for anti-discrimination analysis, including a new inference model based on negated items; a more in depth experimental assessment; and a complexity evaluation of the algorithms proposed. A precise account of the differences between this paper and the KDD 2008 paper is provided in the Related Work section.

1.1 Plan of the Article

The article is organized as follows. In Section 2 we present a scenario for the analysis of direct and indirect discrimination. In Section 3 some standard notions on association and classification rules are recalled, and the measure of extended lift is introduced. In Section 4, we formalize the scenario of Section 2 by introducing the notions of α-protective and α-discriminatory classification rules, where α is a user threshold on the acceptable level of discrimination. The two notions are refined for binary classes to strong α-protection and strong α-discrimination. Direct discrimination checking is presented in Section 5, with experimentation on the German credit dataset. Indirect discrimination is considered in Section 6 and Section 7, where background knowledge is adopted in two inference models. Experimentation on the German credit dataset is reported as well. Further experimentation on the Discovery Challenge 1999 financial dataset is presented in Section 8. Related work is reported in Section 9, while Section 10 summarizes the contribution of the paper. All proofs of theorems are reported in Appendix A, where a conservative extension of the standard notions of association and classification rules is introduced. Computational complexity in time and space of the procedures presented in this paper are discussed in Appendix B.

2. DISCRIMINATION ANALYSIS

The basic problem we are addressing can be stated as follows. Given:

—a dataset of historical decision records,

—a set of potentially discriminated groups,
—and a criterion of unlawful discrimination;

find all pairs consisting of a subset of the decision records, called a context, and a potentially discriminated group within the context for which the criterion of unlawful discrimination hold. In this section, we describe those elements and the process of discrimination analysis in a framework based on itemsets, and on classification rules extracted from the dataset. In the next section, the various elements are formalized and the process is automated.

2.1 Potentially Discriminatory Itemsets

The first natural attempt to formally model potentially discriminated groups 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 approach 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-value pairs that are present either in the data, or in her background knowledge, or in both.

2.2 Modeling the Process of Direct Discrimination Analysis

Discrimination has been identified in law and social study literature as either *direct* or *indirect* (sometimes called systematic) [U.K. Legislation 2009; Australian Legislation 2009; Hunter 1992; Knopff 1986]. Direct discrimination consists of rules or procedures that explicitly impose disproportionate burdens on minority or disadvantaged groups.

We unveil direct discrimination through the extraction from the dataset of historical decision records of *potentially discriminatory (PD) rules* defined as classification rules $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$ that contain potentially discriminatory itemsets $\mathbf{A}$ in their premises. 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

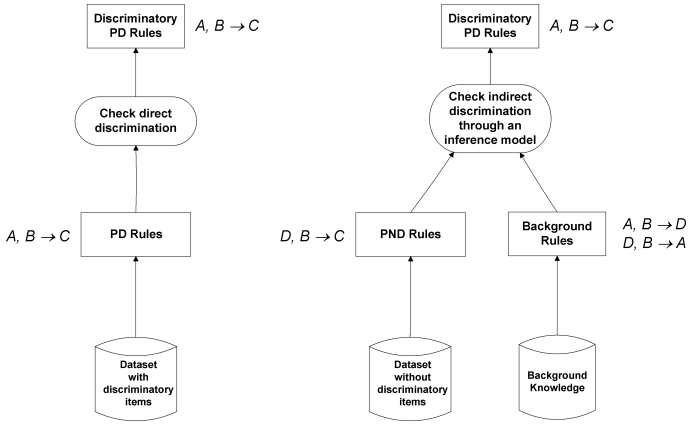

Fig. 1. Modeling the process of direct (left) and indirect (right) discrimination analysis.

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 unveiling the desired level of protection against discrimination or, in other words, for stating the boundary between lawful and unlawful discrimination. PD classification rules are extracted (see Figure 1 left) from a dataset containing potentially discriminatory itemsets. This is the case, for instance, when internal auditors, regulation authorities, or consumer advisor councils want to discover certain information that emerges from the historical decision records such as:

—discrimination malpractices or
—positive policies or affirmative actions [Holzer and Neumark 2006] that tend to favor some disadvantaged categories.

They collect the dataset of past transactions and enrich it, if necessary, with potentially discriminatory itemsets in order to extract discriminatory PD classification rules.

2.3 Modeling the Process of Indirect Discrimination Analysis

Indirect discrimination consists of rules or procedures that, while not explicitly mentioning discriminatory attributes, intentionally or not impose the same disproportionate burdens. For instance, the information on a person's race is typically not available (unless the dataset has been explicitly enriched) or not even collectable. Still, the dataset may unveil discrimination against minority groups.

We unveil indirect discrimination through classification rules $\mathbf{D}, \mathbf{B} \rightarrow \mathbf{C}$ that are potentially nondiscriminatory (PND), that is, that do not contain PD itemsets. They are extracted (see Figure 1 right) from a dataset that may not contain PD itemsets. While apparently unrelated to discriminatory decisions, PND rules may unveil 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 PD rule "rarely give credit to black-race persons from neighborhood 10451 of NYC," which is definitively discriminatory. Summarizing, internal auditors, regulation authorities, and consumer advisory councils can unveil indirect discrimination by identifying discriminatory PD rules through some deduction starting from PND rules and background knowledge. The deduction strategy is called here an *inference model*. In our framework, we assume that background knowledge takes the form of association rules relating a PND itemset **D** to a potentially discriminatory itemset **A** within the context **B**, or, formally, rules of the form $\mathbf{D}, \mathbf{B} \rightarrow \mathbf{A}$ and $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{D}$. 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).

As a final note, this use case resembles the situation described in privacy-preserving data mining [Agrawal and Srikant 2000; Sweeney 2001], where an anonymized dataset coupled with external knowledge might allow for the inference of the identity of individuals.

2.4 An Example of Direct and Indirect Discrimination Analysis

As an example of the processes shown in Figure 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 the time. 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 3 times! 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 directly from the dataset of historical decision records. Given a threshold α of lawful discrimination, all extracted PD rules, including (b), can be checked for α-protection (see Figure 1 left). For instance, if the threshold $\alpha = 3$ is fixed by the analyst, rule (b) would be classified as discriminatory, that is, as unveiling discriminatory decisions.

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 Figure 1 right). Taken in isolation, rule (c) cannot be considered discriminatory or not. Assume now to know from census data that people from neighborhood 10451 are in majority black, that

is, the following association rule holds:

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

Despite rule (c) contains no discriminatory item, it unveils the discriminatory decision of denying credit to a minority subgroup (black people) which has been "redlined" by its 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 an inference model, stated as a formal theorem, that allows us to derive the lower bound $\alpha \geq 3.7$ for α-protection of (e) starting from PND rules (a) and (c) and a lower bound on the confidence of the background rule (d).

3. BASIC DEFINITIONS AND REFERENCE DATASET

3.1 Association and Classification Rules

We recall the notions of itemsets, association rules and classification rules from standard definitions [Agrawal and Srikant 1994; Liu et al. 1998; Yin and Han 2003]. 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 $\mathbf{X}$ with respect to the total number of transactions: $supp_{\mathcal{D}}(\mathbf{X}) = |\{\ T \in \mathcal{D} \mid \mathbf{X} \subseteq T\ \}|/|\mathcal{D}|$, 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 Liu et al. [1998] and Yin and Han [2003] for a discussion of the integration of classification and association rule mining. The support of $\mathbf{X} \rightarrow \mathbf{Y}$ with respect to $\mathcal{D}$ 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 [1994], a number of well-explored algorithms [Goethals 2009] have been designed in order to extract *frequent* itemsets, that is, itemsets with a specified minimum support, and valid association rules, that is, rules with a specified minimum confidence.

The proofs of the formal results presented in the paper suggested a conservative extension of the syntax of rules to boolean expressions over itemsets. The extension, reported in Appendix A.1, allows us to deal uniformly with negation and disjunction of itemsets. As a consequence of the improved expressive power of this language, the formal results of this paper directly extend to association and classification rules over over hierarchies [Srikant and Agrawal 1995] and negated itemsets [Wu et al. 2004].

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* [Tan et al. 2004] (also known as *interest factor*), defined as:

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

The extended lift of $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$ with respect to $\mathbf{B}$ is equivalent to $lift_{\mathcal{B}}(\mathbf{A} \rightarrow \mathbf{C})$ where $\mathcal{B} = \{T \in \mathcal{D} \mid \mathbf{B} \subseteq T\}$ is the set of transactions satisfying the context $\mathbf{B}$. When $\mathbf{B}$ is empty, the extended lift reduces to the standard lift. We refer the reader to Appendix A.2 for proofs of these statements.

3.3 The German Credit Case Study

Throughout the paper, we illustrate the notions introduced by analysing the public domain German credit dataset [Newman et al. 1998], 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.

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 Section 2.1. 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 $2^I \setminus \mathcal{I}_d$ 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, that is 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 nonsensitive 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 itemsets may unveil (indirect) discrimination. The notion of potential (non-)discrimination is now extended to 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 contains PD itemsets or inferred as shown in Figure 1 right. PND rules can be extracted from a dataset which may or may not contain PD itemsets.

[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.

Example 4.4. Consider Example 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.

For a given threshold $\alpha \geq 0$, we say that c is α-protective if its extended lift with respect to $\mathbf{B}$ is lower than α. Otherwise, c is α-discriminatory.

In symbols, given:

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

we write $elift(\gamma, \delta) < \alpha$ as a shorthand for c being α-protective, where:

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

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

Intuitively, the definition assumes that the extended lift of c with respect to $\mathbf{B}$ 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 Example 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: with respect to people asking for credit whose saving status were not known, the bad credit class was assigned in past to nonsingle 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 nonsingle women who want to buy a used car were assigned the bad credit class with a probability more than 6 times the average one for those who asked 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, note that the confidence of the base rule:

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

is $1/6.06 = 16.5\%$.

A general principle in discrimination laws is to consider group representation [Knopff 1986] as a quantitative measure of the qualitative requirement that people in a group are treated "less favorably" [European Union Legislation 2009; U.K. Legislation 2009] than others, or such that "a higher proportion of people without the attribute comply or are able to comply" [Australian Legislation 2009] to a qualifying criteria. We observe that (see Lemma A.9):

$$elift(\gamma, \delta) = \frac{conf(\mathbf{B}, \mathbf{C} \rightarrow \mathbf{A})}{conf(\mathbf{B} \rightarrow \mathbf{A})},$$

namely the extended lift can be defined as the ratio between the proportion of the disadvantaged group **A** in context **B** obtaining the benefit **C** over the overall proportion of **A** in **B**. This makes it clear how extended lift relates to the principle of group over-representation in benefit denying, or, equivalently, of under-representation in benefit granting.

4.3 Strong α-Protection

When the class is a binary attribute, 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%:

```
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 the *good* credit class is assigned to nonsingle 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, as confirmed by 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:*

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

PROOF. See Appendix A.3. □

As an immediate consequence, the (direct) extraction or the (indirect) inference of an α-protective rule $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$ allows for the calculation of the extended lift of the dual rule $\mathbf{A}, \mathbf{B} \rightarrow \neg\mathbf{C}$, and then for unveiling that it is 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 $c' = \mathbf{A}, \mathbf{B} \rightarrow \neg\mathbf{C}$.

For a given threshold $\alpha \geq 1$, we say that c is strongly α-protective if both the extended lifts of c and c' with respect to $\mathbf{B}$ are lower than α. Otherwise, c is strongly α-discriminatory.

In symbols, given

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

we write $glift(\gamma, \delta) < \alpha$ as a shorthand for c being strongly α-protective, where:

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

Analogously, c is strongly α-discriminatory *if* $glift(\gamma, \delta) \geq \alpha$.

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

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

Proofs of these statements are reported in Appendix A.3.

A different way of looking at strong α-discrimination is to consider Lemma 4.8 as the final part of an inference model where (an upper bound on) the confidence of a rule $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$ is inferred first, and then such a value is used to show that the dual rule $\mathbf{A}, \mathbf{B} \rightarrow \neg\mathbf{C}$ is α-discriminatory. Definition 4.9 allows for unveiling that the dual rule is α-discriminatory at the time the rule $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$ is considered, hence checking the final part of the inference model.

Example 4.10. Consider again Example 4.7 and assume the conditions of indirect discrimination as modeled in Figure 1, right. The rule `a-good` cannot be extracted, since the dataset does not include PD itemsets. However, the base rule of `a-good` is PND, and then its confidence 96.3% might be known. Suppose now that by some inference model, such as the ones we will introduce in later sections, an upper bound on the confidence of `a-good` is estimated in 88%. As an immediate consequence of Lemma 4.8, a lower bound on the extended lift of `a-bad` can be calculated as $(1 - 0.88)/(1 - 0.963) = 3.24$. This allows for the conclusion that `a-bad` is 3.24-discriminatory.

5. DIRECT DISCRIMINATION

5.1 Checking α-Protection

Let us consider the case of direct discrimination, as modeled in Figure 1, left. 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 Definition 4.5 (resp., Definition 4.9), provided that the elements of the inequality are available. We define a checking algorithm **CheckAlphaPDCR()** in Figure 3 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 [Goethals 2009]. The procedure **ExtractCR()** in Figure 2 extracts PD and PND classification rules by a single scan over the 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 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 $\mathcal{PND}_{group}$. The **CheckAlphaPDCR()** procedure can then calculate the extended lift of a classification rule $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C} \in \mathcal{PD}_{group}$ from its confidence and the confidence of the base rule $\mathbf{B} \rightarrow \mathbf{C} \in \mathcal{PND}_{group}$.

The computational complexity in both time and space of the procedures presented in this article is discussed in Appendix B.

5.2 The German Credit Case Study

In this section, we analyze the reference dataset in search of direct discrimination. We present the distributions of α-discriminatory PD rules at the variation

ExtractCR()
$\mathcal{C}$ = { class items }
$\mathcal{PD}_{group} = \mathcal{PND}_{group} = \emptyset$ for $group \geq 0$
ForEach k s.t. there exist 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 $|\mathbf{A}| = 0$
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

Fig. 2. Extraction of PD and PND classification rules.

CheckAlphaPDCR(α)
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

Fig. 3. Direct checking of α-discrimination.

of a few parameters that one can use to control the set of extracted rules: minimum support, minimum confidence, class item, and the set I_d of PD itemsets.

Discrimination with respect to support thresholds. The top plot in Figure 4 (resp., Figure 5) shows the distribution of α-discriminatory PD rules (resp., strong α-discriminatory PD rules) for minimum supports of 1%, 0.5% and 0.3%. The figures highlight how lower support values increase the number of PD rules and the maximum α. Notice that, for a same minimum support, α reaches higher values in Figure 5 than in Figure 4, 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 Example 4.7). We report three sample PD rules with decreasing support and increasing extended lift.

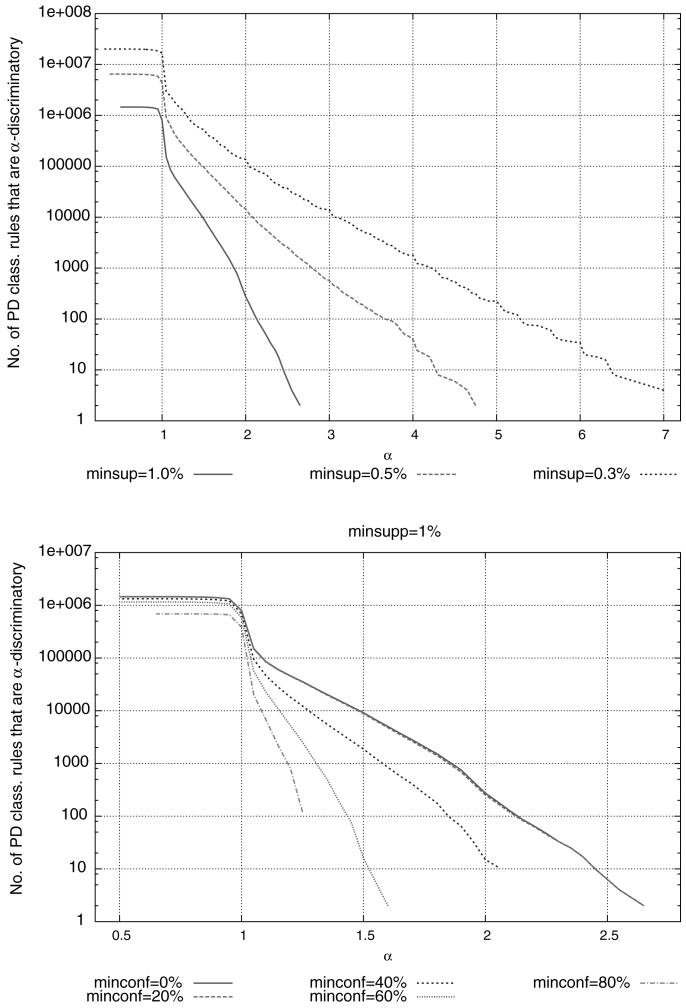

Fig. 4. The German credit dataset. Top: distributions of α-discriminatory PD rules. Bottom: contribution of setting minimum confidence for base rules.

```
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, existing_credits=(1.6-2.2]
    ==> class=bad
    -- supp:(0.005) conf:(1) elift:(3.60)

a3. age=(52.6-inf), employment=1<=X<4, savings_status=>=1000
    ==> class=bad
    -- supp:(0.002) conf:(1) elift:(9)
```

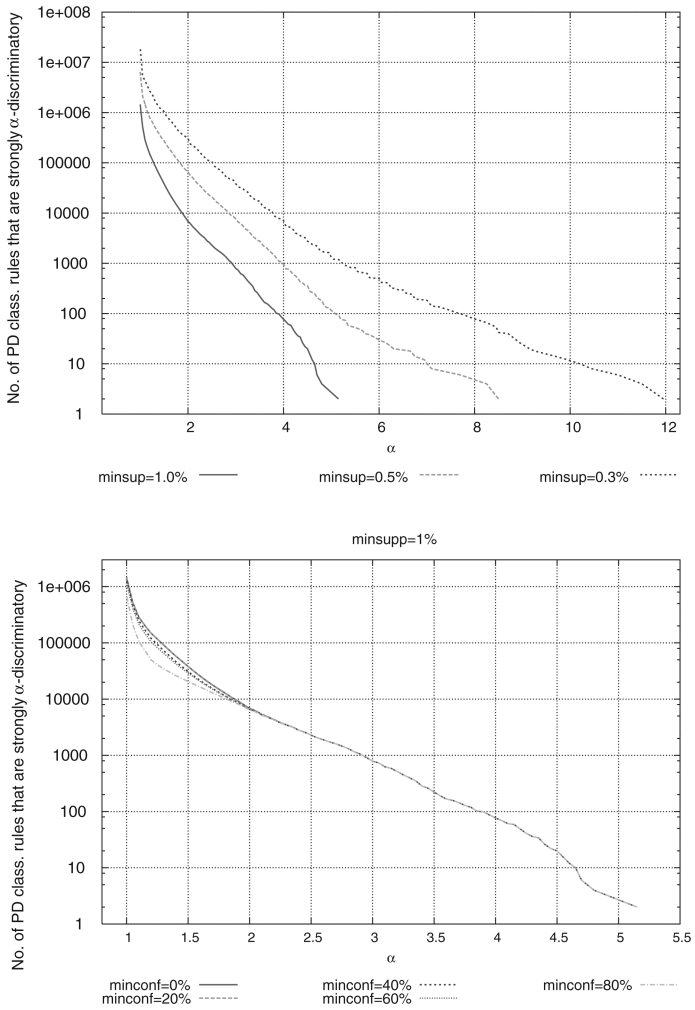

Fig. 5. The German credit dataset. Top: distributions of strongly α-discriminatory PD rules. Bottom: contribution of setting minimum confidence for base rules.

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 states that senior people employed since one to four years, having already two existing credits are assigned the bad credit class 3.6 times more than the average. Finally, rule a3 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)
```

Table I. Execution Times for ExtractCR() and CheckAlphaPDCR()

	No. freq.	ExtractCR()			CheckAlphaPDCR()
minsup	Itemsets	No. PD	No. PND	Time	Time
1%	6.6M	1.45M	1.27M	38s	13.5s
0.5%	26.8M	6.4M	5.3M	163s	55s
0.3%	79.0M	20.0M	15.9M	519s	165s

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.

Table I shows the elapsed times of the procedures **ExtractCR()** and **CheckAlphaPDCR()** on a 32-bit PC with Intel Core 2 Quad 2.4Ghz and 4Gb main memory. We also report the number (M = millions) of frequent itemsets in input to **ExtractCR()** and the number of PD and PND rules yielded in output. For frequent pattern extraction, any system from [Goethals 2009] can be adopted. All procedures reported in this article are implemented in Java 6.

The elapsed times are consistent with the worst-case complexity analysis reported in Appendix B and show good scalability along with the minimum support threshold.

Discrimination with respect to confidence thresholds. Another widely adopted parameter for controlling rule generation is minimum confidence. The bottom plot in Figure 4 shows how the confidence threshold of the base rule affects the distribution of α-discriminatory PD rules. Higher 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.

On the contrary, acting on minimum confidence of the base rule does not result in an effective mechanism for unveiling additional strongly α-discriminatory rules, as shown in Figure 5, bottom plot.

Discrimination with respect to class item. The contribution of the class item to the distribution of discriminatory PD classification rules is shown in Figure 6, where the minimum support is fixed to 1%. The top plot highlights that rules with class item `class=bad` contribute mostly to higher values of extended lift. This confirms that the set of PD itemsets I_d fixed so far (see Example 4.2) characterizes groups of people that are discriminated rather than favored. Also, notice that when $\alpha < 1$, the number of PD rules with class item `class=good` becomes predominant. Since an extended lift lower than 1 means group underrepresentation, this leads to the dual conclusion that people characterized by I_d are underrepresented in benefit granting. Such a dual behavior is explicitly taken into account by strong α-protection, which considers at the same time both under-representation and over-representation—or, in formal terms, extended lifts of both $\mathbf{A}, \mathbf{B} \rightarrow$ `class=good` and $\mathbf{A}, \mathbf{B} \rightarrow$ `class=bad`. As shown at the bottom plot in Figure 6, PD rules $\mathbf{A}, \mathbf{B} \rightarrow$ `class=good` that allow for inferring discrimination of the complementary rule $\mathbf{A}, \mathbf{B} \rightarrow$ `class=bad` are indeed the vast majority.

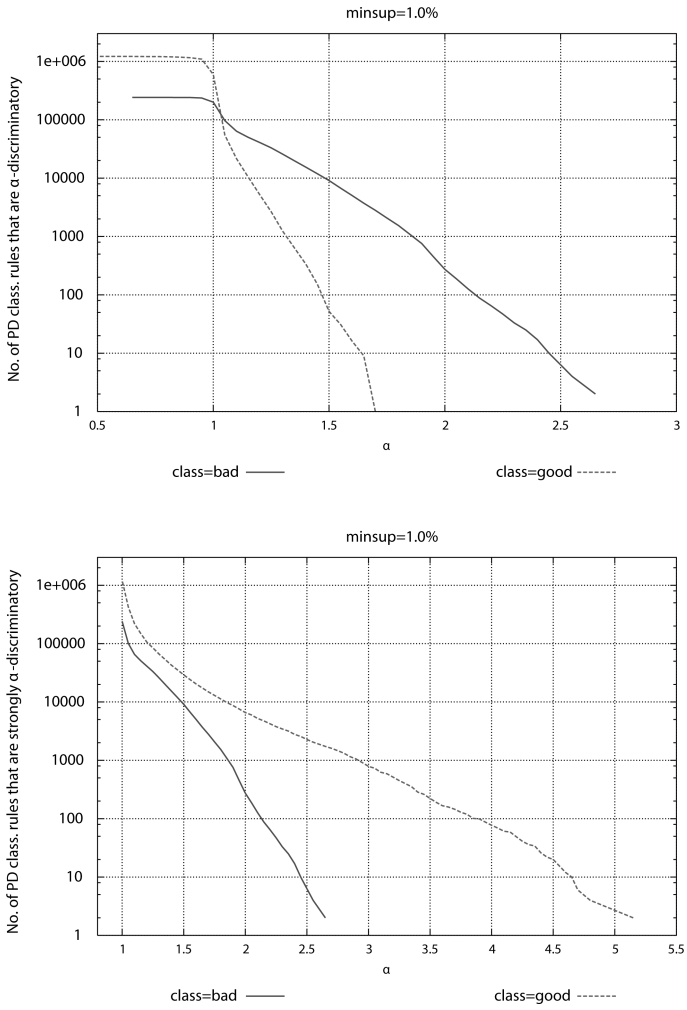

Fig. 6. The German credit dataset. Top: distribution of α-discriminatory PD rules for each class item. Bottom: distribution of strongly α-discriminatory PD rules.

Discrimination with respect to the set of PD discriminatory itemsets. As highlighted by Figure 6, top plot, the set of discriminatory itemsets fixed so far leads mainly to discrimination *against* assigning credit. There are, however, cases where discrimination *in favor* of assigning credit is raised, as in the following:

```
personal_status=female div/sep/mar,
property_magnitude=no known property
employment=<1, other_parties=none
==> class=good
-- supp:(0.005) conf:(1) elift:(2.14)
```

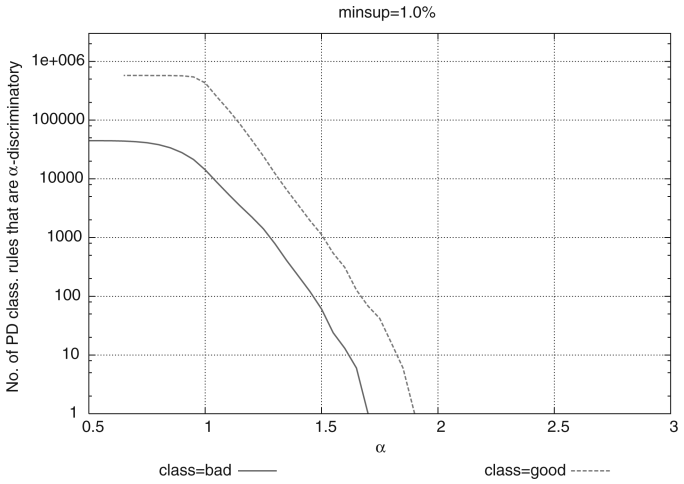

Fig. 7. The German credit dataset. Distributions of α-discriminatory PD classification rules for $I'_d = \{$`personal_status=male single`, `age=(41.4-52.6]`$\}$.

Women who are recently employed, with no known property, and no supporting party are assigned good credit score with a probability of 2.14 times the average one of people in the same conditions. This might reveal a good practice of enforcement of affirmative actions or other policies or laws in support of disadvantaged categories [Holzer and Neumark 2006].

Discrimination *in favor* of assigning credit can also reveal a malpractice of unfair favoritism for certain categories. In order to illustrate this, however, we need to switch to a different set of discriminatory itemsets. Let us fix $I'_d = \{$`personal_status=male single`, `age=(41.4-52.6]`$\}$, namely we are now interested in discrimination in favor of male single and/or people in their 40's. Figure 7 shows the distributions of α-discriminatory PD rules for each class item. Contrasted to Figure 6, top plot, classification rules with class item `class=bad` occur much less frequently and with lower values of extended lift, while rules with class item `class=good` occur slightly more frequently and with slightly higher values of extended lift. This can be interpreted as favoring credit to people which are single man and/or in their 40's.

6. INDIRECT DISCRIMINATION THROUGH BACKGROUND KNOWLEDGE

6.1 Motivating Example

Direct discrimination checking does not take into account PND classification rules, since they do not explicitly contain PD itemsets. Formally, a PND rule is 1-protective. In the case of indirect discrimination, as modelled in Figure 1 right, one assumes the extreme case that PD itemsets are not available at all in the underlying dataset. Hence, only PND classification rules can be extracted. As discussed in Section 2, contexts of discriminatory decisions can still be unveiled in the form of PD classification rules by exploiting some additional background knowledge. Next we highlights an example over the German credit dataset.

Example 6.1. Consider again the German credit dataset, but assume now that PD itemsets have been removed from it. Also, consider the following context:

```
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
     ==> class=bad
     -- conf:(0.167)
```

```
bc. B
    ==> class=bad
    -- conf:(0.027)
```

Rule (dbc) states that young people in the context **B** of people with critical credit history, resident since 2.8 years at least, with savings at most for 100 units, and with no checkings, were 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 Section 2, without any further information, we cannot say whether rule (dbc) unveils any discrimination or not. Assume now to know (by some background knowledge) that in the given context **B**, the set of persons satisfying age=(-inf-30.2] is somewhat related to the set of persons satisfying the PD 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 PD item above. 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 nonsingle women:

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

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

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

By means of the forthcoming Theorem 6.2, we can state that the rule (abc) is at least 3.19-discriminatory. This unveils that nonsingle women in the context were imposed a burden in credit denial of at least 3.19 times the average of people in the context. Since the German credit dataset contains the PD items, we can check how accurate is the lower bound by calculating the actual *glift*() value for (abc): it turns out to be 3.37.

6.2 Inference Model

We formalize the intuitions of the example above in the next result, which derives a lower bound for (strong) α-discrimination of PD classification rules given information available in PND rules (γ, δ) and information available from background rules (β_1, β_2).

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:*

$$conf(\mathbf{A}, \mathbf{B} \rightarrow \mathbf{D}) \geq \beta_1 \qquad conf(\mathbf{D}, \mathbf{B} \rightarrow \mathbf{A}) \geq \beta_2 > 0.$$

Called:

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

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

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

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

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

PROOF. See Appendix A.4. □

Notice that the first two cases of the *glb*() function are mutually exclusive,[2] and that there is no division by zero.[3]

The rule $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{D}$ establishes how much the discriminatory features $\mathbf{A}$ entail $\mathbf{D}$ in the context $\mathbf{B}$, and, on the other side, the rule $\mathbf{D}, \mathbf{B} \rightarrow \mathbf{A}$ says how much the non-discriminatory features $\mathbf{D}$ entail $\mathbf{A}$ in the same context. Together

[2]By conclusion (*i*), $1 - f(1-\gamma) \geq f(\gamma)$. When $f(\gamma) \geq \delta$, this implies $1 - f(1-\gamma) \geq f(\gamma) \geq \delta$ and then $f(1-\gamma) \leq 1 - \delta$. When $f(1-\gamma) > 1 - \delta$, this implies $\delta > 1 - f(1-\gamma) \geq f(\gamma)$.

[3]If $f(\gamma) \geq \delta$ then the divisor is $\delta > 0$. Consider now the case $f(1-\gamma) > 1 - \delta$ and assume, by absurd, that $1 - \delta = 0$. Since $\delta = 1$ implies $\gamma = 1$, we have $f(1-\gamma) = \beta_1/\beta_2(\beta_2 - 1) \neq 0 = 1 - \delta$.

they provide the boundaries within which externally discovered discriminatory features can hide behind the non-discriminatory ones, given a context **B**.

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 **A** and **D** in context **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 with respect to 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 Example 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 $glb(0.833, 0.973) = f(0.167)/0.027 = 3.19$.

Assume that the value of $conf(\mathbf{A}, \mathbf{B} \rightarrow \mathbf{D})$ is known with an approximation of 5%, i.e., $\beta_1 = 0.665$, while β_2 is unchanged. We have $f(x) = 0.665/0.95(0.95 + x - 1)$, and since $f(0.167) = 0.082 > 0.027$, we obtain $glb(0.833, 0.973) = f(0.167)/0.027 = 3.03$, that is, the inferred lower bound is proportionally (5%) lower. Assume now that $conf(\mathbf{D}, \mathbf{B} \rightarrow \mathbf{A})$ is known with an approximation of 5%, that is, $\beta_2 = 0.9$ and β_1 is unchanged. We have $f(x) = 0.7/0.9(0.9 + x - 1)$. Again $f(0.167) = 0.052 > 0.027$ implies $glb(0.833, 0.973) = f(0.167)/0.027 = 1.93$, which is more than proportionally lower than 3.19.

Recalling the redlining example from Section 2, an application of Theorem 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 blacks (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 previous one, stating that black people who live in a certain neighborhood (`race=black, neighborhood=10451`) are discriminated with respect to people in the context (`city=NYC`). Such an inference can be modeled as an instance of Theorem 6.2 that strictly requires a downward closed set of itemsets.

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 Theorem 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 minorities 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 Section 2, which can be rewritten[4] 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 Theorem 6.2. The other expected background rule is $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{D}$, which readily has confidence 100%, that is, $\beta_1 = 1$, 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 Theorem 6.2. By the conclusion of the theorem, we obtain lower bounds for the confidence and the extended lift of $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$, that is, rule `(e)` from Section 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 (with respect to 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 simpler words, `(c)` is a redlining rule unveiling a "disproportionate burden" (of at least 3.75 times than the average of NYC people) over black-race people living in neighborhood 10451.

6.3 Checking the Inference Model

We measure the power of the inference model by defining the *absolute recall* at α as the number of α-discriminatory PD rules that are inferrable by Theorem 6.2 starting from the set of PND classification rules $\mathcal{PND}$ and a set of background rules $\mathcal{BR}$.

In order to test the proposed inference model, we simulate the availability of a large set of background rules under the assumption that the dataset contains the discriminatory items, for example, as in the German credit dataset. We define:

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

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

[4]Notice that, for an association rule $\mathbf{X} \rightarrow \mathbf{Y}$, we admit $\mathbf{X} \cap \mathbf{Y} \neq \emptyset$. The assumption $\mathbf{X} \cap \mathbf{Y} = \emptyset$ is typically required and motivated when considering the issue of rule extraction.

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 Theorem 6.2 do coincide with the confidence values they limit.

A straight implementation of the inference model consists of checking the conditions of Theorem 6.2 for each partition $\mathbf{D}, \mathbf{B}$ of $\mathbf{X}$, where $\mathbf{X} \rightarrow \mathbf{C}$ is a rule in $\mathcal{PND}$. Since there are $2^{|\mathbf{X}|}$ of such partitions, we will be looking for some pruning conditions 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 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 $\beta_2 + \gamma - 1 \geq \alpha\delta\beta_2$; that is,

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

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 *(i)* involves no lookup at rules $\mathbf{B} \rightarrow \mathbf{C}$ to compute $\delta = conf(\mathbf{B} \rightarrow \mathbf{C})$; and checking (iii) involves no lookup at the rule $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{D}$ to compute $\beta_1 = conf(\mathbf{A}, \mathbf{B} \rightarrow \mathbf{D})$. Moreover, conditions *(i,iii)* are monotonic with respect to β_2, hence if we scan the association rules $\mathbf{X} \rightarrow \mathbf{A}$ ordered by descending confidence, we can stop checking for a candidate context as soon as they are 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 Figure 8. As a space optimization, we prevent keeping the whole set of PND rules $\mathcal{PND}$, needed when searching for $\delta = conf(\mathbf{B} \rightarrow \mathbf{C})$, by keeping in $\mathcal{R}_g$ a PND rule $\mathbf{X} \rightarrow \mathbf{C}$ only if there exists some background rule $\mathbf{X} \rightarrow \mathbf{A} \in \mathcal{BR}$. Otherwise, we could not even compute $conf(\mathbf{B} \rightarrow \mathbf{A})$, needed for calculating β_1. Computational complexity in both time and space of the **CheckAlphaPNDCR()** procedure is discussed in Appendix B.

6.4 The German Credit Case Study

With reference to the presented test framework, Figure 9, top plot, shows the distribution of the absolute recall of the proposed inference model, at the variation of α and minimum support. Even for high values of α, the number of indirectly discriminatory rules is considerably high. As an example, for minimum support of 0.3%, 390 PD classification rules that are strongly 4-discriminatory can be inferred from PND rules. As one could expect, the absolute recall heavily depends on the size of the background knowledge. Figure 9, bottom plot, shows the distribution of the absolute recall at the variation of the maximum length of background rules' premise for a fixed minimum support of 0.3%. As an example, if we have background rules $\mathbf{X} \rightarrow \mathbf{A}$ with $|\mathbf{X}| \leq 3$, the best inference

```
CheckAlphaPNDCR(α)
R_g = ∅, for every g ≥ 0
ForEach g s.t. PND_g ≠ ∅
  R_g = {X → C ∈ PND_g | ∃ X → A ∈ BR_g}
  ForEach X → C ∈ R_g
    γ = conf(X → C)
    V = ∅, generate = true    // candidate contexts
    ForEach X → A ∈ BR_g order by conf(X → A) descending
      β2 = conf(X → A)
      s = supp(X → A)
(i)   If β2 > 1 − γ or β2 > γ
        If generate    // lazy generation of candidate contexts
          ForEach B ⊆ X such that B → C ∈ R_g' with g' = |B| ≤ g
            δ = conf(B → C)
(iii)       If β2(1 − αδ) ≥ 1 − γ or β2(1 − α(1 − δ)) ≥ γ
              V = V ∪ {(B, δ)}
            EndIf
          EndForEach
          generate = false
        EndIf
        ForEach (B, δ) ∈ V
(iii)     If β2(1 − αδ) ≥ 1 − γ or β2(1 − α(1 − δ)) ≥ γ
            β1 = s/supp(B → A)   // found in BR_g' with g' = |B| ≤ g
            If glb(γ, δ) ≥ α
              output A, B → C
            EndIf
          Else
            V = V \ {(B, δ)}    // no need to check it anymore
          EndIf
        EndForEach
      EndIf
    EndForEach
  EndForEach
EndForEach
```

Fig. 8. Algorithm for checking indirect strong α-discrimination through background knowledge. Here $\mathcal{BR}_g$ is $\{\mathbf{X} \rightarrow \mathbf{A} \in \mathcal{BR} \mid |\mathbf{X}| = g\}$.

leads to two strongly 2.45-discriminatory rules. Highly discriminatory contexts can then be unveiled only starting from very fine-grained background knowledge.

Table II reports the execution times of the **CheckAlphaPNDCR()** procedure (on a 32-bit PC with Intel Core 2 Quad 2.4Ghz and 4Gb main memory) for rules in $\mathcal{PND}$ and $\mathcal{BR}$ having minimum support of 1% and without/with the optimization checks discussed earlier. The set $\mathcal{PND}$ consists of 1.27 millions of classification rules, and the set $\mathcal{BR}$ consists of 2.1 millions of association rules. Notice that the size of $\mathcal{BR}$ is exceptionally large, since it is obtained starting from a dataset which already contains the PD itemsets. In real cases, only a limited number (in the order of thousands) of background rules are available from statistical sources, surveys, or experts.

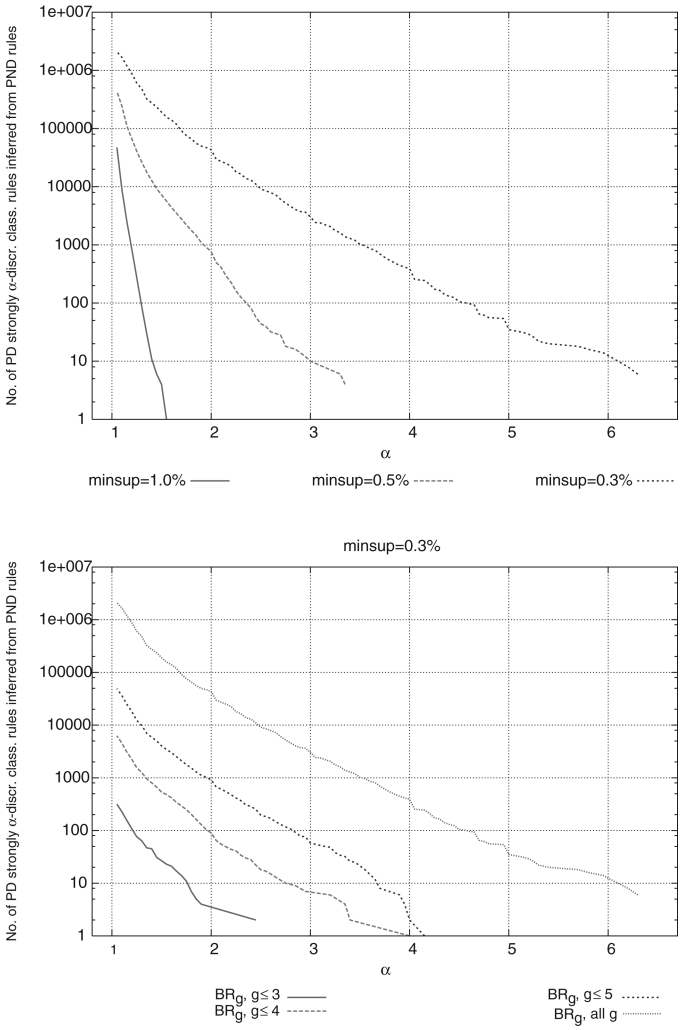

Fig. 9. The German credit dataset. Top: absolute recall of the inference model through background knowledge at the variation of minimum support. Bottom: absolute recall at the variation of maximum length of background rules' premise. Here $\mathcal{BR}_g$ is $\{\mathbf{X} \rightarrow \mathbf{A} \in \mathcal{BR} \mid |\mathbf{X}| = g\}$.

While there is a gain in the execution time in using the optimizations, up to 68.7%, the order of magnitude is the same. This can be explained by observing that condition (i) allows for cutting generation&testing of candidates, but condition (iii) allows for cutting only testing of candidates.

7. INDIRECT DISCRIMINATION THROUGH NEGATED ITEMS

7.1 Motivating Example

A limitation of the inference model based on background knowledge occurs when dealing with a binary attribute a such that $a = v$ is PD and its

Table II. Execution Times for CheckAlphaPNDCR()

	Without Checks	With Checks	Ratio
$\alpha = 2.0$	434s	136s	31.3%
$\alpha = 1.8$	434s	139s	32.0%
$\alpha = 1.6$	434s	144s	33.2%
$\alpha = 1.4$	434s	158s	36.4%

negated item $\neg(a = v)$ is PND. The most common case consists of the PD item `sex=female`, and the PND item `sex=male`. For $\mathbf{D} =$ `sex=male`, and $\mathbf{A} =$ `sex=female`, we have that the assumption $conf(\mathbf{D}, \mathbf{B} \rightarrow \mathbf{A}) = 0 \geq \beta_2 > 0$ of Theorem 6.2 does not hold. As a conclusion, the inference model based on background knowledge cannot derive lower bounds for PND rules involving women starting from PD rules involving men. Such an inference is instead quite natural in practice. Notice that in the German credit dataset this case does not occur, since the attribute `sex` is not binary.

Let us show next an example involving the attribute `foreign_worker`, for which `foreign_worker=no` is PND whilst `foreign_worker=yes` is PD. A rule including `foreign_worker=no` in its premise is considered PND. However, by reasoning as done for binary classes in Section 4.3, such a rule can unveil α-discrimination of the PD rule obtained by replacing `foreign_worker=no` with `foreign_worker=yes`.

Example 7.1. Consider again the German credit dataset, and assume that PD itemsets have been removed from it. Also, consider the following itemset:

```
B = personal_status=male single
    employment=1<=X<4
    purpose=new car
    housing=rent
```

The following PND classification rules can be extracted:

```
nbc. foreign_worker=no, B
     ==> class=good
     -- conf:(1)
```

```
bc. B
    ==> class=good
    -- conf:(0.9)
```

Rule (`nbc`) states that national workers in the context **B** of people that are single male, employed since one to four years, which intend to buy a new car, and have their house for rent, are assigned a good credit scoring with confidence 100%. Rule (`bc`) states that the average confidence of people in context **B** is slightly less, namely 90%. It is quite intuitive that the increasing of confidence from 90% to 100%, yet being a small one, has to be attributed to the omission of foreign workers. Therefore, for the rule:

```
abc. foreign_worker=yes, B
     ==> class=good
```

we expect a decrease in confidence in comparison to (`bc`), or, by reverting to the complementary class `class=bad`, an increase in confidence. In order to estimate

the level of strong α-discrimination of (abc), however, we need to know some further information on the proportion of foreign workers in the context **B**. Assume to know (by some background knowledge) that foreign workers are $50 \pm 0.5\%$ of the people in the context **B** above, that is, the rule:

ba. **B** ==> foreign_worker=yes

has a confidence between 49.5% and 50.5%. By means of the forthcoming Theorem 7.2, we can state that a lower bound for the *glift*() value of (abc) is 1.62. As a consequence, the rule (abc) highlights a burden of 62% more for foreign workers over the average of people in the context above. Since the German credit dataset contains the PD itemsets, we can calculate the actual *glift*() value for (abc), which turns out to be 2.0.

7.2 Inference Model

We formalize the intuitions of the example above 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 rules about the distribution of binary attributes (β_1, β_2).

THEOREM 7.2. *Assume that the attribute of a PD item* **A** *is binary. Let* $\neg\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$ *be a PND classification rule, and let:*

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

and β_1, β_2 *such that:*

$$\beta_2 \geq conf(\mathbf{B} \rightarrow \mathbf{A}) \geq \beta_1 > 0.$$

Called:

$$n_1(x) = \frac{\delta}{\beta_2} + \left(1 - \frac{1}{\beta_1}\right)x \qquad n_2(x) = \frac{\delta}{\beta_1} + \left(1 - \frac{1}{\beta_2}\right)x$$

$$elb(x, y) = \begin{cases} n_1(x)/y & \text{if } n_1(x) \geq y \\ 0 & \text{otherwise} \end{cases}$$

$$glb(x, y) = \begin{cases} n_1(x)/y & \text{if } n_1(x) \geq y \\ (1 - n_2(x))/(1 - y) & \text{if } n_2(x) < y \\ 1 & \text{otherwise} \end{cases}$$

we have:

(*i*) $n_2(\gamma) \geq conf(\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}) \geq n_1(\gamma)$,

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

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

PROOF. See Appendix A.5. □

Notice that, since by (i) $n_2(\gamma) \geq n_1(\gamma)$, the two test conditions in the definition of *glb*() do not overlap. Also, when $\beta_1 = \beta_2$, that is, confidence of $conf(\mathbf{B} \rightarrow \mathbf{A})$

CheckAlphaPNDNegated(α)
$\mathcal{N} = \{a = v_1 \mid a = v_0 \in I_d \text{ and } dom(a) = \{v_0, v_1\}\ \}$
$\mathcal{R}_{group} = \emptyset$, for every $group \geq 0$
ForEach $group$ s.t. $\mathcal{PND}_{group} \neq \emptyset$
  ForEach $\mathbf{X} \rightarrow \mathbf{C} \in \mathcal{PND}_{group}$ s.t. $\mathbf{X} \cap \mathcal{N} \neq \emptyset$
    $\mathbf{N} = \mathbf{X} \cap \mathcal{N}$
    $\gamma = conf(\mathbf{X} \rightarrow \mathbf{C})$
    ForEach $\neg\mathbf{A} \in \mathbf{N}$
      $\mathbf{B} = \mathbf{X} \setminus \neg\mathbf{A}$
      $\delta = conf(\mathbf{B} \rightarrow \mathbf{C})$ // found in $\mathcal{R}_{group-1}$
      $\beta = conf(\mathbf{B} \rightarrow \mathbf{A})$ // found in $\mathcal{BR}_{group-1}$
      $n = \delta/\beta + (1 - 1/\beta)\gamma$
      If $glift(n, \delta) \geq \alpha$
        output $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$
      EndIf
    EndForEach
  EndForEach
  $\mathcal{R}_{group} = \{\mathbf{X} \rightarrow \mathbf{C} \in \mathcal{PND}_{group} \mid \exists\ \mathbf{X} \rightarrow \mathbf{A} \in \mathcal{BR}_{group}\}$
EndForEach

Fig. 10. Algorithm for checking indirect strong α-discrimination through negated items. Here $\mathcal{BR}_g$ is $\{\mathbf{X} \rightarrow \mathbf{A} \in \mathcal{BR} \mid |\mathbf{X}| = g\}$.

is known exactly, then $n_1(\gamma) = n_2(\gamma)$, $elb(\gamma, \delta) = elift(n_1(\gamma), \delta)$ and $glb(\gamma, \delta) = glift\,(n_1(\gamma), \delta)$. This leads to the conclusion that (*ii,iii*) are both necessary and sufficient conditions for (strong) α-discrimination.

Example 7.3. Reconsider Example 7.1. We have: $\gamma = 1, \delta = 0.9, \beta_1 = 0.495$, $\beta_2 = 0.505$. Since $n_2(\gamma) = 0.9/0.495 + (1 - 1/0.505)1 = 0.838 < 0.9 = \delta$, we have $glb(\gamma, \delta) = (1 - 0.838)/(1 - 0.9) = 1.62$. Therefore, the classification rule `(abc)` is at least strongly 1.62-discriminatory. Since the German credit dataset contains the PD items, we can check how accurate is the lower bound by calculating the actual *glift*() value for `(abc)`: it turns out to be 2.

In the case that the confidence of `(ba)` were known exactly, that is, $\beta_1 = \beta_2 = 0.5$, we have $n_2(\gamma) = n_1(\gamma) = 0.9/0.5 + (1 - 1/0.5) = 0.8$ and $glb(\gamma, \delta) = (1 - 0.8)/(1 - 0.9) = 2$, which is the actual *glift*() value of `(abc)`.

7.3 Checking the Inference Model

In order to test the proposed inference model, we simulate the availability of a large and accurate set of background rules for datasets as done in Section 6.3. We recall the definition of the background rule set:

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

The algorithm **CheckAlphaPNDNegated()** in Figure 10 makes a single scan of PND classification rules $\mathbf{X} \rightarrow \mathbf{C}$ ordered by the length of $\mathbf{X}$. For each $\neg\mathbf{A}$ in $\mathbf{X}$ that is the negation of a PD item, the conditions of Theorem 7.2 are checked for $\mathbf{B} = \mathbf{X} \setminus \mathbf{A}$, by looking up the association rule $\mathbf{B} \rightarrow \mathbf{A}$ from the $\mathcal{BR}$ set of background rules. As a space optimization, we prevent keeping the whole set of PND rules $\mathcal{PND}$, needed when searching for $\delta = conf(\mathbf{B} \rightarrow \mathbf{C})$,

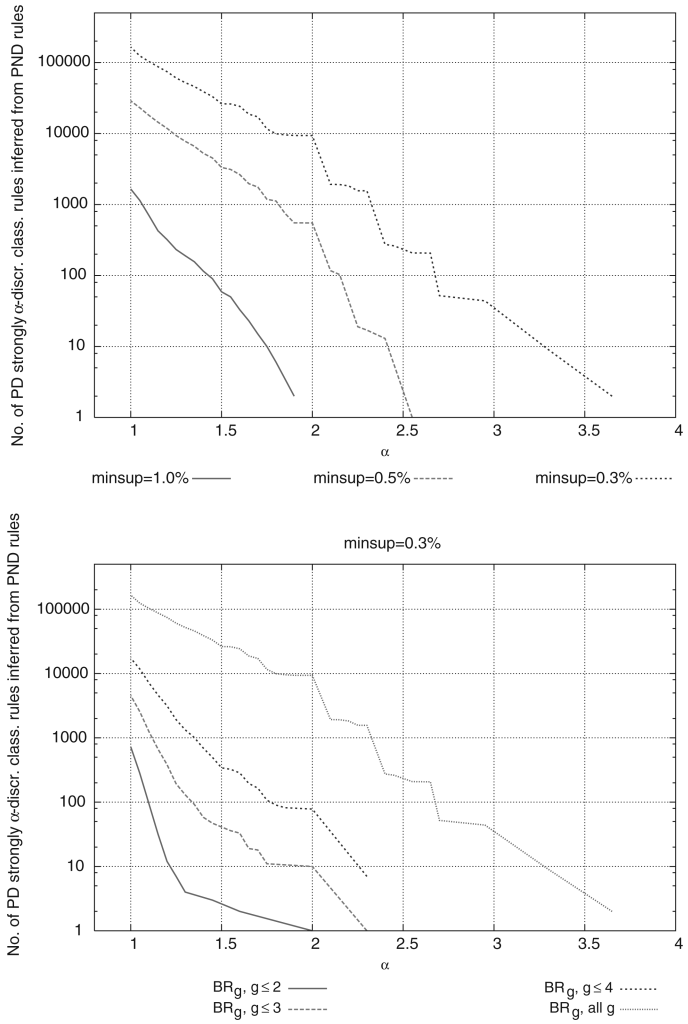

Fig. 11. The German credit dataset. Top: absolute recall of the inference model through negated items at the variation of minimum support. Bottom: absolute recall at the variation of maximum length of background rules' premise. Here $\mathcal{BR}_g$ is $\{\mathbf{X} \rightarrow \mathbf{A} \in \mathcal{BR} \mid |\mathbf{X}| = g\}$.

by keeping in $\mathcal{R}_{group}$ a PND rule $\mathbf{X} \rightarrow \mathbf{C}$ only if there exists some background rule $\mathbf{X} \rightarrow \mathbf{A} \in \mathcal{BR}$. Otherwise, we could not even compute $\beta = conf(\mathbf{B} \rightarrow \mathbf{A})$. Computational complexity in both time and space of the **CheckAlphaPNDNegated()** procedure is discussed in Appendix B.

7.4 The German Credit Dataset

With reference to the presented test framework, Figure 11 shows the distribution of the absolute recall of the inference model of Theorem 7.2, at the variation of minimum support (top plot) and at the variation of the maximum length of background rules' premise (bottom plot). We recall that

Table III. Execution Times for CheckAlphaPNDNegated()

minsup	No. PND Rules	No. Back. Rules	CheckAlphaPNDNegated() Time
1%	1.27M	2.1M	9.6s
0.5%	5.3M	8.4M	47.8s
0.3%	15.9M	24.2M	143s

the only item in I_d satisfying the hypothesis of the inference model is `foreign_worker=yes`. The figure reports then the absolute count of PD rules of the form `foreign_worker` $=$ `yes`, $\mathbf{B} \rightarrow \mathbf{C}$ that can be shown to be strongly α-discriminatory starting from PND rules of the form `foreign_worker=no`, $\mathbf{B} \rightarrow \mathbf{C}$ and background rules of the form $\mathbf{B} \rightarrow$ `foreign_worker` $=$`yes`.

Contrasting the two plots to Figure 9, we observe that for the minimum support of 1% the inference model based on negated items unveils strongly α-discriminatory rules with higher values of α, while for lower minimum support thresholds the inference model based on background knowledge yields stronger conclusions. Moreover, as in Figure 9, the absolute recall of the inference model heavily depends on the size of the background knowledge.

Sample execution times of the **CheckAlphaPNDNegated()** procedure are reported in Table III. They are consistent with the worst-case complexity analysis from Appendix B and show good scalability along with the minimum support threshold.

As for the inference model based on background knowledge, notice that the number of background rules is exceptionally high due to the fact that they are obtained from a dataset which already contains the PD itemsets.

8. CASE STUDY: ANALYSIS OF BAD LOANS

In the German credit case study, the underlying context of analysis is a dataset of historical decisions on granting/denying credit to applicants. The framework proposed in this paper warns us that discriminatory decisions are hidden in such a dataset either directly or indirectly. Concerning the reasons behind those decisions, economists distinguish between "taste-based" discrimination, tracing back to early studies [Becker 1957], and "statistical" discrimination. The former is concerned with dislike against protected-by-law groups. Becker's studies lead to the conclusion that, in a sufficiently competitive market, taste-based discrimination in not employing good black workers is not profitable. Statistical discrimination, also called rational racism in [Harford 2008, Chapter 6], occurs when employers refer directly or indirectly to the average performance of the applicant's racial group as a decision element. Field experiments show [Riach and Rich 2002] that this approach can be profitable, yet illegal.

In this section, we consider a dataset of behaviors, not decisions, namely of persons who, once received a loan, were able or not to repay the loan without problems, such as delays or default. The discrimination analysis still applies in such a different context. Although there is no discriminatory decision to discover here, the extraction of contexts where protected-by-law groups suffered from repaying the loan can help isolating possible sources of statistical discrimination. Business rules built on top of such contexts should be prevented.

The financial dataset from the PKDD 1999 Discovery Challenge [Berka 1999] contains data about the clients of a bank, including personal data (sex and age), demographic data on the district where the client lives (region name, number of cities, rate of urban inhabitants, average salary, unemployment rate, entrepreneur ratio, number of crimes), data on the client bank account (account age, account type, credit card type) and data on the assigned loan (loan amount, loan duration). The class attribute, `status`, assumes the values `ok` or `ko` on the basis whether the loan was fully repaid or not, that is, the bank was in credit at the end of the contract or at the time of the analysis. In summary, the dataset consists of 827 cases and 15 attributes. Continuous attributes have been discretized by equal frequency binning. As done for the German credit dataset in Example 4.2, we fix for the PKDD 1999 dataset $\mathcal{I}_d = 2^{I_d}$, where I_d is now the set of the (discriminatory) items `sex=FEMALE` and `age=(56.5-inf)`.

Figure 12 shows the distributions of α-discriminatory and strongly α-discriminatory PD classification rules. While the absolute counts differ, the distribution shapes look very similar to the ones of Figure 4 and Figure 5. Minimum support threshold turns out to be a mechanism for unveiling the α-level of (strongly) α-discriminatory PD rules. As an example, the following two rules where extracted for minimum support of 0.3%.

```
p1. age=(56.5-inf), crime_n=(5159.25-inf)
    card_type=No, loan_duration=(30-42]
    ==> status=ko
    -- supp:(0.003) conf:(0.5) elift:(6.17)

p2. age=(56.5-inf), sex=FEMALE
    avg_salary=(8764.5-9235], entrepeneurs_ratio=(108.5-125.5]
    ==> status=ko
    -- supp:(0.003) conf:(0.43) elift:(3.47)
```

Rule `(p1)` states that among people living in districts with high crime index, having no credit card, and with a loan of 30 to 42 months, older people had problems with returning the loan 6.17 times the average. Rule `(p2)` states that among people with average salary in the range $8764.5 - 9235$ units and living in a region with entrepreneurs index of $108.5 - 125.5$, older female had problem with their loan 3.47 times the average. New business rules built on top of `(p1)` and `(p2)` could deny loans to prospect applicants satisfying the premises of `(p1)` and `(p2)`. Such rules would be discriminatory for older people and for women.

Of course, even if the PD itemsets were removed from the dataset, indirectly discriminatory contexts can be extracted and, unconsciously, applied in new business rules. Figure 13 shows the distribution of strongly α-discriminatory PD rules obtained by the inference model of Section 6 under the same test condition of Section 6.3. Let us show one of such rules. The following itemset **B**:

```
crime_n=(5159.25-inf), unempl_rate=(3.25-inf), loan_duration=(30-42]
```

describes people living in regions with the highest crime rate and unemployment rate that were granted a loan whose duration is from 31 to 42 months.

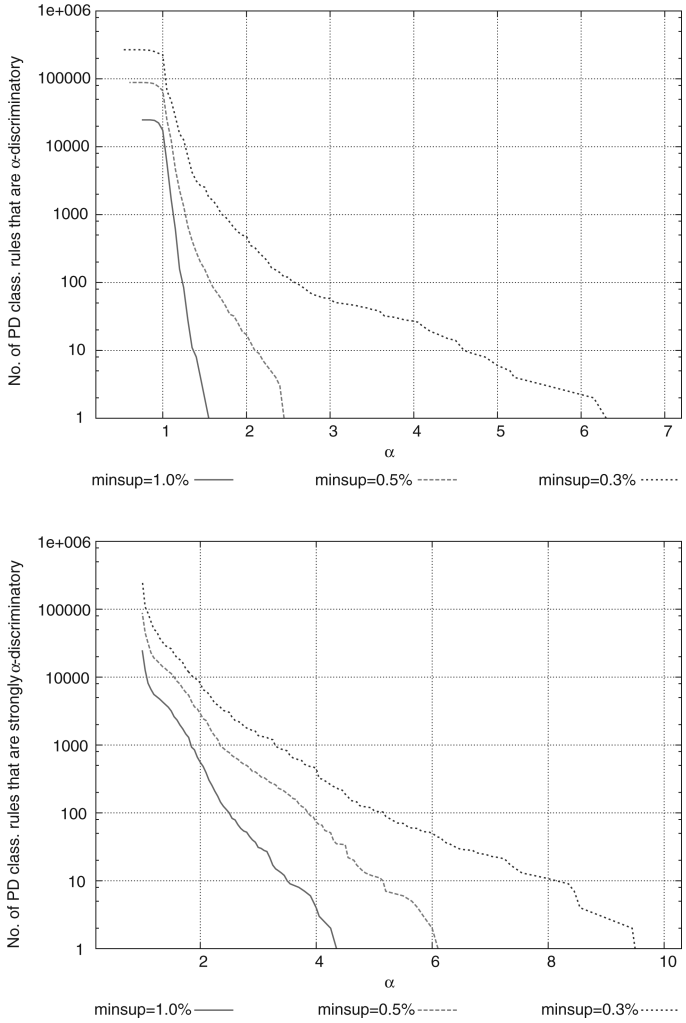

Fig. 12. The PKDD 1999 dataset. Top: distribution of α-discriminatory PD rules. Bottom: distribution of strongly α-discriminatory PD rules.

The following rules can be extracted from the dataset without discriminatory items:

```
dbc. cities_n=(-inf-2.5], B          bc. B
     ==> status=ok                       ==> status=ok
     -- conf:(0.75)                      -- conf:(0.94)
```

Rule (dbc) states that people in the context **B** who live in a region with one or two cities had no problem with returning the loan, with a confidence of 75%. Rule (bc) states that people in the context had no problem in 94% of cases. As a consequence, clients in the context who additionally satisfy cities_n=(-inf-2.5] show problems with returning the loan 25% of times

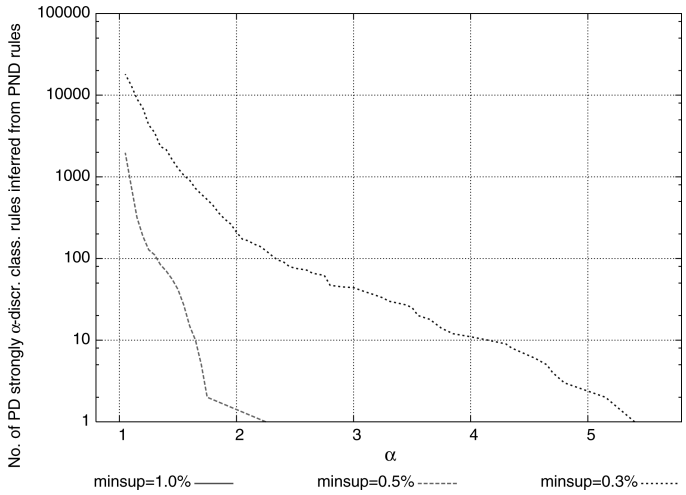

Fig. 13. The PKDD 1999 dataset. Absolute recall of the inference model through background knowledge. Notice that for minimum support of 1% the absolute recall is 0 for $\alpha \geq 1.05$.

$(1 - 0.75)$, which is 4.2 times more than the average problem rate of clients in the context, namely 6% $(1 - 0.94)$. So the bank could be tempted to deny future loans to applicants satisfying **B** and `cities_n=(-inf-2.5]`. By looking at its records, however, the bank could discover that, among all clients in context **B**, people satisfying `cities_n=(-inf-2.5]` approximatively coincide with older people, or, more precisely, that:

```
dba. cities_n=(-inf-2.5], B
     ==> age=(56.5-inf)
     -- conf:(1)
```

```
abd. age=(56.5-inf), B
     ==> cities_n=(-inf-2.5]
     -- conf:(0.8)
```

By Theorem 6.2, the formal conclusion is that the rule:

```
abc. age=(56.5-inf), B
     ==> status=ok
```

is strongly 4.2-discriminatory.

The socially relevant conclusion of this example is that rule `(dbc)` unveils a possible source of statistical indirect discrimination. If a decision system (either automatic, human or mixed) is made aware of such a rule, it could run the risk to impose additional restrictions on prospect loan applicants that would result in a discriminatory action against older people in the context **B**.

9. RELATED WORK

9.1 Discrimination Discovery and Discrimination Prevention

We highlight here the differences between the issues of *discrimination discovery* and *discrimination prevention*. Discrimination discovery, which has been the subject of this paper, consists of supporting the discovery of discriminatory

decisions hidden, either directly or indirectly, in a dataset of historical decision records, possibly built as the result of applying a data mining classifier. Discrimination prevention consists of inducing a classifier that does not lead to discriminatory decisions even if trained from a dataset containing them. Whether or not a set of decisions taken by a classifier is discriminatory can be checked by discrimination discovery methods.

This article is the first to address the discrimination discovery problem by resorting to data mining methods. As a subsequent work, we have studied in Pedreschi et al. [2009] the issue of assessing the statistical significance of the discovered rules. In fact, statistical validation is customary in legal cases before courts [Gastwirth 1992], especially when cases covered by the discovered rules are very few [Piette and White 1999].

Discrimination prevention has been recognized as an issue in the tutorial [Clifton 2003, Slide 19] where the danger of building classifiers capable of racial discrimination in home loans has been put forward, as a common discriminatory behavior of many banks consists of mortgage redlining. The naïve approach of deleting potentially discriminatory itemsets or even whole attributes from the original dataset does not prevent a classifier to learn discriminatory actions, such as the classification rule (c) in Section 2.4, in that it only shields against direct discrimination, not against the indirect one. We foresee three non–mutually exclusive strategies towards discrimination prevention. The first one is to adapt the preprocessing approaches of data sanitization [Hintoglu et al. 2005; Verykios et al. 2004] and hierarchy-based generalization [Sweeney 2002; Wang et al. 2005] from the privacy-preserving literature. Along this line, Kamiran and Calders [2009] adopt a controlled distortion of the training set. The second one is to modify the classification learning algorithm (an in-processing approach), by integrating discrimination measures calculations within it. The third one is to post-process the produced classification model. Along this line, in Pedreschi et al. [2009] a confidence-altering approach is proposed for classification rules inferred by the CPAR algorithm [Yin and Han 2003].

9.2 Comparison with Previous Work

In this article, we extended the preliminary results appeared in Pedreschi et al. [2008] in several directions.

On the methodological side, we clarified the problem of discrimination discovery and the approach of extracting classification rules as a means to discover contexts of discrimination against protected-by-law groups of people.

On the theoretical side, we introduced an inference model based on negated items (see Section 7) which complements the one based on background knowledge (see Section 6) by covering one of the most common cases occurring in practice, namely discrimination against women (and not against men). In Appendix A we introduce a conservative extension of the standard definitions of association and classification rules. The extension allows us to deal uniformly with negation and disjunction of itemsets. As a consequence, the formal results of this article directly extend to the case where PD and PND classification rules

over hierarchies [Srikant and Agrawal 1995] (see Example A.1) and negated itemsets [Wu et al. 2004] (see Example A.2) are extracted.

On the analytical side, we reported the analysis of the PKDD 1999 Discovery Challenge financial dataset (see Section 8) and discussed its application context, which differs from the one of the German credit case study. Moreover, a deeper analysis of the German credit dataset is reported throughout the paper, For example, by studing the distributions of (strong) α-discriminatory PD rules with respect to the class item, with respect to the set I_d of PD itemsets, and with respect to the size of the background knowledge.

On the computational complexity side, in Appendix B we study the space and time worst-case complexity of the procedures proposed in this article. The original procedure **CheckAlphaPNDCR()** from the KDD 2008 paper has been improved for what concerns space requirements. The procedure in this article is linear in the size of background knowledge multiplied by the number of class items. This is typically negligible when compared to the size of PND rules, the space-complexity of the original procedure. Finally, execution times over the German credit dataset are also added throughout the article.

9.3 Extended Lift

Technically, we measured discrimination through generalizations and variants of *lift*, a measure of the significance of an association rule [Tan et al. 2004]. We refined the lift to cope with *contexts*, specified as PND itemsets: how much does a potentially discriminatory condition $\mathbf{A}$ increase/decrease the precision when added to the nondiscriminatory antecedent of a classification rule $\mathbf{B} \rightarrow \mathbf{C}$?

In this sense, there is a relation with the work of Rauch and Simunek [2001], where the notion of *conditional association rules* has been used to analyze a dataset of loans. 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 $conf(\neg\mathbf{A}, \mathbf{B} \rightarrow \neg\mathbf{C}) = 1$. However, we can say nothing about $conf(\mathbf{B} \rightarrow \mathbf{C})$, and, consequently, about the relative strength of the rule with respect to the base classification rule. In addition to $\Leftrightarrow$, the 4ft-Miner system [Rauch and Simunek 2001, 2009] allows for the extraction of conditional rules with other operators. The "above average dependence" operator defines rules $\mathbf{A} \sim^{+} \mathbf{C}/\mathbf{B}$ such that $supp(\mathbf{A}, \mathbf{B}, \mathbf{C}) \geq ms$, where ms is the minimum support threshold, and $lift_{\mathcal{B}}(\mathbf{A} \rightarrow \mathbf{C}) \geq 1 + p$, where $\mathcal{B} = \{T \in \mathcal{D} \mid \mathbf{B} \subseteq T\}$ is the set of transactions verifying $\mathbf{B}$. This is equivalent to check whether the extended lift of $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$ is greater or equal than $1 + p$, that is, whether the rule is $1 + p$-discriminatory. However, the 4ft-Miner system assumes that itemsets $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ are defined starting from specified sets of attributes, not from sets of itemsets. Also, the system adopts a rule extraction algorithm that is general enough to cope with operators defined on the 4-fold contingency table of transactions satisfying or not satisfying $\mathbf{A}$ and/or $\mathbf{C}$. On the one hand, that allows for a general system of rule extraction with respect to several operators. On the other hand, the procedure in Figure 3 exploits the efficiency of the state-of-the art algorithms for the extraction of frequent itemsets.

9.4 Relationship with Privacy-Preserving Data Mining

Finally, some methodological relationships with privacy-preserving data mining [Liu 2009; Vaidya et al. 2006] should be highlighted. A commonality is the attention, in both cases, to key social impacts of data mining which, if not properly tackled, hamper the dissemination and acceptance of the technology: both privacy intrusion and unfair treatment are crucial factors, which call for trustable technologies. On the other hand, the two techniques aim at shielding against two different kinds of threats: privacy-preserving data mining aims at preventing the possibility of learning private personal data by unauthorized (possibly malicious) people, while our method aims at discovering unfair decisions or behaviors and, as a further step, at preventing taking similar decisions by authorized (possibly unaware) people. The issue of indirect discrimination through inference models resembles a privacy-preserving problem, where simply hiding a subset of rules (here, simply having datasets with no PD itemset) does not necessarily guarantee privacy protection from an attacker (here, should not prevent the unveiling of discriminatory decisions). The privacy-preserving literature contains several approaches to tackle this problem, which are all confronted with the trade-off between providing accurate models and preserving the privacy of individuals. It remains an open problem whether some of the existing attack models and privacy-preserving approaches can be effective in our context as well, either for discrimination discovery or for discrimination prevention.

10. CONCLUSIONS

Civil rights laws prohibit discrimination in a number of settings, including credit and insurance scoring, lending, personnel selection and wage, education, and many others. The influence of discriminative behaviors has been the subject of studies in economics, law and social sciences.

In this article, we have introduced the problem of discovering contexts of discriminatory decisions against protected-by-law groups, and provided a knowledge discovery process for solving it. Our approach is based on coding the involved concepts (potentially discriminated groups, contexts of discrimination, measures of discrimination, background knowledge, direct and indirect discrimination) in a coherent framework based on itemsets, association rules, and classification rules extracted from a dataset of historical decision records.

Clearly, many issues in discrimination-aware data mining remain open for future investigation, both on the technical and on the interdisciplinary side. On the technical side, the proposed approach can be extended to deal with continuous attributes, such as age and income; with continuous classes, such as wages and interest rate; with mining models other than classification rules, such as Bayesian models; with additional inference models. On the interdisciplinary side, it is important to pursue the interplay with legislation and regulatory authorities. In our opinion, research in data mining can contribute by providing a methodology for quantitative (self-)assessment and enforcement of discrimination in support of the existing qualitative legislative and regulatory definitions.

REFERENCES

AGRAWAL, R. AND SRIKANT, R. 1994. Fast algorithms for mining association rules in large databases. In *Proceedings of the International Conference on Very Large Databases*. Morgan Kaufmann, 487–499.

AGRAWAL, R. AND SRIKANT, R. 2000. Privacy-preserving data mining. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*. ACM, 439–450.

AUSTRALIAN LEGISLATION. 2009. (a) Equal Opportunity Act—Victoria State, (b) Anti-Discrimination Act—Queensland State. http://www.austlii.edu.au.

BAESENS, B., GESTEL, T. V., VIAENE, S., STEPANOVA, M., SUYKENS, J., AND VANTHIENEN, J. 2003. Benchmarking state-of-the-art classification algorithms for credit scoring. *J. Oper. Resear. Soc. 54*, 6, 627–635.

BECKER, G. S. 1957. *The Economics of Discrimination*. University of Chicago Press.

BERKA, P. 1999. PKDD 1999 discovery challenge. http://lisp.vse.cz/challenge.

CHIEN, C.-F. AND CHEN, L. 2008. Data mining to improve personnel selection and enhance human capital: A case study in high-technology industry. *Exp. Syst. Appl. 34*, 1, 280–290.

CLIFTON, C. 2003. Privacy preserving data mining: How do we mine data when we aren't allowed to see it? In *Procedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Tutorial. http://www.cs.purdue.edu/homes/clifton.

EUROPEAN UNION LEGISLATION. 2009. (a) Racial Equality Directive, (b) Employment Equality Directive. http://ec.europa.eu/employment_social/fundamental_rights.

GASTWIRTH, J. L. 1992. Statistical reasoning in the legal setting. *Amer. Statist. 46*, 1, 55–69.

GOETHALS, B. 2009. Frequent itemset mining implementations repository. http://fimi.cs.helsinki.fi.

HAND, D. J. 2001. Modelling consumer credit risk. *IMA J. Manag. Math. 12*, 139–155.

HAND, D. J. AND HENLEY, W. E. 1997. Statistical classification methods in consumer credit scoring: A review. *J. Royal Statist. Soc. Series A 160*, 523–541.

HARFORD, T. 2008. *Logic of Life*. The Random House.

HINTOGLU, A. A., INAN, A., SAYGIN, Y., AND KESKINÖZ, M. 2005. Suppressing data sets to prevent discovery of association rules. In *Proceedings of the IEEE International Conference on Data Mining*. IEEE Computer Society, 645–648.

HOLZER, H., RAPHAEL, S., AND STOLL, M. 2004. Black job applicants and the hiring officer's race. *Industr. Labor Relat. Rev. 57*, 2, 267–287.

HOLZER, H. J. AND NEUMARK, D. 2006. Affirmative action: What do we know? *J. Policy Anal. Manag. 25*, 463–490.

HUNTER, R. 1992. *Indirect Discrimination in the Workplace*. The Federation Press.

KAMIRAN, F. AND CALDERS, T. 2009. Classification without discrimination. In *Proceedings of the IEEE International Conference on Computer, Control & Communication*. IEEE Press.

KAYE, D. AND AICKIN, M., EDS. 1992. *Statistical Methods in Discrimination Litigation*. Marcel Dekker, Inc.

KNOPFF, R. 1986. On proving discrimination: Statistical methods and unfolding policy logics. *Canad. Pub. Policy 12*, 573–583.

KNUTH, D. 1997. *Fundamental Algorithms*. Addison-Wesley.

KUHN, P. 1987. Sex discrimination in labor markets: The role of statistical evidence. *Amer. Econ. Rev. 77*, 567–583.

LACOUR-LITTLE, M. 1999. Discrimination in mortgage lending: A critical review of the literature. *J. Real Estate Lit. 7*, 15–50.

LIU, B., HSU, W., AND MA, Y. 1998. Integrating classification and association rule mining. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. AAAI Press, 80–86.

LIU, K. 2009. Privacy preserving data mining bibliography. http://www.csee.umbc.edu/~kunliu1/research/privacy_review.html.

MAKKONEN, T. 2007. Measuring discrimination: Data collection and the EU equality law. http://ec.europa.eu/employment_social/fundamental_rights.

NEWMAN, D., HETTICH, S., BLAKE, C., AND MERZ, C. 1998. UCI repository of machine learning databases. http://archive.ics.uci.edu/ml.

PEDRESCHI, D., RUGGIERI, S., AND TURINI, F. 2008. Discrimination-aware data mining. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 560–568.

PEDRESCHI, D., RUGGIERI, S., AND TURINI, F. 2009. Measuring discrimination in socially-sensitive decision records. In *Proceedings of the SIAM International Conference on Data Mining*. SIAM, 581–592.

PIETTE, M. J. AND WHITE, P. F. 1999. Approaches for dealing with small sample sizes in employment discrimination litigation. *J. Foren. Econ. 12*, 43–56.

RAUCH, J. 2005. Logic of association rules. *Appl. Intell. 22*, 1, 9–28.

RAUCH, J. AND SIMUNEK, M. 2001. Mining for association rules by 4ft-Miner. In *Proceedings of the INAP 2001*. Prolog Association of Japan, 285–295.

RAUCH, J. AND SIMUNEK, M. 2009. 4-ft Miner Procedure. `http://lispminer.vse.cz`.

RIACH, P. A. AND RICH, J. 2002. Field experiments of discrimination in the market place. *Econ. J. 112*, 480–518.

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

SRIKANT, R. AND AGRAWAL, R. 1995. Mining generalized association rules. In *Proceedings of the International Conference on Very Large Databases*. Morgan Kaufmann, 407–419.

SWEENEY, L. 2001. Computational disclosure control: A primer on data privacy protection. Ph.D. thesis, MIT, Cambridge, MA.

SWEENEY, L. 2002. Achieving k-anonymity privacy protection using generalization and suppression. *Int. J. Uncertain. Fuzz. Knowl.-Bas. Syst. 10*, 5, 571–588.

TAN, P.-N., KUMAR, V., AND SRIVASTAVA, J. 2004. Selecting the right objective measure for association analysis. *Inform. Syst. 29*, 4, 293–313.

THOMAS, L. C. 2000. A survey of credit and behavioural scoring: Forecasting financial risk of lending to consumers. *Int. J. Forecast. 16*, 149–172.

U.K. LEGISLATION. 2009. (a) Sex Discrimination Act, (b) Race Relation Act. http://www.statutelaw.gov.uk.

U.S. FEDERAL LEGISLATION. 2009. (a) Equal Credit Opportunity Act, (b) Fair Housing Act, (c) Intentional Employment Discrimination, (d) Equal Pay Act, (e) Pregnancy Discrimination Act. http://www.usdoj.gov.

VAIDYA, J., CLIFTON, C. W., AND ZHU, Y. M. 2006. *Privacy Preserving Data Mining*. Advances in Information Security. Springer.

VERYKIOS, V. S., ELMAGARMID, A. K., BERTINO, E., SAYGIN, Y., AND DASSENI, E. 2004. Association rule hiding. *IEEE Trans. Knowl. Data Engin. 16*, 4, 434–447.

VIAENE, S., DERRIG, R. A., BAESENS, B., AND DEDENE, G. 2001. A comparison of state-of-the-art classification techniques for expert automobile insurance claim fraud detection. *J. Risk Insur. 69*, 3, 373–421.

VOJTEK, M. AND KOČENDA, E. 2006. Credit scoring methods. *J. Econ. Finance 56*, 152–167.

WANG, K., FUNG, B. C. M., AND YU, P. S. 2005. Template-based privacy preservation in classification problems. In *Proceedings of the IEEE International Conference on Data Mining*. IEEE Computer Society, 466–473.

WU, X., ZHANG, C., AND ZHANG, S. 2004. Efficient mining of both positive and negative association rules. *ACM Trans. Inform. Syst. 22*, 3, 381–405.

YIN, X. AND HAN, J. 2003. CPAR: Classification based on Predictive Association Rules. In *Proceedings of the SIAM International Conference on Data Mining*. SIAM, 331–335.

Received July 2008; revised June 2009; accepted August 2009