Università di Pisa
Dipartimento di Informatica

Technical Report: TR-07-19

Discrimination-aware data mining

Dino Pedreschi Salvatore Ruggieri Franco Turini

September 28, 2007

Address: Largo B. Pontecorvo 3, 56127 Pisa, Italy. Tel: +39 050 2212700 Fax: +39 050 2212726

Discrimination-aware data mining

Dino Pedreschi Salvatore Ruggieri Franco Turini
Dipartimento di Informatica, Università di Pisa
L.go B. Pontecorvo 3
56127 Pisa, Italy
{pedre,ruggieri,turini}@di.unipi.it

ABSTRACT

In the context of civil rights law, discrimination refers to unfair or unequal treatment of people based on membership to a category or a minority, without regard to individual merit. Rules extracted from databases by data mining techniques, such as classification or association rules, when used for decision tasks such as benefit or credit approval, can be discriminatory, in the above sense. This deficiency of classification and association rules poses ethical and legal issues, as well as obstacles to practical application. In this paper, the notion of discriminatory classification rules is introduced and studied. Examples of potentially discriminatory attributes include gender, race, job, and age. A measure, termed α-protection, of the discrimination power of a classification rule containing a discriminatory item is defined and its properties studied. We show that the introduced notion is non-trivial, in the sense that discriminatory rules can be derived from apparently safe ones under natural assumptions about background knowledge. Finally, we discuss how to check α-protection and provide an empirical assessment on the German credit dataset.

Keywords

Discrimination, knowledge discovery in databases, classification models, classification rules, interestingness measures.

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. According to Wikipedia, for instance, to discriminate is to make a distinction between people on the basis of their membership to a category, without regard to individual merit. Examples of social discrimination include racial, religious, gender, sexual orientation, disability, ethnic, height-related, and age-related discrimination. In the context of civil rights law, discrimination refers to unfair or unequal treatment of an individual (or group) based on certain characteristics. U.S. federal laws guarantee civil rights and prohibit discrimination in a number of settings, including:

- credit/insurance scoring, with the Equal Credit Opportunity Act [4] that prohibits a creditor to discriminate against any applicant on the basis of race, color, religion, national origin, sex or marital status, age, or because all or part of the applicant's income derives from any public assistance program;
- lending, with the Fair Housing Act [6] prohibiting discrimination in the sale, rental, and financing of housing based on race, color, national origin, religion, sex, familial status, and disability;
- personnel selection and wage discrimination, with the Intentional Employment Discrimination [15], the Equal Pay Act [5] and the Pregnancy Discrimination Act [7] prohibiting discrimination in personnel selection and wages based on race, color, religion, sex, national origin or pregnancy.

Other U.S. federal laws exists on discrimination programs or activities concerning public accommodations, education and health care. Services, benefits and aids must be provided in these context in a nondiscriminatory manner. Sample activities and programs covered include for instance academic programs, student services, nursing homes, adoptions, senior citizens centers, hospitals, transportation and open-to-public businesses such as in providing food, lodging, gasoline, and entertainment. Several authorities (regulation boards, consumer advisory councils, commissions) are settled to monitor discrimination compliances in U.S., European Union and many other countries.

Indirect or *systematic* discrimination has also been considered in laws [1, 2, 3]. Indirect discrimination consists of rules, procedures or requirements that, while not explicitly mentioning discriminatory attributes, intentionally or not impose disproportionate burdens on minority or disadvantaged groups. As an example, requiring that a job applicant must be over 1.8 meters tall will have the effect of excluding a disproportionate number of women or of people from some ethnic groups from applying. The legislator has also accounted for the use of scoring systems, which are automatic means of assigning a score to a benefit application based on some predetermined rules. In order to guarantee the fairness of rules used by scoring systems, U.S. law [4] (and also Italian law [13]) requires that adverse actions, such as rejection of a credit application or increase in insurance premium, when based on scoring systems, must be provided with specific and detailed reasons and explanations, in plain, unambiguous language, of the factors that had negatively affected the scoring degree.

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 [10]. The literature in those research fields has given evidence of unfair treatment in racial profiling and redlining [36], mortgage discrimination [25], personnel selection discrimination [19, 21], and wages discrimination [24]. Indirect discrimination has also been investigated [20, 22].

In data mining and machine learning, classification models are constructed on the basis of historical data exactly with the purpose of discrimination in the original Latin sense: i.e. distinguishing between elements of different classes, in order to unveil the reasons of class membership, or to predict it for unclassified samples. In either cases, classification models can be adopted as a support to decision making, clearly also in socially sensitive tasks such as the access of applicant people to benefits, to public services, to credit. As an example, a large body of literature [9, 16, 17, 41, 43, 44] has considered 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 [11] for personnel selection as well. Obviously, the risk of discrimination poses clear ethical and legal issues, as well as real obstacles to the practical application of classification techniques in socially sensitive decision making.

Now the question that naturally arises is the following. While classification models used for decision support can potentially guarantee less arbitrary decisions, can they be discriminating in the social, negative sense? The answer is clearly yes: it is evident that relying on mined models, e.g. classification rules, for decision making does not put ourselves on the safe side, as the rules extracted from the historical data may be discriminatory in the precise sense that disadvantaged groups or minorities can be unfairly classified just on the basis of their own status. Rather dangerously, learning rules 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 are hidden within a classifier. For instance, if it is a current malpractice to deny to pregnant women the access to certain job positions, there is a high chance to find a strong association in the historical data between pregnancy and access denial, and therefore we run the risk of learning a discriminatory rule. Despite its high accuracy and statistical significance, such a rule should be clearly identified: while it is useful as an explanation, which actually reveals the mentioned malpractice, it should absolutely not be used for decision making.

In this paper, we tackle the problem of discrimination in data mining models in a rule-based setting, by introducing the notion of *discriminatory classification rules*, as a criterion to identify the potential risks of discrimination. This criterion is based on two ingredients:

- some selected attribute values are identified as potentially discriminatory, on the basis of domain or background knowledge; examples include female gender, ethnic minority, low-level job, specific age range. It is worth noting that discriminatory items do not necessarily coincide with sensitive attributes with respect to pure privacy protection. For instance, gender is generally considered a non-sensitive attribute, whereas it can be discriminatory in many decision contexts.

- α-protection is introduced as a measure of the discrimination power of a classification rule containing one or more discriminatory items. The idea is to define such a measure as an estimation of the gain in precision of the rule due to the presence of the discriminatory items. The α parameter is the key for tuning the desired level of protection against discrimination.

As an example, consider the classification rules:

```
a. purpose=used car      b. age=senior
   ==> class=bad            gender=female
   -- conf:(0.165)          purpose=used car
                            ==> class=bad
                            -- conf:(1)
```

Rule (a) can be translated into the statement "people that intend to buy a used car are assigned the bad credit class" 16.5% of times. Rule (b) concentrates on "senior women that intend to buy a used car". In this case, the additional (discriminatory!) items in the premise increase the confidence of the rule up to 100%, more than 6 times. α-protection is intended to detect rules where such an increase is higher than a fixed threshold α. We study the properties of α-protection, which motivate its adoption as a criterion to assess the discriminatory power of a classification rule, both theoretically and empirically. The experiments conducted on the real dataset 'German credit' [30] of bank credit approval records show that many discriminatory rules are indeed discovered using our proposed criterion.

As far as indirect discrimination is concerned, we investigate the case when discrimination of a classification rule can be *inferred* by reasoning on a given set of α-protective rules. For instance, this happens in the case of binary classes, e.g. good/bad credit: we show how, in this case, the presence of a rule for the positive class, e.g.,

```
bg. age=senior
    gender=female
    purpose=used car
    ==> class=good
```

may be used under certain conditions to infer that the complementary rule (b) is discriminatory. In order to take into account such inferences, we extend α-protection to *strong* α-protection. Moreover, we show how our approach can be geared to identify rules that are *indirectly* discriminatory, i.e. rules that do not contain discriminatory conditions but, when combined with other rules or with background knowledge, allow to infer discriminatory rules. As an example, consider the classification rule:

```
c. driving_licence=no
   purpose=used car
   ==> class=bad
   -- conf:(1)
```

Assume to know that among the people that intend to buy a used car, people without driving licence are *almost the same* as senior women. Despite rule `(c)` does not contain any discriminatory item, it can bring some information on the discriminatory power of rule `(b)`. We will show a formal result on inferring a lower bound in such cases, and discuss how to prevent inferences that make use of that lower bound.

The paper is organized as follows. In Section 2, we recall standard notions on itemsets, association rules, classification rules, and measures such as support and confidence. Moreover, we introduce the measure of extended lift of an association rule. In Section 3, we introduce the notion of α-discrimination, study its properties on the German credit dataset, and extend it to strong α-discrimination. Inference of discrimination through two *attack models* is considered in Section 4. In Section 5 extensions of the approach and related work are discussed. Finally, in Section 6 the contribution of the paper is summarized. All proofs of theorems are reported in the Appendix A.

2. BASIC DEFINITIONS

2.1 Association and Classification Rules

We recall the notions of itemsets, association rules and classification rules from standard definitions [8, 27, 47]. 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 $\mathcal{I}$ be the set of all items. A transaction is a subset of $\mathcal{I}$, 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 $\mathcal{I}$. As usual in the literature, we write $\mathbf{X}, \mathbf{Y}$ for $\mathbf{X} \cup \mathbf{Y}$. For a transaction T, we say that T verifies $\mathbf{X}$ if $\mathbf{X} \subseteq T$. The support of an itemset $\mathbf{X}$ w.r.t. a non-empty transaction database $\mathcal{D}$ is the ratio of transactions in $\mathcal{D}$ verifying $\mathbf{X}$:

$$supp_{\mathcal{D}}(\mathbf{X}) = |\{\ T \in \mathcal{D} \mid \mathbf{X} \subseteq T\ \}|/|\mathcal{D}|,$$

where $|\ |$ is the cardinality operator. An association rule is an expression $\mathbf{X} \rightarrow \mathbf{Y}$, where $\mathbf{X}$ and $\mathbf{Y}$ are itemsets such that $\mathbf{X} \cap \mathbf{Y} = \emptyset$. $\mathbf{X}$ 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 [27, 47] for a discussion of the integration between classification and association rule mining. The support of $\mathbf{X} \rightarrow \mathbf{Y}$ w.r.t. $\mathcal{D}$ is defined as:

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

The confidence of $\mathbf{X} \rightarrow \mathbf{Y}$, defined when $supp_{\mathcal{D}}(\mathbf{X}) > 0$, is:

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

Support and confidence range over $[0, 1]$. We omit the subscripts in $supp_{\mathcal{D}}()$ and $conf_{\mathcal{D}}()$ when clear from the context. Since the seminal paper by Agrawal and Srikant [8], a number of well explored algorithms [35] have been introduced in order to extract *frequent* itemsets, i.e. itemsets with a specified minimum support, and valid association rules, i.e. rules with a specified minimum confidence.

2.2 Extended Lift

We introduce a key concept for our purposes.

Definition 2.1. *[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 $[0, 1/mc]$. Proofs of these statements are reported in the Appendix A, Lemma A.6. The extended lift can be traced back to the well-known measure of lift [39], defined as:

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

Lemma 2.2. *Let $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$ be an association rule such that $conf_{\mathcal{D}}(\mathbf{B} \rightarrow \mathbf{C}) > 0$. We have:*

$$\frac{conf_{\mathcal{D}}(\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C})}{conf_{\mathcal{D}}(\mathbf{B} \rightarrow \mathbf{C})} \quad = \quad lift_{\mathcal{B}}(\mathbf{A} \rightarrow \mathbf{C})$$

where $\mathcal{B} = \{T \in \mathcal{D} \ |\mathbf{B} \subseteq T\}$.

Notice that when $\mathbf{B}$ is empty, we have $\mathcal{B} = \mathcal{D}$, and then the extended lift reduces to the standard lift.

3. MEASURING DISCRIMINATION

3.1 Discriminatory Itemsets and Rules

Our starting point consists of flagging at syntax level those items which might potentially lead to discrimination in the sense explained in the introduction. Potentially discriminatory itemsets are then itemsets which consists of flagged items only.

Definition 3.1. *[PD/PND itemset] A discriminatory set $\mathcal{I}_d$ is a subset of items, i.e. $\mathcal{I}_d \subseteq \mathcal{I}$. A potentially discriminatory (PD) itemset is a non-empty subset of $\mathcal{I}_d$. A potentially non-discriminatory (PND) itemset is a (possibly empty) subset of $\mathcal{I} \setminus \mathcal{I}_d$.*

Any itemset $\mathbf{X}$ can be uniquely split into a PD part $\mathbf{A} = \mathbf{X} \cap \mathcal{I}_d$ and a PND part $\mathbf{B} = \mathbf{X} \setminus \mathcal{I}_d$.

Example 3.2. *Consider a relation with attributes* `age, gender, education, residence` *and class* `hire`. *Transactions model whether candidates of given age, gender, education and residence were hired or not on past applications. Assume now that $\mathcal{I}_d = \{$*`gender=female`*, *`age=senior`*$\}$. The itemset:*

`gender=female`, `age=young`, `residence=Italy`, *can be split into a PD part, i.e.* `gender=female`*, and a PND part, namely* `age=young`, `residence=Italy`*.*

It is worth noting that we use the adjective *potentially* both for PD and PND itemsets. This may seems strange, since we located PD items as the source of (potential) discrimination. The rationale for using *potentially* in PND itemsets will be clear later on, when we will discuss how discrimination can be hidden in non-discriminatory items as well. The notion of potential (non-)discrimination is now extended to classification rules.

Definition 3.3. *[PD/PND classification rule] A classification rule* $\mathbf{X} \rightarrow \mathbf{C}$ *is called potentially discriminatory (PD) if* $\mathbf{X} = \mathbf{A}, \mathbf{B}$ *with* $\mathbf{A}$ *PD itemset and* $\mathbf{B}$ *PND itemset. It is called potentially non-discriminatory (PND) if* $\mathbf{X}$ *is a PND itemset.*

Example 3.4. *Consider again Example 3.2, and the classification rules:*

```
a.  gender=female              b. age=young
    age=young                  residence=Italy
    residence=Italy            ==> hire=no
    ==> hire=no
```

Rule `(a)` *is potentially discriminatory since its premise contains an item belonging to* $\mathcal{I}_d$*, namely* `gender=female`*. On the contrary, rule* `(b)` *is potentially non-discriminatory.*

3.2 Measuring Discrimination of PDCR

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

$$\begin{aligned} \gamma &= conf(\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}) \\ \delta &= conf(\mathbf{B} \rightarrow \mathbf{C}) > 0. \end{aligned}$$

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

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

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

Intuitively, the definition assumes that the extended lift of c w.r.t. $\mathbf{B}$ is a measure of the degree of discrimination of $\mathbf{A}$ in the context $\mathbf{B}$. This is supported by two basic intuitions. First, α-protection states that the added (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. The second intuition follows from Lemma 2.2, which reduces extended lift to standard lift, and the known properties of standard lift. In this sense, α-protection states that the degree of correlation between $\mathbf{A}$ and $\mathbf{C}$ in the context of $\mathbf{B}$ is bounded by α.

Example 3.6. *Assume that* $\mathcal{I}_d$ *includes the items:*

`personal_status=female div/sep/mar`,

i.e. non-single (divorced, separated or married) female, and

`age=(52.6-inf)`

i.e. senior people. Moreover, fix $\alpha = 3$*. Consider the following classification rules, which are extracted from the German credit dataset [30].*

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

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

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

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

is $0.27/1.52 = 17.8\%$*.*

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

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

is $1/6.06 = 16.5\%$*.*

3.3 Checking α-protection of PDCR

Now that we have set the stage for discussing a discrimination measure of PD classification rules, we need to tackle the problem of checking α-protection, which is implicitly stated by Definition 3.5.

Problem 3.7 (α-protection checking). *Given a set of PD classification rules* $\mathcal{A}$ *and a threshold* α*, find the largest subset of* $\mathcal{A}$ *containing only* α*-protective rules.*

We envisage that checking α-protection can be exploited in at least three different practical purposes:

- check that the set of rules under analysis is made of α-protective rules only, to the purpose of authorizing the safe use of the set of rules as a provably fair model;
- identify all discriminatory rules to the purpose of discovering discrimination malpractices that emerge from the historical transactions in the source dataset;
- identify certain expected discriminatory rules to the purpose of checking that the results of some specific

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

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

Figure 1: Classification rules extraction (left) and checking α-protection (right) algorithms.

positive discrimination policies – or affirmative actions, that tend to favor some disadvantaged category – actually emerge from the historical transactions in the source dataset.

The problem of checking α-protection is solvable by directly checking the inequality of Definition 3.5, provided that the elements of the inequality are available. We define a checking algorithm that starts from the set of frequent itemsets, namely itemsets with a given minimum support. This is the output of any of the several frequent itemset extraction algorithms available at the FIMI repository [35]. The algorithm is reported in Figure 1. On the left hand side of the figure, the extraction of PD and PND classification rules is reported. It requires a single scan of frequent itemsets ordered by the itemset size k. For k-frequent itemsets that include a class item, a single classification rule is produced in output. The confidence of the rule can be computed by looking only at itemsets of length $k-1$. The rules in output are distinguished between PD and PND rules, based on the presence of discriminatory items in their premise. Moreover, the rules are grouped on the basis of the number *group* of non-discriminatory items appearing in their premise. The output is a collection of PD rules $\mathcal{PD}_{group}$ and a collection of PND rules $\mathcal{PND}_{group}$. On the right hand side of Figure 1, the extended lift of a classification rule $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C} \in \mathcal{PD}_{group}$ is computed from its confidence and the confidence of the base rule $\mathbf{B} \rightarrow \mathbf{C} \in \mathcal{PND}_{group}$.

3.4 The German Credit Dataset

We illustrate the notion of α-protection by analysing the German credit dataset [30], which consists of 1000 transactions representing the credit class good/bad of bank account holders. The dataset include the following nominal (or discretized):

- attributes on personal properties: checking account status, duration, savings status, property magnitude, type of housing;
- attributes on past/current credits and requested credit: credit history, credit request purpose, credit request amount, installment commitment, existing credits, other parties, other payment plan;
- attributes on employment status: job type, employment since, number of dependents, own telephone;
- personal attributes: personal status and gender, age, resident since, foreign worker.

In the following, we fix $\mathcal{I}_d$ to consists of the following 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).

Discrimination w.r.t. support thresholds

Figure 2 shows the distribution of α-discriminatory PD classification rules for minimum support thresholds of 1%, 0.5% and 0.3%. The left hand side graph reports the absolute count, while the one at the right hand side reports the relative count w.r.t. the total number of PD classification rules having the minimum support. Figure 2 highlights that lower support values allows for increasing both the maximum extended lift of PD classification rules, the number and the proportion of PD rules with higher extended lift. This confirms the theoretical ranges of the extended lift measure and it is coherent with the intuition that, in smaller and smaller niches of the sampled credit approval transactions, it is possible to find higher discriminatory behavior.

By looking at the extracted classification rules, we report a few example of PD rules with decreasing support and increasing extended lift.

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

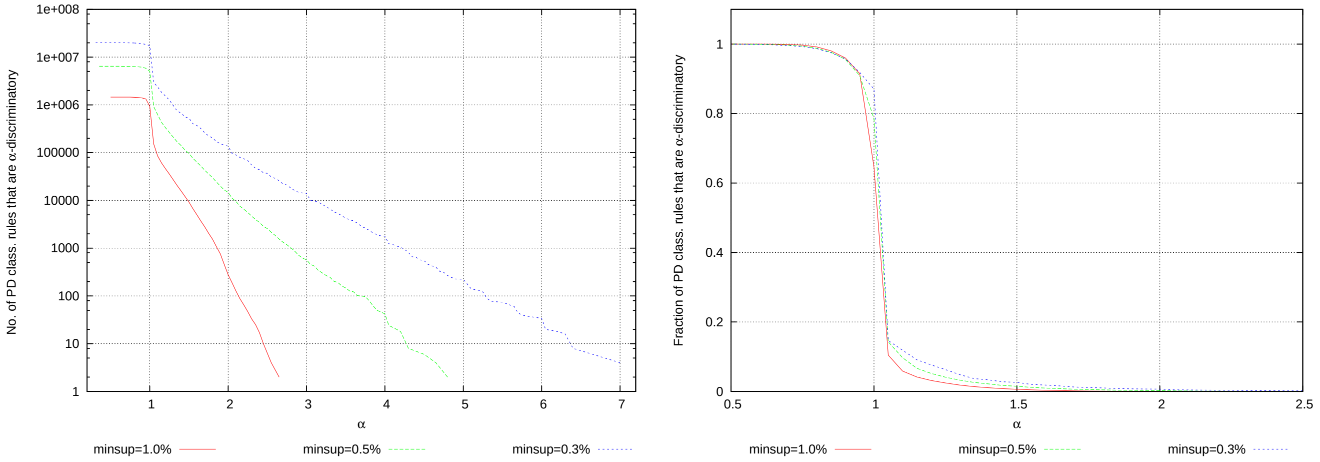

Figure 2: Distributions of discriminatory PD classification rules: absolute count (left) and relative count (right) for minimum support thresholds of 1%, 0.5% and 0.3%.

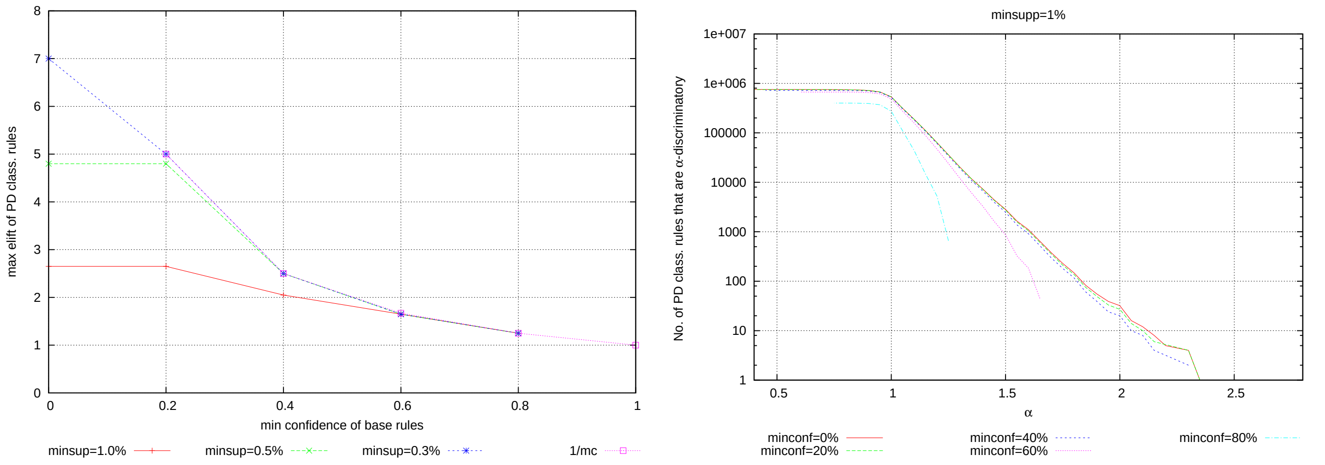

Figure 3: Distributions of discriminatory PD classification rules: contribution of confidence of base rule.

```
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)
```

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 case of not considering the `age`-item. 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)
```

There are 18 cases satisfying the premise of the base rule, which means that people with large savings are usually given good credit. However, only 2 cases out of 18 are assigned `class=bad`. Both of them are senior people!

Discrimination w.r.t. confidence thresholds

In addition to minimum support, a widely adopted parameter for controlling rule generation is minimum confidence. By recalling that extended lift ranges over $[0, 1/mc]$, where mc is the minimum confidence threshold of base rules, Figure 3 shows how the threshold mc affects the distribution of discriminatory classification rules. The left hand side graph reports the maximal extended lift reachable by a PD clas-

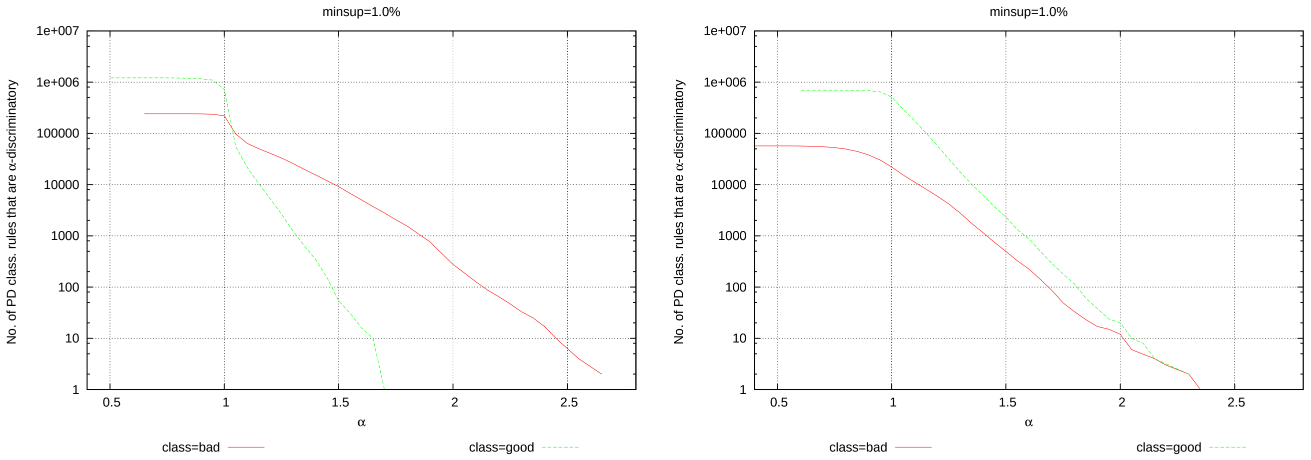

Figure 4: Distributions of discriminatory PD classification rules for two different discriminatory sets. Left: $\mathcal{I}_d = \{$ `personal_status=female div/sep/mar, age=(52.6-inf), job=unemp/unskilled non res, foreign_worker=yes` $\}$. **Right:** $\mathcal{I}'_d = \{$ `personal_status=male single, age=(30.2-41.4]` $\}$.

sification rule with given minimum support and confidence. The graph shows that lower and lower confidence thresholds for base rules lead to higher extended lifts, as the range $[0, 1/mc]$ suggests. Notice that the upper bound $1/mc$ is reached for minimum support of 0.3% and minimum base confidence greater or equal than 20%. The right hand side considers minimum support of 1% and shows the absolute count distribution of discriminatory rules at various minimum confidence thresholds for base rules. Acting on this value allows to reduce both the number of discriminatory classification rules and their maximum extended lift value.

A few sample PD rules with the same support (0.5%) but with increasing extended lift and decreasing confidence are reported below.

```
b1. personal_status=female div/sep/mar
    purpose=used car
    job=high qualif/self emp/mgmt
    ==> class=bad
    -- conf:(0.55) conf_base:(0.17) elift:(3.24)

b2. personal_status=female div/sep/mar
    purpose=used car
    checking_status=0<=X<200
    ==> class=bad
    --  conf:(0.84) conf_base:(0.29) elift:(2.91)

b3. personal_status=female div/sep/mar
    residence_since=(1.6-2.2]
    job=high qualif/self emp/mgmt
    duration=(31.2-inf)
    ==> class=bad
    -- conf:(1) conf_base:(0.48) elift:(2.10)
```

All of the rules consider being woman and not single as the discriminatory item. By comparing rule `b1` against `b2`, we note that the increasing of confidence of the base rule is partly compensated by higher confidence of the classification rule, which leads to a slightly lower extended lift. However, since confidence is at most 1, a higher confidence of the base rule, such as in `b3`, cannot always be compensated, and then the extended lift value tend to decrease considerably.

Discrimination w.r.t. class item

The contribution of the class item to the distribution of discriminatory PD classification rules is shown in Figure 4, where the minimum support of 1% is fixed. The left hand side highlights that rules with class item **`class=bad`** (resp., **`class=good`**) contribute mostly to higher values (resp., lower values) of extended lift. Intuitively, the set of discriminatory items fixed so far leads mainly to discrimination *against* assigning credit. There are, however, cases where discrimination *in favor* of assigning credit is raised. Here there are a few examples.

```
c1. 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)

c2. age=(52.6-inf)
    housing=own
    credit_history=existing paid
    employment=unemployed
    other_parties=none
    ==> class=good
    -- supp:(0.005) conf:(1) elift:(2.13)
```

In rules `c1` and `c2` recently employed women and unemployed (maybe, retired) senior people are assigned good credit score two times more than the average of people in the same conditions. This reveals a good practice of enforcement of affirmative actions or other policies or laws in support of disadvantaged categories. Discrimination *in favor* of assigning credit can also reveal a malpractice of unfair favoritism for certain categories. In order to illustrate this issue, however, we need to switch to a different discriminatory set. Figure 4 shows the distributions of discriminatory PD rules for each class item. The discriminatory set

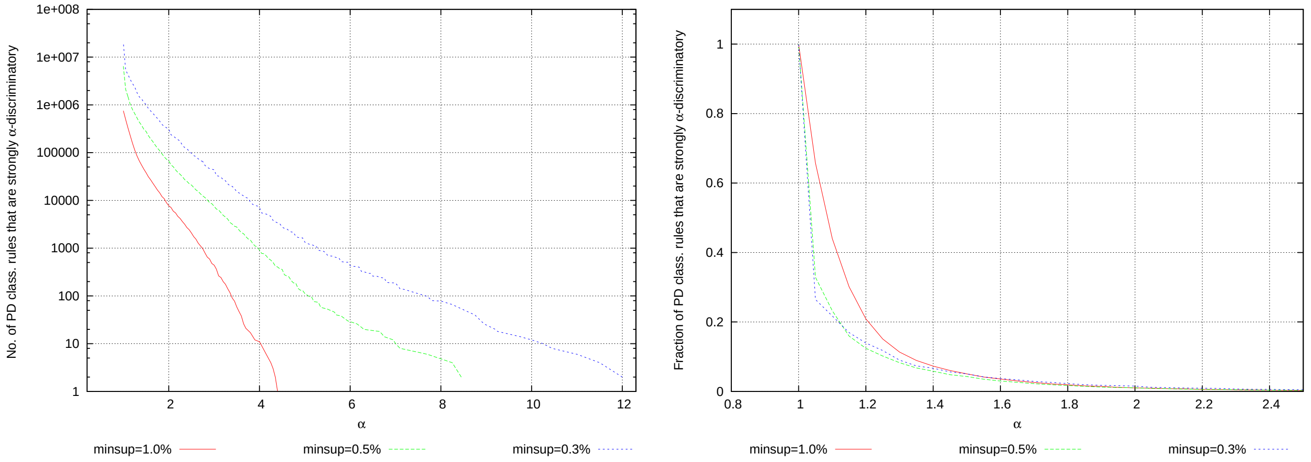

Figure 5: Distributions of strongly discriminatory PD classification rules: absolute count (left) and relative count (right) for minimum support thresholds of 1%, 0.5% and 0.3%.

used used so far, namely $\mathcal{I}_d = \{$ `personal_status=female div/sep/mar`, `age=(52.6-inf)`, `job=unemp/unskilled non res`, `foreign_worker=yes` $\}$, leads to higher extended lift for `class=bad`, rather than for `class=good`. On the contrary, the discriminatory set $\mathcal{I}'_d = \{$ `personal_status=male single`, `age=(30.2-41.4]` $\}$ leads to the distribution shown at the right hand side of Figure 4. Here, classification rules with class item `class=good` occur more frequently for higher extended lift values. Intuitively, $\mathcal{I}'_d$ allows for the extraction of contexts which discriminate *in favor* of assigning credit! Let us report two extracted rules that could be considered as instances of unfair favoritism towards men who are singles and/or in their 30's.

```
d1. age=(30.2-41.4]
    existing_credits=(1.6-2.2]
    duration=(31.2-inf)
    savings_status=<100
    ==> class=good
    -- supp:(0.010) conf:(0.91) elift:(2.14)

d2. age=(30.2-41.4]
    personal_status=male single
    checking_status=0<=X<200
    job=high qualif/self emp/mgmt
    ==> class=good
    -- supp:(0.010) conf:(0.83) elift:(1.90)
```

Rule `d1` discriminates in favor of people in their 30's among those that have had an account for at least 31.2 months, have savings for at most 100 units, and have two existing credits already. Rule `d2` discriminates in favor of males who are single and in their 30's among those with high qualified or self-employed job.

3.5 Strong Discrimination of PDCR

So far, we have considered as discriminatory those PD classification rules whose extended lift is greater or equal than a fixed threshold α. The intuition was to extract contexts where the confidence of the rule with discriminatory items *exceeds* by α times the confidence of the base rule. As a consequence, PD classification rules with low extended lift were implicitly considered non-discriminatory. Hence, shall we not be worried about disclosing them? Next example shows that the answer is not obvious.

Example 3.8. *The following PD classification rule is extracted from the German credit dataset for a 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* good *credit class is assigned to non-single women* less *than the average of people that intend 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. In fact, the following 'dual' rule can be extracted from the dataset.*

```
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)
```

Rule `a-bad` *has an extended lift of* 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. If this property holds in general, the extended lift of a rule $\mathbf{A}, \mathbf{B} \rightarrow$ `class=bad` could be calculated starting from the confidence of $\mathbf{A}, \mathbf{B} \rightarrow$ `class=good` and the confidence of its base rule $\mathbf{B} \rightarrow$ `class=good`. Therefore, α-discrimination of a classification rule could be deduced starting from an α-protective rule (the dual one) and its PND base rule.

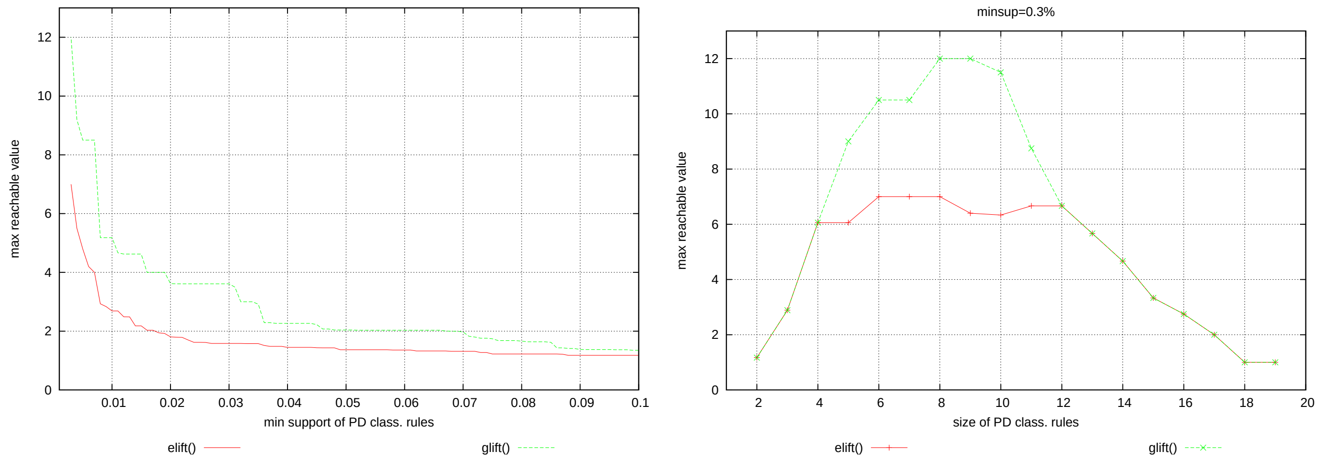

Figure 6: Maximum reachable *elift*() and *glift*() values at the variation of minimum support (left) and maximum rule size (right).

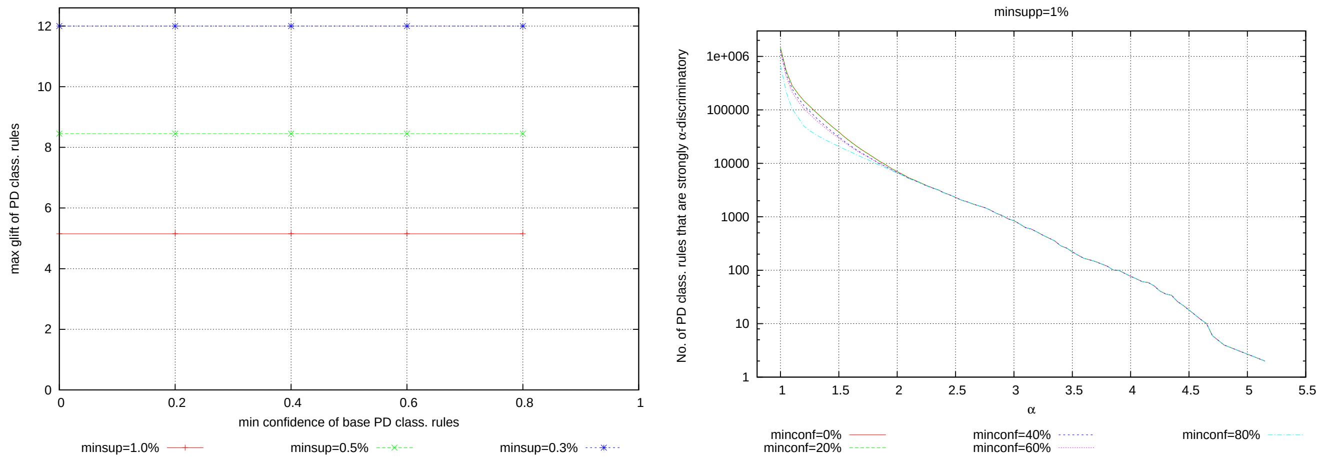

Figure 7: Contribution of setting minimum confidence for base rules to the distribution of strongly discriminatory PD classification rules.

3.5.1 Measuring Discrimination

We generalize the intuition behind the last example to binary classes.

Definition 3.9. *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 3.10. *Assume that the class attribute is binary. Let $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$ be a classification rule, and let:*

$$\begin{aligned} \gamma = & \; conf(\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}) \\ 1 > \quad \delta = & \; conf(\mathbf{B} \rightarrow \mathbf{C}), \end{aligned}$$

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

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

As an immediate consequence, the disclosure of a PD classification rule $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$ with $elift(\gamma, \delta) < \alpha$ might allow to calculate the extended lift of the dual rule $\mathbf{A}, \mathbf{B} \rightarrow \neg\mathbf{C}$ as $(1-\gamma)/(1-\delta)$, and then to decide whether or not it is α-protective. We tackle this leak of the framework by strengthening the notion of α-protection.

Definition 3.11. *[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:*

$$\begin{aligned} \gamma &= conf(\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}) \\ \delta &= conf(\mathbf{B} \rightarrow \mathbf{C}) > 0. \end{aligned}$$

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

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

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

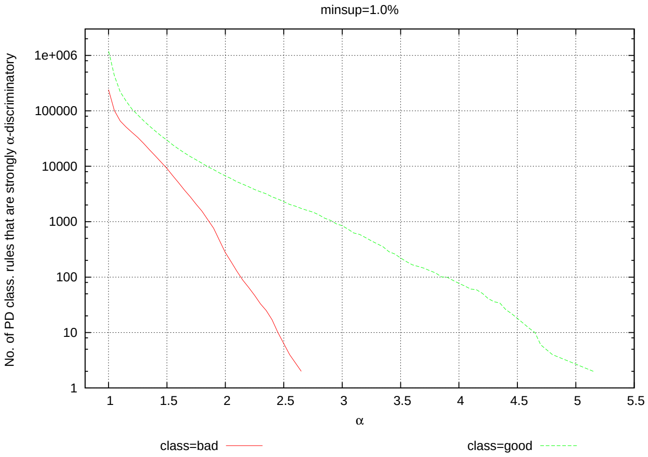

Figure 8: Distributions of strongly discriminatory PD classification rules for different class items.

The $glift()$ function ranges over $[1, \infty[$. If classification rules with a minimum support $ms > 0$ are considered, it ranges over $[1, 1/ms]$. A proof of these statements is reported in Appendix A Lemma A.10, which also states that, for $1 > \delta > 0$:

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

This property makes it clear that $glift()$ overcame the leak of extended lift by considering both the extended lift of a classification rule $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$ and the extended lift of the dual rule $\mathbf{A}, \mathbf{B} \rightarrow \neg\mathbf{C}$.

3.5.2 Checking Strong α-protection

The algorithm at the right hand side of Figure 1 for checking α-protection can be immediately extended to strong α-protection by simply replacing the $elift()$ function with the $glift()$ function.

3.5.3 The German Credit Dataset

Figure 5 shows the absolute and relative distributions of strongly α-discriminatory classification rules for the German credit dataset. Lower minimum support leads to higher values of $glift()$, and to a larger number and proportion of strongly discriminatory rules. Moreover, $glift()$ assumes higher values than those of $elift()$. The left hand side graph in Figure 6 shows the distributions of the maximum reachable value of $glift()$ and $elift()$ w.r.t. minimum support. The higher maximal values of $glift()$ can be exaplained by noting that given a set of PD classification rules $\mathcal{A}$ and a rule $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$ in $\mathcal{A}$, the dual rule $\mathbf{A}, \mathbf{B} \rightarrow \neg\mathbf{C}$ does not necessarily belongs to $\mathcal{A}$. For instance, rule `a-bad` in Example 3.8 has a support of 0.2%, which is strictly lower than the minimum support of 1% fixed for extracting rule `a-good`.

In addition to support, another parameter that can control the maximum reachable $glift()$ and $elift()$ values is the classification rule size, i.e. the number of items appearing in the rule. The right hand side graph of Figure 6 shows the distributions w.r.t. rule size. The "bell" shape of the distributions follows the distribution of the number of rules, which in turn follows the one of frequent itemsets. Rules with intermediate size are the vast majority, and then they exhibits the higher values.

Differently from α-protection, acting on minimum confidence of the base rule does not turn out to be a control mechanism for the $glift()$ measure, since $glift(\gamma, \delta)$ is not monotonic w.r.t. δ. Figure 7 shows that the distribution of strongly α-discriminatory rules is minimally affected by the minimum confidence of base rules.

Finally, let us consider in Figure 8 the distributions of α-discriminatory PD classification rules for the two class items. Contrasting the graph with the one in Figure 4 (left hand side), we can observe that the distribution of rules with class item `class=good` has now larger values and frequencies than the one of rules with class item `class=bad`, which for $\alpha \geq 1$ remains almost unchanged.

4. INDIRECT DISCRIMINATION OF PNDCR

Our aim is to offer decision makers a set of classification rules that do not allow them to excessively exploit discriminatory items in taking decisions. The natural question is we pose now is: the absence of (strongly) α-discriminatory rules is a sufficient guarantee of eliminating all discriminatory actions? Consequently, removing (strongly) α-discriminatory rules is a sufficient guarantee? We provide a strong negative answer to the question. Strong in the precise sense that, even by removing all PD classification rules, it is possible to devise some strategies of inferring (strong) α-discrimination starting from potentially non-discriminatory rules, i.e. rules that do not contain discriminatory items at all, and some background knowledge. Borrowing the terminology from other research fields, we call those strategies *attack models*.

4.1 Attacks Through Negated Items

We start considering a binary attribute such that one of its items is discriminatory and the other is non-discriminatory. Even in the case that all PD classification rules are undisclosed, PND classification rules that contain the non-discriminatory item contain hides some information about the discriminatory one. This is somewhat similar to what has

```
ExtractAR()
    $\mathcal{C} = \{$ class items $\}$
    ForEach $k$ s.t. there exists $k$-frequent itemsets
      $\mathcal{F}_k = \{$ k-frequent itemsets $\}$
      ForEach $\mathbf{X} \in \mathcal{F}_k$ with $\mathbf{X} \cap \mathcal{I}_d \neq \emptyset$ and $\mathbf{X} \cap \mathcal{C} = \emptyset$
        $\mathbf{A} = \mathbf{X} \cap \mathcal{I}_d$
        $\mathbf{B} = \mathbf{X} \setminus \mathbf{A}$
        $group = |\mathbf{B}|$
        $s\_x = supp(\mathbf{X})$
        If exists $\mathbf{B} \rightarrow \mathbf{C} \in \mathcal{PND}_{group}$
          $s = supp(\mathbf{B} \rightarrow \mathbf{C})$
          $c = conf(\mathbf{B} \rightarrow \mathbf{C})$
          $s\_b = s/c$          // equals to $supp(\mathbf{B})$
          $conf = s\_x/s\_b$
          add $\mathbf{B} \rightarrow \mathbf{A}$ to $\mathcal{AR}_{group}$ with confidence $conf$
        EndIf
      EndForEach
    EndForEach
```

```
CheckAlphaPNDNegated($\alpha$)
    $\mathcal{N} = \{a = v_1 \mid a = v_0 \in \mathcal{I}_d$ and $dom(a) = \{v_0, v_1\}\ \}$
    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 $\mathcal{PND}_{group-1}$
          $\beta = conf(\mathbf{B} \rightarrow \mathbf{A})$    // found in $\mathcal{AR}_{group-1}$
          $n = \delta/\beta + (1 - 1/\beta)\gamma$
          If $glift(n, \delta) \geq \alpha$
            output $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$
          EndIf
        EndForEach
      EndForEach
    EndForEach
```

Figure 9: Left: algorithm for extracting association rules as background knowledge. Right: algorithm for checking strong α-discrimination trough PND classification rules containing negated discriminatory items.

been described in Section 3.5 for binary classes.

Example 4.1. *The following PND classification rules are extracted from the German credit dataset by fixing minimum support to 0.5%.*

```
nac. foreign_worker=no
     personal_status=male single
     employment=1<=X<4
     purpose=new car
     housing=rent
     ==> class=good -- conf:(1)

bc.  personal_status=male single
     employment=1<=X<4
     purpose='new car'
     housing=rent
     ==> class=good -- conf:(0.9)
```

Consider the context $\mathbf{B}$ *of people that are single male, employed since one to four years, intend to buy a new car, and have their house for rent. Rule* nac *states that non-foreign workers within the context* $\mathbf{B}$ *are assigned a good credit scoring with confidence 100%. Rule* bc *states that the average confidence of people in context* $\mathbf{B}$ *to get the good scoring is slightly less, namely 90%. While the two rules are apparently safe with respect to discrimination against foreign workers, this is not the case. In fact, it is quite intuitive that the increasing of confidence from 90% to 100% has to be attributed to the omission of foreign workers. In order to estimate how much foreign workers are discriminated, however, we need to know some further information on the proportion of foreign workers in the context* $\mathbf{B}$. *Assume then that it is background knowledge (e.g., public statistics, news, private data) that approximately half of people in the context are foreign workers, i.e.:*

```
ba.  personal_status=male single
     employment=1<=X<4
     purpose=new car
     housing=rent
     ==> foreign_worker=yes -- conf:(0.5)
```

We will show that this information can be used to calculate the confidence of getting the good credit scoring for foreign workers in the context $\mathbf{B}$, *i.e. we can precisely calculate the confidence of:*

```
ac.  foreign_worker=yes
     personal_status=male single
     employment=1<=X<4
     purpose=new car
     housing=rent
     ==> class=good
     -- conf:(0.8) -- elift(0.89) -- glift(2.0)
```

as: $0.9/0.5 + (1 - 1/0.5)1 = 0.8$. *This and confidence of* bc *imply* $elift(0.8, 0.9) = 0.89$ *and* $glift(0.8, 0.9) = 2.0$.

4.1.1 Attack Model

The next result formalizes the attack model behind the last example. It states that the confidence of an undisclosed PD rule containing a binary attribute can be calculated from the confidence of the dual rule (γ), the confidence of its base rule (δ) and some information about the frequency of the binary attribute values in the context of the base rule (β).

Theorem 4.2. *Assume that the attribute of* $\mathbf{A} \in \mathcal{I}_d$ *is binary. Let* $\neg\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$ *be a PND classification rule, and:*

$$\begin{aligned} \gamma &= conf(\neg\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}) \\ \delta &= conf(\mathbf{B} \rightarrow \mathbf{C}) > 0 \end{aligned}$$

and assume that $\beta = conf(\mathbf{B} \rightarrow \mathbf{A}) > 0$ *is known. Let:*

$$n = \frac{\delta}{\beta} + (1 - \frac{1}{\beta})\gamma$$

we have that:

(i) $conf(\mathbf{A} \wedge \mathbf{B} \rightarrow \mathbf{C}) = n$,

(ii) *for* $\alpha \geq 0$, *the PD classification rule* $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$ *is* α-*protective iff* $elift(n, \delta) \leq \alpha$,

(ii) *for* $\alpha \geq 1$, *the PD classification rule* $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$ *is strongly* α-*protective iff* $glift(n, \delta) \leq \alpha$.

While Theorem 4.2 assumes that $conf(\mathbf{B} \rightarrow \mathbf{A})$ is known exactly, it is quite immediate to extend its conclusions to approximated values. In fact, assume that $\beta \in [\beta_1, \beta_2]$. By elementary algebra, it is easy to derive lower and upper bounds for n:

$$n_2 = \frac{\delta}{\beta_1} + (1 - \frac{1}{\beta_2})\gamma \geq n \geq \frac{\delta}{\beta_2} + (1 - \frac{1}{\beta_1})\gamma = n_1,$$

and, for $elift()$ and $glift()$:

$$elift(n, \delta) \geq n_1/\delta,$$

$$glift(n, \delta) \geq \begin{cases} n_1/\delta & \text{if } n_1 \geq \delta \\ (1 - n_2)/(1 - \delta) & \text{if } n_2 < \delta. \end{cases}$$

EXAMPLE 4.3. *Reconsider Example 4.1. Assume to know that the confidence of* ba *is in the range* $[0.495, 0.505]$, *i.e.* $0.5 \pm 1\%$. *Let us calculate:*

$$n_2 = 0.9/0.495 + (1 - 1/0.505)1 = 0.838$$

and then $n_2 < \delta = 0.9$ *implies:*

$$glift(n, \delta) \geq (1 - n_2)/(1 - \delta) = 1.62.$$

We recall that the actual $glift()$ *value is* 2.0.

A question that naturally arises is the following: how can an attacker know (with some approximation) the value of $conf(\mathbf{B} \rightarrow \mathbf{A})$ if the association rule $\mathbf{B} \rightarrow \mathbf{A}$ or the dataset are undisclosed? As it happens in privacy-preserving data mining [28, 38], external, publicly available or private, sources of data can be exploited to derive information on population within a certain context, e.g. about the number of foreign workers living in a certain area. Assuming that the distributions of attribute values for the transaction database $\mathcal{D}$ from which rules are extracted is the same as the one for attribute values of the external sources, at least for the context $\mathbf{B}$, allows for the use of statistics derived from the external sources as an approximation for the value $conf(\mathbf{B} \rightarrow \mathbf{A})$ w.r.t. $\mathcal{D}$. The same argument applies later on for the other attack model that will be presented.

4.1.2 Checking Attack Model

Let us concentrate now on how to compute the set of PND classification rules that allow to infer (strong) discrimination of PD rules. We first extract association rules $\mathbf{B} \rightarrow \mathbf{A}$ using algorithm **ExtractAR()** in Figure 9. Actually, notice that we can restrict to assuming that $\mathbf{B}$ is a PND itemset, $\mathbf{A}$ is PD itemset, and there exists a PND classification rules $\mathbf{B} \rightarrow \mathbf{C}$, i.e. with premise $\mathbf{B}$. All these assumptions are met by Theorem 4.2, and allows us for restricting the set of rules to be extracted[1]. Moreover, we group the rules based on the number of items in their premise. This will improve efficiency of the checking algorithm. The algorithm **CheckAlphaPNDNegated**(α) scan PND classification rules $\mathbf{X} \rightarrow \mathbf{C}$. For each $\neg\mathbf{A}$ in $\mathbf{X}$ that is the negation of a discriminatory item, the conditions of Theorem 4.2 are checked, by looking up the association rule $\mathbf{B} \rightarrow \mathbf{C}$ from the set of extracted rules.

4.1.3 The German Credit Dataset

Figure 10 reports the distributions of the PND classification rules that allow to derive strong discrimination of PD rules using the attack model of Theorem 4.2. More in detail, the only item in the reference discriminatory set $\mathcal{I}_d$ used throughout this paper is $\texttt{foreign\_worker} = \texttt{yes}$. The figure reports the absolute and relative count of PND classification rules of the form $\texttt{foreign\_worker} = \texttt{no}, \mathbf{B} \rightarrow \mathbf{C}$ that allow for deriving strong discrimination of PD rules $\texttt{foreign\_worker} = \texttt{yes}, \mathbf{B} \rightarrow \mathbf{C}$.

4.2 Attacks Through Related Itemsets

As a naive reaction to the results of the last subsection, one could be tempted to discard from the underlying dataset any binary attribute that contain one and only one discriminatory item, such as `personal_status`. Even more radically, one could discard *any attribute that contains any discriminatory item*, such as `age`, `personal_status` and `job` for the German credit dataset. In other words, we would retain PND classification rules only, with no explicit link (such as negated items) to PD rules. Would this solve the discrimination issue? Again, the answer is unfortunately no.

EXAMPLE 4.4. *Consider again the German credit dataset, from which the following PND classification rules are extracted.*

```
bdc. credit_history=critical/other existing credit
     residence_since=(2.8-inf)
     savings_status=<100
     checking_status='no checking'
     age=(-inf-30.2]
     ==> class=good
     -- supp(0.005) conf:(0.833)

bc.  credit_history=critical/other existing credit
     residence_since=(2.8-inf)
     savings_status=<100
     checking_status='no checking'
     ==> class=good
     -- supp(0.036) conf:(0.973)
```

Rule `bdc` *states that young people in the context* $\mathbf{B}$ *of people with critical credit history, residence since 2.8 years at least, with savings at most for 100 units, and with no checkings, are assigned the good credit scoring with a confidence of 83.3%. Rule* `bc` *is obtained from* `bdc` *by discarding the item* `age=(-inf-30.2]` *in the premise, and has a confidence of 97.3%. Both rules are PND with respect to* $\mathcal{I}_d$, *so they are not checked for strong* α*-discrimination, for any* α.

Assume now to know that in the context $\mathbf{B}$ *above, the set of persons satisfying* `age=(-inf-30.2]` *is somewhat related to the set of persons satisfying the discriminatory item* `personal_status=female div/sep/mar`. *If the two sets were exactly the same, we could replace* `age=(-inf-30.2]` *in rule* `bdc` *with the discriminatory item above. This would lead us to a rule:*

```
abc. credit_history=critical/other existing credit
     residence_since=(2.8-inf)
     savings_status=<100
     checking_status='no checking'
     personal\_status=female div/sep/mar
     ==> class=good
```

[1] Actually, we could further restrict to assuming $\mathbf{A}$ to be an item in $\mathcal{I}_d$ such that its attribute is binary and its negated item is not in $\mathcal{I}_d$. However, since the set of extracted association rules will be used later on for other results, we prefer to report only one extraction algorithm.

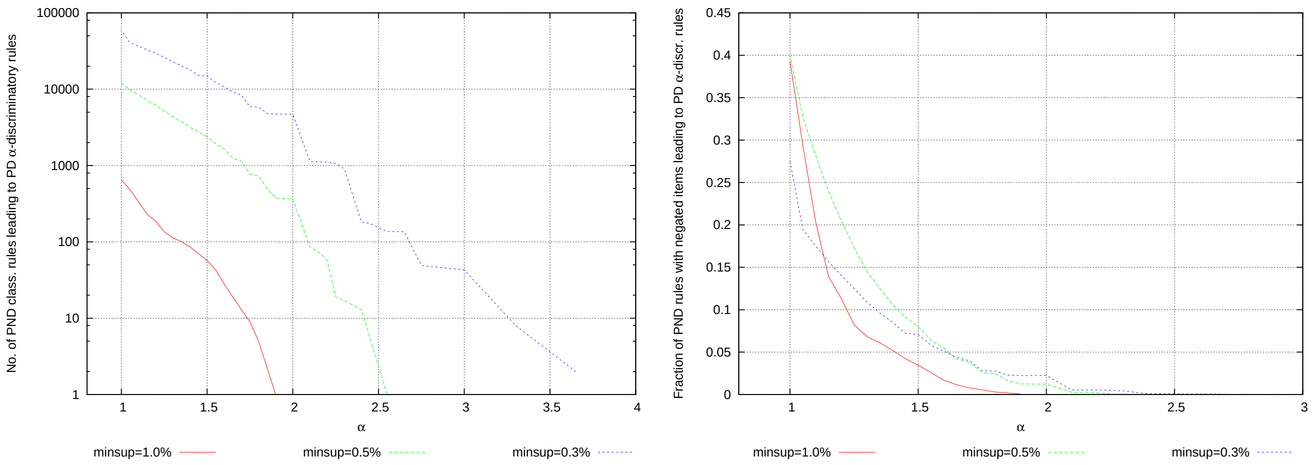

Figure 10: Distributions of PND classification rules containing negation of discriminatory items leading to PD rules which are strongly α-discriminatory: absolute count (left) and relative count (right).

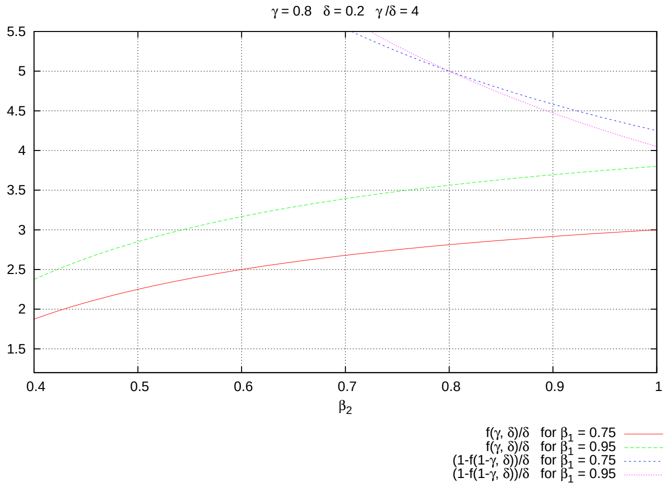

Figure 11: Example of lower and upper bounds for $elift(\gamma, \delta) = \gamma/\delta$ for $\gamma = 0.8$ and $\delta = 0.2$.

with $glift()$ value of $glift(0.833, 0.973) = 6.19$, which is considerably high. In case the two sets coincide only to some extent, we could however obtain some lower bound for the $glift()$ value above. In particular, assume that 100% of young people in the context are women and not single, and that they represents 54.5% of all women not single in the context. More formally, assume to know (e.g., by background knowledge such as public statistics, news, demographics) the confidence of the following two associations rules:

```
abd. credit_history=critical/other existing credit
     residence_since=(2.8-inf)
     savings_status=<100
     checking_status='no checking'
     personal_status=female div/sep/mar
     ==> age=(-inf-30.2]
     -- supp(0.006) conf:(0.545)

dba. credit_history=critical/other existing credit
     residence_since=(2.8-inf)
     savings_status=<100
     checking_status='no checking'
     age=(-inf-30.2]
     ==> personal_status=female div/sep/mar
     -- supp(0.006) conf:(1)
```

We show by means of Theorem 4.5 that a lower bound for the $glift()$ value of `abc` *can be calculated as:*

$$\frac{0.545(1-0.833)}{1-0.973} = 3.36.$$

As a consequence, we can state that the rule `abc` *is at least 3.3-discriminatory. It is worth noting that the actual $glift()$ value for the rule is slightly higher than 3.36, namely 3.37.*

4.2.1 Attack Model

The next result states a lower bound that an attacker could infer for (strong) α-discrimination of PD classification rules given information available in disclosed PND rules (γ, δ)

and information available from public or external sources (β_1, β_2).

THEOREM 4.5. *Let* $\mathbf{D},\mathbf{B} \rightarrow \mathbf{C}$ *be a PND classification rule, and let:*

$$\begin{aligned} \gamma &= conf(\mathbf{D},\mathbf{B} \rightarrow \mathbf{C}) \\ \delta &= conf(\mathbf{B} \rightarrow \mathbf{C}) > 0. \end{aligned}$$

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

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

Called:

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

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

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

we have:

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

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

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

Actually, the first two cases of the *glb*() function are mutually exclusive, since when $f(\gamma) \geq \delta$ by conclusion *(i)* we obtain $1 - f(1-\gamma) \geq f(\gamma) \geq \delta$ and then $f(1-\gamma) \leq 1-\delta$. Also, in the case $f(1-\gamma) > 1-\delta$, the *glb*() function is well-defined since the denominator $1-\delta$ cannot be zero. In fact, if $\delta = 1$, we have that also $\gamma = 1$. Therefore, the case expressions amounts at $\beta_1/\beta_2(\beta_2 - 1) > 0$ which is impossible, since $1 \geq \beta_1, \beta_2 \geq 0$. In order to understand more deeply the roles of β_1 and β_2, we have reported in Figure 11 two sample plots of lower and upper bounds for the $elift(\gamma,\delta)$ function, with $\gamma = 0.8$ and $\delta = 0.2$. The lower bound is $f(\gamma)/\delta$, while the upper bound is $(1 - f(1-\gamma))/\delta$. Both of them are immediate from conclusion *(i)* of Theorem 4.5. The two plots show lower and upper bound when $\beta_1 = 0.75$ and $\beta_1 = 0.95$, for β_2 in $[0.4, 1]$. As both β_1 and β_2 tend to 1, both the lower and upper bounds tend to γ/δ, i.e. to the extended lift. $f(\gamma)$ is monotonic w.r.t both β_1 and β_2: however, an increase of β_1 leads to a proportional improvement of the precision of lower and upper bounds, while an increase of β_2 is less than proportional (in the order of $-1/\beta_2$).

EXAMPLE 4.6. *Reconsider Example 4.4. We have* $\gamma = 0.833, \delta = 0.973, \beta_1 = 0.545$, *and* $\beta_2 = 1$. *The lower bound for the* $glift()$ *value of rule* abc *has been calculated as follows. Called:*

$$f(x) = \frac{0.545}{1}(1 + x - 1) = 0.545x$$

we have $f(1-0.833) = 0.091 > 1 - 0.973 = 0.027$, *and:*

$$glb(0.833, 0.973) = f(1-0.833)/(1-0.973) = 3.36.$$

Assume now that the value of $conf(\mathbf{A},\mathbf{B} \rightarrow \mathbf{D})$ *is known with an approximation of 5%, i.e.* $\beta_1 = 0.518$ *and* $\beta_2 = 1$. *We have* $f(x) = 0.518x$, *and since* $f(1-0.833) = 0.087 > 1 - 0.973 = 0.027$, *we obtain* $glb(0.833, 0.973) = 3.20$, *i.e. the inferred lower bound is proportionally (5%) lower. Finally, assume that* $conf(\mathbf{D},\mathbf{B} \rightarrow \mathbf{A})$ *is known with an approximation of 5%, i.e.* $\beta_1 = 0.545$ *and* $\beta_2 = 0.95$. *We have* $f(x) = 0.545/0.95(0.95 + x - 1) = 0.574(x - 0.05)$. *Again* $f(1-0.833) = 0.067 > 1 - 0.973 = 0.027$ *implies* $glb(0.833, 0.973) = 2.49$, *which is more than proportionally lower than 3.36.*

It is worth noting that the results of Theorem 4.5 are *sufficient* conditions for checking (strong) α-discrimination. If the inferred lower bounds are not as high as α, while the PD classification rule is actually α-discriminatory, we cannot conclude that an attacker has no other mean to infer α-discrimination of the rule.

4.2.2 Checking Attack Model

Consider now the problem of checking which PND classification rules satisfy the sufficient conditions of Theorem 4.5. Basically, for each candidate rule $\mathbf{X} \rightarrow \mathbf{C}$, with $\mathbf{X}$ PND itemset, we have to enumerate all sub-itemsets $\mathbf{D},\mathbf{B} \subseteq \mathbf{X}$ (which are $2^{|\mathbf{X}|}$) such that $\mathbf{X}$ can be written as $\mathbf{D},\mathbf{B}$. What we will be looking for to speed up the enumeration and checking process is some necessary conditions on the inequalities to be checked that restrict the search space. Let us start considering necessary conditions for $elb(\gamma,\delta) \geq \alpha$. If $\alpha = 0$ the expression is always true, so 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:

(i) $\beta_2 > 1 - \gamma$, and

(ii) $\beta_1(\beta_2 + \gamma - 1) \geq \alpha\delta\beta_2$.

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

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

Therefore, *(iii)* is a necessary condition for $elb(\gamma,\delta) \geq \alpha$ as well. The selectivity of conditions *(i,iii)* lies in the fact that checking *(iii)* involves no lookup at the rule $\mathbf{A},\mathbf{B} \rightarrow \mathbf{D}$; and checking *(i)* involves no lookup at rules $\mathbf{B} \rightarrow \mathbf{C}$. Moreover, condition *(iii)* is monotonic w.r.t β_2, hence if we scan association rules $\mathbf{X} \rightarrow \mathbf{A}$ ordered by descending confidence, we can stop checking it as soon as it is false. Finally, we observe that similar necessary conditions can be derived for $glb(\gamma,\delta) \geq \alpha$. The generate&test algorithm that incorporates the necessary conditions is shown in Figure 12.

The algorithm scans PND classification rules $\mathbf{X} \rightarrow \mathbf{C}$ for each $\mathcal{PND}_{group}$. We first lookup association rules $\mathbf{X} \rightarrow \mathbf{A}$ from $\mathcal{AR}_{group}$ such that the necessary condition *(ii)* holds. We scan such rules by descending confidence. If there is at least one of such rules, the candidate contexts $\mathbf{B} \subseteq \mathbf{X}$ are generated such that they satisfy condition *(iii)*. For each such rule, the candidate contexts are tested again for condition *(iii)*. Due to monotonicity of *(iii)* w.r.t. β_2, if the condition does not hold, the context are removed from the set of candidate. Otherwise, $glb(\gamma,\delta)$ (or $elb(\gamma,\delta)$) must be calculated. To this purpose, we still need to retrieve

```
        CheckAlphaPNDCR($\alpha$)
          ForEach $group$ s.t. $\mathcal{PND}_{group} \neq \emptyset$
            ForEach $\mathbf{X} \rightarrow \mathbf{C} \in \mathcal{PND}_{group}$
              $\gamma = conf(\mathbf{X} \rightarrow \mathbf{C})$
              $generateContexts = true$
(o)           ForEach $\mathbf{X} \rightarrow \mathbf{A} \in \mathcal{AR}_{group}$ order by $conf(\mathbf{X} \rightarrow \mathbf{A})$ descending
                $\beta_2 = conf(\mathbf{X} \rightarrow \mathbf{A})$
                $s = supp(\mathbf{X} \rightarrow \mathbf{A})$
(i)             If $\beta_2 > 1 - \gamma$ or $\beta_2 > \gamma$
                  If $generateContexts$
                    $generateContexts = false$
                    $\mathcal{V} = \emptyset$
                    ForEach $\mathbf{B} \subseteq \mathbf{X}$
                      $\delta = conf(\mathbf{B} \rightarrow \mathbf{C})$   // found in $\mathcal{PND}_g$ with $g = |\mathbf{B}| \leq group$
(iii)                 If $\beta_2(1 - \alpha\delta) \geq 1 - \gamma$ or $\beta_2(1 - \alpha(1 - \delta)) \geq \gamma$
                        $\mathcal{V} = \mathcal{V} \cup \{(\mathbf{B}, \delta)\}$
                      EndIf
                    EndForEach
                  EndIf
                  ForEach $(\mathbf{B}, \delta) \in \mathcal{V}$
(iii)               If $\beta_2(1 - \alpha\delta) \geq 1 - \gamma$ or $\beta_2(1 - \alpha(1 - \delta)) \geq \gamma$
                      $\beta_1 = s/supp(\mathbf{B} \rightarrow \mathbf{A})$      // found in $\mathcal{AR}_g$ with $g = |\mathbf{B}| \leq group$
                      If $glb(\gamma, \delta) \geq \alpha$
                        output $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$
                      EndIf
                    Else
                      $\mathcal{V} = \mathcal{V} \setminus \{(\mathbf{B}, \delta)\}$
                    EndIf
                  EndForEach
                EndIf
              EndForEach
            EndForEach
          EndForEach
```

Figure 12: Algorithm for checking strong α-discrimination inferrable from PND classification rules.

$\beta_1 = conf(\mathbf{A}, \mathbf{B} \rightarrow \mathbf{D})$. However, the rule $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{D}$ is not of the form stored in the $\mathcal{AR}_{group}$ sets. By noting that:

$$\beta_1 = \frac{supp(\mathbf{A}, \mathbf{B}, \mathbf{D})}{supp(\mathbf{A}, \mathbf{B})} = \frac{supp(\mathbf{X} \rightarrow \mathbf{A})}{supp(\mathbf{B} \rightarrow \mathbf{A})}$$

we can calculate β_1 using rules $\mathbf{X} \rightarrow \mathbf{A}$, already looked up, and $\mathbf{B} \rightarrow \mathbf{A}$, which is in the form stored in $\mathcal{AR}_g$, with $g = |\mathbf{B}|$. We report below the execution times (on a PC with Xeon 2.4Ghz and 2Gb main memory) of the checking algorithm running on rules for the German credit dataset with minimum support of 1% and without/with the necessary condition checkings.

	necessary cond. checks		ratio
	no	yes	
$\alpha = 2.0$	10m21s	3m12s	31.0%
$\alpha = 1.8$	10m21s	3m15s	31.4%
$\alpha = 1.6$	10m21s	3m23s	32.7%
$\alpha = 1.4$	10m21s	3m49s	36.9%

Whilst there is a gain in the execution time, up to 69%, 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.

4.2.3 The German Credit Dataset

Using the notation of Theorem 4.5, we say that a PD classification rule $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$ is *inferred* by the PND classification rule $\mathbf{D}, \mathbf{B} \rightarrow \mathbf{C}$, or, conversely, that $\mathbf{D}, \mathbf{B} \rightarrow \mathbf{C}$ *leads to* $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$. Figure 13 shows the absolute and relative distributions of PND classification rules that lead to strongly α-discriminatory PD rules for the German credit dataset. The right hand side graph highlights that those PND rules are a small percentage of total PND classification rules. The left hand side graph, however, warns us that the lower bounds inferrable for $glift()$ can reach considerably high values.

By observing that a PD rule $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$ can be inferred by more than one PND rule (e.g. from $\mathbf{D_1}, \mathbf{B} \rightarrow \mathbf{C}$ and $\mathbf{D_2}, \mathbf{B} \rightarrow \mathbf{C}$), it is also interesting to study the distribution of strongly discriminatory PD rules inferred by PD rules. The absolute and relative (w.r.t. the total number of strongly α-discriminatory PD rules) distributions are shown in Figure 14. While the absolute count does not differ sensibly from the distribution of Figure 13, the relative count highlights that the proportion of strongly α-discriminatory rules that can be inferred through the attack model of Theorem 4.5 is not negligible even for relatively high values of α. As an example, for minimum support of 0.3%, more than the 10% of the PD classification rules that are 2.2-discriminatory can be inferred from PND rules.

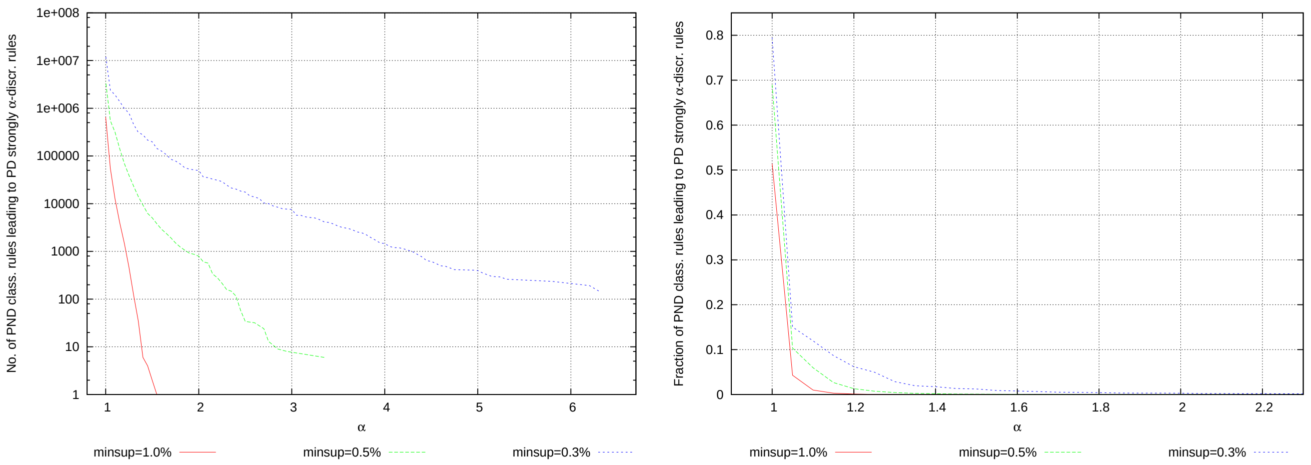

Figure 13: Distribution of PND classification rules leading to PD rules that are strongly α-discriminatory: absolute (left) and relative (right) count.

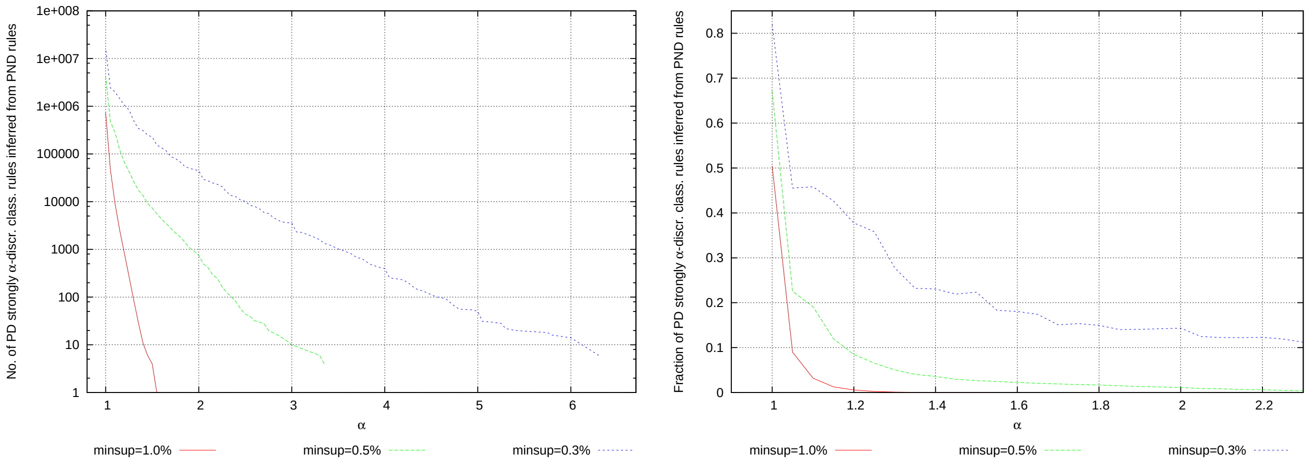

Figure 14: Distribution of strongly α-discriminatory PD classification rules inferred from PD rules: absolute (left) and relative (right) count.

5. DISCUSSION AND RELATED WORK

5.1 Other Classification Rule Sets

As a consequence of the generality of the rule-based approach, the theoretical framework introduced in this paper applies to a variety of classification models having classification rules at their basis, such as decision trees [31], rule-based classifiers [14], and association rule-based classifiers [27, 47]; or that can be translated into classification rules, such as support vector machines [29].

As far as the algorithms proposed in this paper are concerned, however, we observe that the procedure **CheckAlphaPDCR()** of Figure 1 for checking (strong) α-protection works for any base-closed set of classification rules.

Definition 5.1. *A set $\mathcal{A}$ of classification rules is* base-closed *if for every PD rule* $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$ *in $\mathcal{A}$ with* $\mathbf{A}$ *PD itemset and* $\mathbf{B}$ *PND itemset, the rule* $\mathbf{B} \rightarrow \mathbf{C}$ *is in $\mathcal{A}$.*

The set of classification rules extracted from frequent itemsets using the **ExtractCR()** procedure of Figure 1 is base-closed. Other base-closed sets are the sets of classification rules with a specified minimum coverage, where the coverage [40] of a classification rule $\mathbf{X} \rightarrow \mathbf{C}$ is defined as $supp(\mathbf{X})$. Coverage and support have similar uses: they allow to specify a lower bound on the significativeness of a rule.

The attack checking procedures **CheckAlphaPNDNegated()** of Figure 9 and **CheckAlphaPNDCR()** of Figure 12 require a slightly stronger property.

Definition 5.2. *A set $\mathcal{A}$ of classification rules is* downward base-closed *if for every PD rule* $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$ *in $\mathcal{A}$ with* $\mathbf{A}$ *PD itemset and* $\mathbf{B}$ *PND itemset, for every* $\mathbf{B}' \subseteq \mathbf{B}$ *the rule* $\mathbf{B}' \rightarrow \mathbf{C}$ *is in $\mathcal{A}$.*

The sets of classification rules having a specified minimum support or a specified minimum coverage are downward base-closed.

5.2 Related Work

To the best of our knowledge, this paper is the first to address the discrimination problem from the point of view of knowledge discovery from databases. Nevertheless, discrimination has been recognized as an issue in the tutorial [12, 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, i.e. of drawing lines around high risk minority neighborhoods.

Technically, we measured discrimination through generalizations and variants of *lift*, a measure of the statistical significance of a rule [39]. We extended lift to cope with *contexts*, specified as non-discriminatory itemsets: how much does a potentially discriminatory condition $\mathbf{A}$ increase/decrease the precision when added to the non-discriminatory antecedent of a classification rule $\mathbf{B} \rightarrow \mathbf{C}$? In this sense, there is a relation with the work of [33], 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 rule with respect to the base classification rule. In addition to $\Leftrightarrow$, the 4ft-Miner system [33, 34] allows 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 $1 + p$, i.e. 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 items. Also, the system adopts a rule extraction algorithm that is general enough to extract rules for 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 w.r.t. several operators. On the other hand, the procedure in Figure 1 exploits the efficiency of the state-of-the art algorithms for the extraction of frequent patterns.

Finally, we mention that the issue of indirect discrimination through attack models resembles a privacy-preserving problem [26, 28], where simply hiding a subset of rules – e.g., those with very low support – does not necessarily guarantee privacy protection from an attacker. The privacy-preserving literature contains several approaches to tackle this problem, that are all confronted with the trade-off between providing accurate models (rules) and preserving the privacy of individuals. In our specific framework, the problem is not privacy protection, but providing instead that decisions are taken without any bias given by discriminatory items. Therefore, it remains an open problem whether some of the existing approaches for privacy-preserving can be effective in our context as well, including data sanitization by distorting [18] or by blocking the data sets [42], or by hierarchy-based generalization approaches [38, 45].

6. CONCLUSION

Civil rights laws prohibit discrimination in a number of settings, including credit/insurance scoring, lending, personnel selection and wage, education and many others. The influence of discriminative behaviors has been the subject of studies in economics and social sciences. In this paper, we have shown that discrimination may be hidden in knowledge discovery models extracted from databases, and we have considered classification rule models.

Our main contribution is in the introduction of the notion of (strong) α-protection, a formal measure of the discriminatory power of a classification rule, which proves to be a powerful tool for reasoning about discrimination. On this basis, we can formalize and solve the problem of focussing on potentially discriminatory rules, for the purpose of explaining the practices that emerge from the historical training set, and to verify the absence of harmful rules before putting a model at work for decision-making or prediction tasks. We could appreciate from our experiments that the combination of the minimum support, of the rule size and of the α threshold helps in keeping the number of the resulting discriminatory rules small – enough to enable interactive exploration of the set of solutions. This is a promising situation, as the most natural setting for our approach is that of an interactive analytical tool, supporting the browsing of the space of discriminatory rules.

In addition, we have considered *indirect discrimination* by dealing with two attack models that allow the discovery of discriminatory classification rules starting from rules that do not contain discriminatory items at all. The first attack model exploits information on discriminatory items conveyed by their negated item. The second one consists of exploiting known relations from background knowledge. This possibility is relevant in tackling subtle issues, such as redlining in credit approval: a rule stating that the residents in a specific neighborhood asking for a car loan are given bad credit is not a discriminatory rule per se, but it entails a discriminatory rule if the background knowledge tells us that being a resident in that neighborhood is strongly correlated to with some discriminatory condition, such as being a member of an ethnic minority. We have proposed an approach, based on Theorem 4.2 and Theorem 4.5, that allows to identify such situations.

Clearly, many issues in discrimination-aware data mining remain open for future investigation, including a more comprehensive assessment over realistic datasets. We mention here the problem of automatic enforcement of α-protection, i.e. how to transform the training dataset in such a way that the discriminatory rules are removed from the output of the mining task, by means of controlled distortion of the source data.

7. REFERENCES

[1] Sex Discrimination Act 1975. U.K. Legislation, http://www.statutelaw.gov.uk.

[2] Race Relation Act 1976. U.K. Legislation, http://www.statutelaw.gov.uk.

[3] Equal Opportunity Act 1995. Victoria State Legislation, http://www.austlii.edu.au.

[4] Equal Credit Opportunity Act. U.S. Federal Legislation, http://www.usdoj.gov.

[5] Equal Pay Act. U.S. Federal Legislation, http://www.usdoj.gov.

[6] Fair Housing Act. U.S. Federal Legislation, http://www.usdoj.gov.

[7] Pregnancy Discrimination Act. U.S. Federal Legislation, http://www.usdoj.gov.

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

[9] B. Baesens, T. Van Gestel, S. Viaene, M. Stepanova, J. Suykens, and J. Vanthienen. Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society*, 54(6):627–635, 2003.

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

[11] C.-F. Chien and L.F. Chen. Data mining to improve personnel selection and enhance human capital: A case study in high-technology industry. *Expert Systems with Applications*, 34(1):280–290, 2008.

[12] C. Clifton. Privacy preserving data mining: How do we mine data when we aren't allowed to see it? In *Proc. of the 9th Int.'l Conf. on Knowledge Discovery and Data Mining (KDD 2003), Tutorial*, 2003. http://www.cs.purdue.edu/-homes/clifton/DistDM/Clifton_PPDM.ppt.

[13] Code of conduct and professional practice applying to information systems managed by private entities with regard to consumer credit, reliability, and timeliness of payments. Italian authority for personal data privacy, 2004. http://www.garanteprivacy.it.

[14] W. W. Cohen. Fast effective rule induction. In *Proc. of the 12th Int.'l Conference on Machine Learning (ICML 1995)*, pages 115–123. Morgan Kaufmann, 1995.

[15] Intentional Employment Discrimination. U.S. Federal Legislation, http://www.usdoj.gov.

[16] D. J. Hand. Modelling consumer credit risk. *IMA Journal of Management mathematics*, 12:139–155, 2001.

[17] D. J. Hand and W. E. Henley. Statistical classification methods in consumer credit scoring: a review. *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, 160:523–541, 1997.

[18] A. A. Hintoglu, A. Inan, Y. Saygin, and M. Keskinöz. Suppressing data sets to prevent discovery of association rules. In *Proc. of the 5th IEEE Int.'l Conference on Data Mining (ICDM 2005)*, pages 645–648. IEEE Computer Society, 2005.

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

[20] R. Hunter. *Indirect Discrimination in the Workplace.* The Federation Press, 1992.

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

[22] R. Knopff. On proving discrimination: Statistical methods and unfolding policy logics. *Canadian Public Policy*, 12:573–583, 1986.

[23] D.E. Knuth. *Fundamental Algorithms.* Addison-Wesley, 1997.

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

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

[26] N. Li, T. Li, and S. Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *Proc. of the 23nd Int.'l Conference on Data Engineering (ICDE 2007)*, pages 106–115. IEEE Computer Society, 2007.

[27] B. Liu, W. Hsu, and Y. Ma. Integrating classification and association rule mining. In *Proc. of the 4th Int.'l Conference on Knowledge Discovery and Data Mining (KDD 1998)*, pages 80–86, 1998.

[28] K. Liu. Privacy preserving data mining bibliography, 2006. http://www.csee.umbc.edu/∼kunliu1/research/privacy-_review.html.

[29] D. Martens, B. Baesens, T. Van Gestel, and J. Vanthienen. Comprehensible credit scoring models using rule extraction from support vector machines. *European Journal of Operational Research*, 183(3):1466–1476, 2007.

[30] D.J. Newman, S. Hettich, C.L. Blake, and C.J. Merz. UCI repository of machine learning databases, 1998. http://www.ics.uci.edu/∼mlearn/MLRepository.html.

[31] J. R. Quinlan. *C4.5: Programs for Machine Learning.* Morgan Kaufmann, San Mateo, CA, 1993.

[32] J. Rauch. Logic of association rules. *Appl. Intell.*, 22(1):9–28, 2005.

[33] J. Rauch and M. Simunek. Mining for association rules by 4ft-Miner. In *Proc. of the 14th International Conference on Applications of Prolog (INAP 2001)*, pages 285–295, 2001.

[34] J. Rauch and M. Simunek. 4-ft Miner Procedure, Visited on September 2007. http://lispminer.vse.cz.

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

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

[37] R. Srikant and R. Agrawal. Mining generalized association rules. In *VLDB'95, Proceedings of 21th International Conference on Very Large Data Bases*, pages 407–419. Morgan Kaufmann, 1995.

[38] L. Sweeney. Achieving k-anonymity privacy protection using generalization and suppression. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, 10(5):571–588, 2002.

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

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

[41] L. C. Thomas. A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. *International Journal of Forecasting*, 16:149–172, 2000.

[42] V. S. Verykios, A. K. Elmagarmid, E. Bertino, Y. Saygin, and E. Dasseni. Association rule hiding. *IEEE Trans. Knowl. Data Eng.*, 16(4):434–447, 2004.

[43] S. Viaene, R. A. Derrig, B. Baesens, and G. Dedene. A comparison of state-of-the-art classification techniques for expert automobile insurance claim fraud detection. *Journal of Risk & Insurance*, 69(3):373–421, 2001.

[44] M. Vojtek and E. Kočenda. Credit scoring methods. *Journal of Economics and Finance*, 56:152–167, 2006.

[45] K. Wang, B. C. M. Fung, and P. S. Yu. Template-based privacy preservation in classification problems. In *Proc. of the 5th IEEE International Conference on Data Mining (ICDM 2005)*, pages 466–473. IEEE Computer Society, 2005.

[46] X. Wu, C. Zhang, and S. Zhang. Efficient mining of both positive and negative association rules. *ACM Trans. Inf. Syst.*, 22(3):381–405, 2004.

[47] X. Yin and J. Han. CPAR: Classification based on Predictive Association Rules. In *Proc. of the 3rd SIAM International Conference on Data Mining*, 2003.

APPENDIX

A. PROOFS

Proofs are provided in a general setting, which extends the standard definition of [8] beyond itemsets. We refer the reader to [32] for a general logic calculi of association rules.

A.1 Association and Classification Rules

A pattern expression $\mathbf{X}$ is a boolean expression over items. We allow conjunction ($\wedge$) and disjunction ($\vee$) operators, and the constants `true` and `false`. For a transaction T, we say that T verifies $\mathbf{X}$, and write $T \models \mathbf{X}$, iff:

- $\mathbf{X}$ is $a = v$ and $a = v$ belongs to T;
- $\mathbf{X}$ is `true`;
- $\mathbf{X}$ is $\mathbf{X}_1 \wedge \mathbf{X}_2$ and both $T \models \mathbf{X}_1$ and $T \models \mathbf{X}_2$;
- $\mathbf{X}$ is $\mathbf{X}_1 \vee \mathbf{X}_2$ and $T \models \mathbf{X}_1$ or $T \models \mathbf{X}_2$.

With this semantics in mind, an itemset $\{i_1, \ldots, i_n\}$ can be interpreted as the pattern expression $i_1 \wedge \ldots, i_n$ (which reduces to `true` when $n = 0$). Moreover, since the domains of attributes are finite, negated items such as $\neg(a = v)$ (also written, $a \neq v$) can be introduced as a shorthand for $a = v_1 \vee \ldots \vee a = v_n$, where $\{v_1, \ldots, v_n\} = \mathit{dom}(a) \setminus \{v\}$. This is consistent with Definition 3.9. Negation can be extended to expressions by the De Morgan's laws:

- $\neg(\mathbf{X} \vee \mathbf{Y})$ is $\neg\mathbf{X} \wedge \neg\mathbf{Y}$,
- $\neg(\mathbf{X} \wedge \mathbf{Y})$ is $\neg\mathbf{X} \vee \neg\mathbf{Y}$,
- $\neg$`true` is `false` and $\neg$`false` is `true`.

For any transaction T, we have that $T \models \neg\mathbf{X}$ iff $T \models \mathbf{X}$ does not hold. A pattern expression $\mathbf{X}$ is *valid* if $T \models \mathbf{X}$ for every transaction T. The support of a pattern expression $\mathbf{X}$ w.r.t. a non-empty transaction database $\mathcal{D}$ is the ratio of transactions in $\mathcal{D}$ verifying $\mathbf{X}$:

$$\mathit{supp}_{\mathcal{D}}(\mathbf{X}) = |\{\, T \in \mathcal{D} \mid T \models \mathbf{X} \,\}|/|\mathcal{D}|$$

where $|\,|$ is the cardinality operator. An association rule is an expression $\mathbf{X} \rightarrow \mathbf{Y}$, where $\mathbf{X}$ and $\mathbf{Y}$ are pattern expressions. $\mathbf{X}$ is called the *premise* and $\mathbf{Y}$ is called the *consequence* of the association rule. We say that it is a *classification rule* if $\mathbf{Y}$ is a class item and no class item appears in $\mathbf{X}$. We say that $\mathbf{X} \rightarrow \mathbf{Y}$ is a *positive association rule* if $\mathbf{X}$ and $\mathbf{Y}$ are conjunctions of items with no item appearing in both. In other words, association and classification rules for itemsets are a special case of association and classification rules for pattern expressions.

The support of $\mathbf{X} \rightarrow \mathbf{Y}$ w.r.t. $\mathcal{D}$ is defined as:

$$\mathit{supp}_{\mathcal{D}}(\mathbf{X} \rightarrow \mathbf{Y}) = \mathit{supp}_{\mathcal{D}}(\mathbf{X} \wedge \mathbf{Y}).$$

The confidence of $\mathbf{X} \rightarrow \mathbf{Y}$, defined when $\mathit{supp}_{\mathcal{D}}(\mathbf{X}) > 0$, is:

$$\mathit{conf}_{\mathcal{D}}(\mathbf{X} \rightarrow \mathbf{Y}) = \mathit{supp}_{\mathcal{D}}(\mathbf{X} \wedge \mathbf{Y})/\mathit{supp}_{\mathcal{D}}(\mathbf{X}).$$

Support and confidence range over $[0, 1]$. We omit the subscripts in $\mathit{supp}_{\mathcal{D}}()$ and $\mathit{conf}_{\mathcal{D}}()$ when clear from the context. The formulation of association rules in terms of pattern expressions (instead of itemsets, i.e. conjunctions of items) allows us to model in a unified framework several extensions of standard association rules, including negative association rules [46] and hierarchies [37].

EXAMPLE A.1. (MODELLING HIERARCHIES) *Consider a hierarchy on attribute* `age` *with a level including values* `young`, `adult` *and* `elder`*; and a second level mapping* `young` *in the domain values* `0` ... `18`, `adult` *in the domain values* `19` ... `60`, *and* `elder` *into* `61` ... `99`. *While* `age = adult` *is not an item (since* `adult` *is not in the domain of* `age`*), it can be considered as a shorthand for the expression* `age = 19` $\vee \ldots \vee$ `age = 60`.

EXAMPLE A.2. (NEGATED ITEMS) *Consider the attribute* `gender` *and assume that its domain is binary, i.e.:*

$$\mathit{dom}(\texttt{gender}) = \{\texttt{male}, \texttt{female}\}.$$

The expression $\neg$ `gender = female` *is a syntactic abbreviation for* `gender = male`. *Assume now that:*

$$\mathit{dom}(\texttt{gender}) = \{\texttt{male}, \texttt{female}, \texttt{null}\}.$$

This assumption is realistic when transactions admit unknown/unspecified values, or for variable-length transactions, i.e. transactions that include at-most *one a-item for every attribute a. In this case,* $\neg$ `gender = female` *is a shorthand for* `gender = male` $\vee$ `gender = null`.

We start by stating a general relation which comes from the third-excluded principle of boolean logic.

LEMMA A.3. *For* $\mathbf{X}, \mathbf{Y}$ *pattern expressions, we have:*

(i) $\mathit{supp}(\mathbf{X}) = \mathit{supp}(\mathbf{X} \wedge \mathbf{Y}) + \mathit{supp}(\mathbf{X} \wedge \neg\mathbf{Y})$

(ii) $\mathit{supp}(\mathbf{X}) \geq \mathit{supp}(\mathbf{X} \wedge \mathbf{Y})$.

PROOF. *(i)* follows directly from observing that, for a transaction T, it turns out that $T \models \mathbf{X}$ iff $T \models \mathbf{X} \wedge \mathbf{Y}$ or $T \models \mathbf{X} \wedge \neg\mathbf{Y}$. *(ii)* is an immediate consequence of *(i)*. □

Let us now relate confidence of an association rule to the confidence of the rule with negated consequence.

LEMMA A.4. *Let* $\mathbf{X} \rightarrow \mathbf{Y}$ *be an association rule. We have:*

$$\mathit{conf}(\mathbf{X} \rightarrow \mathbf{Y}) = 1 - \mathit{conf}(\mathbf{X} \rightarrow \neg\mathbf{Y})$$

PROOF. Let us calculate:

$$\begin{array}{cl} & \mathit{conf}(\mathbf{X} \rightarrow \mathbf{Y}) \\ = & \mathit{supp}(\mathbf{X} \wedge \mathbf{Y})/\mathit{supp}(\mathbf{X}) \\ = & \{\text{ Lemma A.3 } \mathit{(i)} \text{ }\} \\ & (\mathit{supp}(\mathbf{X}) - \mathit{supp}(\mathbf{X} \wedge \neg\mathbf{Y}))/\mathit{supp}(\mathbf{X}) \\ = & 1 - \mathit{supp}(\mathbf{X} \wedge \neg\mathbf{Y})/\mathit{supp}(\mathbf{X}) \\ = & 1 - \mathit{conf}(\mathbf{X} \rightarrow \neg\mathbf{Y}). \end{array}$$

□

We conclude by reformulating the well-known principle of Inclusion-Exclusion [23] in the context of pattern expressions.

LEMMA A.5 (INCLUSION-EXCLUSION PRINCIPLE). *Let* $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, *and* $\mathbf{D}$ *be pattern expressions. Then:*

$$\begin{array}{rcl} \mathit{supp}(\mathbf{A}) & \geq & \mathit{supp}(\mathbf{A} \wedge \mathbf{B}) + \mathit{supp}(\mathbf{A} \wedge \mathbf{C}) \\ & & -\mathit{supp}(\mathbf{A} \wedge \mathbf{B} \wedge \mathbf{C}). \end{array}$$

PROOF. We have:

$$\begin{array}{ll} & supp(\mathbf{A} \wedge \mathbf{B}) \\ = & \{ \text{ Lemma A.3 } \textit{(i)} \} \\ & supp(\mathbf{A} \wedge \mathbf{B} \wedge \mathbf{C}) + supp(\mathbf{A} \wedge \mathbf{B} \wedge \neg\mathbf{C}) \\ \leq & \{ \text{ Lemma A.3 } \textit{(ii)} \} \\ & supp(\mathbf{A} \wedge \mathbf{B} \wedge \mathbf{C}) + supp(\mathbf{A} \wedge \neg\mathbf{C}) \\ = & \{ \text{ Lemma A.3 } \textit{(i)} \} \\ & supp(\mathbf{A} \wedge \mathbf{B} \wedge \mathbf{C}) + supp(\mathbf{A}) - supp(\mathbf{A} \wedge \mathbf{C}). \end{array}$$

□

A.2 Extended Lift

The following lemma states the range of variability of extended lift.

LEMMA A.6. *Let* $\mathbf{A} \wedge \mathbf{B} \rightarrow \mathbf{C}$ *be an association rule such that:*

$$\begin{array}{rcl} s & = & supp(\mathbf{A} \wedge \mathbf{B} \rightarrow \mathbf{C}) \geq ms > 0 \\ \gamma & = & conf(\mathbf{A} \wedge \mathbf{B} \rightarrow \mathbf{C}) \\ \delta & = & conf(\mathbf{B} \rightarrow \mathbf{C}) > 0. \end{array}$$

We have that:

(i) γ/δ *belongs to the range* $[0, 1/ms]$*,*

(ii) if $\delta \geq mc$ *then* γ/δ *belongs to* $[0, 1/mc]$*.*

PROOF. The lower-bounds of zero are immediate due to the fact that $\gamma, \delta \geq 0$. As for the upper-bounds, for *(i)* we calculate:

$$\begin{array}{rcl} \gamma/\delta & = & \dfrac{supp(\mathbf{A} \wedge \mathbf{B} \wedge \mathbf{C})/supp(\mathbf{A} \wedge \mathbf{B})}{supp(\mathbf{B} \wedge \mathbf{C})/supp(\mathbf{B})} \\ & = & \dfrac{supp(\mathbf{A} \wedge \mathbf{B} \wedge \mathbf{C})}{supp(\mathbf{B} \wedge \mathbf{C})} \dfrac{supp(\mathbf{B})}{supp(\mathbf{A} \wedge \mathbf{B})} \\ & \leq & \{ \text{ Lemma A.3 } \textit{(ii)} \} \\ & & supp(\mathbf{B})/supp(\mathbf{A} \wedge \mathbf{B}) \\ & \leq & 1/supp(\mathbf{A} \wedge \mathbf{B}) \\ & \leq & \{ \text{ Lemma A.3 } \textit{(ii)} \} \\ & \leq & 1/s \leq 1/ms. \end{array}$$

Concerning *(ii)*, we have: $\gamma/\delta \leq 1/\delta \leq 1/mc$. □

Next result relates support and confidence of conjuncts in pattern expressions and association rules to the underlying database of transactions.

LEMMA A.7. *Let* $\mathbf{X}, \mathbf{Y}$ *and* $\mathbf{B}$ *be pattern expressions. Then:*

(i) $supp_{\mathcal{D}}(\mathbf{X} \wedge \mathbf{B}) = supp_{\mathcal{B}}(\mathbf{X})\ supp_{\mathcal{D}}(\mathbf{B})$*,*

(ii) $conf_{\mathcal{D}}(\mathbf{X} \wedge \mathbf{B} \rightarrow \mathbf{Y}) = conf_{\mathcal{B}}(\mathbf{X} \rightarrow \mathbf{Y})$*.*

where $\mathcal{B} = \{T \in \mathcal{D} \mid T \models \mathbf{B}\}$*.*

PROOF. For *(i)*, we calculate:

$$\begin{array}{ll} & supp_{\mathcal{D}}(\mathbf{X} \wedge \mathbf{B}) \\ = & \dfrac{|\{\, T \in \mathcal{D} \mid T \models \mathbf{X} \wedge \mathbf{B} \,\}|}{|\mathcal{D}|} \\ = & \dfrac{|\{\, T \in \mathcal{B} \mid T \models \mathbf{X} \,\}||\mathcal{B}|}{|\mathcal{D}||\mathcal{B}|} \\ = & supp_{\mathcal{B}}(\mathbf{X})\ supp_{\mathcal{D}}(\mathbf{B}). \end{array}$$

For *(ii)*, we calculate:

$$\begin{array}{ll} & conf_{\mathcal{D}}(\mathbf{X} \wedge \mathbf{B} \rightarrow \mathbf{Y}) \\ = & \dfrac{supp_{\mathcal{D}}(\mathbf{X} \wedge \mathbf{B} \wedge \mathbf{Y})}{supp_{\mathcal{D}}(\mathbf{X} \wedge \mathbf{B})} \\ = & \{ \textit{ (i) } \} \\ = & \dfrac{supp_{\mathcal{B}}(\mathbf{X} \wedge \mathbf{Y})supp_{\mathcal{D}}(\mathbf{B})}{supp_{\mathcal{B}}(\mathbf{X})supp_{\mathcal{D}}(\mathbf{B})} \\ = & conf_{\mathcal{B}}(\mathbf{X} \rightarrow \mathbf{Y}). \end{array}$$

□

We are now in the position to relate extended lift to standard lift [39], which we define for pattern expressions as follows:

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

LEMMA A.8. *Let* $\mathbf{A} \wedge \mathbf{B} \rightarrow \mathbf{C}$ *be an association rule such that* $conf_{\mathcal{D}}(\mathbf{B} \rightarrow \mathbf{C}) > 0$*. We have:*

$$\frac{conf_{\mathcal{D}}(\mathbf{A} \wedge \mathbf{B} \rightarrow \mathbf{C})}{conf_{\mathcal{D}}(\mathbf{B} \rightarrow \mathbf{C})} \quad = \quad lift_{\mathcal{B}}(\mathbf{A} \rightarrow \mathbf{C})$$

where $\mathcal{B} = \{T \in \mathcal{D} \mid T \models \mathbf{B}\}$*.*

PROOF.

$$\begin{array}{ll} & \dfrac{conf_{\mathcal{D}}(\mathbf{A} \wedge \mathbf{B} \rightarrow \mathbf{C})}{conf_{\mathcal{D}}(\mathbf{B} \rightarrow \mathbf{C})} \\ = & \{ \text{ Lemma A.7 (i) } \} \\ & \dfrac{conf_{\mathcal{D}}(\mathbf{A} \wedge \mathbf{B} \rightarrow \mathbf{C})}{supp_{\mathcal{B}}(\mathbf{C})} \\ = & \{ \text{ Lemma A.7 (ii) } \} \\ = & \dfrac{conf_{\mathcal{B}}(\mathbf{A} \rightarrow \mathbf{C})}{supp_{\mathcal{B}}(\mathbf{C})} = lift_{\mathcal{B}}(\mathbf{A} \rightarrow \mathbf{C}). \end{array}$$

□

Lemma 2.2 is a special case of this result.

A.3 Strong Discrimination of PDCR

Let us show next (a generalization of) Lemma 3.10.

LEMMA A.9. *Let* $\mathbf{A} \wedge \mathbf{B} \rightarrow \mathbf{C}$ *be an association rule, and let:*

$$\begin{array}{rrl} & \gamma = & conf(\mathbf{A} \wedge \mathbf{B} \rightarrow \mathbf{C}) \\ 1 > & \delta = & conf(\mathbf{B} \rightarrow \mathbf{C}). \end{array}$$

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

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

PROOF. By Lemma A.4,

$$\mathit{conf}(\mathbf{B} \to \neg\mathbf{C}) = 1 - \mathit{conf}(\mathbf{B} \to \mathbf{C}) = 1 - \delta > 0,$$

since $\delta < 1$. Moreover:

$$\begin{aligned}
& \frac{\mathit{conf}(\mathbf{A} \wedge \mathbf{B} \to \neg\mathbf{C})}{\mathit{conf}(\mathbf{B} \to \neg\mathbf{C})} \\
= \ & \{ \text{ Lemma A.4 } \} \\
& \frac{1 - \mathit{conf}(\mathbf{A} \wedge \mathbf{B} \to \neg\mathbf{C})}{1 - \mathit{conf}(\mathbf{B} \to \neg\mathbf{C})} \\
= \ & \frac{1-\gamma}{1-\delta}.
\end{aligned}$$

□

The following result shows that the $glift()$ function of Definition 3.11 ranges over $[1, 1/ms]$, where ms is the minimum support of a rule.

LEMMA A.10. *Let* $\mathbf{A} \wedge \mathbf{B} \to \mathbf{C}$ *be an association rule such that:*

$$\begin{aligned}
s &= \mathit{supp}(\mathbf{A} \wedge \mathbf{B} \to \mathbf{C}) \geq ms > 0 \\
\gamma &= \mathit{conf}(\mathbf{A} \wedge \mathbf{B} \to \mathbf{C}) \\
\delta &= \mathit{conf}(\mathbf{B} \to \mathbf{C}) > 0.
\end{aligned}$$

We have that: $glift(\gamma, \delta) \in [1, 1/ms]$, *and for* $1 > \delta$*:*

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

PROOF. First, we observe that $\delta = 1$ implies $\gamma = 1$. In fact, when $\mathit{supp}(\mathbf{B} \wedge \mathbf{C}) = \mathit{supp}(\mathbf{B})$, i.e. all transaction verifying $\mathbf{B}$ also verify $\mathbf{C}$, we have $\mathit{supp}(\mathbf{A} \wedge \mathbf{B} \wedge \mathbf{C}) = \mathit{supp}(\mathbf{A} \wedge \mathbf{B})$, i.e. $\gamma = 1$. As a consequence, when $\delta = 1$, we have $glift(\gamma, \delta) = \gamma/\delta = 1 \in [1, 1/ms]$. Consider now the remaining case $1 > \delta > 0$. Since the following property holds by elementary algebra:

$$\gamma/\delta \geq 1 \quad \text{iff} \quad (1-\gamma)/(1-\delta) \leq 1.$$

we obtain $glift(\gamma, \delta) \geq 1$ and:

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

Let us show now the upper bound. By Lemma A.6 *(i)*, $elift(\gamma, \delta) \leq 1/ms$. Moreover:

$$\begin{aligned}
elift(1-\gamma, 1-\delta) &= \{ \text{ Lemma A.9 } \} \\
&= \frac{\mathit{supp}(\mathbf{A} \wedge \mathbf{B} \wedge \mathbf{C})/\mathit{supp}(\mathbf{A} \wedge \mathbf{B})}{\mathit{supp}(\mathbf{B} \wedge \mathbf{C})/\mathit{supp}(\mathbf{B})} \\
&= \frac{\mathit{supp}(\mathbf{A} \wedge \mathbf{B} \wedge \mathbf{C})}{\mathit{supp}(\mathbf{B} \wedge \mathbf{C})} \frac{\mathit{supp}(\mathbf{B})}{\mathit{supp}(\mathbf{A} \wedge \mathbf{B})} \\
&\leq \{ \text{ Lemma A.3 } \textit{(ii)} \} \\
& \ \mathit{supp}(\mathbf{B})/\mathit{supp}(\mathbf{A} \wedge \mathbf{B}) \\
&\leq 1/\mathit{supp}(\mathbf{A} \wedge \mathbf{B}) \\
&= \{ \text{ Lemma A.7 (ii) } \} \\
&\leq 1/s \leq 1/ms.
\end{aligned}$$

Therefore, $glift(\gamma, \delta) \leq 1/ms$. □

A.4 Indirect Discrimination of PNDCR

The extended lift is a symmetric measure, in the sense that the roles of $\mathbf{A}$ and $\mathbf{C}$ are dual, as it is made clear by the following lemma.

LEMMA A.11. *Let* $\mathbf{A} \wedge \mathbf{B} \to \mathbf{C}$ *be an association rule with* $\mathit{supp}(\mathbf{B} \to \mathbf{C}) > 0$ *and* $\mathit{supp}(\mathbf{B} \to \mathbf{A}) > 0$. *Then:*

$$\frac{\mathit{conf}(\mathbf{A} \wedge \mathbf{B} \to \mathbf{C})}{\mathit{conf}(\mathbf{B} \to \mathbf{C})} = \frac{\mathit{conf}(\mathbf{B} \wedge \mathbf{C} \to \mathbf{A})}{\mathit{conf}(\mathbf{B} \to \mathbf{A})}$$

PROOF.

$$\begin{aligned}
& \frac{\mathit{conf}(\mathbf{A} \wedge \mathbf{B} \to \mathbf{C})}{\mathit{conf}(\mathbf{B} \to \mathbf{C})} \\
= \ & \frac{\mathit{supp}(\mathbf{A} \wedge \mathbf{B} \wedge \mathbf{C})/\mathit{supp}(\mathbf{A} \wedge \mathbf{B})}{\mathit{supp}(\mathbf{B} \wedge \mathbf{C})/\mathit{supp}(\mathbf{B})} \\
= \ & \frac{\mathit{supp}(\mathbf{A} \wedge \mathbf{B} \wedge \mathbf{C})/\mathit{supp}(\mathbf{B} \wedge \mathbf{C})}{\mathit{supp}(\mathbf{A} \wedge \mathbf{B})/\mathit{supp}(\mathbf{B})} \\
= \ & \frac{\mathit{conf}(\mathbf{B} \wedge \mathbf{C} \to \mathbf{A})}{\mathit{conf}(\mathbf{B} \to \mathbf{A})}.
\end{aligned}$$

□

Theorem 4.2 *(i)* is an instance of the following result, by considering $\mathbf{C}$ to be a class item, $\mathbf{A}$ a PD item over a binary predicate, and $\neg\mathbf{A}$ a PND item.

THEOREM A.12. *Let* $\mathbf{A} \wedge \mathbf{B} \to \mathbf{C}$ *be an association rule such that:*

$$\begin{aligned}
\gamma &= \mathit{conf}(\neg\mathbf{A} \wedge \mathbf{B} \to \mathbf{C}) \\
\delta &= \mathit{conf}(\mathbf{B} \to \mathbf{C}) > 0 \\
\beta &= \mathit{conf}(\mathbf{B} \to \mathbf{A}) > 0.
\end{aligned}$$

We have that:

$$\mathit{conf}(\mathbf{A} \wedge \mathbf{B} \to \mathbf{C}) = \frac{\delta}{\beta} + (1 - \frac{1}{\beta})\gamma.$$

PROOF. We calculate:

$$\begin{aligned}
& \frac{\delta}{\beta} + (1 - \frac{1}{\beta})\gamma \\
= \ & \{ \text{ Definition of } \beta \ \} \\
& \delta \frac{\mathit{supp}(\mathbf{B})}{\mathit{supp}(\mathbf{B} \wedge \mathbf{A})} + (\frac{\mathit{supp}(\mathbf{B} \wedge \mathbf{A}) - \mathit{supp}(\mathbf{B})}{\mathit{supp}(\mathbf{B} \wedge \mathbf{A})})\gamma \\
= \ & \{ \text{ Lemma A.3 } \textit{(i)} \ \} \\
& \delta \frac{\mathit{supp}(\mathbf{B})}{\mathit{supp}(\mathbf{B} \wedge \mathbf{A})} - \frac{\mathit{supp}(\mathbf{B} \wedge \neg\mathbf{A})}{\mathit{supp}(\mathbf{B} \wedge \mathbf{A})}\gamma \\
= \ & \{ \text{ Definition of } \delta, \gamma \ \} \\
& \frac{\mathit{supp}(\mathbf{B} \wedge \mathbf{C})}{\mathit{supp}(\mathbf{B} \wedge \mathbf{A})} - \frac{\mathit{supp}(\neg\mathbf{A} \wedge \mathbf{B} \wedge \mathbf{C})}{\mathit{supp}(\mathbf{B} \wedge \mathbf{A})} \\
= \ & \frac{\mathit{supp}(\mathbf{B} \wedge \mathbf{C}) - \mathit{supp}(\neg\mathbf{A} \wedge \mathbf{B} \wedge \mathbf{C})}{\mathit{supp}(\mathbf{B} \wedge \mathbf{A})} \\
= \ & \{ \text{ Lemma A.3 } \textit{(i)} \ \} \\
= \ & \frac{\mathit{supp}(\mathbf{A} \wedge \mathbf{B} \wedge \mathbf{C})}{\mathit{supp}(\mathbf{B} \wedge \mathbf{A})} \\
= \ & \mathit{conf}(\mathbf{A} \wedge \mathbf{B} \to \mathbf{C}).
\end{aligned}$$

□

Theorem 4.2 follows immediately.

Proof of Theorem 4.2.

PROOF. *(i)* is an instance of Theorem A.12. Conclusions *(ii,iii)* are immediate consequences of *(i)* and the definition of (strong) α-protection. □

Next result provides upper and lower bounds for confidence of an association rule $\mathbf{D} \rightarrow \mathbf{C}$ given the confidence of $\mathbf{A} \rightarrow \mathbf{C}$ and some (background) knowledge about the relations between expressions $\mathbf{D}$ and $\mathbf{A}$.

Lemma A.13. *Let $\mathbf{A}, \mathbf{D}, \mathbf{C}$ be pattern expressions, and:*

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

We have that:

$$1 - \frac{\beta_1}{\beta_2}(\beta_2 - \gamma) \geq conf(\mathbf{A} \rightarrow \mathbf{C}) \geq \frac{\beta_1}{\beta_2}(\beta_2 + \gamma - 1).$$

Proof. Let us call: $\overline{\beta}_1 = conf(\mathbf{A} \rightarrow \mathbf{D})$ and $\overline{\beta}_2 = conf(\mathbf{D} \rightarrow \mathbf{A})$. We have:

$$\begin{aligned} & conf(\mathbf{A} \rightarrow \mathbf{C}) = \frac{supp(\mathbf{A} \wedge \mathbf{C})}{supp(\mathbf{A})} \\ \geq\ & \frac{supp(\mathbf{A} \wedge \mathbf{C} \wedge \mathbf{D})}{supp(\mathbf{A})} \\ =\ & \{\ \overline{\beta}_1/\overline{\beta}_2 = supp(\mathbf{D})/supp(\mathbf{A})\ \} \\ =\ & \frac{\overline{\beta}_1}{\overline{\beta}_2} \frac{supp(\mathbf{A} \wedge \mathbf{C} \wedge \mathbf{D})}{supp(\mathbf{D})} \\ \geq\ & \{\ \text{Inclusion-Exclusion Lemma A.5}\ \} \\ & \frac{\overline{\beta}_1}{\overline{\beta}_2} \frac{(supp(\mathbf{D} \wedge \mathbf{A}) + supp(\mathbf{D} \wedge \mathbf{C}) - supp(\mathbf{D}))}{supp(\mathbf{D})} \\ =\ & \frac{\overline{\beta}_1}{\overline{\beta}_2}(\overline{\beta}_2 + \gamma - 1) \\ \geq\ & \{\ \overline{\beta}_1 \geq \beta_1,\ \overline{\beta}_2 \geq \beta_2,\ \gamma \leq 1\ \} \\ & \frac{\beta_1}{\beta_2}(\beta_2 + \gamma - 1). \end{aligned}$$

Moreover:

$$\begin{aligned} & conf(\mathbf{A} \rightarrow \mathbf{C}) = \frac{supp(\mathbf{A} \wedge \mathbf{C})}{supp(\mathbf{A})} \\ \leq\ & \{\ \text{Inclusion-Exclusion Lemma A.5}\ \} \\ =\ & \frac{supp(\mathbf{A}) + supp(\mathbf{A} \wedge \mathbf{D} \wedge \mathbf{C}) - supp(\mathbf{A} \wedge \mathbf{D})}{supp(\mathbf{A})} \\ =\ & 1 - \frac{supp(\mathbf{A} \wedge \mathbf{D}) - supp(\mathbf{A} \wedge \mathbf{D} \wedge \mathbf{C})}{supp(\mathbf{A})} \\ =\ & \{\ \overline{\beta}_1/\overline{\beta}_2 = supp(\mathbf{D})/supp(\mathbf{A})\ \} \\ =\ & 1 - \frac{\overline{\beta}_1}{\overline{\beta}_2} \frac{supp(\mathbf{A} \wedge \mathbf{D}) - supp(\mathbf{A} \wedge \mathbf{D} \wedge \mathbf{C})}{supp(\mathbf{D})} \\ =\ & \{\ \text{Lemma A.3 } \textit{(i)}\ \} \\ \leq\ & 1 - \frac{\overline{\beta}_1}{\overline{\beta}_2} \frac{supp(\mathbf{A} \wedge \mathbf{D}) - supp(\mathbf{D} \wedge \mathbf{C})}{supp(\mathbf{D})} \\ =\ & 1 - \frac{\overline{\beta}_1}{\overline{\beta}_2}(\overline{\beta}_2 - \gamma) \\ \leq\ & \{\ \overline{\beta}_1 \geq \beta_1,\ \overline{\beta}_2 \geq \beta_2,\ \gamma \leq 1\ \} \\ & 1 - \frac{\beta_1}{\beta_2}(\beta_2 - \gamma). \end{aligned}$$

□

Theorem 4.5 follows from Lemma A.13.

Proof of Theorem 4.5.

Proof. Let $\mathcal{B} = \{T \in \mathcal{D} \mid T \models \mathbf{B}\}$. By Lemma A.7 *(ii)*, we have:

$$\begin{aligned} \gamma &= conf_{\mathcal{B}}(\mathbf{D} \rightarrow \mathbf{C}) \\ conf_{\mathcal{B}}(\mathbf{A} \rightarrow \mathbf{D}) &\geq \beta_1 \\ conf_{\mathcal{B}}(\mathbf{D} \rightarrow \mathbf{A}) &\geq \beta_2 > 0. \end{aligned}$$

By Lemma A.13, we obtain:

$$1 - \frac{\beta_1}{\beta_2}(\beta_2 - \gamma) \geq conf_{\mathcal{B}}(\mathbf{A} \rightarrow \mathbf{C}) \geq \frac{\beta_1}{\beta_2}(\beta_2 + \gamma - 1).$$

which, by definition of f, can be rewritten as:

$$1 - f(1-\gamma) \geq conf_{\mathcal{B}}(\mathbf{A} \rightarrow \mathbf{C}) \geq f(\gamma)$$

and, again by Lemma A.7 *(ii)*, yields:

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

This shows conclusion *(i)*. Consider now conclusion *(ii)*. By dividing by $\delta = conf(\mathbf{B} \rightarrow \mathbf{C})$ both sides of the inequality $conf(\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}) \geq f(\gamma)$ from (1), we can conclude that $elb(\gamma, \delta)$ is a lower bound for the extended lift of $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$. Finally, consider conclusion *(iii)*. Let us call:

$$\gamma' = conf(\mathbf{A} \wedge \mathbf{B} \rightarrow \mathbf{C}).$$

If $f(\gamma) \geq \delta$, by (1) we have:

$$\gamma' \geq f(\gamma) \geq \delta$$

and then:

$$glift(\gamma', \delta) = \gamma'/\delta \geq f(\gamma)/\delta = glb(\gamma, \delta) \geq \alpha$$

which implies that $\mathbf{A} \wedge \mathbf{B} \rightarrow \mathbf{C}$ is not strongly α-protective. If $f(1-\gamma) > 1 - \delta$, again by (1) we have:

$$\gamma' \leq 1 - f(1-\gamma) < \delta$$

and then:

$$glift(\gamma', \delta) = (1-\gamma')/(1-\delta) \geq f(1-\gamma)/(1-\delta) = glb(\gamma, \delta) \geq \alpha$$

which implies that $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$ is not strongly α-protective. The last case of $glb()$ is trivial since $glift(\gamma', \delta)$ is always greater or equal than 1. □