Measuring Discrimination in Socially-Sensitive Decision Records

Dino Pedreschi Salvatore Ruggieri Franco Turini
Dipartimento di Informatica, Università di Pisa
{pedre,ruggieri,turini}@di.unipi.it

Abstract

Discrimination in social sense (e.g., against minorities and disadvantaged groups) is the subject of many laws worldwide, and it has been extensively studied in the social and economic sciences. We tackle the problem of determining, given a dataset of historical decision records, a precise measure of the degree of discrimination suffered by a given group (e.g., an etnic minority) in a given context (e.g., a geographic area) with respect to the decision (e.g. credit denial). In our approach, this problem is rephrased in a classification rule based setting, and a collection of quantitative measures of discrimination is introduced, on the basis of existing norms and regulations. The measures are defined as functions of the contingency table of a classification rule, and their statistical significance is assessed, relying on a large body of statistical inference methods for proportions. Based on this basic method, we are then able to address the more general problems of: (1) unveiling all discriminatory decision patterns hidden in the historical data, combining discrimination analysis with association rule mining, (2) unveiling discrimination in classifiers that learn over training data biased by discriminatory decisions, and (3) in the case of rule-based classifiers, sanitizing discriminatory rules by correcting their confidence. Our approach is validated on the German credit dataset and on the CPAR classifier.

1 Introduction

In social sense, discrimination refers to an action based on prejudice resulting in unfair treatment of people, where the distinction between people is operated on the basis of their membership to a category or minority, without regard to individual merit or circumstances. Examples of social discrimination include racial/ethnic, religious, gender, sexual orientation, disability, and age-related discrimination; a large body of international laws and regulations [4, 5, 20, 21] prohibit discrimination in socially-sensitive decision making tasks, including credit scoring/approval, house lending, and personnel selection. In order to prove (or disprove) a discrimination charge before a court, or to perform a social analysis of discrimination in a given context, it is clearly needed to rely on quantitative measures of the phenomenon under study: for this reason, discrimination has been the subject of a large body of research in legal, economic and social sciences, as well as the subject of empirical analysis in a large number of juridical cases [12].

In this paper, we propose a systematic framework for measuring discrimination, based on the analysis of the historical decision records stored out of a socially-sensitive decision task, e.g., credit approval. We generalize the approach of [16] and show how a comprehensive repertoire of discrimination measures, encompassing all the notions that we found in the juridical literature, can be defined in terms of the confidence (or probability) of the decision rules that describe the phenomenon under analysis, namely the potential discrimination of a given group within a certain context. Clearly, the confidence of such rules can be estimated with reference to the available historical decision records; on this basis, we address the crucial issue of the statistical significance of the proposed discrimination measures, trying to find an answer not only to: "what is an adequate estimation of the degree of discrimination?" but also "how confident are we on such an estimation, given the available data?"

We take a different approach with respect to our earlier work in [16]: we now investigate whether evidence of discrimination can be found in a given set of decisions, by measuring the degree of discrimination of a rule that formalizes an expert's hypothesis – e.g., a suspicious pattern that an anti-discrimination body is interested to verify. The next natural step is to repeat such a procedure for all the classification rules that emerge from the historical data, thus unveiling all the discriminatory patterns hidden in the data. As a further contribution, we investigate how the newly introduced discrimination measures and significance tests can be used to reason about classifier themselves: first, to assess whether or not a classifier built over the historical decision records is biased by the discrimination behavior hidden in the data; second, to sanitize rule-based classifiers by means of corrections to their potentially discriminatory rules. Given that discrimination can be

induced either directly, that is on the basis of sensitive data available to the classifier (such as gender or age), or, much more subtly, indirectly, that is by attribute values related to the sensitive ones (such as ZIP area and minority race), the task of sanitizing the rules of the classifier is quite complex. We propose and study an approach consisting of just modifying the confidence of the rules in the classifier.

Summarizing the contribution of this paper: (1) we define a family of formal measures of discrimination for classification rules (Sects. 3-4), including a notion of statistical significance (Sect. 5); (2) we combine the discrimination measures with association rule mining to unveil direct and indirect discrimination in datasets of decisions (Sect. 6) or in the output of classifiers (Sect. 7); (3) for rule-based classifiers, we propose a discrimination correction based on the measures (Sect. 8); and, finally, (4) we experiment the theoretical definitions and results on the publicly available German credit granting dataset [15] and on the CPAR rule-based classifier [22] (Sect. 9).

2 Preliminaries

We recall the notions of itemsets, association rules and classification rules from standard definitions [1]. Consider 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 the values of a. We assume that $dom(a)$ is finite for every attribute a. A c-item is called a class item. An item is any a-item. Let I be the set of all items. A transaction is a subset of I, with exactly one a-item for every attribute a. A database of transactions, denoted by $\mathcal{D}$, is a set of transactions. An itemset $\mathbf{X}$ is a subset of I. We denote by 2^I the set of all itemsets. As usual in the literature, we write $\mathbf{X}, \mathbf{Y}$ for $\mathbf{X} \cup \mathbf{Y}$, that is the set of items including both $\mathbf{X}$ and $\mathbf{Y}$. It is worth noting, however, that $\mathbf{X}, \mathbf{Y}$ characterizes a set of individuals who have both the attributes in $\mathbf{X}$ and in $\mathbf{Y}$. For a transaction T, we say that T verifies $\mathbf{X}$ if $\mathbf{X} \subseteq T$. The support of an itemset $\mathbf{X}$ w.r.t. a non-empty transaction database $\mathcal{D}$ is the ratio of transactions in $\mathcal{D}$ verifying $\mathbf{X}$ with respect to the total number of transactions: $supp_{\mathcal{D}}(\mathbf{X}) = |\{\ T \in \mathcal{D} \mid \mathbf{X} \subseteq T\ \}|/|\mathcal{D}|$, where $|\ |$ is the cardinality operator. An association rule is an expression $\mathbf{X} \to \mathbf{Y}$, where $\mathbf{X}$ and $\mathbf{Y}$ are itemsets. $\mathbf{X}$ is called the *premise* (or the *body*) and $\mathbf{Y}$ is called the *consequence* (or the *head*) of the association rule. We say that $\mathbf{X} \to \mathbf{C}$ is a *classification rule* if $\mathbf{C}$ is a class item and $\mathbf{X}$ contains no class item. The support of $\mathbf{X} \to \mathbf{Y}$ w.r.t. $\mathcal{D}$ is defined as: $supp_{\mathcal{D}}(\mathbf{X} \to \mathbf{Y}) = supp_{\mathcal{D}}(\mathbf{X}, \mathbf{Y})$. The confidence of $\mathbf{X} \to \mathbf{Y}$, defined when $supp_{\mathcal{D}}(\mathbf{X}) > 0$, is: $conf_{\mathcal{D}}(\mathbf{X} \to \mathbf{Y}) = supp_{\mathcal{D}}(\mathbf{X}, \mathbf{Y})/supp_{\mathcal{D}}(\mathbf{X})$. Support and confidence range over $[0, 1]$. We omit the subscripts in $supp_{\mathcal{D}}()$ and $conf_{\mathcal{D}}()$ when clear from the context. Also, the notation readily extends to negated itemsets $\neg\mathbf{X}$. Nevertheless, when using negated itemsets in the paper we will be able to calculate support and/or confidence by formulas that involve itemsets without negations. Since the seminal paper [1], a number of well explored algorithms [10] have been designed in order to extract *frequent* itemsets, i.e., itemsets with a specified minimum support.

3 Potentially Discriminated Groups

Civil rights laws explicitly identify the groups to be protected against discrimination, e.g., females or black people or minorities. With the syntax of Sect. 2, those groups can be represented as items, e.g., `sex=female` or `race=black`. However, discrimination typically occurs for subgroups rather than for the whole group. The intersection of two disadvantaged minorities is a, possibly empty, smaller (even more disadvantaged) minority as well. As an example, we could be interested in discrimination against elder females. With our syntax, this group would be represented by the itemset `sex=female, age=elder`. We then fix a set $\mathcal{I}_d$ of potentially discriminatory (PD) itemsets, which will be the object of the discrimination analysis. Following [16], the only formal property we require for $\mathcal{I}_d$ is downward closure.

DEFINITION 3.1. *A set of itemsets $\mathcal{I}$ is downward closed if when $\mathbf{A}_1 \in \mathcal{I}$ and $\mathbf{A}_2 \in \mathcal{I}$ then $\mathbf{A}_1, \mathbf{A}_2 \in \mathcal{I}$.*

The property is sufficient for uniquely splitting an itemset $\mathbf{X}$ into a PD itemset $\mathbf{A} \in \mathcal{I}_d$ and a potentially non-discriminatory (PND) itemset $\mathbf{B} = \mathbf{X} \setminus \mathbf{A} \notin \mathcal{I}_d$ by setting $\mathbf{A}$ to the largest subset of $\mathbf{X}$ that belongs to $\mathcal{I}_d$. Actually, defining $\mathbf{A}$ in such a way is equivalent[1] to require the downward closure property of $\mathcal{I}_d$.

4 Measuring Discrimination

The basic problem in the analysis of discrimination, given a dataset of historical decision records, is precisely to quantify the degree of discrimination suffered by a given group (say, an ethnic group) in a given context (say, a geographic area and/or an income range) with respect to the decision (say, credit approval). In our approach, we rephrase this problem in a rule based setting: if $\mathbf{A}$ is the condition (i.e., the itemset) that characterizes the group which is suspected of being discriminated, $\mathbf{B}$ is the itemset that chacterizes the context and $\mathbf{C}$ is the decision (class) item, then the analysis of discrimination is pursued by studying the rule $\mathbf{A}, \mathbf{B} \to \mathbf{C}$, together with its *confidence* with respect

[1] Assume that $\mathbf{A}_1 \in \mathcal{I}_d$ and $\mathbf{A}_2 \in \mathcal{I}_d$. Since $\mathbf{A}_1, \mathbf{A}_2$ includes both $\mathbf{A}_1$ and $\mathbf{A}_2$, then the largest subset of $\mathbf{A}_1, \mathbf{A}_2$ that is in $\mathcal{I}_d$ must be $\mathbf{A}_1, \mathbf{A}_2$ itself.

to the underlying decision dataset - namely, how often such a rule is true in the dataset itself. On the basis of this idea, we introduce in this section a family of measures of the degree of discrimination of a potentially discriminatory (PD) rule $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$, where $\mathbf{A}$ is a non-empty PD itemset and $\mathbf{B}$ is a PND itemset. Our approach in defining the family of measures consists of translating the qualitative statements of existing laws, regulations and legal cases into quantitative formal counterparts over classification rules.

4.1 Ratio Measures Unfortunately, there is no uniformity nor general agreement on a standard definition of discrimination by legislations. A general principle is to consider group under-representation as a quantitative measure of the qualitative requirement that people in a group are treated "less favorably" [5, 20] than others, or such that "a higher proportion of people without the attribute comply or are able to comply" [4] to a qualifying criteria. As a first proposal, we recall from [16] the notion of extended lift, a measure of the increased confidence in concluding an assertion $\mathbf{C}$ resulting from adding (potentially discriminatory) information $\mathbf{A}$ to a rule $\mathbf{B} \rightarrow \mathbf{C}$ where no PD itemset appears.

DEFINITION 4.1. *Let $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$ be a classification rule with $conf(\mathbf{B} \rightarrow \mathbf{C}) > 0$. The extended lift of the rule is:*

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

A rule `sex=female, car=own` $\rightarrow$ `credit=no` with an extended lift of 3 means that being a female increases 3 times the probability of having refused credit with respect to the average confidence of people owning a car. An alternative way, yet equivalent[2], of defining the extend lift is as the ratio between the proportion of the disadvantaged group $\mathbf{A}$ in context $\mathbf{B}$ obtaining the benefit $\mathbf{C}$ over the overall proportion of $\mathbf{A}$ in $\mathbf{B}$:

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

This makes it clear how extended lift relates to the principle of group representation. In addition to extended lift, other measures can be formalized starting from different definitions of discrimination provided by laws. According to the Anti-discrimination Act of the Queensland State [4], discrimination on the basis of an attribute happens if "a person treats, or proposes to treat, a person with an attribute less favorably than another person without the attribute". Since the term of comparison is another person *without* the attribute, the ratio should now consider "people with" over "people without" the attribute.

DEFINITION 4.2. *Let $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$ be a classification rule with $conf(\neg\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}) > 0$. The selection lift of the rule is:*

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

It is immediate to observe that the selection lift is equivalent to:

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

A special case of selection lift occurs when contrasting the sex items, i.e., $\mathbf{A}$ is `sex = female` and $\neg\mathbf{A}$ is `sex = male`. This is the form stated in the Sex Discrimination Act of U.K. [20]. In the literature and jurisprudence, such a contrast is generalized to non-binary attributes as, for instance, when comparing the credit denial ratio of blacks to the one of whites. This yields a third measure, which given $\mathbf{A}$ as a single item $\mathtt{a} = \mathtt{v}_1$ (e.g., black race) compares it to the most favored item $\mathtt{a} = \mathtt{v}_2$ (e.g., white race).

DEFINITION 4.3. *Let $\mathtt{a} = \mathtt{v}_1, \mathbf{B} \rightarrow \mathbf{C}$ be a classification rule, and $\mathtt{v}_2 \in dom(a)$ with $conf(\mathtt{a} = \mathtt{v}_2, \mathbf{B} \rightarrow \mathbf{C})$ minimal and non-zero. The contrasted lift of the rule is:*

$$clift(\mathtt{a} = \mathtt{v}_1, \mathbf{B} \rightarrow \mathbf{C}) = \frac{conf(\mathtt{a} = \mathtt{v}_1, \mathbf{B} \rightarrow \mathbf{C})}{conf(\mathtt{a} = \mathtt{v}_2, \mathbf{B} \rightarrow \mathbf{C})}.$$

The formulation above is substantiated by the Racial Equality Directive of E.U. [5], where discrimination "shall be taken to occur where one person is treated less favorably than another is in a comparable situation on grounds of racial or ethnic origin". Here the comparison appears to be done between two races (the disadvantaged one and the favored one). The U.S. legislation goes further [21, (d) Section 4D] by stating that "a selection rate for any race, sex, or ethnic group which is less than four-fifths (or eighty percent) of the rate for the group with the highest rate will generally be regarded as evidence of adverse impact". Since we are considering benefit refusal (denial rate), the four-fifths rule turns out to fix a maximum threshold value for $clift()$ of $5/4 = 1.25$.

Let us introduce a final measure based on odds ratios. In the gambling terminology, the odds 2/3 (2 to 3) means that for every 2 cases an event may occur there are 3 cases the event may not occur. Stated in terms of the probability p of the event, the odds ratio is $p/(1-p)$. Therefore, a fair bet would offer \$3 for every \$2 one wager on the occurrence of the event. In the

[2] $\frac{conf(\mathbf{A},\mathbf{B} \rightarrow \mathbf{C})}{conf(\mathbf{B} \rightarrow \mathbf{C})} = \frac{supp(\mathbf{A},\mathbf{B},\mathbf{C})\ supp(\mathbf{B})}{supp(\mathbf{A},\mathbf{B})\ supp(\mathbf{B},\mathbf{C})} = \frac{conf(\mathbf{B},\mathbf{C} \rightarrow \mathbf{A})}{conf(\mathbf{B} \rightarrow \mathbf{A})}$.

Classification rule: $c = \mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$

$\mathbf{B}$	$\mathbf{C}$	$\neg\mathbf{C}$
$\mathbf{A}$	a_1	$n_1 - a_1$
$\neg\mathbf{A}$	a_2	$n_2 - a_2$

$$p_1 = a_1/n_1 \quad p_2 = a_2/n_2 \quad p = (a_1 + a_2)/(n_1 + n_2)$$

$$elift(c) = \frac{p_1}{p}, \quad slift(c) = \frac{p_1}{p_2}, \quad olift(c) = \frac{p_1(1-p_2)}{p_2(1-p_1)}$$

$$elift_d(c) = p_1 - p, \quad slift_d(c) = p_1 - p_2$$

Figure 1: Contingency table for a classification rule

employment discrimination literature [8], the "event" modelled is promotion or hiring of a person. The odds of a classification rule $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$ can then be defined as:

$$odds(\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}) = \frac{conf(\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C})}{1 - conf(\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C})},$$

or, since $1 - conf(\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}) = conf(\mathbf{A}, \mathbf{B} \rightarrow \neg\mathbf{C})$, as:

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

The odds ratio in employment hiring is the ratio between the odds of hiring a person belonging to a minority group over the odds of hiring a person not belonging to that group. Let us extend the concept to rules.

DEFINITION 4.4. *Let $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$ be a classification rule with $conf(\neg\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}) > 0$ and $conf(\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}) < 1$. The odds lift of the rule is:*

$$olift(\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}) = \frac{odds(\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C})}{odds(\neg\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C})}.$$

It is immediate to observe that the odds lift is equivalent to:

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

An alternative view of the measures introduced so far can be given starting from the contingency table of $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$ shown in Fig. 1. Each cell in the table is filled in by the number of tuples in the transactions database $\mathcal{D}$ (i.e., the absolute support) satisfying $\mathbf{B}$ and the coordinates. Using the notation of the figure, confidence of $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$ is $p_1 = a_1/n_1$. Analogously, extended, selection and odds lifts can be defined as shown in the figure.

The next result relates the four measures.

LEMMA 4.1. *Let c be a classification rule $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$. Then either $\{olift(c),\ clift(c)\} \geq slift(c) \geq elift(c) \geq 1$, or $\{olift(c),\ clift(c)\} \leq slift(c) \leq elift(c) \leq 1$.*

4.2 Difference Measures Although the measures introduced so far are defined in terms of ratios, measures based on the difference of confidences have been considered on the legal side as well. For instance, in the U.K., a difference of 5% in confidence between female ($\mathbf{A}$ is `sex=female`) and male ($\neg\mathbf{A}$ is `sex=female`) treatment is assumed by courts as significant of discrimination against women. We define next a version of extended and selection lift using differences.

DEFINITION 4.5. *Let $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$ be a classification rule. We define:*

$$\begin{aligned} elift_d(\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}) &= conf(\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}) - \\ &\quad conf(\mathbf{B} \rightarrow \mathbf{C}) \\ slift_d(\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}) &= conf(\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}) - \\ &\quad conf(\neg\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}) \end{aligned}$$

Difference-based measures range over $[-1, 1]$. Lemma 4.1 readily extends to them.

LEMMA 4.2. *Let c be a classification rule $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$. Then either $slift_d(c) \geq elift_d(c) \geq 0$ or $slift_d(c) \leq elift_d(c) \leq 0$.*

4.3 Maximum Measures For a classification rule $\texttt{sex} = \texttt{female}, \mathbf{B} \rightarrow \texttt{credit} = \texttt{yes}$ with a ratio measure lower than 1, the complementary decision rule $\texttt{sex} = \texttt{female}, \mathbf{B} \rightarrow \texttt{credit} = \texttt{no}$ has a measure greater than 1, and viceversa. In general, with reference to Fig. 1, let $c' = \mathbf{A}, \mathbf{B} \rightarrow \neg\mathbf{C}$. By the property: $conf(c) + conf(c') = 1$, we have: $elift(c') = (1 - p_1)/(1 - p)$, $slift(c') = (1 - p_1)/(1 - p_2)$ and $olift(c') = 1/olift(c)$. Similarly, for difference-based measures, we have $elift_d(c') = -elift_d(c)$ and $slift_d(c') = -slift_d(c)$. Since one is interested in detecting classification rules with high measure values, the general approach of measuring both $f(c)$ and $f(c')$ can be weakened, when the class is a binary attribute, to checking $f^m(c)$ for the adjusted measure $f^m()$ defined as follows.

DEFINITION 4.6. *Let $f()$ be one of the measures from Definitions 4.1-4.5, c a classification rule $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$, and c' its complementary-decision rule $\mathbf{A}, \mathbf{B} \rightarrow \neg\mathbf{C}$. We define the measure $f^m()$ as:*

$$f^m(c) = \begin{cases} max\{f(c), f(c')\} & \text{if } f(c),\ f(c') \text{ are defined,} \\ f(c) & \text{if only } f(c) \text{ is defined,} \\ f(c') & \text{if only } f(c') \text{ is defined.} \end{cases}$$

Intuitively, $f^m(\texttt{sex} = \texttt{female}, \mathbf{B} \rightarrow \texttt{credit} = \texttt{no})$ measures the degree of discrimination against or of favoritism in assigning credit to females in context $\mathbf{B}$. Notice that favoritism versus minorities, such as reserving quotas in employment selection procedures, is often

supported by the law, with the name of affirmative actions, as a means to compensate for past discrimination against the minority. $f^m(c)$ is greater or equal than 1 (resp., 0) for ratio measures (resp., difference measures) when both $f(c)$ and $f(c')$ are defined. Lemma 4.1 simplifies to the following form.

LEMMA 4.3. *Let c be a classification rule $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$, with contingency table as in Fig. 1. If $p_1, p_2 \neq 0, 1$, then: $olift^m(c) \geq slift^m(c) \geq elift^m(c) \geq 1$.*

4.4 Discriminatory Classification Rules We generalize the notion of discriminatory classification rules from [16] as follows.

DEFINITION 4.7. (a-PROTECTION) *Let $f()$ be one of the measures from Definitions 4.1-4.6, and $a \in \mathbb{R}$ a fixed threshold. A classification rule $c = \mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$ is a-protective w.r.t. $f()$ if $f(c) < a$. Otherwise, c is a-discriminatory.*

Intuitively, a is a fixed threshold stating an acceptable level of discrimination accordingly to laws, regulations, and jurisprudence. Classification rules below such a level are considered safe, whilst rules whose measure is greater or equal than such a level are considered a *prima facie* evidence of discrimination. The notion of a-protection is parametric to a measure $f()$. When considering $f()$ as $elift()$, it falls down to the original notion of α-protection[3] [16].

5 Statistical Significance of the Measures

While a high value of a discrimination measure for a classification rule can represent a *prima-facie* evidence of discrimination against a minority, the statistical significance of such a value has to be considered. This approach is customary in legal cases before courts [8, 17]. A confidence interval for a statistical parameter θ (in our case, difference, ratio or odds of two proportions) is an interval $[L_1, L_2]$ that reasonably contains the true value for the parameter. Typically the interval is stated in the form $\hat{\theta} \pm d$, where $\hat{\theta}$ is a point estimate and d is the margin of error. Given an observed contingency table, a confidence interval $[L_1, L_2]$ returned by some method at $100(1-\alpha)\%$ level of significance is such that $[L_1, L_2]$ contains the true value of θ in at least $100(1-\alpha)\%$ of cases. Stated in terms of statistical tests, this means that the null hypothesis $\theta = \theta_0$ cannot be rejected at the significance level of $100(1-\alpha)\%$ for every θ_0 in $[L_1, L_2]$.

In our context, we are given a classification rule $c = \mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$, and a reference measure $f()$. We can then interpret the contingency table of c (Fig. 1) as the result of an experiment, which returned a value $f(c)$ for the data at hand (the historical decision records). What is the chance that past decisions were affected by randomness rather than explicit discrimination against minority $\mathbf{A}$? A confidence interval provides us with a range for the true value of $f(c)$ over the entire population (of decisions), at a certain significance level. We will exploit this parallel to revise the definition of a-discrimination.

5.1 From Measures to Tests on Proportions Consider the contingency table for a classification rule $c = \mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$ in Fig. 1. We observe that the ratio and difference measures introduced in Sec. 4.1-4.2 have been the subject of extensive studies in the field of statistical inference:

- $slift(c)$ is the ratio p_1/p_2 of two proportions, also known as the risk ratio or relative risk (RR) [2];
- $slift_d(c)$ is the difference $p_1 - p_2$ of two proportions, also known as the risk difference (RD) [2, 7];
- $olift(c)$ is the odds ratio (OR) $p_1(1-p_2)/(p_2(1-p_1))$ of two proportions [2, 7];
- $elift(c)$ is the ratio p_1/p related to the population attributable risk (PAR) defined as $PAR = (p - p_1)/p$ by the formula $elift(c) = 1 - PAR$;
- $elift_d(c)$ is the difference $p_1 - p$ related to the attributable risk (AR) [7, 9] defined as $AR = p - p_1$ by the formula $elift_d(c) = -AR$.

Statistical tests and confidence intervals for the difference, ratio, and odds of proportions have been proposed throughout the last 50 years. Let us denote by π_1 and π_2 the true proportions of p_1 and p_2. Difference, ratio and odds of π_1 and π_2 follow discrete distribution probabilities. However, when numbers in the contingency table are large, the distributions can be asymptotically approximated by a normal or a log-normal distribution. Based on this, Wald confidence intervals can be calculated [2, 6, 7] as follows.

Let Z_α denote the critical value of the normal distribution cutting off probability α, namely $\Phi(Z_\alpha) = \alpha$ where $\Phi()$ is the cumulative normal distribution.

RD: Called $\hat{p} = p_1 - p_2$, the confidence interval for $\pi_1 - \pi_2$ is $[\hat{p} - d, \hat{p} + d]$ where:

$$d = Z_{1-\alpha/2}\sqrt{p(1-p)(\frac{1}{n_1} + \frac{1}{n_2})}.$$

RR: Called $\hat{r} = p_1/p_2$, the confidence interval for π_1/π_2 is $[\hat{r}/e^d, \hat{r}e^d]$ where:

$$d = Z_{1-\alpha/2}\sqrt{\frac{1}{a_1} - \frac{1}{n_1} + \frac{1}{a_2} - \frac{1}{n_2}}. \tag{5.1}$$

[3] We use the name "a-protection" instead of "α-protection" in order not to generate confusion later on when confidence intervals at the significance level of $100(1-\alpha)\%$ will be considered.

We refer the reader to [7, page 102] for Wald intervals of OR; to [7, page 132] for the ones of PAR; and to [13] for the ones of AR. In addition to the Wald confidence intervals outlined before, other asymptotic methods have been proposed in the statistical inference literature. We refer the reader to survey and comparison papers [6, 13, 14, 19]. Moreover, in order to improve the approximation of a discrete distribution by the normal or log-normal distribution several corrections for continuity have been proposed, such as Yates's correction and the Mid-p method [2]. Later on, we will consider the simple but effective plus-4 method [3], consisting of adding $Z_\alpha^2/4$ cases to each cell in the contingency table.

When numbers in a contingency table are very low, the approximation to normal distribution becomes imprecise. This is a critical issue not only from a theoretical point of view, but also in practice under a legal profile (see [17] for a discussion). Exact methods have been proposed in the statistic literature, where "exact" means that the actual discrete distribution of the statistical parameter is adopted in computing the confidence intervals. The original work on the subject traces back to Fisher's exact method for a single proportion, and it is currently a research topic in the statistical inference area. The issues here are twofold and contrasting. On the one hand, one looks for intervals whose width is as strict as possible. On the other hand, calculations of discrete distributions are computationally expensive. We anticipate that our use of confidence intervals will mostly be independent from the method used to derive them. Nevertheless, the more precise intervals we have the more significative discrimination conclusions we can derive. In the experiments (see Sect. 9), we will use the Wald confidence intervals corrected with the plus-4 method when $n_1 + n_2 > 30$ (see Fig. 1), and a recent exact method based on an extension of the Sterne's test [18] otherwise.

5.2 Revisiting a-protection We revisit the notions of a-protection and a-discrimination by relativizing them to a significance level. We assume that a method for computing the confidence interval for a measure $f()$ is fixed, and we write $[L_1^f(c), L_2^f(c)]$ to denote the confidence interval for the contingency table of rule c.

DEFINITION 5.1. (a-PROTECTION) *Let $f()$ be one of the measures from Definitions 4.1-4.6, and $a \in \mathbb{R}$ a fixed threshold. A classification rule $c = \mathbf{A}, \mathbf{B} \to \mathbf{C}$ is a-protective w.r.t. $f()$ at the significance level of $100(1-\alpha)\%$ if $L_2^f(c) < a$. c is a-discriminatory at the significance level of $100(1-\alpha)\%$ if $L_1^f(c) \geq a$.*

At the significance level of 0%, we have $Z_{1-\alpha/2} = Z_{1/2} = 0$ and then the (Wald) confidence intervals fall down to $L_1^f(c) = L_2^f(c) = f(c)$. Therefore, the definition above is a conservative extension of Def. 4.7. In general, the higher the significance level is, the wider is the confidence interval. At 100% significance level, the confidence interval is the whole set of reals. Certainly the true value of the measure belongs to this interval, but this information is of no use.

Finally, notice that when $L_1^f(c) < a \leq L_2^f(c)$ the rule c is neither a-discriminatory nor a-protective. Intuitively, there is no sufficient statistical evidence to conclude anything. On the basis of the analysis objectives, one would treat such rules as those a-discriminatory (and then, for instance, start conducting further investigation on them) or as those a-protective (and then stop the analysis since statistical significance could not be used as an evidence before a court).

6 Unveiling Discrimination in Datasets

We introduced in Sect. 4 and 5 various measures of discrimination and their associated significance tests, with the purpose of analyzing the discriminatory power of a specific classification rule; we are now in the position of extending this method to the broader problem of unveiling all discriminatory decision patterns hidden in a decision dataset. This goal can be achieved combining a-discrimination and association rule mining, i.e., by extracting classification rules that are a-discriminatory (at some fixed confidence level). When the dataset contains itemsets in $\mathcal{I}_d$, such as for gender and age items, checking a-discrimination can be done directly on extracted PD classification rules. When the dataset does not contain PD itemsets, as in the case of the race attribute, the check can be done indirectly. The next two subsections discuss the two cases.

6.1 Direct Discrimination Classification rules can be extracted from a dataset as a post-processing phase of frequent pattern extraction, a task largely studied and with a large number of competing algorithms [10]. Fig. 2 shows how to extract PD classification rules and check for a-discrimination using their contingency tables. The **DirectDiscriminationCheck()** procedure scans frequent patterns (having a specified minimum support threshold) of size k. Each pattern $\mathbf{R}$ give rises to a rule $c = \mathbf{A}, \mathbf{B} \to \mathbf{C}$. If $\mathbf{A}$, the PD part, is not empty we can build the contingency table of c. The values a_1 and n_1 in Fig. 1 are readily available by looking up at frequent patterns of size $k-1$. Concerning a_2 and n_2, we have to compute $supp(\neg\mathbf{A}, \mathbf{B}, \mathbf{C})$ and $supp(\neg\mathbf{A}, \mathbf{B})$. However, negated itemsets are typically not extracted by frequent pattern mining algorithms. Still, we can resort to support and confidence of $\mathbf{B} \to \mathbf{C}$ by noting that $supp(\neg\mathbf{A}, \mathbf{B}, \mathbf{C}) = supp(\mathbf{B} \to \mathbf{C}) - supp(\mathbf{A}, \mathbf{B}, \mathbf{C})$ and $supp(\neg\mathbf{A}, \mathbf{B}) =$

DirectDiscriminationCheck()

$N = |\mathcal{D}|$, $\mathcal{C} = \{$ class items $\}$, $\mathcal{L} = \emptyset$
ForEach k s.t. there exists k-frequent itemsets
$\mathcal{F}_k = \{$ k-frequent itemsets $\}$
delete from $\mathcal{L}$ unmarked elements and unmark all the marked ones
ForEach $\mathbf{R} \in \mathcal{F}_k$ with $\mathbf{R} \cap \mathcal{C} \neq \emptyset$
$\mathbf{C} = \mathbf{R} \cap \mathcal{C}$, $\mathbf{X} = \mathbf{R} \setminus \mathbf{C}$
$a_1 = supp(\mathbf{R})$
$n_1 = supp(\mathbf{X})$ // $\mathbf{X}$ found in $\mathcal{F}_{k-1}$
$\mathbf{A}$ = largest subset of $\mathbf{X}$ in $\mathcal{I}_d$
$\mathbf{B} = \mathbf{X} \setminus \mathbf{A}$
If $|\mathbf{A}| = 0$
add marked $\mathbf{B} \rightarrow \mathbf{C}$ to $\mathcal{L}$ with $supp = a_1$ and $conf = a_1/n_1$
Else
mark $\mathbf{B} \rightarrow \mathbf{C}$ in $\mathcal{L}$ // $\mathbf{B} \rightarrow \mathbf{C}$ found in $\mathcal{L}$
$a_2 = supp(\mathbf{B} \rightarrow \mathbf{C}) - a_1$
$n_2 = supp(\mathbf{B} \rightarrow \mathbf{C})/conf(\mathbf{B} \rightarrow \mathbf{C}) - n_1$
called $c = \mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$, check $L_1^f(c) \geq a$
using the contingency table $\begin{pmatrix} a_1 N & (n_1 - a_1)N \\ a_2 N & (n_2 - a_2)N \end{pmatrix}$
EndIf
EndForEach
EndForEach

Figure 2: Extraction and a-discrimination checking of PD classification rules.

$supp(\mathbf{B}) - supp(\mathbf{A}, \mathbf{B}) = supp(\mathbf{B} \rightarrow \mathbf{C})/conf(\mathbf{B} \rightarrow \mathbf{C}) - supp(\mathbf{A}, \mathbf{B})$. To this end, during the scans we maintain the set $\mathcal{L}$ of rules of the form $\mathbf{B} \rightarrow \mathbf{C}$ such that either $\mathbf{B}, \mathbf{C}$ is a frequent pattern of size k or $\mathbf{A}', \mathbf{B}, \mathbf{C}$ is a frequent pattern of size k for some $\mathbf{A}'$. Rules $\mathbf{B}, \mathbf{C}$ in the set $\mathcal{L}$ which do not satisfy this invariant property (they remain unmarked during the pass) are deleted at the beginning of iteration $k + 1$.

6.2 Indirect Discrimination Assume that the dataset does not contain some PD itemsets or it does not contain any PD itemset at all. For instance, the information on a person's race is typically not available (unless the dataset has been explicitly enriched with such an information in order to check for discrimination). Can the set of rules $\mathbf{D}, \mathbf{B} \rightarrow \mathbf{C}$, where $\mathbf{D}, \mathbf{B}$ is a PND itemset, extracted from such a dataset unveil, at least partially, discriminatory patterns? This issue has been considered in [16], where those rules are called potentially non-discriminatory (PND) and an inference model is proposed exploiting background knowledge (e.g., census data) with respect to the $elift()$ measure. Let us now generalize the approach. Consider the following contingency tables for a PND classification rule $\mathbf{D}, \mathbf{B} \rightarrow \mathbf{C}$ (left-hand side) and for the PD rule $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$ (right-hand side), where $\mathbf{A}$ is a PD itemset:

B	**C**	$\neg$**C**
D	b_1	$m_1 - b_1$
$\neg$**D**	b_2	$m_2 - b_2$

B	**C**	$\neg$**C**
A	a_1	$n_1 - a_1$
$\neg$**A**	a_2	$n_2 - a_2$

Given the left-hand side contingency table, we want to derive lower and upper bounds for $p_1 = a_1/n_1$ and $p_2 = a_2/n_2$, and then for their difference, ratio and odds. The idea is to consider itemsets $\mathbf{A}$ that are approximatively equivalent to $\mathbf{D}$ in the context $\mathbf{B}$, namely such that:

$$\beta_1 = conf(\mathbf{A}, \mathbf{B} \rightarrow \mathbf{D}) \quad \beta_2 = conf(\mathbf{D}, \mathbf{B} \rightarrow \mathbf{A})$$

are near to 1. β_1 and β_2 are typically provided as background knowledge (e.g., census data on distribution of races over the territory). A lower bound for a_1 is obtained by considering that, in the worst case, there are at least $\beta_2 m_1$ tuples satisfying $\mathbf{A}, \mathbf{B}$ (those satisfying $\mathbf{D}, \mathbf{B}$ multiplied by β_2), of which at most $m_1 - b_1$ do not satisfy $\mathbf{C}$. Summarizing, $a_1 \geq \beta_2 m_1 - (m_1 - b_1)$, and then:

$$p_1 \geq \beta_2 m_1/n_1 - (m_1/n_1 - b_1/n_1).$$

Since $\beta_1/\beta_2 = supp(\mathbf{D}, \mathbf{B})/supp(\mathbf{A}, \mathbf{B}) = m_1/n_1$, the inequality can be rewritten as:

$$p_1 \geq \beta_1/\beta_2(\beta_2 + b_1/m_1 - 1).$$

Analogously, $n_1 - a_1 \geq \beta_2 m_1 - b_1$, which leads to:

$$1 - \beta_1 + (b_1/m_1)(\beta_1/\beta_2) \geq p_1.$$

Similarly, we derive upper and lower bounds for p_2:

$$1 - \beta_1' + (b_2/m_2)(\beta_1'/\beta_2') \geq p_2 \geq \beta_1'/\beta_2'(\beta_2' + b_2/m_2 - 1),$$

where: $\beta_1' = conf(\neg\mathbf{A}, \mathbf{B} \rightarrow \neg\mathbf{D})$ and $\beta_2' = conf(\neg\mathbf{D}, \mathbf{B} \rightarrow \neg\mathbf{A})$. Notice that these two values can

be calculated from β_1, β_2 as follows. First, calculate:

$$n_1 = m_1\beta_2/\beta_1 \quad n_2 = m_1 + m_2 - n_1, \tag{6.2}$$

then:

$$\begin{aligned}
\beta_1' &= conf(\neg\mathbf{A}, \mathbf{B} \rightarrow \neg\mathbf{D}) \\
&= 1 - conf(\neg\mathbf{A}, \mathbf{B} \rightarrow \mathbf{D}) \\
&= 1 - \frac{supp(\mathbf{D}, \mathbf{B}) - supp(\mathbf{D}, \mathbf{B}, \mathbf{A})}{supp(\neg\mathbf{A}, \mathbf{B})} \\
&= 1 - (1 - \beta_2)m_1/n_2, \qquad (6.3)\\
\beta_2' &= conf(\neg\mathbf{D}, \mathbf{B} \rightarrow \neg\mathbf{A}) \\
&= 1 - conf(\neg\mathbf{D}, \mathbf{B} \rightarrow \mathbf{A}) \\
&= 1 - \frac{supp(\mathbf{A}, \mathbf{B}) - supp(\mathbf{D}, \mathbf{B}, \mathbf{A})}{supp(\neg\mathbf{D}, \mathbf{B})} \\
&= 1 - (1 - \beta_1)n_1/m_2. \qquad (6.4)
\end{aligned}$$

Using the derived upper and lower bounds for p_1 and p_2, one can easily derive upper and lower bounds for $p_1 - p_2$, p_1/p_2 and for the other measures introduced. The approach applies to Wald confidence intervals as well. As an example, observing that (5.1) can be rewritten as $d = Z_{1-\alpha/2}\sqrt{(1/p_1 - 1)/n_1 + (1/p_2 - 1)/n_2}$, an upper bound for d is:

$$d \leq Z_{1-\alpha/2}\sqrt{(1/LB_1 - 1)/n_1 + (1/LB_2 - 1)/n_2},$$

where LB_1 is any lower bound for p_1 and LB_2 is any lower bound for p_2.

7 Unveiling Discrimination in Classifiers

Since a classifier is trained on past decision records, possibly including discriminatory decisions, it may learn to behave discriminatorily. Let us introduce some terminology and notation. An attribute is called predictive if it is not the class attribute. For our purposes, a classifier is a (automatically built/trained) procedure that implements a function $cl()$ assigning to a transaction T over predictive attributes a class item $cl(T)$, the predicted class. We define the output of a classifier over an input set of transactions $\mathcal{T}$ as $\{(T, cl(T)) \mid T \in \mathcal{T}\}$.

We say that a classifier is discriminatory over an input set $\mathcal{T}$ if the classifier output over $\mathcal{T}$ contains discriminatory decisions. Such decisions can be unveiled as described in Sect. 6.

8 Correcting Rule-based Classifiers

In rule based classifiers, $cl(T)$ is computed from the confidences of a set of classification rules $\mathcal{R}$ extracted from a training set. As an example, the CPAR [22] classifier first considers for each class item the average confidence of the top (with respect to confidence) k rules in $\mathcal{R}$ such that T satisfies the rule premise. The class value for which the averaged confidence is the highest is returned as $cl(T)$.

For rule-based classifiers, we are in the position to calculate the degree of discrimination of rules in $\mathcal{R}$. Obviously, adopting a set of discriminatory rules lead the classifier to yield outputs that contain discriminatory decisions. The converse is not necessarily true, as we will discuss in Sect. 8.3, namely an output can contain discriminatory decision even if $\mathcal{R}$ does not contain any discriminatory rule. In this section, we propose a simple approach for correcting rules in $\mathcal{R}$ as a means to prevent discriminatory decisions in the output of a rule-based classifier using $\mathcal{R}$. The approach consists of correcting (reducing or increasing) the confidence of rules in $\mathcal{R}$. This is a minimal modification of the structure of the classifier: we do not add nor remove rules in $\mathcal{R}$, we do not change the logics of computing $cl(T)$, we do not act on the learning algorithm nor on the training set. The change in confidence, however, must be kept as low as possible in order not to degrade the accuracy of the classifier. By setting confidence of rules with class item `credit=no` to 100%, and confidence of rules with class item `credit=yes` to 0%, we end up with a classifier denying credit to everybody. It is not discriminatory, but it is of no use.

We mention here that an orthogonal approach, based on massaging the training set, is presented in [11].

8.1 Correcting Direct Discrimination Consider a PD classification rule $c = \mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$ in $\mathcal{R}$ and its contingency table (see Fig. 1) calculated on the training set used to build the classifier. Assume that for a reference discrimination measure f, c is a-discriminatory at the level of $100(1-\alpha)\%$, i.e., $a \leq L_1^f(c)$. In order to correct the discriminatory behavior of the rule, we first modify its contingency table as follows:

B	**C**	$\neg$**C**
A	$a_1 - \Delta$	$n_1 - a_1 + \Delta$
$\neg$**A**	a_2	$n_2 - a_2$

where $abs(\Delta)$ is the minimum integer such that $a > L_1^f(c)$ when considering such a revised contingency table for c. Then, we change the confidence of c to $(a_1 - \Delta)/n_1$. Let us detail the method for $slift()$. The approach is similar for the other measures introduced.

Using the Wald interval for ratio of proportions (RR), we set Δ to the minimum integer such that:

$$\frac{(a_1 - \Delta)/(n_1 p_2)}{e^{Z_{1-\alpha/2}\sqrt{\frac{1}{(a_1-\Delta)} - \frac{1}{n_1} + \frac{1}{a_2} - \frac{1}{n_2}}}} < a,$$

which, after elementary algebra, is:

$$a_1 - \Delta < a n_1 p_2 e^{Z_{1-\alpha/2}\sqrt{\frac{1}{(a_1-\Delta)} - \frac{1}{n_1} + \frac{1}{a_2} - \frac{1}{n_2}}}.$$

Albeit the solutions of the inequality cannot be expressed in solved form, we observe that the function $a_1 - \Delta$ is monotonic decreasing in Δ, while $an_1p_2e^{Z_{1-\alpha/2}\sqrt{\frac{1}{(a_1-\Delta)}-\frac{1}{n_1}+\frac{1}{a_2}-\frac{1}{n_2}}}$ is monotonic increasing in Δ. For $\Delta = 0$, the inequality is false (otherwise, the rule would not be a-discriminatory). For $\Delta = \Delta_0 = a_1 - (a-1)n_1p_2$, we have:

$$a_1 - \Delta_0 = (a-1)n_1p_2 < an_1p_2e^x$$

since $e^x \geq 1$ for any $x \geq 0$, and $p_2 > 0$ by definition of $slift()$. Hence the inequality above holds[4]. Summarizing, there is one and only one solution of the inequality above in the range $[0, \Delta_0]$ (or $[\Delta_0, 0]$ depending on the sign of Δ_0]). Since we are interested in the ceiling integer of such a solution, a simple binary search over integers in the range $[0, \lceil\Delta_0\rceil]$ is sufficient.

8.2 Correcting Indirect Discrimination PND classification rules in $\mathcal{R}$ have to be corrected as well. In fact, a PND rule $\mathbf{D}, \mathbf{B} \rightarrow \mathbf{C}$ such that $\mathbf{D}$ is "almost equivalent" to a PD itemset $\mathbf{A}$ has "almost the same effect" on people from the minority group $\mathbf{A}$ as the PD rule $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$. A typical example is the redlining situation, where $\mathbf{A}$ is a minority group, and $\mathbf{D}$ is a specific zip code locating people of the minority group among the people in $\mathbf{B}$. We distinguish two cases. In semi-indirect discrimination, we correct PND rules that are related to itemsets $\mathbf{A}$ which occur in the dataset (e.g., itemsets over gender and age), while in indirect discrimination the PD information $\mathbf{A}$ is not available in the data (e.g., itemsets over race). In semi-indirect discrimination, we are in the position to calculate:

$$\beta_1 = conf(\mathbf{A}, \mathbf{B} \rightarrow \mathbf{D}) \quad \beta_2 = conf(\mathbf{D}, \mathbf{B} \rightarrow \mathbf{A}),$$

since the information $\mathbf{A}$ is available in the training set. In indirect discrimination, this is not possible and then β_1, β_2 have to be provided by external background knowledge (as in the case discussed in Sect. 6.2). The correction proceeds in both cases as follows. Consider the following contingency table (w.r.t. the training set) for $\mathbf{D}, \mathbf{B} \rightarrow \mathbf{C}$:

$\mathbf{B}$	$\mathbf{C}$	$\neg\mathbf{C}$
$\mathbf{D}$	b_1	$m_1 - b_1$
$\neg\mathbf{D}$	b_2	$m_2 - b_2$

and let $p'_1 = b_1/m_1$, $p'_2 = b_2/m_2$. What would the effect of this rule over a minority $\mathbf{A}$ be? In the extreme case that $\mathbf{D}$ and $\mathbf{A}$ are equivalent over the people satisfying context $\mathbf{B}$, we could replace above $\mathbf{D}$ by $\mathbf{A}$. In such a case, we should correct $\mathbf{D}, \mathbf{B} \rightarrow \mathbf{C}$ in the same way as we did for $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$ in the last subsection. In general, however, $\mathbf{D}$ can be only approximatively equivalent to a PND itemset $\mathbf{A}$. With reference to Fig. 1, we can estimate the probability that an element in $\mathbf{B}$ satisfying $\mathbf{A}$ (say, the minority group) is assigned class $\mathbf{C}$ (say, credit denial) by the rule $\mathbf{D}, \mathbf{B} \rightarrow \mathbf{C}$ as $p_1 = a_1/n_1 \approx \beta_1 p'_1 + (1-\beta_1)p'_2$, since the element satisfies $\mathbf{D}$ with probability β_1 (and then it has chance p'_1 of being assigned class $\mathbf{C}$) and it satisfies $\neg\mathbf{D}$ with probability $1 - \beta_1$ (and then it has chance p'_2 of being assigned class $\mathbf{C}$).

Concerning an element in $\mathbf{B}$ satisfying $\neg\mathbf{A}$ (say, the advantaged group), we assume instead $p_2 = a_2/n_2 \approx p'_2$, i.e., a member of an advantaged group is assigned class $\mathbf{C}$ with the advantageous probability p'_2. Using the calculations (6.2) for n_1, and n_2, the contingency table estimated as the effect of applying the classification rule $\mathbf{D}, \mathbf{B} \rightarrow \mathbf{C}$ over elements in $\mathbf{B}$ that may/may not satisfy $\mathbf{A}$ is:

$\mathbf{B}$	$\mathbf{C}$	$\neg\mathbf{C}$
$\mathbf{A}$	$n_1(\beta_1 p'_1 + (1-\beta_1)p'_2)$	$n_1(1 - \beta_1 p'_1 - (1-\beta_1)p'_2)$
$\neg\mathbf{A}$	$n_2 p'_2$	$n_2(1-p'_2)$

We can apply the reasoning for correcting PD classification rules to this contingency table by setting $p'_1 = (b_1 - \Delta)/m_1$. The value of Δ for which the contingency table above leads to non a-discrimination (at a fixed confidence level) will determine the revised confidence for classification rule $\mathbf{D}, \mathbf{B} \rightarrow \mathbf{C}$.

Finally, since a same rule $\mathbf{D}, \mathbf{B} \rightarrow \mathbf{C}$ can be written in many ways (just fix $\mathbf{B}$ to any subset of the premise, and let $\mathbf{D}$ be the rest of the premise), the most conservative correction should be chosen, for any possible split of the premise and for any $\mathbf{A}$. Practically, however, we restrict to $\mathbf{A}$ such that $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{D}$ satisfy a minimum support threshold over the training set.

8.3 Discussion On the theoretical side, should we expect that correcting confidence of rules in $\mathcal{R}$ will completely remove discriminatory decisions in the output of the rule-based classifier? The answer is negative. Since we act on each rule in isolation, the combined effects of two or more rules is ignored. As an example, assume $\mathcal{R}$ consisting of two rules: *(i)* `own car = yes` $\rightarrow$`credit = no`; and *(ii)* `driver = yes` $\rightarrow$`credit = yes`. The actual confidence of the rules will be irrelevant for what follows, so we can assume they have been already corrected w.r.t. the threshold $a = 2$ and the $slift()$ measure. Using the CPAR classification algorithm, the following output would be produced:

[4]Notice that Δ_0 is the value for which $L_1^{slift}(c) \leq (a_1 - \Delta_0)/(n_1p_2) = a - 1$, i.e., $L_1^{slift}(c) < a$, and then such that c is certainly non a-discriminatory.

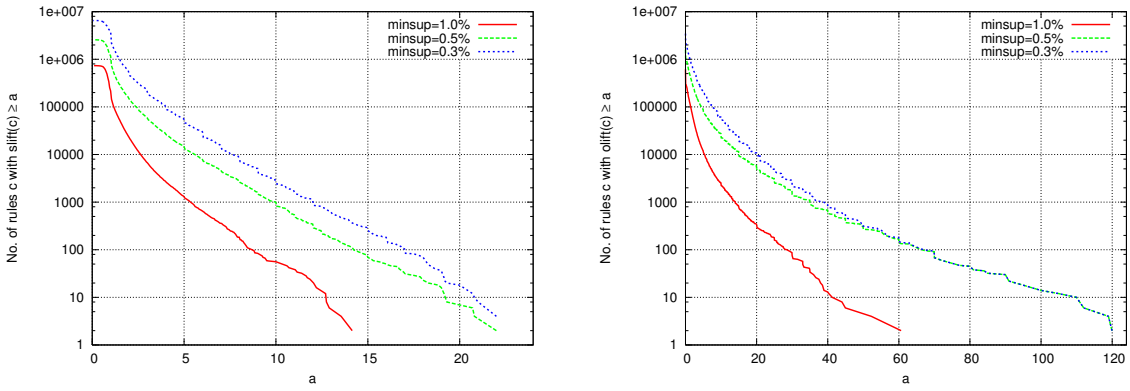

Figure 3: Distributions of $slift()$ and $olift()$ measures vs. classification rule minimum support.

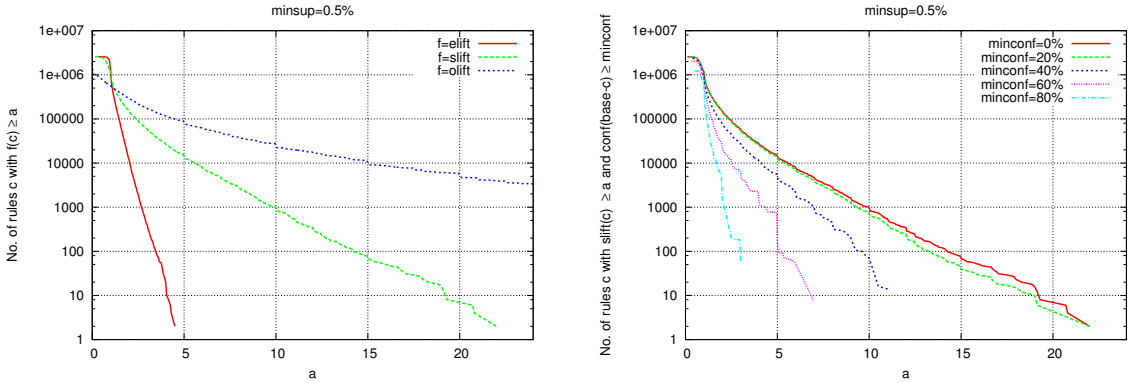

Figure 4: Left: distributions of $elift()$, $slift()$ and $olift()$ measures. Right: distribution of $slift()$ vs. base rule minimum confidence. Using the notation of Fig. 1, $conf(base\text{-}c) = p$.

sex	driver	own car	ZIP	credit
male	no	yes	101	no
female	no	yes	101	no
female	yes	no	100	yes
male	yes	no	101	yes

The first two cases are handled by rule *(i)*, and the last two cases by rule *(ii)*. However, the output dataset contains discriminatory decisions, since the following 2-discriminatory (w.r.t. $slift()$) classification rule can be extracted from it: `sex = female, ZIP = 101` $\rightarrow$ `credit = no`. This is just one of the theoretical open problems that have to be further investigated.

9 Experimental Results

We will report some analyses over the public domain German credit dataset [15], consisting of 1000 transactions representing the good/bad credit class of bank account holders. The dataset includes nominal (or discretized) attributes on *personal properties*: checking account status, duration, savings status, property magnitude, type of housing; on *past/current credits and requested credit*: credit history, credit request purpose, credit request amount, installment commitment, existing credits, other parties, other payment plan; on *employment status*: job type, employment since, number of dependents, own telephone; and on *personal attributes*: personal status and gender, age, resident since, foreign worker. We fix $\mathcal{I}_d$ to include all itemsets built on 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). High values of the discrimination measures will occur when people in one or more of those categories is denied credit more often than people not in those categories.

9.1 Unveiling Discrimination in the German Credit Dataset The German Credit dataset contains discriminatory decisions w.r.t. all of the measures introduced in this paper. By means of the **DirectDiscriminationCheck()** procedure of Fig. 2, we proceed by extracting patterns of discrimination in the form of PD classification rules (with a fixed minimum support threshold) with high values of the measures. Fig. 4 (left) shows the distributions of a-protective rules w.r.t. extended, selection and odds lift. We observe that, if clas-

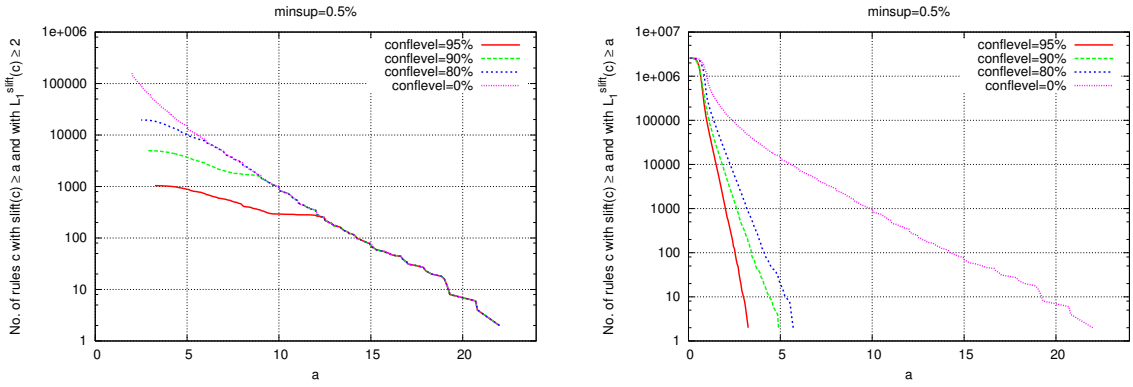

Figure 5: Left: distributions of $slift()$ statistically greater or equal than 2 at various confidence levels. Right: distributions of $slift()$ greater or equal than a at various confidence levels.

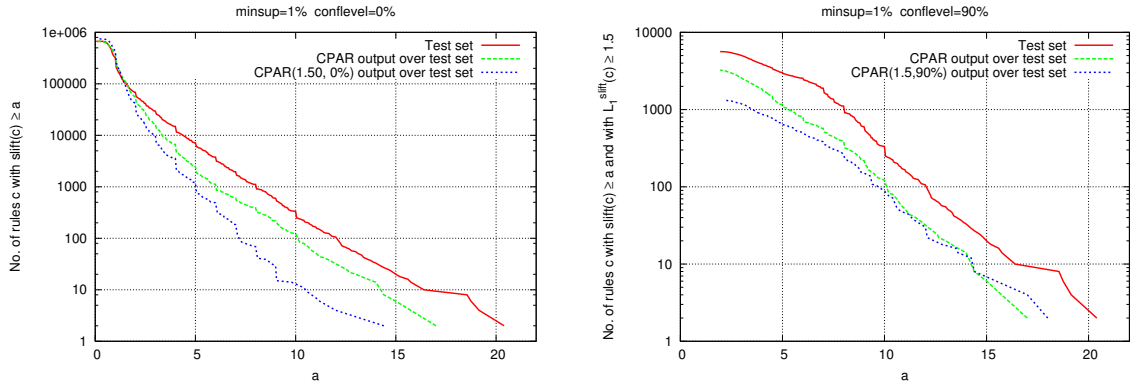

Figure 6: Distributions of $slift()$ for German credit test set, CPAR output, and corrected CPAR output.

sification rules with a minimum support ms are considered, the extended lift ranges over $[0, 1/ms]$. This property does not extend to selection lift nor to odds lift, which in general are unbound from above. Nevertheless, Fig. 3 shows that, in practice, the lower minimum support threshold the more niches of discrimination can be unveiled. A similar reasoning can be done for the minimum confidence threshold of the base rule, where the base rule of $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$ is $\mathbf{B} \rightarrow \mathbf{C}$. Fig. 4 (right) shows that we should look for contexts $\mathbf{B}$ where the credit denial rate is low in order to find the most discriminatory decisions. Consider now the notion of a-discrimination at a certain confidence level. Fig. 5 (left) shows the distributions of the PD rules that are 2-discriminatory w.r.t. their $slift()$ value at various confidence levels. Intuitively, the higher the confidence level, the lower is the number of 2-discriminatory rules. Rules with very high selection lift will remain 2-discriminatory until high confidence levels. Fig. 5 (right) shows the total number of PD rules that are a-discriminatory at various confidence levels. At 90% confidence level, no rule is 5-discriminatory. Intuitively, higher confidence levels greatly reduce the number of statistically significant discriminatory rules.

9.2 Unveiling Discrimination in CPAR In order to study discrimination in the output of CPAR, we split the German credit dataset into a 60% training set, used to train a CPAR classifier, and into a 40% test set. The trained classifier consists of a set $\mathcal{R}$ of 189 classification rules. Given a transaction T in the test set, where the actual class is omitted, the CPAR classification algorithm consists of finding for each possible class the top k (with $k = 5$) rules in $\mathcal{R}$ whose premise is satisfied by T. The class with the highest average confidence of the k rules is returned as the predicted class $cl(T)$. The CPAR output over the test set is then the set of transactions $(T, cl(T))$. In order to unveil discrimination in the test set and/or in the CPAR output, we proceed as in the last subsection by extracting from those sets patterns of discrimination in the form of PD classification rules. Fig. 6 (left) shows the distribution of the $slift()$ measure on the test set, unveiling PD rules with a selection lift of up to 20, and on the CPAR output over the test set, unveiling PD rules with lower selection lift. Intuitively, the output of CPAR is "less discriminatory" than the test set. We motivate this by two reasons. First, as expected by any classifier, CPAR try to be accurate but not to overfit the

data. This leads to a model that does not reproduce all niches of discrimination. Second, more specific to CPAR internals, since classification averages the confidence of 5 rules, the effects of one discriminatory rule out of 5 is mitigated. Fig. 6 (right) shows the distributions of PD rules that are 1.5-discriminatory at the 90% confidence level both on the test set and on the CPAR output over the test set. The beneficial effect of CPAR is homogeneous along the whole range of $slift()$ values.

9.3 Correcting CPAR We have implemented the correction of direct and semi-indirect discrimination described in Sect. 8. Let us denote by CPAR(a, CL) the CPAR classification algorithm on a set of rules corrected for a-discrimination at the CL confidence level. Correcting the set of rules of CPAR for 1.5-discrimination at the confidence level of 0% changed the accuracy of 87 rules in $\mathcal{R}$ (46% of the total), and it resulted in no loss of classification accuracy w.r.t. the test set. Fig. 6 (left) shows the distribution of the $slift()$ measure for the CPAR(1.5, 0%) output over the test set. We observe that the improvement over the raw CPAR output is considerable. For $a = 10$, we have 13 a-discriminatory PD rules extracted from the output of CPAR(1.5, 0%), which is a good improvement over the 122 extracted from the raw CPAR output. Nevertheless, we point out that, for the reasons discussed in Sect. 8.3, the discriminatory decisions in the output dataset are not totally removed. Finally, consider CPAR(1.5, 90%), and its output over the test set. The distribution of PD rules that are 1.5-discriminatory at 90% confidence level is shown in Fig. 6 (right). Compared to the raw CPAR, the total number of discriminatory rules is lowered from 3236 to 1317, and the improvement is constant almost over the whole range of selection lift values.

10 Conclusions

On the basis of a review of existing laws, we have formalized and studied a family of discrimination measures for classification rules, including a notion of statistical significance. Discriminatory classification rules are the basic tool for unveiling direct and indirect discriminatory decisions in datasets of historical records, and in the output of classifiers. Also, the measures of discrimination provided the basis for a correction method for rule-based classifiers.

References

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

[2] A. Agresti. *Categorical Data Analysis*. Wiley-Interscience, 2002.

[3] A. Agresti and C. Brian. Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures. *The American Statistician*, 54(4):280–288, 2000.

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

[5] European Union Legislation. (a) Racial Equality Directive, (b) Employment Equality Directive, 2008. `http://ec.europa.eu/employment_social`.

[6] C. Farrington and G. Manning. Test statistics and sample size formulae for comparative binomial trials with null hypothesis of non-zero risk difference or non-unity relative risk. *Stat. in Medic.*, 9:1447-1454, 1990.

[7] J. L. Fleiss, B. Levin, and M. C. Paik. *Statistical Methods for Rates and Proportions*. Wiley, 2003.

[8] J. L. Gastwirth. Statistical reasoning in the legal setting. *The American Statistician*, 46(1):55–69, 1992.

[9] O. Gefeller. An annotated bibliography on the attributable risk. *Biometrical J.*, 34:1007–1012, 1992.

[10] B. Goethals. Frequent itemset mining implementations repository, 2008. `http://fimi.cs.helsinki.fi`.

[11] F. Kamiran and T. Calders. Classification without discrimination. In *Proc. of IEEE-IC4*. IEEE press, 2009. Accepted for publication.

[12] N. Lerner. *Group Rights and Discrimination in International Law*. Martinus Nijhoff Publishers, 1991.

[13] H. M. Leung and L. L. Kupper. Comparisons of confidence intervals for attributable risk. *Biometrics*, 37(2):293–302, 1981.

[14] R. G. Newcombe. Interval estimation for the difference between independent proportions: comparison of eleven methods. *Stat. in Medic.*, 17:873–89, 1998.

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

[16] D. Pedreschi, S. Ruggieri, and F. Turini. Discrimination-aware data mining. In *Proc. of KDD 2008*, pages 560–568. ACM, 2008.

[17] M. J. Piette and P. F. White. Approaches for dealing with small sample sizes in employment discrimination litigation. *J. of Forensic Economics*, 12:43–56, 1999.

[18] J. Reiczigel, Z. Abonyi-Tóth, and J. Singer. An exact confidence set for two binomial proportions and exact unconditional confidence intervals for the difference and ratio of proportions. *Computational Statistics & Data Analysis*, 52(11):5046–5053, 2008.

[19] M. Tian, M. L. Tang, H. K. T. Ng, and P. S. Chan. Confidence intervals for the risk ratio under inverse sampling. *Statistics in Medicine*, 27:3301–3324, 2008.

[20] U.K. Legislation. (a) Sex Discrimination Act, (b) Race Relation Act, 2008. `http://www.statutelaw.gov.uk`.

[21] U.S. Federal Legislation. (a) Equal Credit Opportunity Act, (b) Fair Housing Act, (c) Intentional Employment Discrimination, (d) Uniform guidelines on employee selection procedure, 2008. `http://www.usdoj.gov`.

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