Artificial Intelligence and Law manuscript No.
(will be inserted by the editor)

Integrating Induction and Deduction for Finding Evidence of Discrimination

Salvatore Ruggieri ·
Dino Pedreschi ·
Franco Turini

Abstract We present a reference model for finding (prima facie) evidence of discrimination in datasets of historical decision records in socially sensitive tasks, including access to credit, mortgage, insurance, labor market and other benefits. We formalize the process of direct and indirect discrimination discovery in a rule-based framework, by modelling protected-by-law groups, such as minorities or disadvantaged segments, and contexts where discrimination occurs. Classification rules, extracted from the historical records, allow for unveiling contexts of unlawful discrimination, where the degree of burden over protected-by-law groups is evaluated by formalizing existing norms and regulations in terms of quantitative measures. 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. Key legal concepts and reasonings are then used to drive the analysis on the set of classification rules, with the aim of discovering patterns of discrimination, either direct or indirect. Analyses of affirmative actions, favoritism and argumentation against discrimination allegations are also modelled in the proposed framework. Finally, we present an implementation, called LP2DD, of the overall reference model that integrates induction, through data mining classification rule extraction, and deduction, through a computational logic implementation of the analytical tools. The LP2DD system is put at work on the analysis of a dataset of credit decision records.

Keywords Direct Discrimination · Indirect Discrimination · Affirmative Actions · Classification rules · Data mining · Knowledge Discovery · Logic Programming

1 Introduction

Civil right laws prohibit discrimination on the basis of race, color, religion, nationality, sex, marital status, age and pregnancy in a number of settings, including: credit and insurance; sale, rental, and financing of housing; personnel selection and wages; access to public accommodations, education, nursing homes, adoptions, and health care.

S. Ruggieri (✉) · D. Pedreschi · F. Turini,
Dipartimento di Informatica, Università di Pisa,
Largo B. Pontecorvo 3, 56127 Pisa, Italy.
E-mail: {ruggieri,pedre,turini}@di.unipi.it

For the key legal references, we refer the reader to the Australian Legislation (2010); European Union Legislation (2010); United Nations Legislation (2010); U.K. Legislation (2010); U.S. Federal Legislation (2010). Several authorities (regulation boards, consumer advisory councils, commissions) monitor and report on discrimination compliances. For instance, the European Commission publishes an annual report on the progress in implementing the Equal Treatment Directives by the member states (see Bell et al (2007)); and in the U.S.A. the Attorney General reports to the Congress about the annual referrals to the Equal Credit Opportunity Act. Also, jurisprudence accounts for a large body of cases, as reported by Ellis (2005); Lerner (1991); Schiek et al (2007). From the research side, the literature in economics and social sciences has given evidence of unfair treatment in racial profiling and redlining in Calem et al (2004); Squires (2003); mortgage lending in LaCour-Little (1999); consumer market in Riach and Rich (2002); Yinger (1998); credit and housing in Dymski (2006); personnel selection in Hunter (1992); and wages in Kuhn (1987).

Given the current state of the art of decision support systems (DSS), socially sensitive decisions may be taken by automatic systems, e.g., for screening or ranking applicants to a job position, to a loan, to school admission and so on. For instance, data mining and machine learning classification models are constructed on the basis of historical data exactly with the purpose of learning the distinctive elements of different classes, such as good/bad debtor in credit/insurance scoring systems (see Baesens et al (2003); Hand and Henley (1997); Thomas (2000)) or good/bad worker in personnel selection (see Chien and Chen (2008)). When applied for automatic decision making, DSS can potentially guarantee more uniform decisions, but still they can be discriminating in the social, negative sense. Moreover, the decisions taken by those systems may be hard to be stated in intelligible terms, even if their internals are disclosed as in a case before a court. In fact, a DSS is often the result of merging/weighting several hand-coded business rules and routinely built predictive models which are black-box software due to technical (e.g., neural networks), legacy (e.g., programming languages), or proprietary reasons. Currently, what the state of the art can offer is the verification of an hypothesis of possible discrimination by means of statistical analysis of past decision records. On the contrary, we aim at *extracting* contexts of possible discrimination supported by legally-grounded measures of the degree of discrimination suffered by protected-by-law groups in such contexts. Reasoning on the extracted contexts can support all the actors in an argument about possible discriminatory behaviors. The DSS owner can use them both to prevent incurring in discriminatory decisions, and as a means to argument against allegations of discriminatory behavior. A complainant in a case can use them to find specific situations in which there is a prima facie evidence of discrimination against groups she belongs to. Finally, control authorities can base the fight against discrimination on a formalized process of intelligent data analysis.

However, the actual discovery of discriminatory situations and practices, hidden in the decision records under analysis, may reveal an extremely difficult task. The reason for this difficulty is twofold. On the one side, a huge number of possible contexts may, or may not, be the theater for discrimination. To see this point, consider the case of gender discrimination in credit approval: although an analyst may observe that no discrimination occurs in general, i.e., when considering the whole available decision records, it may turn out that it is extremely difficult for aged women to obtain car loans. Many small or large niches may exist that conceal discrimination, and therefore all possible specific situations should be considered as candidates, consisting of all possible combinations of variables and variable values: personal data, demographics,

social, economic and cultural indicators, etc. Clearly, the anti-discrimination analyst is faced with a huge range of possibilities, which make her work hard: albeit the task of checking some known suspicious situations can be conducted using available statistical methods, the task of discovering niches of discrimination in the data is unsupported. We call this issue the *inductive* problem in discrimination discovery.

On the other side, discrimination is rarely defined in rigorous and universal terms. First, protected-by-law groups, such as minorities and disadvantaged people, are sometimes not fully identified, leaving space for ambiguous issues such as in the debate about multiple, intersectional and compound discrimination discussed in ENAR (2007). Second, the interpretation of existing legislations lead to different quantitative measures of discrimination and, a fortiori, to different thresholds between what is legal and illegal. Third, discrimination can be hidden behind apparently neutral practices, known as indirect discrimination, that must be unveiled by some deductive reasoning exploiting additional knowledge, which we call background knowledge. Fourth, a few policies, known as affirmative actions, that favor minorities are allowed, encouraged or even enforced by laws. Finally, in case a prima-facie evidence of discrimination is found in the data, the anti-discrimination analyst has still to consider possible argumentations of the respondent, e.g., in opposing a genuine occupational requirement justification. We call these issues the *deductive* problem in discrimination discovery.

In this paper, we propose a reference model for the process of discrimination analysis and discovery in DSS. We assume that a DSS is a black-box predictive model, whose input is a case consisting of attribute-value pairs (e.g., applicant data) and the output is a class value (e.g., a yes/no decision). The discovery of contexts of discrimination is formalized by an "inductive+deductive" approach. The inductive part consists of extracting classification rules from the set of historical decision records. We generalize the approach of Pedreschi et al (2008) and show how a comprehensive repertoire of discrimination measures, encompassing all the notions that we found in the juridical literature, and of their statistical significance can be defined in terms of the contingency table of the extracted classification rules. The deductive part consists of rule meta-reasoning over the set of extracted rules and, possibly, additional background knowledge. We show how the anti-discrimination analyst can reason uniformly about the concepts of direct discrimination, indirect discrimination, affirmative actions, favoritism, and genuine occupational requirement argumentations. Notice that the use of a combination of deduction and induction is ubiquitous in the Artificial Intelligence field; in applications to legal reasoning, it is employed, for example, in Stranieri et al (1999) where production rules are combined with case-based reasoning, and in Zeleznikow et al (1994) where rules are combined with neural networks. Our proposed approach is implemented in the LP2DD system (Logic Programming to Discover Discrimination), which is intended as a tool supporting discrimination analysis and discovery. Despite its name, LP2DD also integrates algorithms for frequent pattern mining and R procedures for computing statistical confidence intervals. We describe the architecture of the LP2DD system and show how the various elements of our approach are coded by analyzing the public domain German credit dataset (see Newman et al (1998)).

1.1 Plan of the Paper

The paper is organized as follows. After setting up the basic concepts on classification rules, logic programming notation and the experimental data, the proposed reference

model is discussed in Sec. 3. Then, ratio-based and difference-based families of quantitative measures of discrimination are studied in Sect. 4. The statistical significance of each measure is accounted for in Sect. 5. The induction of classification rules for discrimination analysis is presented in Sect. 6. Rule meta-reasoning follows for tackling: direct discrimination in Sect. 7, indirect discrimination in Sect. 8, argumentation of the respondent in Sect. 9, affirmative actions and favoritism in Sect. 10. The LP2DD system architecture and sample analyses over the German credit dataset are presented in Sect. 11. Finally, Sect. 12 reports related issues, open directions and conclusions.

2 Preliminaries

2.1 Frequent Classification Rules

We recall the notions of itemsets, association rules and classification rules from standard definitions by Agrawal and Srikant (1994); Tan et al (2006). 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. Continuous domain can be accounted for by first discretizing values into ranges. A c-item is called a class item. An item is any a-item. Let I be the set of all items.

A transaction is a subset of I, with exactly one a-item for every attribute a. A database of transactions, denoted by $\mathcal{D}$, is a set of transactions. An itemset $\mathbf{X}$ is a subset of I. We denote by 2^I the set of all itemsets. As usual in the literature, we write $\mathbf{X}, \mathbf{Y}$ for $\mathbf{X} \cup \mathbf{Y}$. For a transaction T, we say that T verifies $\mathbf{X}$ if $\mathbf{X} \subseteq T$. It is worth noting, then, that $\mathbf{X}, \mathbf{Y}$ characterizes a set of transactions that satisfy both $\mathbf{X}$ and $\mathbf{Y}$. The absolute support of an itemset $\mathbf{X}$ w.r.t. a non-empty transaction database $\mathcal{D}$ is the number of transactions in $\mathcal{D}$ verifying $\mathbf{X}$: $asupp(\mathbf{X}) = |\{\ T \in \mathcal{D} \mid \mathbf{X} \subseteq T\ \}|$, where $|\ |$ is the cardinality operator. The (relative) support of $\mathbf{X}$ is the ratio of transactions verifying $\mathbf{X}$ over the total number of transactions: $supp(\mathbf{X}) = asupp(\mathbf{X})/|\mathcal{D}|$.

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 $\mathbf{X} \to \mathbf{Y}$ is a *classification rule* if $\mathbf{Y}$ is a class item and $\mathbf{X}$ contains no class item. The support of $\mathbf{X} \to \mathbf{Y}$ is defined as: $supp(\mathbf{X} \to \mathbf{Y}) = supp(\mathbf{X}, \mathbf{Y})$. The coverage of $\mathbf{X} \to \mathbf{Y}$ is: $cov(\mathbf{X} \to \mathbf{Y}) = supp(\mathbf{X})$. The confidence of $\mathbf{X} \to \mathbf{Y}$, defined when $supp(\mathbf{X}) > 0$, is:

$$conf(\mathbf{X} \to \mathbf{Y}) = supp(\mathbf{X}, \mathbf{Y})/supp(\mathbf{X}).$$

Support, coverage and confidence range over $[0, 1]$. 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 only itemsets without negations. Since the seminal paper by Agrawal and Srikant (1994), many well explored algorithms have been designed in order to extract the set of *frequent* itemsets, i.e., itemsets with a specified minimum support. A survey on frequent pattern mining is due to Han et al (2007); a survey on interestingness measures for association rules is reported by Geng and Hamilton (2006); and, finally, a repository of implementations is maintained by Goethals (2010).

2.2 Logic Programming

We use standard notation for Prolog programs as in the textbook of Sterling and Shapiro (1994). A (Horn) clause A `:-` $B_1, \ldots, B_n$`.`, with $n \geq 0$, is a first order formula where $A, B_1, \ldots, B_n$ are literals, "`:-`" is the reverse implication connective, and "`,`" is the conjunction connective. Negation is denoted by `\+`. When $n = 0$, the program clause is called a fact, and it is written as A`.` A goal is `:-` $B_1, \ldots, B_n$`.`, where $B_1, \ldots, B_n$ are literals. Variable names start with capital letter. "_" denotes an anonymous variable.

A logic program is a finite set of clauses. A Prolog programs is a logic program whose operational semantics is SLDNF-resolution via the leftmost selection rule (see Apt (1997) for technical details). Non-logical predicates include arithmetic assignment (`is`) and comparison predicates (`<`, `=<`, `=\=`, `>=`, `>`). The empty list is denoted by `[]`. The list constructor is `[.|.]`.

2.3 The German credit case study

We will report some analyses over the public domain German credit dataset, publicly available from the UCI repository of machine learning datasets maintained by Newman et al (1998). The dataset consists of 1000 records over bank account holders. It 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. Finally, the class attribute takes values representing the good/bad creditor classification of the bank account holder.

3 Reference Model

The main goal of our research is to provide DSS owners and control authorities, from now on the *users*, with a general framework in support of discrimination analysis and discovery. In this section, we introduce a reference model for the overall process. Fig. 1 depicts our proposal.

3.1 Input Pool

The discrimination analysis starts from an *input pool* provided by the user. The input pool is a set of cases, e.g., application forms, credit requests, and skill tests, which are described by a collection of attribute values. Cases include the attributes taken as input by the DSS, e.g., age of applicant, amount requested, and job type, and, possibly, other attributes providing additional information which is not (or cannot legally be) input for the DSS, such as the race of applicants, their ethnic origin or disability. As an example, the input pool for the German credit case study is a table:

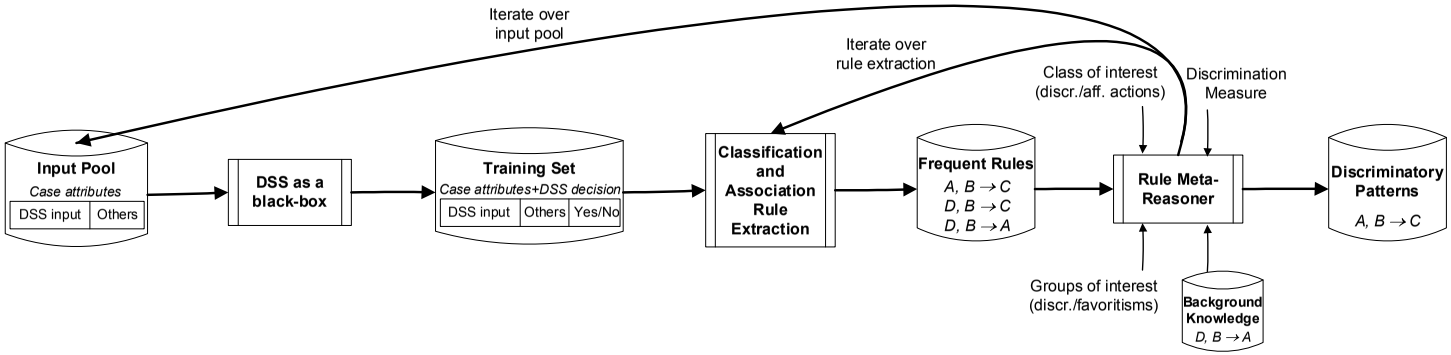

Fig. 1 Reference Model for Analysing and Reasoning on Discrimination in DSS.

pers_status	age	job	purpose	credit_amnt	...
female_single	gt_52	self_emp	furniture	5_10_k	...
male_married	30_to_35	unemp	new_car	10_15_k	...
...	...	...	...	...	...

recording a row with the data about each credit application. Most of the data are used by a DSS for screening applications. The attribute `personal_status` is an example of an attribute that is not or cannot legally be input of the DSS. Still, it can be reasonably available for discrimination analysis, because it is recorded in the bank registry of clients, or it is retrieved from external sources of data accessible to an anti-discrimination authority. Notice, however, that there exist others attributes, such as race, that might not be available at all at the detail of single persons (but only at an aggregate level).

It is worth noting that DSS are typically built only for application areas with abundance of data. On the contrary, the literature on discrimination analysis (see e.g., Riach and Rich (2002); Rorive (2009)) has typically faced the problem of collecting input pools through controlled experimentation known as *situation testing*, where a set of individuals is matched for all relevant characteristics other than those expected to lead to discrimination.

One way of defining the input pool is to collect data from the set of applicants (called "application pool" by Kaye and Aickin (1992)) to a position, benefit, etc, as for the German credit dataset. In this case, the reference population for discrimination analysis is the set of applicants. The application pool allows for comparing the protected-by-law groups against a homogeneous population with respect to interest in obtaining the benefit. In some cases, however, the application pool approach may be biased. Kaye and Aickin (1992) mention the *Dothard vs Rawlinson* case, where a woman who applied for a job as prison guard was rejected because she could not meet the minimum requirements of weight and height. In the pool of applicants, women were not under-represented in obtaining a job position. However, the court accepted the argument that those requirements "would exclude 41.13% of the female [labour] population of the U.S. while excluding less than 1% of the male population". Since women not meeting the requirements were discouraged to apply, the application pool approach is biased. While it would be impossible to build an input pool of the whole population, it is the case that a representative sample can be selected starting from the application pool, or that a syntectic dataset can be built starting from available data and following the approach of situation testing. In both cases, the resulting dataset would then be the input pool in the reference model of Fig. 1.

3.2 DSS Decisions

The DSS under analysis is supposed to be a *black-box* predictive software, yielding a decision for each case in the input pool. The input pool enriched with the DSS decision represents the output of the DSS, which we call the *training set*. An example of the training set for the German credit case study is the following:

pers_status	age	job	purpose	credit_amnt	...	class
female_single	gt_52	self_emp	furniture	5_10_k	...	bad
male_married	30_to_35	unemp	new_car	10_15_k	...	good
...	...	...	...	...	...	...

where the first application is assigned the "bad" credit class label, i.e., credit is denied, and the second application is assigned the "good" credit class, i.e., credit is granted.

Why do we assume the DSS to be a black-box? This view is general enough to deal with a DSS without any intelligible or symbolic representation, as in the case of neural networks and legacy programming languages. However, one could object that when the DSS internals are intelligible and they can be disclosed (e.g., when the owner is forced to by a court) the discrimination analysis should be given the DSS logic itself as an input. We argue that this is not the case. As an example, consider an hypothetical DSS whose logic consists of the following rules:

```
IF own car = yes THEN credit = no
                 ELSE IF driver = yes THEN credit = yes
                                      ELSE credit = no
```

These rules seem not to discriminate in any way against women. For the following contrived input pool, they lead to the decisions reported in the last column.

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

Here, `driver` and `own car` are attributes used by the DSS, whilst `sex` and `ZIP` are additional attributes added to the input pool for discrimination discovery. By looking at the decisions, we observe that women living in the area with `ZIP = 101` are assigned no credit with frequency 100%, while men living in the same area are assigned no credit with frequency 50%. The ratio of the two frequencies, namely 2, will be later on defined as a measure of discrimination. If a ratio of 2 would be deemed unacceptable by the law, and the provided input pool would be representative of the underlying population, we could conclude that the DSS decisions have discriminatory *effects* for women living in the area `ZIP = 101`. Although the DSS logic has no explicit discriminatory intent, its analyses are not complete enough to prevent what is known in the literature as *indirect* or *systematic discrimination*. It is a general principle that the law prohibits not only explicit discrimination, but also any practice whose *effects* (intentional or not) are discriminatory. We maintain that, in order to unveil discriminatory effects, the user has to reason on the training set, not over the DSS logic.

3.3 Inductive Component

Starting from the training set, the inductive part of the reference model consists of extracting the set of *frequent classification and association rules*, namely those classification and association rules whose support is greater or equal than a user-specified minimum threshold. The minimum support threshold allows for considering rules that apply in a sufficiently large number of cases, accordingly to some requirements stated by law, regulations and past sentences. The value of the threshold is a parameter of the reference model, and, as shown in Fig. 1, the extraction of rules can be iterated when searching for smaller niches of discriminatory contexts.

In our approach, we distinguish three sets of extracted rules.

Potentially discriminatory (PD) classification rules have the form: $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$, where $\mathbf{A}$ is an itemset denoting protected-by-law groups, and called a PD itemset; $\mathbf{B}$ is an itemset denoting a context of discrimination; and $\mathbf{C}$ is a class item, denoting a decision. As an example, `race=black, purpose=new_car` $\rightarrow$ `class=bad` is a PD rule about denying credit to blacks (the potentially discriminated group) among those applying for the purpose of buying a new car. On the one side, PD rules explicitly mention groups potentially subject to discrimination, under the assumption that such groups can be denoted by attributes available in the input pool dataset. On the other side, the extraction of rules allows for solving the *inductive issue* mentioned in the introduction, by letting contexts $\mathbf{B}$ of possible discrimination emerge.

Potentially non-discriminatory (PND) classification rules have the form: $\mathbf{D}, \mathbf{B} \rightarrow \mathbf{C}$, where now the premise of the rule does not explicitly denote any protected-by-law group, but only PND itemsets $\mathbf{D}$ and $\mathbf{B}$. While PND rules seem unrelated to discrimination analysis, we will show that they can unveil indirect discrimination, where an apparently neutral condition turns out to have effects on protected-by-law groups. As an example, `zip=1234, purpose=new_car` $\rightarrow$ `class=bad` is a PND rule about denying credit to people from a certain neighborhood applying for the purpose of buying a new car. However, if people from that neighborhood are mostly blacks, then the PND involves approximatively the same individuals as `(race=black, zip=1234), purpose=new_car` $\rightarrow$ `class=bad`.

Finally, association rules of the form $\mathbf{D}, \mathbf{B} \rightarrow \mathbf{A}$ are extracted as well, where $\mathbf{A}$ is a PD itemset, and $\mathbf{D}$, $\mathbf{B}$ are not. Such rules are useful in the deductive component (see later on) to relate the distribution of protected-by-law groups to an apparently neutral condition $\mathbf{D}$ in a context $\mathbf{B}$. As an example, `zip=1234, purpose=new_car` $\rightarrow$ `race=black` is a PND rule about the proportion of blacks over people from a certain neighborhood and applying for the purpose of buying a new car.

3.4 Deductive Component

We base discrimination analysis on a few measure of discrimination for a classification rule, defined starting from its contingency table and, possibly, including a test of its statistical significance. The measures of discrimination are introduced by formalizing existing laws and regulations. The deductive part of the reference model consists of unveiling PD classification rules $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$ whose measure is above a threshold modelling the boundary between what is legal and what is illegal. We will translate several legal concepts and reasonings into rule filtering and deduction:

- *direct discrimination*, is unveiled by looking at PD rules $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$, where $\mathbf{C}$ denies some benefit;
- *indirect discrimination*, is unveiled by looking at PND rules $\mathbf{D}, \mathbf{B} \rightarrow \mathbf{C}$, where $\mathbf{C}$ denies some benefit, and by relating the apparently neutral condition $\mathbf{D}$ to some (unknown) protected-by-law group $\mathbf{A}$;
- *argumenting against discrimination allegations* supported by a PD rule $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$, where $\mathbf{C}$ denies some benefit, is modelled by searching for PND rules $\mathbf{D}, \mathbf{B} \rightarrow \mathbf{C}$ such that $\mathbf{D}$ is a legitimate requirement, having the same effects of the PD rule;
- *affirmative actions* are unveiled by looking at PD rules $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$, where $\mathbf{C}$ grants some benefit;

– *favoritism* is unveiled by looking at PD rules $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$, where $\mathbf{C}$ grants some benefit and $\mathbf{A}$ models advantaged groups.

The *rule meta-reasoner* component, described in depth in Sects. 7-10, supports the user in the discrimination analysis by providing various measures of discrimination and meta-rule deductions. The rule meta-reasoner is an interactive analytical tool for exploring and reasoning about classification rules, either the extracted ones or others that can be inferred from them, in search of *prima facie* evidence of discrimination. As the exploration may end up into a niche of the input pool (e.g., applicants from a specific region), the user can iterate the process over a different input pool and/or lower minimum support.

In addition to the set of extracted rules, the analysis may also need to refer to *background knowledge*, namely information from external sources or common sense, such as census data, household surveys, administrative records. We assume that also background knowledge is provided in the form of association rules of the form $\mathbf{D}, \mathbf{B} \rightarrow \mathbf{A}$ or $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{D}$.

The output of the discrimination analysis is a set of *discriminatory patterns*, namely PD classification rules that hold over the training set and such that they unveil groups subject to discrimination and contexts where discrimination took place. Discriminatory patterns are required to overcome admissible legal argumentations such as minimum number of involved individuals, statistical significance of the conclusion, and legitimate requirement justifications.

4 Measures of Discrimination

The basic problem in the analysis of discrimination 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 a decision (say, credit denial). 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 $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$, together with its confidence with respect to the underlying decision dataset - namely, how often such a rule is true in the dataset itself. In this section, we first discuss in Sect. 4.1 how to denote the potentially discriminated (PD) groups that are protected by the law against discrimination, and, consequently, how to locate them in a classification rule premise. Then, we introduce a family of measures of the degree of discrimination of a potentially discriminatory rule $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$, where $\mathbf{A}$ is a non-empty PD itemset and $\mathbf{B}$ is not a PD 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 Potentially Discriminated Groups

Civil rights laws explicitly identify the groups to be protected against discrimination, e.g., women or black people. With our syntax, those groups can be represented as items, e.g., `sex=female` or `race=black`. Therefore, we can assume that the laws provide us with a set I_d of items, which we call potentially discriminatory (PD) items, denoting

groups of people that could be potentially discriminated. Given a classification rule `sex=female, car=own` → `credit=no`, it is straightforward to separate in its premise `sex=female` from `car=own`, in order to reason about potential discrimination against women with respect to people owining a car.

However, discrimination typically occurs for subgroups rather than for the whole group. For instance, we could be interested in discrimination against older women. With our syntax, this group would be represented as the itemset `sex=female, age=older`. The intersection of two disadvantaged minorities (here, `sex=female` and `age=older`) is a, possibly empty, smaller (even more disadvantaged) minority as well. As a consequence, we have to generalize from the notion of potentially discriminatory *item* to the one of potentially discriminatory (PD) *itemset*. We denote by $\mathcal{I}_d$ the set of PD itemsets. As a first proposal, $\mathcal{I}_d$ could be defined by admitting itemsets built on PD items only, i.e., $\mathcal{I}_d$ is of the form 2^{I_d} for a set of items I_d.

Definition 1 A set of itemsets $\mathcal{I}$ is generated by the set of items I if $\mathcal{I} = \{\mathbf{A} \mid \mathbf{A} \subseteq I\}$.

Again, provided with a classification rule `sex=female, age=older, car=own` → `credit=no` we are in the position to isolate the potentially discriminated group in the premise by selecting those items that belong to I_d.

Generated itemsets are not general enough. Consider the case of "gender-plus" allegations, an expression coined by the U.S. courts to describe conducts breaching the law on the ground of sex-plus-something-else, e.g. discrimination against women working in the army in obtaining a new job position. With our syntax, this group would be represented as the itemset `sex=female, job=army`. Provided with a rule `sex=female, job=army` → `hire=no` we have now the problem of separating the PD group in the premise. In fact, using the definition of PD itemset, since `job=army` is not a PD item, we would separate `sex=female` from `job=army`, i.e., we would consider discrimination against females over the people working in the army. This is not what we were originally looking for. An even worse case is concerned with the definition of minorities. Assume to be interested in discrimination against white people living in a specific neighborhood (because they are minorities there) albeit neither being white nor living in some neighborhood are groups of interest for discrimination. In other words, discrimination may be the result of several joint characteristics that are not necessarily discriminatory in isolation. This case is called intersectional or compound discrimination in ENAR (2007).

Stated formally, fixing $\mathcal{I}_d$ to a generated set may be not enough general to cover all possible groups of interest in the discrimination analysis. Thus, the only formal property we require for $\mathcal{I}_d$ is that the intersection of two itemsets belonging to it (two disadvantaged groups) belongs to it as well (it is a disadvantaged group as well). This property is called downward closure by Pedreschi et al (2008).

Definition 2 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 $\mathbf{A}_1, \mathbf{A}_2 \in \mathcal{I}$.

Downward closure allows us to account for multiple causes of discrimination, called multiple discrimination in ENAR (2007). As an example, an older woman can be subject to discrimination against elder people (itemset `age=elder`), against women (itemset `sex=female`) or against both (itemset `age=elder, sex=female`). On the technical side, the downward closure property is sufficient for separating PD itemsets in the premise of a classification rule. In fact, given $\mathbf{X} \rightarrow \mathbf{C}$, the itemset $\mathbf{X}$ can be uniquely

split 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 $\mathcal{I}_d$.

Definition 3 A classification rule $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$ is called potentially discriminatory (PD rule) if $\mathbf{A}$ is non-empty, and potentially non-discriminatory (PND rule) otherwise.

PD rules explicitly state conclusions involving potentially discriminated groups. The next subsections are devoted to the proposal of a few quantitative measures of the "burden" imposed over such groups and unveiled by a discovered PD rule.

4.2 Ratio Measures

Unfortunately, there is no uniformity nor general agreement on a standard quantification of discrimination by legislations. A general principle mentioned by Knopff (1986) is to consider group under-representation as a quantitative measure of the qualitative requirement that people in a group are treated "less favorably" (see European Union Legislation (2010); U.K. Legislation (2010)) than others, or such that "a higher proportion of people without the attribute comply or are able to comply" (see Australian Legislation (2010)) to a qualifying criterium. As a first proposal, we recall from Pedreschi et al (2008) 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 Let $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$ be a PD 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. While this means that women are discriminated among car owners, notice that we cannot conclude that being a woman is the actual reason of discrimination (see Sect. 9 for a discussion). An alternative way, yet equivalent[1], 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 over-representation in benefit denying, or, equivalently, of under-representation in benefit granting.

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 Australian Legislation (2010)(b), 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.

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

Definition 5 Let $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$ be a PD 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. Legislation (2010). 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 6 Let $\mathtt{a} = \mathtt{v}_1, \mathbf{B} \rightarrow \mathbf{C}$ be a PD 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 European Union Legislation (2010)(a), where discrimination "shall be taken to occur where one person is treated less favorably than another who 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. Federal Legislation (2010)(d) goes further 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 employment discrimination literature (see Gastwirth (1992)), 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.

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	$n_2 - a_2$

$$p_1 = a_1/n_1 = conf(\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}) \qquad p_2 = a_2/n_2 = conf(\neg\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C})$$

$$p = (a_1 + a_2)/(n_1 + n_2) = conf(\mathbf{B} \rightarrow \mathbf{C})$$

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

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

Fig. 2 Contingency table for a classification rule.

Definition 7 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. 2. Each cell in the table is filled in by the number of transactions in the transaction database $\mathcal{D}$ satisfying $\mathbf{B}$ and the coordinates (a.k.a., their absolute support). 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 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 $\{olift(c),\ clift(c)\} \leq slift(c) \leq elift(c) \leq 1$.*

Proof Let us show the first conclusion, namely $\{olift(c),\ clift(c)\} \geq slift(c) \geq elift(c)$ when $elift(c) \geq 1$. The other conclusion follows by similar reasonings.

[$slift(c) \geq elift(c)$].

Consider the contingency table from Fig. 2. $elift(\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}) \geq 1$ can be rewritten as $a_1/n_1 \geq (a_1 + a_2)/(n_1 + n_2)$, i.e., $a_1 n_2 \geq a_2 n_1$.

By definition, $slift(\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}) \geq elift(\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C})$ iff $conf(\neg\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}) \leq conf(\mathbf{B} \rightarrow \mathbf{C})$ iff $a_2/n_2 \leq (a_1 + a_2)/(n_1 + n_2)$. By elementary algebra, this equals to $a_2 n_1 \leq a_1 n_2$ which holds by hypothesis.

[$olift(c) \geq slift(c)$].

Let $p_1 = a_1/n_1$ and $p_2 = a_2/n_2$. By the previous step, $p_1/p_2 = slift(\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}) \geq elift(\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}) \geq 1$. This implies $1 - p_1 \leq 1 - p_2$ and then $(1 - p_2)/(1 - p_1) \geq 1$. Therefore, $p_1/p_2 \leq (p_1/p_2)(1 - p_2)/(1 - p_1) = olift(\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C})$.

[$clift(c) \geq slift(c)$].

Let $\mathbf{A}$ be $\mathtt{a} = \mathtt{v_1}$, and consider the following contingency table:

B	**C**	$\neg$**C**
$\mathtt{a} = \mathtt{v_1}$	a_1	$n_1 - a_1$
$\mathtt{a} = \mathtt{v_2}$	a_2	$n_2 - a_2$
...	...	...
$\mathtt{a} = \mathtt{v_k}$	a_k	$n_k - a_k$

where $\mathtt{a} = \mathtt{v_2}$ is as in Def. 6, i.e., $0 \neq a_2/n_2 \leq a_i/n_i$ for $i = 2 \ldots k$. This can be rewritten as $a_2 n_i \leq n_2 a_i$ for $i = 2 \ldots k$. By summing up all inequalities, we have $a_2 \Sigma_{i=2}^k n_i \leq n_2 \Sigma_{i=2}^k a_i$, i.e., $a_2/n_2 \leq \Sigma_{i=2}^k a_i / \Sigma_{i=2}^k n_i$ which is $conf(\mathtt{a} = \mathtt{v_2}, \mathbf{B} \rightarrow \mathbf{C}) \leq conf(\neg\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C})$. This readily implies $clift(\mathtt{a} = \mathtt{v_1}, \mathbf{B} \rightarrow \mathbf{C}) \geq slift(\mathtt{a} = \mathtt{v_1}, \mathbf{B} \rightarrow \mathbf{C})$. □

4.3 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 (**A** is `sex=female`) and male ($\neg$**A** is `sex=female`) treatment is assumed by courts as significant of discrimination against women. Therefore, we define next a version of extended and selection lift using differences.

Definition 8 Let $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$ be a classification rule. We define:

$$elift_d(\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}) = conf(\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}) - conf(\mathbf{B} \rightarrow \mathbf{C})$$
$$slift_d(\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}) = conf(\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}) - conf(\neg\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}).$$

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

Lemma 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 $slift_d(c) \leq elift_d(c) \leq 0$.*

Proof Let us show the first conclusion, namely $slift_d(c) \geq elift_d(c)$ when $elift_d(c) \geq 0$. The other conclusion follows by similar reasonings.

Consider the contingency table from Fig. 2. $elift_d(c) \geq 0$ can be rewritten as $p_1 - p \geq 0$, namely $p_1/p \geq 1$. By Lemma 1, this implies $slift(c) \geq elift(c)$, i.e., $p_1/p_2 \geq p_1/p$, which is equivalent to $p \geq p_2$. This implies $p_1 - p_2 \geq p_1 - p$, namely $slift_d(c) \geq elift_d(c)$, which is our conclusion. □

4.4 Discriminatory Classification Rules

Once we are provided with a quantitative measure of discrimination and a threshold between "legal" and "illegal" degree, we are in the position to isolate classification rules whose measure is below/above the threshold. The following notion generalizes α-protection[2] introduced by Pedreschi et al (2008), which boils down to a-protection w.r.t. $elift()$.

[2] 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 introduced.

Definition 9 (a-protection) Let $f()$ be one of the measures from Definitions 4-8, and $a \in \mathbb{R}$ a fixed threshold. A PD 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. As an example, the four-fifth rule of the U.S. Federal Legislation (2010)(d) sets $a = 1.25$ for the contrasted lift measure. Classification rules denying credit and with a measure below such a level are considered safe, whilst rules whose measure is greater or equal than such a level can then be considered a *prima facie* evidence of discrimination. As we will see in Sects. 7-10, specialisations of a-protection allow for modelling not only discrimination, but also affirmative actions and favoritism.

5 Statistical Significance

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 must be considered. This approach is customary in legal cases before courts, as reported by Gastwirth (1984, 1992); Piette and White (1999). 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. 2) 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. 2. If we interpret the records covered by c as a statistical sample of the overall population, we observe that the ratio and difference measures introduced in Sec. 4.2-4.3 can be interpreted as ratios and differences of two proportions, a subject of extensive studies in the field of statistical inference. From the textbook by Agresti (2002) and the research book by Fleiss et al (2003), we have that:

- $slift(c)$ is the ratio p_1/p_2 of two proportions, also known as the risk ratio or relative risk (RR);

- $slift_d(c)$ is the difference $p_1 - p_2$ of two proportions, also known as the risk difference (RD);
- $olift(c)$ is the odds ratio (OR) $p_1(1-p_2)/(p_2(1-p_1))$ of two proportions;
- $elift(c)$ 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) 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 as follows (see Agresti (2002); Farrington and Manning (1990); Fleiss et al (2003) for details).

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:

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

OR: Called $\hat{o} = p_1(1-p_2)/(p_2(1-p_1))$, the confidence interval for $\pi_1(1-\pi_2)/(\pi_2(1-\pi_1))$ is $[\hat{o}/e^d, \hat{o}e^d]$ where:

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

We refer the reader to Fleiss et al (2003) for Wald intervals of PAR; and to Leung and Kupper (1981) 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 the survey and comparison papers of Farrington and Manning (1990); Leung and Kupper (1981); Newcombe (1998); Tian et al (2008). 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 (see Agresti (2002)). Later on, we will consider the simple but effective plus-4 method from Agresti and Brian (2000), 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 the 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 Piette and White (1999) 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. 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.

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 $[L_1^f(\alpha, c), L_2^f(\alpha, c)]$ to denote the confidence interval for the contingency table of rule c at the significance level of $100(1-\alpha)\%$.

Definition 10 (a-protection) Let $f()$ be one of the measures from Definitions 4-8, and $a \in \mathbb{R}$ a fixed threshold. A PD classification rule $c = \mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$ is a-protective w.r.t. $f()$ at the significance level of $100(1-\alpha)\%$ if $L_2^f(\alpha, c) < a$. c is a-discriminatory at the significance level of $100(1-\alpha)\%$ if $L_1^f(\alpha, c) \geq a$.

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 $L_1^f(1, c) = L_2^f(1, c) = f(c)$. Therefore, the definition above is a conservative extension of Def. 9. In general, the higher the significance level is, the wider is the confidence interval. At 100% significance level, , we have $Z_{1-\alpha/2} = Z_1 = \infty$ and then the confidence intervals cover the 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 $L_1^f(\alpha, c) < a \leq L_2^f(\alpha, c)$ the rule c is neither a-discriminatory nor a-protective. Intuitively, there is no sufficient statistical evidence to draw a conclusion.

6 Induction of PD and PND rules

In this section, we discuss the algorithmic aspects of extracting classification rules, association rules and their contingency tables from the training set. This is the inductive step in the overall process of Fig. 1. We rely on frequent itemset mining algorithms from the knowledge discovery literature (see Agrawal and Srikant (1994); Han et al (2007)), with many efficient-in-practice implementations (see Goethals (2010)). Frequent itemset mining consists of extracting the itemsets having a support greater or equal than a specified minimum threshold. As discussed in Sect. 3, fixing a minimum support threshold for the classification and association rules under consideration is not a restriction, but rather it models the natural requirement that a rule is concerned with a sufficiently large number of (protected-by-law) individuals. How large it is a parameter of the reference model, and, as shown in Fig. 1, it can be changed when searching for smaller niches of discriminatory contexts.

The set of k-frequent itemsets and their support values, where k is the length of (a.k.a., the number of items in) the itemset, is denoted by $\mathcal{F}_k$. Most of the algorithms calculate frequent itemsets by yielding $\mathcal{F}_k$ for increasing $k = 1, 2, \ldots$. Fig. 3 reports the procedure **ExtractClassificationRules()** for extracting classification rules and their contingency tables by scanning k-frequent itemsets in increasing order of k. The

ExtractClassificationRules()
$N = |\mathcal{D}|$, $\mathcal{C} = \{$ class items $\}$, $\mathcal{L} = \emptyset$
ForEach k s.t. there exist k-frequent itemsets
$\mathcal{F}_k = \{$ k-frequent itemsets $\}$
$(\star)$ 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 $\mathbf{B} \rightarrow \mathbf{C}$ to $\mathcal{L}$ with $supp = a_1$ and $cov = n_1$
output $\mathbf{B} \rightarrow \mathbf{C}$ in $\mathcal{PND}$,
with contingency table $\begin{pmatrix} a_1 N & (n_1 - a_1)N \\ 0 & 0 \end{pmatrix}$
Else
$a_2 = supp(\mathbf{B} \rightarrow \mathbf{C}) - a_1$ // $\mathbf{B} \rightarrow \mathbf{C}$ found in $\mathcal{L}$
$n_2 = cov(\mathbf{B} \rightarrow \mathbf{C}) - n_1$
output $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$ in $\mathcal{PD}$,
with contingency table $\begin{pmatrix} a_1 N & (n_1 - a_1)N \\ a_2 N & (n_2 - a_2)N \end{pmatrix}$
EndIf
$(\star)$ mark $\mathbf{B} \rightarrow \mathbf{C}$ in $\mathcal{L}$
EndForEach
EndForEach

Fig. 3 Extraction of PD and PND classification rules, and their contingency tables. Lines annotated with $(\star)$ apply only when PD itemsets satisfy Def. 1.

procedure maintains the set of $(k-1)$-frequent itemsets. During the scan, an itemset $\mathbf{R}$ that includes a class item gives rise to a rule $c = \mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$. If the PD part $\mathbf{A}$ is empty, the rule is PND – otherwise it is a PD rule. In the former case, the contingency table of Fig. 2 falls down to computing a_1 and n_1, which can be calculated from the support values of $\mathbf{R}$ and $\mathbf{B}$ (with $\mathbf{B}$ itemset in $\mathcal{F}_{k-1}$). In the latter case, we have in addition to compute a_2 and n_2. We can resort to support and coverage of $\mathbf{B} \rightarrow \mathbf{C}$ by noting that $supp(\neg\mathbf{A}, \mathbf{B}, \mathbf{C}) = supp(\mathbf{B} \rightarrow \mathbf{C}) - supp(\mathbf{A}, \mathbf{B}, \mathbf{C})$ and $supp(\neg\mathbf{A}, \mathbf{B}) = supp(\mathbf{B}) - supp(\mathbf{A}, \mathbf{B}) = cov(\mathbf{B} \rightarrow \mathbf{C}) - supp(\mathbf{A}, \mathbf{B})$. To this end, during the scans we maintain the set $\mathcal{L}$ of PD rules of the form $\mathbf{B} \rightarrow \mathbf{C}$. When PD itemsets are generated (see Def. 1), memory can be saved by pruning $\mathcal{L}$ at an iteration $k+1$ by removing $\mathbf{B} \rightarrow \mathbf{C}$ if there is no PD itemset $\mathbf{A}$ such that $\mathbf{A}, \mathbf{B}, \mathbf{C}$ is k-frequent. In fact, $\mathbf{B} \rightarrow \mathbf{C}$ cannot be looked up any more: if there exists a $(k+h)$-frequent itemset $\mathbf{A'}, \mathbf{B}, \mathbf{C}$, with $\mathbf{A'}$ PD itemset, we would have that, for some $\mathbf{A} \subseteq \mathbf{A'}$, the itemset $\mathbf{A}, \mathbf{B}, \mathbf{C}$ is k-frequent; but since $\mathbf{A} \subseteq \mathbf{A'}$ implies that $\mathbf{A}$ is a PD itemset (from Def. 1), we would conclude the absurd that there exists $\mathbf{A}, \mathbf{B}, \mathbf{C}$ that is k-frequent.

Fig. 4 reports the procedure **ExtractAssociationRules()** for extracting association rules of the form $\mathbf{X} \rightarrow \mathbf{A}$. As before, we scan $\mathcal{F}_k$ for increasing k. This time we are interested in itemsets $\mathbf{R}$ not containing a class item. If the PD part $\mathbf{A}$ of $\mathbf{R}$ is not empty, then the association rule $\mathbf{X} \rightarrow \mathbf{A}$ can be produced in output, where $\mathbf{X} = \mathbf{R} \setminus \mathbf{A}$ is the PND part of $\mathbf{R}$. The support a_1 of the rule is the support of $\mathbf{R}$. The coverage n_1 is retrieved from the support of $\mathbf{X}$. To this end, during the scans we maintain the set $\mathcal{L}$ of PND itemsets $\mathbf{X}$. As in the case of the **ExtractClassificationRules()** pro-

ExtractAssociationRules()
$\mathcal{C}$ = { class items }, $\mathcal{L} = \emptyset$
ForEach k s.t. there exist k-frequent itemsets
$\mathcal{F}_k$ = { k-frequent itemsets }
($\star$) delete from $\mathcal{L}$ unmarked elements and unmark all the marked ones
ForEach $\mathbf{R} \in \mathcal{F}_k$ with $\mathbf{R} \cap \mathcal{C} = \emptyset$
$a_1 = supp(\mathbf{R})$
$\mathbf{A}$ = largest subset of $\mathbf{X}$ in $\mathcal{I}_d$
$\mathbf{X} = \mathbf{R} \setminus \mathbf{A}$
If $|\mathbf{A}| = 0$
add $\mathbf{X}$ to $\mathcal{L}$ with $supp = a_1$
Else
$n_1 = supp(\mathbf{X})$ // $\mathbf{X}$ found in $\mathcal{L}$
output $\mathbf{X} \rightarrow \mathbf{A}$
with contingency table $\begin{pmatrix} a_1 N & (n_1 - a_1)N \\ 0 & 0 \end{pmatrix}$
EndIf
($\star$) mark $\mathbf{X}$ in $\mathcal{L}$
EndForEach
EndForEach

Fig. 4 Extraction of association rules of the form $\mathbf{X} \rightarrow \mathbf{A}$, and their contingency tables. Lines annotated with ($\star$) apply only when PD itemsets satisfy Def. 1.

cedure, memory can be saved by pruning $\mathcal{L}$ under the assumption that PD itemsets are generated (see Def. 1).

As a final note, we point out that **ExtractClassificationRules()** and **ExtractAssociationRules()** are specialised procedures for extracting PD and PND classification rules, and association rules of the form $\mathbf{X} \rightarrow \mathbf{A}$. This solution is more efficient than first extracting all association rules and then filtering the ones we are interested in. Under some conditions, the approach can go further and be integrated within the phase of frequent itemset mining. For instance, Webb (2000) proposes a solution to extract classification rules with lift (whose $elift()$ is a generalization) above a minimum threshold. Another approach dealing with generic operators over a contingency table is proposed by Rauch and Simunek (2005) and implemented in the 4ft-Miner system by Rauch and Simunek (2010).

7 Direct Discrimination

From this section on, we formalize various legal concepts in discrimination analysis and discovery as reasonings over the set of extracted classification and association rules. This covers the rule meta-reasoner in Fig. 1.

We start by considering direct discrimination, which, accordingly to Ellis (2005), occurs "where one person is treated less favorably than another". For the purposes of making a *prima facie* evidence in a case before the court, it is enough to show that only one individual has been treated unfairly in comparison to another. However, this may be difficult to prove. The complainant may then use aggregate analysis to establish a regular pattern of unfavorable treatment of the disadvantaged group she belongs to. This is also the approach that control authorities and internal auditing may undertake in analysing historical decisions in search of contexts of discrimination against protected-by-law groups.

In direct discrimination, we assume that the input pool dataset contains attributes to denote potentially discriminated groups. This is a reasonable assumption for attributes such as sex and age, or for attributes that can be explicitly added by control authorities, such as pregnancy status. The next section will consider the case of attributes not available at all or not even collectable. Under our assumption, regular patterns of discrimination can then be identified by looking at PD classification rules of the form:

$$\mathbf{A}, \mathbf{B} \rightarrow \texttt{benefit} = \texttt{no},$$

i.e., where the consequent consists of denying a benefit (a loan, school admission, a job, etc.). 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. Rules of the form above can then be screened by searching for values of the adopted measure greater than a fixed, legally grounded, threshold a.

Definition 11 (direct discrimination) Let $f()$ be one of the measures from Definitions 4-8, and $a \in \mathbb{R}$ a fixed threshold. A PD classification rule $c = \mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$, where $\mathbf{C}$ denies some benefit and $\mathbf{A}$ refers to a disadvantaged group, is an a-directly discriminatory rule w.r.t. $f()$ if $f(c) \geq a$.

The notion of a-direct discrimination boils down to "a-discrimination of PD classification rules denying benefit". Also, statistical significance of the measure $f()$ can be readily taken into account, as done for a-discrimination, to define a-directly discriminatory rules at a significance level of $100(1 - \alpha)\%$.

8 Indirect Discrimination

The E.U. Directives (see European Union Legislation (2010); Tobler (2008)) provide a broad definition of indirect (also known as systematic) discrimination as occurring "where an apparently neutral provision, criterion or practice would put persons of a racial or ethnic origin at a particular disadvantage compared with other persons". In other words, the actual result of the apparently neutral provision is the same as an explicitly discriminatory one. In our framework, the "actual result" is modelled by a PD rule $c = \mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$ that is a-directly discriminatory, while an "apparently neutral provision" is modelled by a PND rule $\mathbf{B} \rightarrow \mathbf{C}$ where PD itemsets do not occur at all. The issue with unveiling indirect discrimination is that the actual result c is unavailable, e.g., because the input pool does not contain attributes to denote the potentially discriminated groups. For instance, the information on a person's race is typically not available and, in many countries, not even collectable. In indirect discrimination, the problem consists of deducting the actual result c starting from the set of PND rules, and, possibly, from additional knowledge.

A typical legal case study of indirect discrimination is concerned with redlining (see e.g., the *Hussein vs Saints Complete House Furniture* case reported by Makkonen (2006)), which inspires the following example. Assume that a Liverpool furniture store refuse to consider 99% of applicants to a job from a particular postal area `zip=1234` which had a high rate of unemployment. An extracted classification rule `zip=1234, city=Liverpool` $\rightarrow$ `app=no` with confidence 99% is apparently neutral with respect to race discrimination, though the average refusal rate in the Liverpool area is much lower, say 9%. With our notation, the rule `city=Liverpool` $\rightarrow$ `app=no` has then confidence

9%. Assume now to know that 50% of the population in the postal area `zip=1234` is black, i.e., that the association rule `zip=1234, city=Liverpool` $\rightarrow$ `race=black` has confidence 50%. It is now legitimate to ask ourselves whether from such rules, one can conclude that blacks in the postal area are discriminated; or, formally, that a discrimination measure (e.g., the extended lift) of the rule:

$$\texttt{(zip=1234, race=black), city=Liverpool} \rightarrow \texttt{app=no}, \tag{1}$$

is particularly high, where the PD itemset is the one written in parenthesis. Consider the following contingency tables for a known PND classification rule $\mathbf{D}, \mathbf{B} \rightarrow \mathbf{C}$ (left-hand side) and for an unknown PD rule $(\mathbf{A}, \mathbf{D}), \mathbf{B} \rightarrow \mathbf{C}$ (right-hand side):

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

$\mathbf{B}$	$\mathbf{C}$	$\neg\mathbf{C}$
$\mathbf{A}, \mathbf{D}$	a_1	$n_1 - a_1$
$\neg(\mathbf{A}, \mathbf{D})$	a_2	$n_2 - a_2$

Given the left-hand side contingency table, we want to derive a lower bound for $p_1 = a_1/n_1 = conf((\mathbf{A}, \mathbf{D}), \mathbf{B} \rightarrow \mathbf{C})$. The idea is to consider PD itemsets $\mathbf{A}$ that are approximatively equivalent to $\mathbf{D}$ in the context $\mathbf{B}$, namely such that:

$$\beta = conf(\mathbf{D}, \mathbf{B} \rightarrow \mathbf{A})$$

is near to 1. β is 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 βm_1 transactions satisfying $(\mathbf{A}, \mathbf{D}), \mathbf{B}$ (those satisfying $\mathbf{D}, \mathbf{B}$ multiplied by β), of which at most $m_1 - b_1$ do not satisfy $\mathbf{C}$. Summarizing, $a_1 \geq \beta m_1 - (m_1 - b_1)$, and then $p_1 \geq \beta m_1/n_1 - (m_1/n_1 - b_1/n_1)$. Since $\beta = supp(\mathbf{D}, \mathbf{B}, \mathbf{A})/supp(\mathbf{D}, \mathbf{B}) = n_1/m_1$, the inequality can be rewritten as:

$$p_1 \geq (\beta + \gamma - 1)/\beta,$$

where $\gamma = b_1/m_1 = conf(\mathbf{D}, \mathbf{B} \rightarrow \mathbf{C})$. In our previous example:

$\mathbf{D}$ is `zip=1234` , $\mathbf{B}$ is `city=Liverpool`,
$\mathbf{C}$ is `app=no` $\mathbf{A}$ is `race=black`.

We have $\beta = 0.5$ since 50% of population in the postal area is black, $\gamma = 0.99$ since 99% of people in the postal area is refused application, and $p = (a_1 + a_2)/(n_1 + n_2) = 0.09$ since 9% of people from Liverpool is refused application on average. Summarizing, a lower bound for the extended lift p_1/p of the classification rule (1) is:

$$p_1/p \geq (\beta + \gamma - 1)/(\beta \cdot p) = 1/0.5(0.5 + 0.99 - 1)/0.09 = 10.89.$$

In general, an inference model consists of deriving lower and upper bounds for the values of the contingency table of an unknown PD classification rule, and a fortiori for the various discrimination measures, starting from:

- assumptions on the form of the premise of the rule;
- background knowledge, which in our framework consists of association rules.

As a result, indirect discrimination inference strategies boil down to "rule inference" strategies. The following definition formalizes the redlining strategy.

Definition 12 A PD classification rule $c = (\mathbf{A}, \mathbf{D}), \mathbf{B} \rightarrow \mathbf{C}$ such that $elift(c) \geq lb$ is inferred by the *redlining strategy* if there exist PND rules $\mathbf{D}, \mathbf{B} \rightarrow \mathbf{C}$ and $\mathbf{B} \rightarrow \mathbf{C}$, and a background knowledge rule $\mathbf{D}, \mathbf{B} \rightarrow \mathbf{A}$ such that, called:

$$\gamma = conf(\mathbf{D}, \mathbf{B} \rightarrow \mathbf{C}) \quad p = conf(\mathbf{B} \rightarrow \mathbf{C}) > 0 \quad \beta = conf(\mathbf{D}, \mathbf{B} \rightarrow \mathbf{A}) > 0,$$

we have: $lb = (\beta + \gamma - 1)/(\beta \cdot p)$.

A PD rule $(\mathbf{A}, \mathbf{D}), \mathbf{B} \rightarrow \mathbf{C}$ inferred by the redlining strategy, with $\mathbf{C}$ denying a benefit, is an lb-directly discriminatory rule w.r.t. $elift()$, where lb is the inferred lower bound for $elift()$. Notice that we implicitly make the assumption that $\mathbf{D}$ (e.g., `zip=1234`) is a PND itemset and $\mathbf{A}, \mathbf{D}$ (e.g., `zip=1234, race=black`) is a PD itemset. This is not in contrast with Def. 2, since the following downward closed set can be used to specify PD itemsets for redlining analysis:

$$\mathcal{I}_d = \{\texttt{zip=}z_1\texttt{, ..., zip=}z_n\texttt{, race=black} \mid z_1, \ldots, z_n \text{ are zip codes } \}.$$

An itemset `zip=`z_1`, ..., zip=`z_n`, race=black` boils down to `zip=`z_1`, race=black` if $z_1 = \ldots = z_n$; and it boils down to an empty itemset if there exists $z_i \neq z_j$, with $1 \leq i < j \leq n$.

In addition to redlining, other inference strategies are investigated by Pedreschi et al (2008), including:

- for binary classes, a simple inference based on the property $conf(\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}) = 1 - conf(\mathbf{A}, \mathbf{B} \rightarrow \neg\mathbf{C})$ has been directly integrated within the definition of a-discrimination (w.r.t. $elift()$) by introducing the notion of strong a-discrimination;
- a generalization of the redlining inference strategy, where the extended lift of $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$ is inferred starting from PND classification rules $\mathbf{D}, \mathbf{B} \rightarrow \mathbf{C}$ and $\mathbf{B} \rightarrow \mathbf{C}$, and from background association rules $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{D}$ and $\mathbf{D}, \mathbf{B} \rightarrow \mathbf{A}$;
- and an inference strategy that derives the extended lift of $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$ starting from the extended lift of $\neg\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$, where $\mathbf{A}$ is an item over a binary attribute – typically the gender attribute, and from background association rules $\mathbf{D}, \mathbf{B} \rightarrow \mathbf{A}$ – typically the gender distribution over the territory;

Finally, notice that the notion of inference strategy readily extends in presence of statistical significance of a measure. Upper and lower bounds for the Wald confidence intervals introduced in Sect. 5.1 can be obtained starting from upper and lower bounds for the values in the contingency tables of the PD rules being inferred.

9 Argumenting against Discrimination Allegations

Consider a PD classification rule denying some benefit:

$$\mathbf{A}, \mathbf{B} \rightarrow \texttt{benefit} = \texttt{no},$$

that has been unveiled, either directly or indirectly. In a case before a court, such a rule supports the complainant position if she belongs to the disadvantaged group $\mathbf{A}$, she satisfies the context conditions $\mathbf{B}$ and the rule is a-directly discriminatory (w.r.t. one of the definitions of Sect. 4) where a is a threshold stated in law, regulations or past sentences. Showing that no rule satisfies those conditions supports the respondent position. However, this is an exceptional case. When one or more such rules exist, the respondent

is then required to prove that the "provision, criterion or practice is objectively justified by a legitimate aim and the means of achieving that aim are appropriate and necessary" (see Ellis (2005)). A typical example in the literature is the one of the "genuine occupational requirement", also called "business necessity" by the U.S. Federal Legislation (2010)(f). For instance, assume that the complainant claims for discrimination against women among applicants for a job position. A classification rule `sex=female, city=NYC` $\rightarrow$ `hire=no` with high selection lift supports her position. The respondent might argue that the rule is an instance of the more general one `drive_truck=false, city=NYC` $\rightarrow$ `hire=no`. Such a rule is legitimate, since the requirement that prospect workers are able to drive trucks can be considered a genuine occupational requirement (for some specific job).

Let us formalize the argumentation of the respondent by saying that a PD classification rule $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$ is an instance of a PND rule $\mathbf{D}, \mathbf{B} \rightarrow \mathbf{C}$ if: the rule $\mathbf{D}, \mathbf{B} \rightarrow \mathbf{C}$ holds at the same or higher confidence, namely $conf(\mathbf{D}, \mathbf{B} \rightarrow \mathbf{C}) \geq conf(\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C})$; and, a case satisfying $\mathbf{A}$ in context $\mathbf{B}$ satisfies condition $\mathbf{D}$ as well, namely $conf(\mathbf{A}, \mathbf{B} \rightarrow \mathbf{D}) = 1$. These two conditions can be relaxed as follows.

Definition 13 Let p be in $[0, 1]$. A PD classification rule $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$ is a p-instance of a PND rule $\mathbf{D}, \mathbf{B} \rightarrow \mathbf{C}$ if:

(1) $conf(\mathbf{D}, \mathbf{B} \rightarrow \mathbf{C}) \geq p \cdot conf(\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C})$; and,
(2) $conf(\mathbf{A}, \mathbf{B} \rightarrow \mathbf{D}) \geq p$.

A respondent argumenting against discriminatory allegations supported by a PD rule $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$ must show that the rule is a p-instance of some PND rule $\mathbf{D}, \mathbf{B} \rightarrow \mathbf{C}$ with p as near to 1 as possible, and with $\mathbf{D}$ modelling a genuine occupational requirement. This task can be accomplished in the reference model. Given a classification rule $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$, we have to search for PND classification rules of the form $\mathbf{D}, \mathbf{B} \rightarrow \mathbf{C}$ with confidence satisfying (1); and, for each of such rules, we have to check that the association rule $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{D}$ satisfies condition (2). By noting that:

$$conf(\mathbf{A}, \mathbf{B} \rightarrow \mathbf{D}) = \frac{supp(\mathbf{D}, \mathbf{B} \rightarrow \mathbf{A})}{cov(\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C})},$$

we can restrict to consider association rules of the form $\mathbf{D}, \mathbf{B} \rightarrow \mathbf{A}$, which are extracted from the training set as described in Sect. 6. This trick has the advantage that the search for $supp(\mathbf{D}, \mathbf{B} \rightarrow \mathbf{A})$ is over a much smaller set of association rules[3], and that the coverage of $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$ is available given its contingency table.

On the contrary, a complainant or a control authority can prevent respondent's argumentation by showing that, for a sufficiently high p, the PD rule $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$ is not a p-instance of any PND rule $\mathbf{D}, \mathbf{B} \rightarrow \mathbf{C}$.

10 Discrimination in Favor of

10.1 Affirmative Actions

Many legislations account for affirmative actions (see Sowell (2005); Holzer and Neumark (2004); ENAR (2008)), sometimes called positive actions or reverse discrimination, as a range of policies to overcome and to compensate for past and present discrimination by providing opportunities to those traditionally denied for. Policies range

[3] For a rule $\mathbf{X} \rightarrow \mathbf{A}$, there are $2^{|\mathbf{X}|}$ rules $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{D}$ obtained by splitting $\mathbf{X}$ into $\mathbf{D}$ and $\mathbf{B}$.

from the mere encouragement of under-represented groups to quotas in favor of those groups. For instance, U.S. federal contractors are required to identify and set goals for hiring under-utilized minorities and women: Holzer and Neumark (2006) analyse the implications of such a requirement. Also, universities have voluntarily implemented admission policies that give preferential treatment to women and minority candidates: Lerner and Nagai (2000) study the impact of those policies. Affirmative action policies "shall in no case entail as a consequence the maintenance of unequal or separate rights for different racial groups after the objectives for which they were taken have been achieved" United Nations Legislation (2010)(a). It is therefore important to assess and to monitor the application of affirmative actions.

In our proposed reference model, affirmative actions can be unveiled from the training set by proceedings in a similar way as for discriminatory actions. The basic idea is to search, either directly or indirectly, for a-discriminatory PD rules of the form:

$$\mathbf{A}, \mathbf{B} \rightarrow \texttt{benefit} = \texttt{yes},$$

i.e., where the consequent consists of granting a benefit (a loan, school admission, a job, etc.). Rules of this form having a value of the adopted measure greater than a fixed threshold a highlight contexts $\mathbf{B}$ where the disadvantaged group $\mathbf{A}$ is actually favored.

Definition 14 (affirmative action) Let $f()$ be one of the measures from Definitions 4-8, and $a \in \mathbb{R}$ a fixed threshold. A PD classification rule $c = \mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$, where $\mathbf{C}$ grants some benefit and $\mathbf{A}$ refers to a disadvantaged group, is an a-affirmative action rule w.r.t. $f()$ if $f(c) \geq a$.

The approaches for unveiling direct and indirect discrimination presented in Sects. 7-8 are directly applicable to unveil affirmative actions by simply switching the role of the two classes `benefit=yes` and `benefit=no`.

10.2 Favoritism

Favoritism refers to when someone appears to be treated better than others for reasons not related to individual merit, business necessity or affirmative actions. For instance, favoritism in the workplace might result in a person being promoted faster than others unfairly or being paid more to do the same job as others. Kim (2007) studies political decisions, such as distributing income across regions or groups, that can be discretionary and favor the group or district to which a politician belongs to. The difference between affirmative actions and favoritism lies then in the group which is favored: in affirmative actions, the group is an historically disadvantaged one and the practice is suggested or required by the law; in favoritism, the group is favored for reasons that are not supported by explicit rules or legislation.

In the proposed reference model, favoritism can be analysed by switching to a set of PD itemsets that denotes the favored groups and by checking for rules of the form:

$$\mathbf{A}, \mathbf{B} \rightarrow \texttt{benefit} = \texttt{yes},$$

as in the case of affirmative actions.

Definition 15 (favoritism) Let $f()$ be one of the measures from Definitions 4-8, and $a \in \mathbb{R}$ a fixed threshold. A PD classification rule $c = \mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$, where $\mathbf{C}$ grants some benefit and $\mathbf{A}$ refers to a favored group, is an a-favoritism action rule w.r.t. $f()$ if $f(c) \geq a$.

The approaches for unveiling direct and indirect discrimination presented in Sects. 7-8 are directly applicable to unveil favoritism:

- by simply switching the role of the two classes `benefit=yes` and `benefit=no`; and
- by defining $\mathcal{I}_d$, the set of PD itemsets, to denote the set of favored groups.

As an example, by fixing PD items to include `personal_status=male single` and `age=40-50`, we can analyse favoritism versus single male and/or people in their 40's.

10.3 Symmetry in Discrimination Measures

Let us point out a symmetry property of the discrimination measures introduced in Sect. 4 when the class attribute is binary.

For a classification rule such as $\texttt{sex} = \texttt{female}, \mathbf{B} \rightarrow \texttt{benefit} = \texttt{no}$ with a ratio measure lower than 1, the complementary decision rule $\texttt{sex} = \texttt{female}, \mathbf{B} \rightarrow \texttt{benefit} = \texttt{yes}$ has a measure greater than 1, and viceversa. In general, with reference to Fig. 2, 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 $olift(c') = 1/olift(c)$.

This symmetry of discrimination measures can be interpreted as follows. A PD rule $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$ is a-directly discriminatory, for some $a < 1$ iff the complementary decision rule $\mathbf{A}, \mathbf{B} \rightarrow \neg\mathbf{C}$ is a b-affirmative action, for some $b > 1$. Intuitively, a burden lower than the average highlighted by $\texttt{sex} = \texttt{female}, \mathbf{B} \rightarrow \texttt{benefit} = \texttt{no}$ means a favor greater than the average highlighted by $\texttt{sex} = \texttt{female}, \mathbf{B} \rightarrow \texttt{benefit} = \texttt{yes}$.

As a natural consequence, a-direct discrimination, a-affirmative actions and a-favoritism are worth investigating only when $a > 1$ for ratio measures, and only when $a > 0$ for difference measures.

11 The LP2DD Analytical System

The proposed reference model provides us with a framework for discrimination analysis by translating key concepts from the legal viewpoint into quantitative measures and deduction rules over classification and association rules extracted from a training set and/or from background knowledge. The rule meta-reasoner in Fig. 1 exploits such translations as building blocks in support of iterative and interactive discrimination pattern discovery. In this section, we present the LP2DD system (Logic Programming to Discover Discrimination), an intuitive implementation of the reference model in a computational logic language.

The overall architecture of LP2DD is shown in Fig. 5. The LP2DD system relies on data mining algorithms for the inductive part. Any frequent pattern extraction algorithm from the Frequent Itemset Mining Implementations repository of Goethals (2010) can be plugged-in the system. Classification and association rule extraction is performed through the procedures **ExtractClassificationRules()** and **ExtractAssociationRules()** devised in Sect. 6. As a result, Prolog facts for PD rules, PND rules

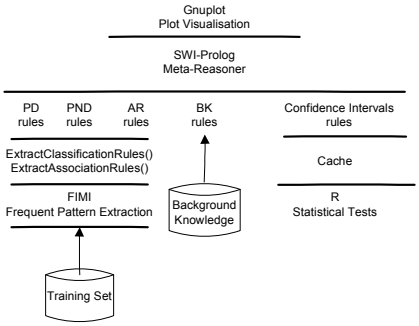

Fig. 5 Architecture of the LP2DD system.

and association rules are produced. Background knowledge can be added by asserting Prolog facts using the syntax that will be described later on.

The deductive part, i.e., the meta-reasoner, of the LP2DD system is written in SWI-Prolog (see Wielemaker (2009)), and it will be presented in depth later on. The user-interface is in SWI-Prolog as well, calling GNUplot (see Williams and Kelley (2010)) for rendering distribution plots. Finally, a module written in the R statistical software language (see R Development Core Team (2010)) is part of the system for computing confidence intervals of the discrimination measures. While Wald confidence intervals can be easily calculated in Prolog, sophisticated approaches for exact methods are typically implemented in statistical programming languages, such as R. More precisely, LP2DD adopts Wald confidence intervals corrected with the plus-4 method when $n_1 + n_2 > 30$ (see Fig. 2), and a recent exact method based on an extension of the Sterne's test due to Reiczigel et al (2008) otherwise. Since exact method calculations can be time consuming, a cache module is in between the meta-reasoner and the R module.

In the rest of this section, we present the details of rule extraction and of the meta-reasoner using the German credit dataset as a case study.

11.1 Rule Extraction and Representation

The following log of Prolog goals to the LP2DD system show how the user can:

1 locate the German credit training set, in comma-separated-values format or in ARFF format; notice that obtaining the training set from the input pool is not part of the LP2DD system, since it is very specific of the DSS at hand.
2 fix the class items for which rules have to be extracted;
3 fix the PD items of interest for the analysis; in the log: senior people and/or non-single women; in LP2DD, PD itemsets are generated (see Def. 1) from PD items;
4 extract association and classification rules having a minimum absolute support threshold (10 in the log, equivalent to relative support of 1%);
5 load from the cache, and possibly calculate from scratch, confidence intervals for contingency tables of PD rules w.r.t. a measure and a confidence level (in the log, selection lift and 95% respectively).

```
% load training set items
1 ?- arff_load('german_credit').
true .

% fix class items of interest
2 ?- set_class([class=good,class=bad]).
true .

% fix PD items
3 ?- set_pd([age=52-inf, personal_status=female_div_or_sep_or_mar]).
true .

% extract PD and PND classification rules
% with minimum absolute support of 10
4 ?- extract(10).
true .

% load from cache confidence intervals for slift at 95% confidence
5 ?- ci_load(slift, 0.95).
true .
```

The following facts are defined as the result of the previous steps. Items are represented by the predicate `item(`n`, `i`)`, where n is an integer code for item i. Coding is necessary for computational efficiency reasons. Class items are modelled by `item_class(`i`)` atoms, and PD items by `item_pd(`i`)` atoms. Extracted PND rules are stored in facts `pndrule(`b`, `c`, ct(`a_1`, `b_1`))`, where b is the list of (codes of) items in the premise, c is the class item code in the conclusion, and `ct(`a_1`, `b_1`)` is the contingency table of the rule (with reference to Fig. 2, since $\mathbf{A}$ is empty, $a_2 = n_2 = 0$ and then it is not necessary to record the second row). Extracted PD rules are stored in facts `pdrule(`a`, `b`, `c`, ct(`a_1`, `b_1`, `a_2`, `b_2`))`, where a is the list of PD items and b is the list of PND items in the premise. Also, the whole contingency table is now recorded. Association rules of the form $\mathbf{D}, \mathbf{B} \rightarrow \mathbf{A}$ are stored in the `arule` predicate with contingency table as in the case of PND rules. Confidence intervals for selection lift, also known as risk ratio in the statistics literature, are cached in `ci_rr(`ct`, `lb`, `ub`)`, where ct is a contingency table, and lb and ub are the boundaries of the confidence interval. Finally, we mention that all lists of items are ordered (w.r.t. item code), so that the representation of an itemset is unique.

```
% items
item(1,checking_status=negative).
item(2,checking_status=0-200).
item(4,checking_status=200-inf).
...

% class items
item_class(class=bad).
item_class(class=good).

% PD items
item_pd(personal_status=female_div_or_sep_or_mar).
item_pd(age=52-inf).

% PND classification rules
pndrule([1], 78, ct(139,135) ).
pndrule([3,15,62,75], 78, ct(22,3) ).
...
```

```
% PD classification rules
pdrule([55], [51,62], 78, ct(25,4,157,40) ).
pdrule([42,55], [23,57,72], 78, ct(20,2,51,11) ).
...

% association rules
arule([72], [42,55], ct(30,815) ).
arule([36,59], [42], ct(22,28) ).
...

% cached confidence intervals
ci_rr(ct(12,7,1,2),0.744,28.97).
ci_rr(ct(14,13,1,1),0.51,9.534).
...
```

Association rules modelling background knowledge are stored in the `background` predicate in the same form as in the `arule` predicate. The user can load or assert them as Prolog facts. For testing purposes, we simulate the availability of a large set of background rules of the form $\mathbf{D}, \mathbf{B} \rightarrow \mathbf{A}$ under the assumption that the dataset contains the PD discriminatory itemsets, e.g., as for the German credit dataset. We then define the program clause:

```
background(DB, A, CT) :-  arule(DB, A, CT).
```

Predicates are provided in the LP2DD system for decoding itemsets (`itemset_decode`); for splitting an itemset into its PD and PND parts (`itemset_split`); for counting the number of answers to a goal (`count`, `distribution`). For readability reasons, we omit explicit coding/decoding of items for the rest of the paper. Next, we report three sample goals related to counting PND rules and PD rules, and to splitting a rule premise into its PD and PND parts.

```
% counting number of PND rules
6 ?- item_class(C), count( pndrule(B, C, CT), N).
C = (class=good),
N = 2102339 ; % no of PND rules with class=good
C = (class=bad),
N = 341867 ; % no of PND rules with class=bad

7 ?- item_class(C), count( pdrule(A, B, C, CT), N).
C = (class=good),
N = 215819 ; % no of PD rules with class=good
C = (class=bad),
N = 72394 ; % no of PD rules with class=bad

% splitting AB into PD part A and PND part B
8 ?- AB = [checking_status=negative, age=52-inf], itemset_split(AB, A, B).
A = [age=52-inf],
B = [checking_status=negative] .
```

11.2 Meta-Reasoner, Part I

The core of the meta-rule reasoner is shown in Fig. 6 for the part concerning discrimination measures. A few measures are defined for a given contingency table, including confidence of PND rules (clause `(cn1)`) and PD rules `(cn2)`, coverage `(cv1,cv2)`, extended lift `(el)`, selection lift `(sl)`, and odds lift `(ol)`. PD classification rules with a

```
(cn1) confidence(ct(A,B), CN) :-
      AB is A + B,
      AB =\= 0,
      CN is A/(A+B).

(cn2) confidence(ct(A,B,_,_), CN) :-
      AB is A + B,
      AB =\= 0,
      CN is A/(A+B).

(cv1) coverage(ct(A,B), CV) :-
      CV is A+B.

(cv2) coverage(ct(A,B,_,_), CV) :-
      CV is A+B.

(el) elift(ct(A,B,C,D), EL) :-
      AC is A + C,
      AC =\= 0,
      AB is A + B,
      AB =\= 0,
      N is A+B+C+D,
      EL is (A*N)/(AB*AC).
```

```
(sl) slift(ct(A,B,C,D), SL) :-
      C =\= 0,
      AB is A + B,
      AB =\= 0,
      CD is C+D,
      SL is (A*CD)/(AB*C).

(ol) olift(ct(A,B,C,D), OL) :-
      C =\= 0,
      B =\= 0,
      OL is (A*D)/(C*B).

(c) check(slift, T, A, B, C, CT) :-
      pdrule(A, B, C, CT),
      slift(CT, EL),
      EL >= T.

(d) discrimination(M, T, A, B, C, CT) :-
      item(C, class=bad),
      check(M, T, A, B, C, CT).

(a) affirmative(M, T, A, B, C, CT) :-
      item(C, class=good),
      check(M, T, A, B, C, CT).

(f) favoritism(M, T, A, B, C, CT) :-
      affirmative(M, T, A, B, C, CT).
```

Fig. 6 Core Meta-Reasoner of the LP2DD system, Part I.

discrimination measure greater or equal than a given threshold, are detected by predicate `check`, whose first parameter is the measure to be used. Clause `(c)` shows its definition for the extended lift. As stated in Sect. 7 and in Sect. 10, checking direct discrimination and affirmative actions is modelled by searching for PD classification rules denying credit (see predicate `discrimination` in clause `(d)`) and granting credit (see predicate `affirmative` in clause `(a)`) to protected-by-law groups. Also, favoritism is modelled as affirmative actions (see `favoritism` in clause `(f)`) but with reference to groups that are not protected-by-law. The following log of Prolog goals to the LP2DD system show how the user can:

`1` count the number of PND rules denying credit having selection lift greater or equal than 10, or, in more intuitive words, count the number of 10-directly discriminatory rules w.r.t. $slift()$;

`2` enumerate the PND rules having a selection lift of at least 3 and a context of length 2;

`3` do the same analysis as in `(1-2)` for rules granting credit to disadvantaged people, namely for checking affirmative actions;

`4-6` do the analysis as in `(1-2)` for rules granting credit to advantaged people (single males and/or people in their 40's), namely for checking favoritism; this requires the re-extraction of classification rules since the set of PD items is changed.

```
% count no. of PND rules with a minimum measure
1 ?- count( discrimination(slift, 10, A, B, C, CT), N).
N = 52 .

% enumerate PND rules with a minimum measure
```

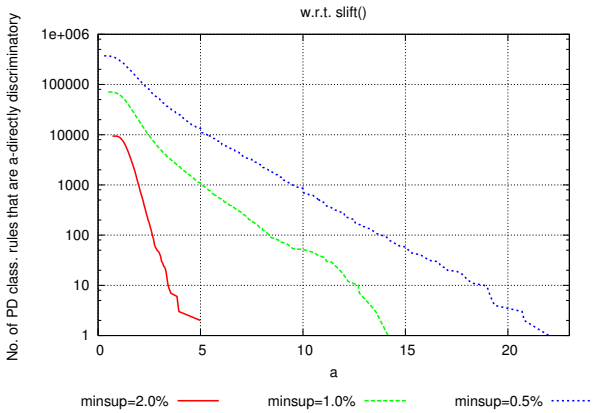

Fig. 7 Distributions of a-directly discriminatory rules w.r.t. $slift()$ vs. minimum support.

```
2 ?- discrimination(slift, 3, A, B, C, CT), length(B, 2).
A = [personal_status=female_div_or_sep_or_mar],
B = [employment=1-4, age=0-31],
C = (class=bad)
CT = ct(11, 9, 1, 21) .

% enumerate PND rules for affirmative actions
3 ?- affirmative(slift, 3, A, B, C, CT).
A = [personal_status=female_div_or_sep_or_mar],
Bs = [duration=17-31, property_magnitude=life_insurance, housing=rent],
C = (class=good)
CT = ct(10, 3, 1, 3) .

% change PD items
4 ?- set_pd([personal_status=male_single, age=41-52]).
true

% extract PD and PND classification rules
5 ?- extract(10).
true

% enumerate PND rules for favoritism
6 ?- favoritism(slift, 4, A, B, C, CT), length(B, 2).
A = [personal_status=male_single],
B = [property_magnitude=life_insurance, num_dependents=2-inf],
C = (class=good),
CT = ct(24, 6, 1, 4) .
```

A few plots can be produced showing distributions of interesting subsets of classification rules.

Fig. 7 reports the distribution of a-directly discriminatory rules w.r.t. $slift()$ at the variation of the minimum support threshold in rule extraction. We observe that, if classification 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[4], which

[4] With reference to Fig. 2, consider a rule c with $a_1 = x, n_1 = x + 1, a_2 = 1, n_2 = y$, for x, y natural numbers. Fixed $x = ms|\mathcal{D}|$ to satisfy the minimum support requirement, we have

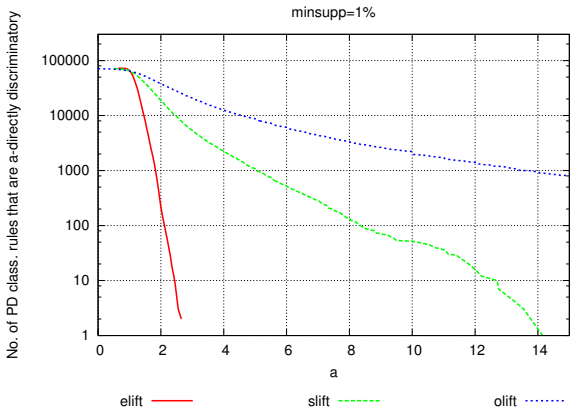

Fig. 8 Distributions of a-directly discriminatory rules w.r.t. $elift()$, $slift()$ and $olift()$.

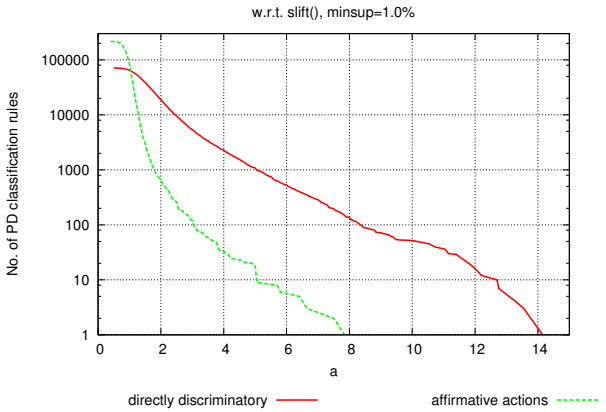

Fig. 9 Distributions of a-directly discriminatory and a-affirmative actions rules w.r.t. $slift()$.

in general are unbound from above. Nevertheless, Fig. 7 shows that, in practice, the lower minimum support threshold the more niches of discrimination can be unveiled.

Fig. 8 compares the distributions of a-directly discriminatory rules w.r.t. $elift()$, $slift()$ and $olift()$. Accordingly to Lemma 1, when $a \geq 1$, the odds lift assumes values greater than the selection lift that, in turn, is greater than the extended lift. It is worth noting, however, that, in general, the order imposed by those measures is not the same, namely $elift(c_1) > elift(c_2)$ does not imply $slift(c_1) > slift(c_2)$ for two PND rules c_1 and c_2. Interestingly, none of the top 10 rules w.r.t. $elift()$ is a top 10 rule w.r.t. $slift()$ or w.r.t. $olift()$ and vice-versa. Therefore, the choice of the reference measure turns out to heavily affect the relevance of a context of discrimination.

$slift(c) = (xy)/(x+1) \geq y/2$, which is unbound. The reasoning is analogous for the odds lift, which is $olift(c) = x(y-1)$.

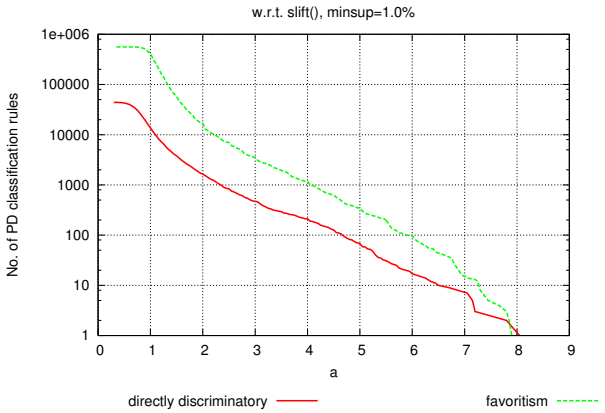

Fig. 10 Distributions of a-directly discriminatory and a-favoritism action rules w.r.t. $slift()$ for $\mathcal{I}_d$ generated by {`personal_status=male single`, `age=(41.4-52.6]`}.

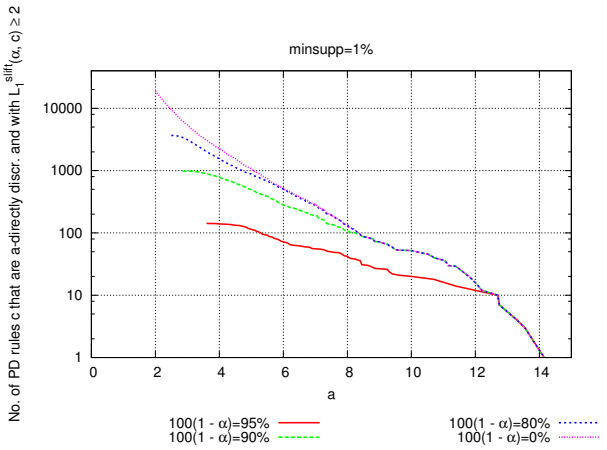

Fig. 11 Distributions of a-directly discriminatory rules c w.r.t. $slift()$ and such that $L_1^{slift}(\alpha, c) \geq 2$ at various confidence levels.

Fig. 9 compares the distributions of a-directly discriminatory rules and a-affirmative actions for the $slift()$ measure. It is immediate to see that, for the fixed set of PD itemsets, directly discriminatory rules occurs more in number and with higher measure values than affirmative actions. This is an indicator that the groups represented by the PD itemsets are actually unfavored rather than favored in the dataset of historical decisions. The situation is reversed if we change the set of PD items to include single males and/or people in their 40's. In the case of favored groups, disproportionate benefit granting has been called favoritism. Fig. 10 contrasts direct discrimination, namely benefit denial, with favoritism.

Consider now the notion of a-discrimination at a certain confidence level. Fig. 11 shows the distributions of PD rules c that are a-directly discriminatory w.r.t. $slift()$ and 2-discriminatory at various confidence levels, namely such that $L_1^{slift}(\alpha, c) \geq 2$.

```
(in) pinstance(A,B,C,CT,MinP,D,P) :-
       coverage(CT, SBA),
       confidence(CT, CN),
       arule(BD, A, CT1),
       remove(BD, B, D),
       support(CT1, SBDA),
       P1 is SBDA/SBA,
       P1 >= MinP,
       pndrule(BD, C, CT2),
       confidence(CT2, CN1),
       P2 is CN1/CN,
       P2 >= MinP,
       P is min(P1, P2).

(ni) pnoinstance(A,B,C,CT,MinP) :-
       \+ pinstance(A,B,C,CT,MinP,_,_).
```

```
(i) redlining(elift,BMin,AD,B,C,LB,A) :-
       background(DB, A, CT_BDA),
       confidence(CT_BDA, BETA),
       BETA >= BMin,
       split(DB, D, B),
       pndrule(B, C, CT_BC),
       confidence(CT_BC, P),
       P > 0,
       pndrule(DB, C, CT_DBC),
       confidence(CT_DBC, GAMMA),
       LB is (BETA+GAMMA-1)/(P*BETA),
       merge(A, D, AD).
```

Fig. 12 Core Meta-Reasoner of the LP2DD system, Part II.

Intuitively, the higher the confidence level, the lower is the number of directly discriminatory rules. Rules with very high selection lift will remain 2-discriminatory until high confidence levels. As one can expect, higher confidence levels greatly reduce the number of statistically significant discriminatory rules.

11.3 Meta-Reasoner, Part II

The core of the meta-rule reasoner is shown in Fig. 12 for the part concerning argumentation against discriminatory rules and inference strategies.

Predicate `pinstance` defined by clause `(in)` checks whether a PND classification rule is a p-instance of some PD rule, according to Def. 13. A goal `:- pinstance(A, B, C, CT, MinP, D, P)` instantiates `D` to an itemset $\mathbf{D}$ and `P` to a value p greater or equal than `MinP` such that $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$ is a p-instance of $\mathbf{D}, \mathbf{B} \rightarrow \mathbf{C}$. Predicate `pnoinstance` defined by clause `(ni)` succeeds when there is no such PD rule $\mathbf{D}, \mathbf{B} \rightarrow \mathbf{C}$. The following Prolog goals to the LP2DD system show how the user can:

1 count the number of PND rules denying credit having selection lift greater or equal than 3, and such that they are not 0.80-instances of any PD rule;
2 enumerate PND rules that are 0.8-instances of PD rules $\mathbf{D}, \mathbf{B} \rightarrow \mathbf{C}$ where $\mathbf{D}$ is some legally grounded requirement encoded as an itemset including `age=0-31` and/or `credit_history=critical`;
3 enumerate PND rules that are not 0.8-instances of PD rules.

```
% Discriminatory PND rules that are not instances
% of PD rules
1 ?- count( (discrimination(slift, 3, A, B, C, CT),
             pnoinstance(A, B, C, CT, 0.80)), N).
N = 38 .

% PND rules that are instances of PD rules
2 ?- discrimination(slift, 3, A, B, C, CT),
       pinstance(A, B, C, CT, 0.8, D, P),
       subset(D, [age=0-31,credit_history=critical]).
A = [personal_status=female_div_or_sep_or_mar],
```

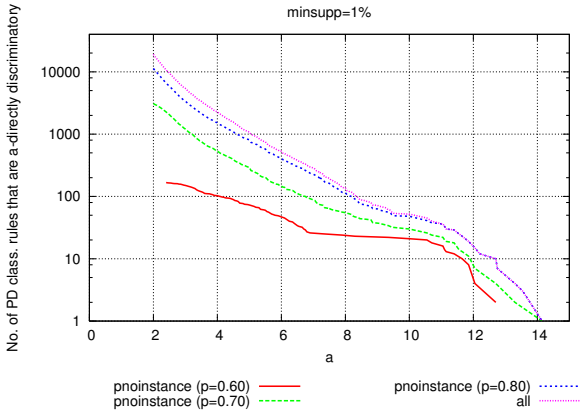

Fig. 13 Distributions of a-directly discriminatory rules w.r.t. $slift()$ that are not p-instances of a PND rule, for $a \geq 2$.

```
B = [duration=17-31, residence_since=2-inf, housing=rent, num_dependents=0-1],
C = (class=bad),
D = [age=0-31],
CT = ct(21, 20, 2, 11),
P = 0.829268 .

% PND rules that are not instances of any PD rule
3 ?- discrimination(slift, 3, A, B, C, CT),
        pnoinstance(A, B, C, CT, 0.8).
A = [personal_status=female_div_or_sep_or_mar],
B = [property_magnitude=real_estate, other_payment_plans=none,
        num_dependents=0-1, own_telephone=none],
C = (class=bad),
CT = ct(20, 36, 9, 75) .
```

Fig. 13 shows the distributions of a-directly discriminatory rules w.r.t. $slift()$ that are not p-instances of a PND rule, for sample $p = 0.6$, $p = 0.7$ and $p = 0.8$. The number of a-directly discriminatory rules that are not instances of PND rules decreases as p decreases. Rules occurring at lower values of p should be given higher attention in the discrimination analysis, since there is no immediate (i.e., in the data) justification for them, according to the formalization of the genuine occupational requirement/business necessity principle provided in Def. 13.

Let us consider now indirect discrimination. The redlining inference strategy of Def. 12 is implemented by the `redlining` predicated (see clause `(i)` in Fig. 12). The search for PD rules is driven by background knowledge association rules $\mathbf{DB} \rightarrow \mathbf{A}$ having some minimum confidence `BMin`, namely stating that the protected group $\mathbf{A}$ represents at least a fraction `BMin` of people in $\mathbf{DB}$. For each possible split of the itemset $\mathbf{DB}$ into $\mathbf{D}$ and $\mathbf{B}$ the lower bound lb is calculated as in Def. 12. Finally, the PD itemset in the inferred PD rule $\mathbf{AD}, \mathbf{B} \rightarrow \mathbf{C}$ is built as $\mathbf{AD} = \mathbf{A}, \mathbf{D}$.

The following Prolog goals over the German credit dataset exploit the redlining strategy in searching for a 1.5-directly discriminatory rule and a 2-affirmative action w.r.t. $elift()$ respectively. In order to run the goal, we have simulated the availability of background knowledge by defining facts for the `background` predicate starting from

association rules extracted from the training set and stored in the `arule` predicate (see Sect. 11.1).

```
% Searching for indirect discrimination
4 ?- redlining(elift, 0.9, AD, B, class=bad, LB, A), LB >= 1.5.
LB = 1.52308,
AD = [personal_status=female_div_or_dep_or_mar, housing=rent],
B = [employment=0-1, installment_commitment=2.8-inf, own_telephone=none]
A = [personal_status=female_div_or_dep_or_mar] .


% Searching for indirect affirmative actions
5 ?- redlining(elift, 0.8, AD, B, class=good, LB, A), LB >= 2.
LB = 2.40625,
AD = [purpose=furniture_or_equipment, personal_status=female_div_or_sep_or_mar],
B = [employment=0-1, housing=rent, own_telephone=none],
A = [personal_status=female_div_or_sep_or_mar] .
```

In the first answer, the context **B** consists of people employed by at most one year, with large installment commitment and not owing a phone. In such a context, at least 90% of people having an house for rent (i.e., **D**) are non-single women (i.e., **A**), where the threshold of 90% has been specified as a parameter in the goal. Having denied credit to people in the context having an house for rent had the effect of denying credit mainly to women in the context. Formally, the rule $(\mathbf{A}, \mathbf{D}), \mathbf{B} \rightarrow \mathbf{C}$ has an extended lift of 1.52308 or higher. Since we are simulating the absence of PD itemsets, and we have them actually, we can calculate the extended lift of the inferred PD rule, which turn out to be exactly 1.52308.

The second goal and answer cover a similar reasoning, but now for the class item `class=good`, hence inferring PD rules that unveil indirect affirmative actions.

12 Conclusions

12.1 Discrimination discovery and discrimination prevention

The subject of this paper consists of supporting discrimination discovery, namely the unveiling of discriminatory decisions hidden, either directly or indirectly, in a dataset of historical decision records, possibly built as the result of applying a classifier. Traditionally, classification models from the machine learning and the data mining literature are constructed on the basis of historical data with the purpose of distinguishing between elements of different classes. Learning from historical data may mean to discover traditional prejudices that are endemic in reality, and to assign to such practices the status of general rules, maybe unconsciously, as these rules can be deeply hidden within a classifier. For instance, if it is a current malpractice to deny pregnant women 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 to discriminate. Discrimination prevention consists of inducing a classifier that does not lead to discriminatory decisions even if trained from a dataset containing them. The naïve approach of deleting potentially discriminatory itemsets or even whole attributes from the original dataset does not prevent a classifier to learn discriminatory actions in that it only shields against direct discrimination, not against the indirect one. We foresee three non mutually-exclusive strategies towards discrimination prevention. The first one is to adapt the preprocessing approaches of data

sanitization (see e.g., Hintoglu et al (2005); Verykios et al (2004)) and hierarchy-based generalization (see e.g., Sweeney (2002); Wang et al (2005)) from the privacy-preserving literature. Along this line, Kamiran and Calders (2009) adopt a controlled distortion of the training set. The second one is to modify the classification learning algorithm (an in-processing approach), by integrating discrimination measures calculations within it. The third one is to post-process the produced classification model. Along this line, Pedreschi et al (2009) propose a confidence-altering approach for classification rules inferred by the CPAR algorithm of Yin and Han (2003).

12.2 Statistical discrimination or rational racism

In the German credit case study, the underlying context of analysis is a dataset of historical decisions on granting/denying credit to applicants. The framework proposed in this paper warns us that discriminatory decisions are hidden in such a dataset either directly or indirectly. Concerning the reasons behind those decisions, economists distinguish between "taste-based" discrimination, tracing back to the early studies of Becker (1957), and "statistical" discrimination. The former is concerned with dislike against protected-by-law groups. Becker's studies lead to the conclusion that, in a sufficiently competitive market, taste-based discrimination in not employing good black workers is not profitable. Statistical discrimination, also called rational racism by Harford (2008), occurs when employers refer directly or indirectly to the average performance of the applicant's racial group as a decision element. Field experiments reported by Riach and Rich (2002) show that this approach can be profitable, yet illegal. The discrimination analysis framework of this paper can help contrasting rational racism. As an example, assume a database of delays or defaults in repaying a loan. Although there is no discriminatory decision to discover here, the extraction of contexts where protected-by-law groups suffered from repaying the loan can help isolating possible sources of statistical discrimination and, a fortiori, preventing such an information be made public or hand-coded in a DSS.

12.3 The Importance of Data Collection

We performed several experiments with the German credit dataset to assess the functionalities of the LP2DD system. The quality of the answers obviously depends both on the quality of the dataset and the appropriateness of the formalization we provide for the legislation. The importance of data collection for the fight against discrimination is emphasized in studies promoted by the European Commission (see Makkonen (2006, 2007)). The construction of a "gold" dataset from real cases of direct discrimination, indirect discrimination, affirmative actions and all other concepts discussed in this paper should be pursued as a means to evaluate the quality of the patterns of discrimination discovered by our or by other approaches, according to some general evaluation strategy (see e.g., Stranieri and Zeleznikow (1999)).

12.4 Extensions of the Approach

The approach presented can be refined in several directions. First, attributes with continuous values are now required to be discretized a priori. The approach could

then be refined to account for continuous values, both in decisions (e.g., wage amount, mortgage interest rate) and in attributes (e.g., age, income). Statistical and economic theories of discrimination largely consider the continuum case rather than the discrete case. Second, the bias due to the use of frequent classification rules should be compared with the bias due to the use of other classification models, e.g., Bayesian models or defeasible logic (see Johnston and Governatori (2003)). Finally, the LP2DD system could be integrated with computational logic models of legal argument, such as those based on logic meta-programming surveyed by Prakken and Sartor (2002).

12.5 Conclusion

This paper introduced a reference model for the analysis and discovery of discrimination in socially-sensitive decisions taken by DSS. The approach consists first of extracting frequent classification rules, and then of analysing them on the basis of quantitative measures of discrimination and their statistical significance. The key legal concepts of protected-by-law groups, direct discrimination, indirect discrimination, genuine occupational requirement, affirmative actions and favoritism are formalized as reasonings over the set of extracted rules and, possibly, additional background knowledge. We have presented the LP2DD system, implementing the reference model, and a few analyses on the German credit dataset. LP2DD is intended as an analytical tool supporting DSS owners and control authorities in the interactive and iterative process of discrimination discovery.

References

Agrawal R, Srikant R (1994) Fast algorithms for mining association rules in large databases. In: Proc. of VLDB 1994, Morgan Kaufmann, pp 487–499

Agresti A (2002) Categorical Data Analysis. Wiley-Interscience

Agresti A, Brian C (2000) 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

Apt KR (1997) From Logic Programming to Prolog. Prentice Hall

Australian Legislation (2010) (a) Equal Opportunity Act – Victoria State, (b) Anti-Discrimination Act – Queensland State. `http://www.austlii.edu.au`

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

Becker GS (1957) The Economics of Discrimination. University of Chicago Press

Bell M, Chopin I, Palmer F (2007) Developing Anti-Discrimination Law in Europe. European Network of Legal Experts in Anti-Discrimination, `http://ec.europa.eu/employment_social/fundamental_rights`

Calem PS, Gillen K, Wachter S (2004) The neighborhood distribution of subprime mortgage lending. Journal of Real Estate Finance and Economics 29:393–410

Chien CF, Chen L (2008) 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

Dymski GA (2006) Discrimination in the credit and housing markets: Findings and challenges. In: Rodgers WM (ed) Handbook on the Economics of Discrimination, pp 215–259

Ellis E (2005) EU Anti-Discrimination Law. Oxford University Press

ENAR (2007) European Network Against Racism, Fact Sheet 33: Multiple Discrimination. http://www.enar-eu.org

ENAR (2008) European Network Against Racism, Fact Sheet 35: Positive Actions. http://www.enar-eu.org

European Union Legislation (2010) (a) Racial Equality Directive, (b) Employment Equality Directive. http://ec.europa.eu/employment_social/fundamental_rights

Farrington CP, Manning G (1990) Test statistics and sample size formulae for comparative binomial trials with null hypothesis of non-zero risk difference or non-unity relative risk. Statistics in Medicine 9:1447–1454

Fleiss JL, Levin B, , Paik MC (2003) Statistical Methods for Rates and Proportions. Wiley

Gastwirth JL (1984) Statistical methods for analyzing claims of employment discrimination. Industrial and Labor Relations Review 38:75–86

Gastwirth JL (1992) Statistical reasoning in the legal setting. The American Statistician 46(1):55–69

Geng L, Hamilton HJ (2006) Interestingness measures for data mining: A survey. ACM Computing Surveys 38(3)

Goethals B (2010) Frequent itemset mining implementations repository. http://fimi.cs.helsinki.fi

Han J, Cheng H, Xin D, Yan X (2007) Frequent pattern mining: Current status and future directions. Data Mining and Knowledge Discovery 15(1):55–86

Hand DJ, Henley WE (1997) Statistical classification methods in consumer credit scoring: a review. Journal of the Royal Statistical Society, Series A 160:523–541

Harford T (2008) The Logic of Life. The Random House Publishing Group

Hintoglu AA, Inan A, Saygin Y, Keskinöz M (2005) Suppressing data sets to prevent discovery of association rules. In: Proc. of IEEE ICDM 2005, IEEE Computer Society, pp 645–648

Holzer HJ, Neumark D (eds) (2004) The Economics of Affirmative Action. Cheltenham: Edward Elgar

Holzer HJ, Neumark D (2006) Affirmative action: What do we know? Journal of Policy Analysis and Management 25:463–490

Hunter R (1992) Indirect Discrimination in the Workplace. The Federation Press

Johnston B, Governatori G (2003) Induction of defeasible logic theories in the legal domain. In: Proc. of ICAIL 2003, ACM, pp 204–213

Kamiran F, Calders T (2009) Classification without discrimination. In: IEEE Int.'l Conf. on Computer, Control & Communication (IEEE-IC4), IEEE press

Kaye D, Aickin M (eds) (1992) Statistical Methods in Discrimination Litigation. Marcel Dekker, Inc.

Kim KH (2007) Favoritism and reverse discrimination. European Economic Review 51:101–123

Knopff R (1986) On proving discrimination: Statistical methods and unfolding policy logics. Canadian Public Policy 12:573–583

Kuhn P (1987) Sex discrimination in labor markets: The role of statistical evidence. The American Economic Review 77:567–583

LaCour-Little M (1999) Discrimination in mortgage lending: A critical review of the literature. Journal of Real Estate Literature, 7:15–49

Lerner N (1991) Group Rights and Discrimination in International Law. Martinus Nijhoff Publishers

Lerner R, Nagai AK (2000) Reverse discrimination by the numbers. Journal Academic Questions 13:71–84

Leung HM, Kupper LL (1981) Comparisons of confidence intervals for attributable risk. Biometrics 37(2):293–302

Makkonen T (2006) Measuring Discrimination: Data Collection and the EU Equality Law. European Network of Legal Experts in Anti-Discrimination, http://ec.europa.eu/employment_social/fundamental_rights

Makkonen T (2007) European handbook on equality data. European Network of Legal Experts in Anti-Discrimination, http://ec.europa.eu/employment_social/fundamental_rights

Newcombe RG (1998) Interval estimation for the difference between independent proportions: comparison of eleven methods. Statistics in Medicine 17:873–89

Newman D, Hettich S, Blake C, Merz C (1998) UCI repository of machine learning databases. http://archive.ics.uci.edu/ml

Pedreschi D, Ruggieri S, Turini F (2008) Discrimination-aware data mining. In: Proc. of ACM KDD 2008, ACM, pp 560–568, Extended version to appear in ACM Trans. on Knowledge Discovery from Data

Pedreschi D, Ruggieri S, Turini F (2009) Measuring discrimination in socially-sensitive decision records. In: Proc. of the SIAM SDM 2009, SIAM, pp 581–592

Piette MJ, White PF (1999) Approaches for dealing with small sample sizes in employment discrimination litigation. Journal of Forensic Economics 12:43–56

Prakken H, Sartor G (2002) The role of logic in computational models of legal argument: A critical survey. In: Kakas AC, Sadri F (eds) Computational Logic. Logic Programming and Beyond, Springer, Lecture Notes in Computer Science, vol 2408, pp 342–381

R Development Core Team (2010) R: A language and environment for statistical computing. Version 2.7.2, http://www.R-project.org

Rauch J, Simunek M (2005) An alternative approach to mining association rules. In: Foundations of Data Mining and knowledge Discovery, Studies in Computational Intelligence, vol 6, Springer, pp 211–231

Rauch J, Simunek M (2010) 4-ft Miner Procedure. http://lispminer.vse.cz

Reiczigel J, Abonyi-Tóth Z, Singer J (2008) 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

Riach PA, Rich J (2002) Field experiments of discrimination in the market place. The Economic Journal 112:480–518

Rorive I (2009) Proving discrimination cases - the role of situation testing. Centre For Equal Rights & Migration Policy Group, http://www.migpolgroup.com/publications.php

Schiek D, Waddington L, Bell M (2007) Cases, Materials and Text on National, Supranational and International Non-Discrimination Law. IUS Commune Casebooks for the Common Law of Europe

Sowell T (ed) (2005) Affirmative Action Around the World: An Empirical Analysis. Yale University Press

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

Sterling L, Shapiro E (1994) The Art of Prolog, 2nd edn. The MIT Press

Stranieri A, Zeleznikow J (1999) The evaluation of legal knowledge based systems. In: Proc. of ICAIL 1999, ACM, pp 18–24

Stranieri A, Zeleznikow J, Gawler M, Lewis B (1999) A hybrid rule - neural approach for the automation of legal reasoning in the discretionary domain of family aaw in australia. Artificial Intellence and Law 7(2-3):153–183

Sweeney L (2002) Achieving k-anonymity privacy protection using generalization and suppression. International Journal on Uncertainty and Fuzziness in Knowledge-Based Systems 10(5):571–588

Tan PN, Steinbach M, Kumar V (2006) Introduction to Data Mining. Addison-Wesley

Thomas LC (2000) A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. International Journal of Forecasting 16:149–172

Tian M, Tang ML, Ng HKT, Chan PS (2008) Confidence intervals for the risk ratio under inverse sampling. Statistics in Medicine 27:3301–3324

Tobler C (2008) Limits and potential of the concept of indirect discrimination. European Network of Legal Experts in Anti-Discrimination, `http://ec.europa.eu/employment_social/fundamental_rights`

UK Legislation (2010) (a) Sex Discrimination Act, (b) Race Relation Act. `http://www.statutelaw.gov.uk`

United Nations Legislation (2010) (a) Convention on the Elimination of All forms of Racial Discrimination, (b) Convention on the Elimination of All forms of Discrimination Against Women. `http://www.ohchr.org`

US Federal Legislation (2010) (a) Equal Credit Opportunity Act, (b) Fair Housing Act, (c) Intentional Employment Discrimination, (d) Equal Pay Act, (e) Pregnancy Discrimination Act, (f) Civil Right Act. `http://www.usdoj.gov`

Verykios VS, Elmagarmid AK, Bertino E, Saygin Y, Dasseni E (2004) Association rule hiding. IEEE Transactions on Knowledge and Data Engineering 16(4):434–447

Wang K, Fung BCM, Yu PS (2005) Template-based privacy preservation in classification problems. In: Proc. of IEEE ICDM 2005, IEEE Computer Society, pp 466–473

Webb GI (2000) Efficient search for association rules. In: Proc. of ACM KDD 2000, ACM, pp 99–107

Wielemaker J (2009) SWI-Prolog. University of Amsterdam, Version 5.6, `http://www.swi-prolog.org`

Williams T, Kelley C (2010) Gnuplot. Version 4.0, `http://www.gnuplot.info`

Yin X, Han J (2003) CPAR: Classification based on Predictive Association Rules. In: Proc. of SIAM SDM 2003, SIAM, pp 331–335

Yinger J (1998) Evidence on discrimination in consumer markets. The Journal of Economic Perspectives 12:23–40

Zeleznikow J, Vossos G, Hunter D (1994) The IKBALS project: Multi-modal reasoning in legal knowledge based system. Artificial Intellence and Law 2(3):169–203