Integrating Induction and Deduction for Finding Evidence of Discrimination

Dino Pedreschi Salvatore Ruggieri Franco Turini

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

ABSTRACT

Automatic Decision Support Systems (DSS) are widely adopted for screening purposes in socially sensitive tasks, including access to credit, mortgage, insurance, labor market and other benefits. While less arbitrary decisions can potentially be guaranteed, automatic DSS can still be discriminating in the socially negative sense of resulting in unfair or unequal treatment of people. We present a reference model for finding (prima facie) evidence of discrimination in automatic DSS which is driven by a few key legal concepts. First, frequent classification rules are extracted from the set of decisions taken by the DSS over an input pool dataset. Key legal concepts are then used to drive the analysis of the set of classification rules, with the aim of discovering patterns of discrimination. We present an implementation, called LP2DD, of the overall reference model integrating induction, through data mining classification rule extraction, and deduction, through a computational logic implementation of the analytical tools.

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Data Mining; I.2.3 [**Artificial Intelligence**]: Deduction and theorem proving

General Terms

Algorithms, Design, Legal Aspects

Keywords

Direct and systematic discrimination, Classification, Data mining, Logic Programming, Scoring systems.

1. INTRODUCTION

Civil right laws [2, 8, 28, 29, 30] 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. Several authorities (regulation boards, consumer advisory councils, commissions) are settled to monitor and report on discrimination compliances in the United States, European Union and many other countries. For instance, the European Commission publishes an annual report [4] on the progress in implementing the Equal Treatment Directives by the E.U. member states. Also, jurisprudence accounts for a large body of cases [7, 17]. From a research point of view, the literature in economics and social sciences has given evidence of unfair treatment in racial profiling and redlining [5, 23], mortgage lending [16], consumer market [21], personnel selection [11], and wages [15].

With 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 [3, 10, 27] or good/bad worker in personnel selection [6]. When applied for automatic decision making, DSS can potentially guarantee less arbitrary 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. 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. Since the burden of the proof is on the respondent, it becomes a priority for the DSS owner to provide confidence on non-discrimination of decisions taken by the DSS. Analogously, regulation authorities must be provided with methods and tools for unveiling discriminatory decisions taken by DSS owners under their surveillance.

In this paper, we propose a reference model for 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 (applicant data) and the output is a class value (a yes/no decision). Our approach consists in first extracting frequent classification rules from the set of decisions taken by the DSS over an input pool dataset, with an inductive approach based on data mining. We consider this set of rules as a model of the historical decisions

ICAIL-2009 Barcelona, Spain

of the DSS. The induced rules are then loaded as part of a meta-reasoner, where we code the key legal measures and reasonings for discovering patterns of direct and systematic discrimination. This is the deductive part of the approach. We present the LP2DD system, written in a computational logic formalism, which implements the proposed reference model as a tool in support of discrimination analysis and discovery.

2. PRELIMINARIES

2.1 Frequent Classification Rules

We recall the notions of itemsets, association rules and classification rules [26]. Let $\mathcal{R}$ be a non-empty relation over attributes $a_1, \ldots, a_n$, namely $\emptyset \subset \mathcal{R} \subseteq dom(a_1) \times \ldots \times dom(a_n)$. We assume that $dom(a)$, the domain of values of a, is finite for every attribute a. Continuous domain can be accounted for by first discretizing values into ranges. Also, an attribute c is fixed and called the class attribute.

An a-item is an expression $a = v$, where a is an attribute and $v \in dom(a)$. An item is any a-item. A c-item is called a class item. Let I be the set of all items. An itemset $\mathbf{X}$ is a subset of I. We denote by 2^I the set of all itemsets. As usual in the literature, we write $\mathbf{X}, \mathbf{Y}$ for $\mathbf{X} \cup \mathbf{Y}$, that is the set of items including both $\mathbf{X}$ and $\mathbf{Y}$. For a tuple $\sigma \in \mathcal{R}$, we say that σ verifies $\mathbf{X}$ if $\sigma \models \mathbf{X}$, namely for every $a = v$ in $\mathbf{X}$, $\sigma(a) = v$. The absolute support of an itemset $\mathbf{X}$ is the number of tuples in $\mathcal{R}$ verifying $\mathbf{X}$: $asupp(\mathbf{X}) = |\{\ \sigma \in \mathcal{R} \mid \sigma \models \mathbf{X}\ \}|$, where $|\ |$ is the cardinality operator. The (relative) support of $\mathbf{X}$ is the ratio of tuples verifying $\mathbf{X}$ over the cardinality of $\mathcal{R}$: $supp(\mathbf{X}) = asupp(\mathbf{X})/|\mathcal{R}|$.

An association rule is an expression $\mathbf{X} \rightarrow \mathbf{Y}$, where $\mathbf{X}$ and $\mathbf{Y}$ are itemsets. $\mathbf{X}$ is called the *premise* (or the *body*) and $\mathbf{Y}$ is called the *consequence* (or the *head*) of the association rule. We say that $\mathbf{X} \rightarrow \mathbf{C}$ is a *classification rule* if $\mathbf{C}$ is a class item and $\mathbf{X}$ contains no class item. The support of $\mathbf{X} \rightarrow \mathbf{Y}$ is defined as: $supp(\mathbf{X} \rightarrow \mathbf{Y}) = supp(\mathbf{X}, \mathbf{Y})$. The coverage of $\mathbf{X} \rightarrow \mathbf{Y}$ is: $cov(\mathbf{X} \rightarrow \mathbf{Y}) = supp(\mathbf{X})$. The confidence, defined when $supp(\mathbf{X}) > 0$, is: $conf(\mathbf{X} \rightarrow \mathbf{Y}) = supp(\mathbf{X}, \mathbf{Y})/supp(\mathbf{X})$. Support, coverage and confidence range over $[0, 1]$. Also, the notation readily extends to negated itemsets $\neg\mathbf{X}$. Many well explored algorithms have been designed in order to extract the set of *frequent* itemsets, i.e., itemsets with a specified minimum support. Starting from them, the association and classification rules with a specified minimum support are readily computable. They are called frequent association and classification rules.

2.2 Logic Programming

We use standard notation for Prolog programs [24]. 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 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. Non-logical predicates include arithmetic assignment (`is`) and comparison predicates (`=\=`, `>=`, `>`). The empty list is denoted by `[]`. The list constructor functor is `[.|.]`.

2.3 The German credit case study

We will report some analyses over the public domain German credit dataset (available from the UCI repository of machine learning datasets, `http://archive.ics.uci.edu/ml`), consisting of 1000 tuples over bank account holders. The dataset includes nominal (or discretized) attributes on *personal properties*: checking account status, duration, savings status, property magnitude, type of housing; on *past/current credits and requested credit*: credit history, credit request purpose, credit request amount, installment commitment, existing credits, other parties, other payment plan; on *employment status*: job type, employment since, number of dependents, own telephone; and on *personal attributes*: personal status and gender, age, resident since, foreign worker. 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 Overall Description

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. The input pool can be built starting from the historical records of applications, possibly enriched with external data, or, as it happens in situation testing [21], from a set of individuals matched for all relevant characteristics other than those expected to lead to discrimination.

The DSS 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*. Starting from it, the set of *frequent classification and association rules* is induced, or, more precisely, the set of 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.

The discrimination analysis relies on the definition of the groups of interest, on a measure of discrimination for a classification rule, computable starting from frequencies extracted from the training set, and on a few legal principles, that can be formalised through meta-rule deductions over the set of extracted rules. The *rule meta-reasoner* component, described in depth in Sect. 4, 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

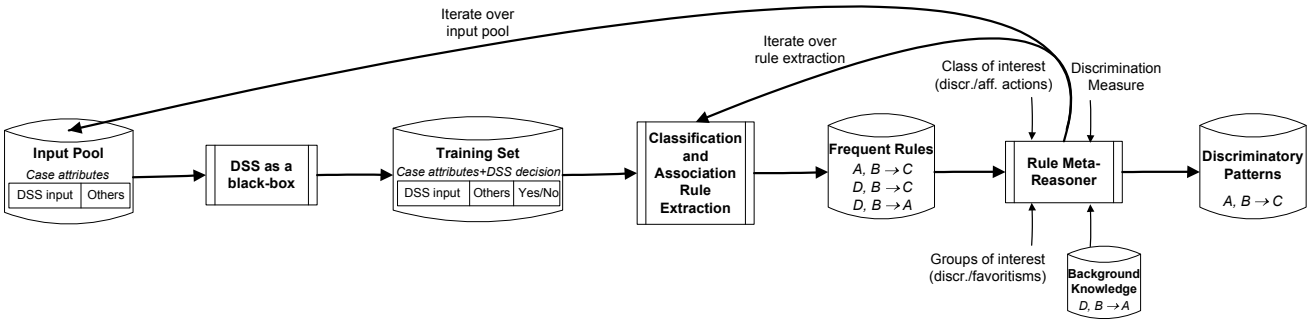

Figure 1: Reference Model for Analysing and Reasoning on Discrimination in DSS.

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 or classification rules.

The output of the discrimination analysis is a set of *discriminatory patterns*, namely classification rules that hold over the training set and such that their discrimination measure is beyond a legally admissible value. Such rules unveil contexts where a protected-by-law group is discriminated.

3.2 DSS as a black-box

One basic assumption in the proposed reference model is that the DSS is supposed to be a black-box, and a form of *reverse engineering* is done through classification rule induction to reconstruct its underlying logic. This is general enough to work with a DSS that has no intelligible nor symbolic representation, as in the case of neural networks and legacy programming languages. However, one could object that the classification rule induction step is unnecessary in the case that the DSS logic is disclosed and intelligible, as when the owner is forced to by a court or a control authority. The following example shows that the rule induction step is instead fundamental. Consider a DSS whose logic consists of the following decisions:

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

These decisions 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 of the cases in the input pool. 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 [14], 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 that (intentionally or not) has discriminatory effects, namely indirect discrimination.

With the approach of the proposed reference model, the classification rule `sex = female, ZIP = 101` $\rightarrow$ `credit = no` is extracted from the training set in the above table. Rule meta-reasoning over extracted rules allows for checking that a discrimination measure for the rule is beyond an acceptable value, thus unveiling a context (`ZIP = 101`) where a protected group (`sex = female`) is discriminated.

4. RULE ANALYSIS AND REASONING

Let us describe in detail the key component of the reference model, namely the rule meta-reasoner. First, we discuss in Sect. 4.1 how to denote the groups that are protected by the law against discrimination, and, consequently, how to locate them in a classification rule premise. In Sect. 4.2 a few discrimination measures are introduced, by formalizing existing laws and regulations. On this basis, key legal reasonings about direct discrimination evidence, respondent defence, affirmative actions, and indirect discrimination are formalized as deductions of discriminatory rules from the set of extracted rules and, possibly, from background knowledge. These deductions are discussed in Sects. 4.3-4.6.

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 immediate 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 is a, possibly empty, smaller (even more disadvantaged) minority as well. As a consequence, we have to generalize the notion of PD item to the one of potentially discriminatory (PD) itemset $\mathcal{I}_d$, which can be defined as those itemsets built on PD items only, i.e., $\mathcal{I}_d = 2^{I_d}$. 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.

Consider now the case of 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. Stated formally, setting $\mathcal{I}_d = 2^{I_d}$ for some set of PD items I_d is not an enough general definition for PD itemsets. 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 [19].

DEFINITION 4.1. *A set of itemsets $\mathcal{I}$ is downward closed if when $\mathbf{A}_1 \in \mathcal{I}$ and $\mathbf{A}_2 \in \mathcal{I}$ then $\mathbf{A}_1, \mathbf{A}_2 \in \mathcal{I}$.*

This property is the formal counterpart of "gender-plus" allegations [7], 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 part-time female workers. The downward closure property is sufficient for separating PD itemsets in the premise of a classification rule. In fact, given $\mathbf{X} \to \mathbf{C}$, 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 $\mathcal{I}_d$. A classification rule $\mathbf{A}, \mathbf{B} \to \mathbf{C}$ is called potentially discriminatory (PD rule) if $\mathbf{A}$ is non-empty, and potentially non-discriminatory (PND rule) otherwise.

4.2 Measures of Discrimination

In this section, we will consider a few measures of the degree of discrimination of a rule $\mathbf{A}, \mathbf{B} \to \mathbf{C}$, where $\mathbf{A}$ is the PD itemset and $\mathbf{B}$ is the PND itemset in its premise. Also, we call $\mathbf{B}$ the context. The definition of a quantitative measure is the building block for monitoring and finding evidence of an unacceptable level of discrimination. However, there is no uniformity or general agreement on a standard definition by legislations and, within a same country, by cases of jurisprudence. A general principle is to consider group under-representation [14] as a quantitative measure of the qualitative requirement that people in a group are treated "less favorably" [8, 28] than others, or such that "a higher proportion of people without the attribute comply or are able to comply" [2] to a qualifying criteria.

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

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

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

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

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

Figure 2: Contingency table for a classification rule

We recall from [19] 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 $\mathbf{B} \to \mathbf{C}$ where no PD itemset appears. We introduce its definition by starting from the contingency table of $\mathbf{A}, \mathbf{B} \to \mathbf{C}$ shown in Fig. 2. Each cell in the table is filled in by the number of tuples (i.e., the absolute support) in the training set satisfying $\mathbf{B}$ and the coordinates. Using the notation of the figure, we have $conf(\mathbf{A}, \mathbf{B} \to \mathbf{C}) = p_1 = a_1/n_1$ and $conf(\mathbf{B} \to \mathbf{C}) = p = (a_1 + a_2)/(n_1 + n_2)$.

DEFINITION 4.2. *Let $c = \mathbf{A}, \mathbf{B} \to \mathbf{C}$ be a classification rule with contingency table as in Fig. 2.*

The extended lift of c is defined when $p > 0$ as: $elift(c) = p_1/p = conf(\mathbf{A}, \mathbf{B} \to \mathbf{C})/conf(\mathbf{B} \to \mathbf{C})$.

A rule `sex=female, car=own` → `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. By trivial algebra, we observe that:

$$elift(c) = conf(\mathbf{B}, \mathbf{C} \to \mathbf{A})/conf(\mathbf{B} \to \mathbf{A}),$$

namely the extended lift can be defined 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 $\mathbf{B}$. This makes it clear how extended lift relates to the principle of group representation.

In addition to extended lift, other measures can be formalized starting from different definitions of discrimination provided in laws. The Racial Equality Directive of E.U. [8] states that discrimination "shall be taken to occur where one person is treated less favorably than another is in a comparable situation on grounds of racial or ethnic origin". Here the comparison appears to be done between two races (the disadvantaged one and the favored one). The U.S. legislation goes further [30, (d) Section 4D] by stating that "a

selection rate for any race, sex, or ethnic group which is less than four-fifths (or eighty percent) of the rate for the group with the highest rate will generally be regarded as evidence of adverse impact". We model the contrast between a disadvantaged group $\mathbf{A}$ and the rest of the population $\neg\mathbf{A}$ by defining the selection lift of c as $slift(c) = p_1/p_2$, using the notation of Fig. 2. Since we are considering benefit refusal (denial rate), the four-fifths rule turns out to fix a maximum acceptable value for $slift()$ of $5/4 = 1.25$.

Also, in the employment discrimination literature [9], the measure of odds ratio has been considered, defined as $odds(c) = p_1 \cdot (1 - p_2)/(p_2 \cdot (1 - p_1))$. 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 $p/(1 - p)$. In employment discrimination, the "event" modelled is promotion or hiring of a person. The odds ratio is then 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.

Although the measures introduced so far are defined in terms of ratios, measures based on the difference of confidences have been considered on the legal side as well. For instance, in the U.K., a difference of 5% in confidence between female ($\mathbf{A}$ is `sex=female`) and male ($\neg\mathbf{A}$ is `sex=female`) treatment is assumed by courts as significant of discrimination against women. Two difference-based measures are reported in Fig. 2. Finally, we mention that tests of statistical significance are customary in legal cases before courts [1, 9] as a means to rule out (classification rules whose) contingency tables that lead to high measure values but that are statistically nonsignificant at some level of confidence.

4.3 Direct Discrimination

Direct discrimination occurs "where one person is treated less favorably than another" [7]. 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. The burden of proof will then shift to the respondent who must prove that there has been no breach of the principle of equal treatment. Again, the respondent may use aggregate analysis to show balanced decisions over groups. The proposed reference model can support both positions.

In direct discrimination, we assume that the input pool dataset contains attributes to denote the group of interest $\mathbf{A}$ (e.g., `sex=female`) for each case in the input pool. A PD classification rule denying benefit, i.e., of the form:

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

supports the complainant position if she belongs to the disadvantaged group $\mathbf{A}$, she satisfies the context conditions $\mathbf{B}$ and the discrimination measure of the rule (w.r.t. one of the definitions of Sect. 4.2) is above an acceptable level with reference to what is stated in law, regulations or past sentences, e.g., the four-fifths rule.

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" [7]. A typical example in the literature is the one of the "genuine occupational requirement". 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 a more general rule `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). Formally, we say 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 $conf(\mathbf{A}, \mathbf{B} \rightarrow \mathbf{D}) = 1$. The two conditions can be relaxed as follows.

DEFINITION 4.3. *Let p be in $[0, 1]$. A classification rule $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$ is a p-instance of $\mathbf{D}, \mathbf{B} \rightarrow \mathbf{C}$ if:*
(1) $conf(\mathbf{D}, \mathbf{B} \rightarrow \mathbf{C}) \geq p \cdot conf(\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C})$; and,
(2) $conf(\mathbf{A}, \mathbf{B} \rightarrow \mathbf{D}) \geq p$.

The issue for the respondent, however, consists now of finding out a suitable itemset $\mathbf{D}$ and a factor $p \approx 1$. 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 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 the search to frequent association rules of the form $\mathbf{D}, \mathbf{B} \rightarrow \mathbf{A}$, which are extracted from the training set (see Fig. 1). This has the advantage that the search is over a much smaller set of association rules[1].

4.4 Affirmative Actions

Many legislations account for affirmative actions [22], 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 from the mere encouragement of under-represented groups to quotas in favor of those groups. Citing Article 2.2 from [29](a), these 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". It is therefore important to assess and to monitor the application of affirmative actions. In the proposed reference model, this can be achieved by analysing PD classification rules granting benefit:

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

Rules of this form having a value of the adopted measure greater than a fixed threshold highlight contexts $\mathbf{B}$ where the disadvantaged group $\mathbf{A}$ is actually favored.

[1]For every 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 $\mathbf{B}$.

4.5 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. 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/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. 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.

4.6 Indirect Discrimination

The E.U. Directives 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" [7]. The counterpart to this broad definition is that the type and number of admissible inferences of indirect discrimination is left open. In our reference model, indirect discrimination occurs when the input pool does not contain attributes to denote the group under analysis. For instance, the information on a person's race is typically not available and, in many countries, not even collectable. In the classification rule terminology, only PND rules $\mathbf{D}, \mathbf{B} \rightarrow \mathbf{C}$, where $\mathbf{D}, \mathbf{B}$ is a PND itemset, can be extracted from the training set. Can we use PND rules to unveil, at least partially, discriminatory patterns?

A typical example of indirect discrimination is concerned with redlining. We mention here the *Hussein vs Saints Complete House Furniture* case [18], where a Liverpool furniture store refused to consider (almost all) applicants for jobs from a particular postal area 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%. However, the Labour Force Surveys indicated that 50% of the population in the postal area were 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 the extend lift (or another measure) of the rule:

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

is particularly high, where the PD itemset is the one written in parenthesis. This issue has been considered in [19], where an inference strategy exploiting background knowledge is proposed. Let us recall here the approach. 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 $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$ (right-hand side):

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

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

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

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

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

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

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

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

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

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

We have $\beta_1 = 1$ by definition of $\mathbf{A}$, $\beta_2 = 0.5$ since 50% of population in the postal area is black, 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:

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

In general, lower and upper bounds for the various discrimination measures can be devised, starting from extracted PND classification rules and background knowledge relating potentially discriminated groups $\mathbf{A}$ to other groups of people $\mathbf{D}$ in specific contexts $\mathbf{B}$. In our reference model, background knowledge is modelled in the syntax of association rules of the form $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{D}$ and $\mathbf{D}, \mathbf{B} \rightarrow \mathbf{A}$. As a result, indirect discrimination inference strategies boil down to "rule inference" strategies, where PD rules are inferred starting from background knowledge and PND rules. The following definition formalizes the redlining strategy.

DEFINITION 4.4. *A 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 exists a background knowledge rule* $\mathbf{D}, \mathbf{B} \rightarrow \mathbf{A}$ *such that, called:*

$$\gamma = conf(\mathbf{D}, \mathbf{B} \rightarrow \mathbf{C}) \quad \delta = conf(\mathbf{B} \rightarrow \mathbf{C})$$

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

we have: $lb = 1/\beta_2(\beta_2 + \gamma - 1)/\delta$.

5. THE LP2DD ANALYTICAL SYSTEM

The proposed reference model provides 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 LP2DD system relies on data mining algorithms for the inductive part (classification and association rule extraction) and in SWI-Prolog (`http://www.swi-prolog.org`) for the deductive part (the meta-reasoner). Any frequent pattern extraction algorithm from the Frequent Itemset Mining Implementations repository (`http://fimi.cs.helsinki.fi`) can be plugged-in the system. The user-interface is in SWI-Prolog, calling external modules when required.

5.1 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 non-single women;

4 extract frequent association and classification rules having a minimum support threshold (10 in the log).

```
% 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 .
```

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. 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) ).
...
```

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. Predicates are provided in the LP2DD system for decoding (`itemset_decode`) itemsets; 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 two sample queries related to counting PND rules and to splitting a rule premise into its PD and PND parts.

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

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

5.2 Meta-Reasoner

The core of the meta-rule reasoner is shown in Fig. 3. A few measures are defined for a given contingency table, including confidence of PND rules (clause `(cn1)`) and PD rules `(cn2)`, coverage (clauses `(cv1,cv2)`), extended lift `(el)`,

```
(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).
```

```
(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) inference(B2Min, D, B, C, LB, A) :-
      background(DB, A1, CT_BDA),
      confidence(CT_BDA, B2),
      B2 >= B2Min,
      split(DB, D, B),
      pndrule(DB, C, CT_DBC),
      pndrule(B, C, CT_BC),
      confidence(CT_BC, DELTA),
      confidence(CT_DBC, GAMMA),
      LB is 1/B2*(B2+GAMMA-1)/DELTA,
      merge(A1, D, A).
```

Figure 3: A Core Meta-Reasoner over Extracted Classification Rules based on Computational Logic.

selection lift (`sl`), and odds lift (`ol`). PD classification rules with a 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. 4.3-4.4, checking for discrimination and affirmative actions is modelled by searching for classification rules denying credit (see predicate `discrimination` in clause (`d`)) and granting credit (see `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 intuitive words, for checking discriminatory patterns w.r.t. the selection lift measure;

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 changes.

```
% 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
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) .
```

Predicate `pinstance` defined by clause (`in`) checks whether a PND classification rule is a p-instance of some PD rule, according to Def. 4.3. A goal `:- pinstance(A, B, C, CT, MinP, D, P)` instantiates `D` to an itemset $\mathbf{D}$ and `P` to a

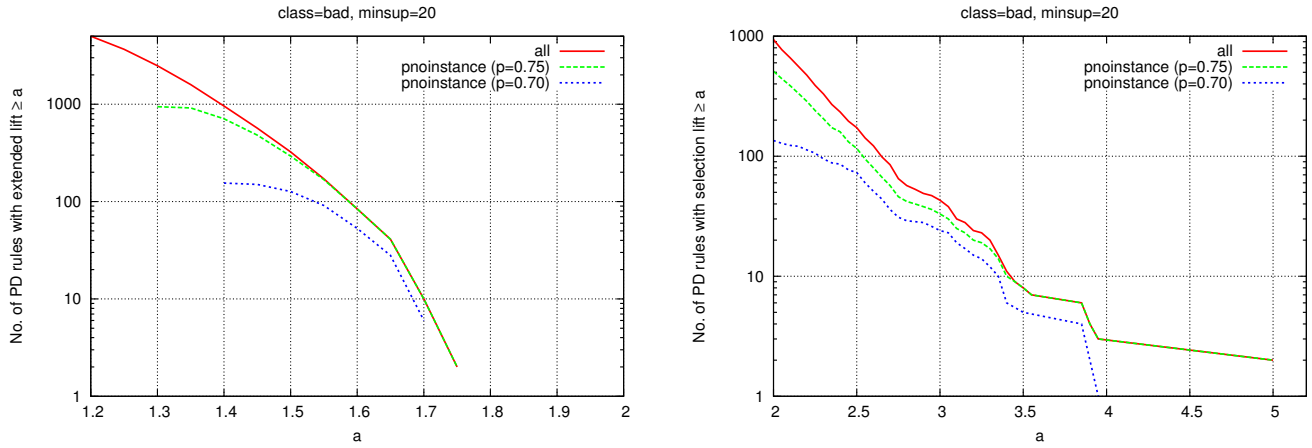

Figure 4: Distributions of extended lift (left) and selection lift (right) for all PD rules and for PD rules that are not p-instances of any PND rule.

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:

7 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;

8 enumerate PND rules that are 0.8-instances of PD rules;

9 enumerate PND rules that are not 0.8-instances of PD rules.

```
% Discriminatory PND rules that are not instances
% of PD rules
7 ?- 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
8 ?- discrimination(slift, 3, A, B, C, CT),
        pinstance(A, B, C, CT, 0.8, D, P).
A = [personal_status=female_div_or_sep_or_mar],
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
9 ?- 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) .
```

We performed extensive experimentations with the German credit dataset to assess the functionalities of the meta-reasoner. 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 construction of a "gold" dataset from real cases of direct discrimination, indirect discrimination and genuine occupational requirements should be pursued as a means to evaluate the quality of discovered patterns of discrimination, according to some evaluation strategy [25]. Fig. 4 shows the distributions of extended and selection lifts for all PD classification rules, and for PD rules that are not p-instances of any PND rule, for sample $p = 0.7$ and $p = 0.75$. The plots are obtained as outputs of the LP2DD system. The number of classification 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 principle provided in Def. 4.3.

Let us consider now indirect discrimination. The redlining inference strategy of Def. 4.4 is implemented by the `inference` predicated (see clause `(i)` in Fig. 3). The search for PD rules is driven by background knowledge association rules $\mathbf{DB} \rightarrow \mathbf{A}_1$ having some minimum confidence `B2Min`, namely stating that the protected group $\mathbf{A}_1$ represents at least a fraction `B2Min` 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. 4.4. Finally, the PD itemset in the inferred PD rule $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$ is built as $\mathbf{A} = \mathbf{A}_1, \mathbf{D}$.

The following Prolog goal over the German credit dataset searches for a PD rule with a selection lift of at least 2. 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. 5.1).

```
% Searching for indirect discrimination
10 ?- inference(0.8, D, B, C, LB, A), LB >= 2.
LB = 2.40625,
D = [purpose=furniture_or_equipment],
B = [employment=0-1, housing=rent, own_telephone=none],
C = (class=good),
A = [purpose=furniture_or_equipment,
        personal_status=female_div_or_sep_or_mar] .
```

In the answer, the context $\mathbf{B}$ consists of people employed by at most one year, renting an house, and not owning a phone. In such a context, at least 80% of people asking for credit to buy furniture or equipment (i.e., $\mathbf{D}$) are non-single women (i.e., $\mathbf{A}_1$), where the threshold of 80% has been specified as a parameter in the above goal. Having denied credit to people in the context that intended to buy furniture or equipment had the effect of denying credit mainly to women. Formally, the rule $(\mathbf{A}_1, \mathbf{D}), \mathbf{B} \rightarrow \mathbf{C}$ has a selection lift of 2.40625 or higher.

6. CONCLUSIONS

This paper introduced a reference model for the analysis and reasoning of discrimination in DSS. The approach consists first of extracting frequent classification rules, and then on analysing them on the basis of quantitative measures of discrimination. Key legal concepts are formalized into reasonings on the set of extracted rules and background knowledge. We have developed a logic programming system, called LP2DD, implementing the reference model, that is intended as an analytical tool supporting DSS owners and control authorities in the interactive and iterative process of discrimination analysis.

The approach presented can be refined in several directions. First, albeit a black-box view is enough for *unveiling* discrimination, we observe that the owner of a DSS may be interested in *building* DSS that take no discriminatory decision. The problem of *discrimination preventing* DSS, however, cannot be tackled without entering the details of the internal representation of the DSS. A first approach, dealing with data mining classifiers, is reported in [13]. Second, the approach based on classification rules could be extended to account for continuous decisions (e.g., wage amount, mortgage interest rate) and for continuous attributes (e.g., age, income) without resorting to apriori discretization. Third, 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 [26] or defeasible logic [12]. Finally, the LP2DD system could be integrated with computational logic models of legal argument, such as those based on logic meta-programming [20].

7. REFERENCES

[1] O. Ashenfelter and R. Oaxaca. The economics of discrimination: Economists enter the courtroom. *The American Economic Review*, 77:321–325, 1987.

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

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

[4] M. Bell, I. Chopin, and F. Palmer. *Developing Anti-Discrimination Law in Europe*. 2007. `http://ec.europa.eu/employment_social/fundamental_rights`.

[5] P. S. Calem, K. Gillen, and S. Wachter. The neighborhood distribution of subprime mortgage lending. *J. of Real Estate Finance and Economics*, 29:393–410, 2004.

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

[7] E. Ellis. *EU Anti-Discrimination Law*. Oxford University Press, 2005.

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

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

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

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

[12] B. Johnston and G. Governatori. Induction of defeasible logic theories in the legal domain. In *Proc. of ICAIL 2003*, pages 204–213. ACM, 2003.

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

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

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

[16] M. LaCour-Little. Discrimination in mortgage lending: A critical review of the literature. *J. of Real Estate Literature,*, 7:15–49, 1999.

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

[18] T. Makkonen. *Measuring Discrimination: Data Collection and the EU Equality Law*. 2006. `http://ec.europa.eu/employment_social/fundamental_rights`.

[19] D. Pedreschi, S. Ruggieri, and F. Turini. Discrimination-aware data mining. In *Proc. of the Int.'l Conf. on Knowledge Discovery and Data Mining (KDD 2008)*, pages 560–568. ACM, 2008.

[20] H. Prakken and G. Sartor. The role of logic in computational models of legal argument: A critical survey. In A. C. Kakas and F. Sadri, editors, *Computational Logic. Logic Programming and Beyond*, volume 2408 of *Lecture Notes in Computer Science*, pages 342–381. Springer, 2002.

[21] P. A. Riach and J. Rich. Field experiments of discrimination in the market place. *The Economic Journal*, 112:480–518, 2002.

[22] T. Sowell, editor. *Affirmative Action Around the World: An Empirical Analysis*. Yale University Press, 2005.

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

[24] L. Sterling and E. Shapiro. *The Art of Prolog*. The MIT Press, 1986.

[25] A. Stranieri and J. Zeleznikow. The evaluation of legal knowledge based systems. In *Proc. of ICAIL 1999*, pages 18–24. ACM, 1999.

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

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

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

[29] United Nations Legislation. (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`.

[30] U.S. Federal Legislation. (a) Equal Credit Opportunity Act, (b) Fair Housing Act, (c) Intentional Employment Discrimination, (d) Equal Pay Act, (e) Pregnancy Discrimination Act. `http://www.usdoj.gov`.