Chapter 5
The Discovery of Discrimination

Dino Pedreschi and Salvatore Ruggieri and Franco Turini

Abstract Discrimination discovery from data consists in the extraction of discriminatory situations and practices hidden in a large amount of historical decision records. We discuss the challenging problems in discrimination discovery, and present, in a unified form, a framework based on classification rules extraction and filtering on the basis of legally-grounded interestingness measures. The framework is implemented in the publicly available DCUBE tool. As a running example, we use a public dataset on credit scoring.

5.1 Introduction

Human right laws (European Union Legislation, 2011; United Nations Legislation, 2011; U.S. Federal Legislation, 2011) prohibit discrimination against protected groups on the grounds of race, color, religion, nationality, sex, marital status, age and pregnancy; and 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) 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 Chopin & Do, 2010); and in the US the Attorney General reports to the Congress on the annual referrals to the Equal Credit Opportunity Act.

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. Classical approaches adopted in legal cases (Finkelstein & Levin, 2001) are limited to the

D. Pedreschi and S. Ruggieri and F. Turini
Dipartimento di Informatica, Università di Pisa, Italy.
Email: {pedre,ruggieri,turini}@di.unipi.it

verification of an hypothesis of possible discrimination by means of statistical analysis of past decision records. However, they reveals to be inadequate to cope with the problem of *searching for* niches of discriminatory decisions hidden in a large dataset of decisions.

Discrimination discovery from data consists in the actual discovery of discriminatory situations and practices hidden in a large amount of historical decision records. The aim is to *extract 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 future 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. Control authorities can base the fight against discrimination on a formalized process of intelligent data analysis.

However, discrimination discovery from data may reveal itself an extremely difficult task. The reason is twofold. First, personal data in decision records are typically highly dimensional: as a consequence, a huge number of possible contexts may, or may not, be the theater for discrimination. To see this point, consider the case of gender discrimination in credit approval: although an analyst may observe that no discrimination occurs in general, it may turn out that foreign worker women obtain loans to buy a new car only rarely. 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. The anti-discrimination analyst is thus faced with a combinatorial explosion of possibilities, which make her work hard: albeit the task of checking some known suspicious situations can be conducted using available statistical methods and known stigmatized groups, the task of discovering niches of discrimination in the data is unsupported. The second source of complexity is *indirect discrimination* (see e.g., Tobler, 2008), namely apparently neutral practices that take into account personal attributes correlated with indicators of race, gender, and other protected grounds and that result in discriminatory effects on such protected groups. Even when the race of a credit applicant is not directly recorded in the data, racial discrimination may occur, e.g., as in the practice of *redlining*: people living in a certain neighborhood are frequently denied credit; while not explicitly mentioning race, this fact can be an indicator of discrimination, if from demographic data we can learn that most of people living in that neighborhood belong to the same ethnic minority. Once again, the anti-discrimination analyst is faced with a large space of possibly discriminatory situations: how can she highlight all interesting discriminatory situations that emerge from the data, both directly and in combination with further background knowledge in her possession (e.g., census data)?

We present a classification rule mining approach for the discrimination discovery problem, based on the following ideas. Decision policies are induced from past decision records as classification rules of the form: PREMISES → DECISION, where each rule comes with a confidence measure, stating the probability of the decision given the premises of the rule; for instance, the rule RACE=BLACK, CITY=NYC → CLASS=BAD with confidence 0.75 states that black people from NYC are assigned bad credit with a 75% probability.

Three kinds of facts (items) are used in decision rules: (potentially) discriminatory items, such as RACE=BLACK, (potentially) non-discriminatory items, such as CITY=NYC, and decision items, such as CLASS=BAD. The potentially discriminatory items are specified by a reference legal framework, to denote some designated groups of people protected by the anti-discrimination laws. The non-discriminatory items define the context where a discriminatory decision may take place - here, the set of applicants from the city of NYC.

Given an historical dataset of decision records, the decision rules hidden in the dataset can be found using *association rule mining*, which allows to extract all the classification rules of the desired form that, in the source dataset, are supported by a specified minimum number of decisions. Continuing the example, the rule RACE=BLACK, CITY=NYC → CLASS=BAD is automatically found by association rule mining, if the number of black people in NYC receiving the bad credit is above a minimum threshold value. Such a threshold, known as the minimum support, is meaningful from a legal viewpoint, since it accounts for a minimum number of possibly discriminated persons.

In which circumstances does an extracted rule reveal a (possibly unintentional) discriminatory decision strategy? The idea here is to weight the discrimination of a rule by the gain of confidence due to the presence of the potentially discriminatory items in the premise of the rule. In the example, we compare the 0.75 confidence of the rule RACE=BLACK, CITY=NYC → CLASS=BAD with the confidence of the rule obtained removing the first item, i.e., CITY=NYC → CLASS=BAD. If, e.g., the confidence of the latter rule is 0.25, then we conclude that black people in NYC have a probability of being assigned bad credit which is 3 times larger than that of the general population of NYC. In this case, a measure called *elift* is used to quantify discrimination risk, which is defined as the ratio of the confidence of the two rules above (with and without the discriminatory item). Whether the rule in the example is to be considered discriminatory or not can now be assessed by thresholding the *elift* measure - possibly according to a value specified in the reference legislation, that limits the acceptable disproportion of treatment. While we use *elift* to illustrate examples throughout the chapter, it is worth noting that several other measures of discrimination (see Section 5.2.2) have been considered in the legal and economic literature, none of which is superior to the others. Actually, our approach is parametric in the definition of a reference measure.

By considering all classification rules with a value of the *elift* higher than the threshold, we can find all the contexts where a discriminatory decision has been taken: in the example, by enumerating *all rules* of the form RACE=BLACK, **B** → CLASS=BAD an anti-discrimination analyst discovers all situations **B** where black people suffered a discriminatory credit decision, whatever the complexity of the context **B** and in compliance with the reference legal framework.

So far, we have assumed that discriminatory items are recorded in the source data. This is not always the case, e.g., race may be not available or even collectable. What if the discriminatory variables are not directly available? In this case, indirect discrimination may occur. Consider the rule ZIP=10451, CITY=NYC → CLASS=BAD, with confidence 0.95, stating that the residents of a given neighborhood of NYC are assigned bad credit with a 95% chance. Apparently, this rule does not unveil any discriminatory practice. However, assume that the following other rule can be coded from available information, such as census data: ZIP=10451, CITY=NYC → RACE=BLACK, with confidence 0.80, stating that 80% of residents of that particular neighborhood of NYC are black. Then it is possible to prove a theoretical lower bound of 0.94 for the confidence of the combined rule ZIP=10451, CITY=NYC, RACE=BLACK → CLASS=BAD, stating that 94% of black people in that neighborhood are assigned bad credit, around 3.7 times the general population of NYC. This reasoning shows that the original rule unveils a case of redlining.

Different measures of the discrimination power of the mined decision rules can be defined, according to the provision of different anti-discrimination regulations: e.g., the EU Directives (European Union Legislation, 2011) state that discrimination on a given attribute occurs when “a higher proportion of people without the attribute comply or are able to comply” (which we will code as the *risk ratio* measure), while the US Equal Pay Act (U.S. Federal Legislation, 2011) states that: “a selection rate for any race, sex, or ethnic group which is less than four-fifths of the rate for the group with the highest rate will generally be regarded as evidence of adverse impact” (which we will code as the *selection ratio* measure).

Our discrimination discovery approach opens a promising avenue for research, based on an apparently paradoxical idea: data mining, which is typically used to create potentially discriminatory profiles and classifications, can also be used the other way round, as a powerful aid to the anti-discrimination analyst, capable of automatically discovering the patterns of discrimination that emerge from the available data with the strongest *prima facie* evidence. The preliminary experiments on a dataset of credit decisions operated by a German bank show that this method is able to pinpoint evidence of discrimination: the cited highly discriminatory rule that “foreign worker women are assigned bad credit among those who intend to buy a new car” is actually discovered from such a database.

The rest of the chapter is organized as follows. Section 5.2 introduces the technicalities of classification rules and measures of discrimination defined over them. Using those tools, we show how the anti-discrimination analyst can go through the analysis of direct discrimination (Section 5.3), indirect discrimination (Section 5.4), respondent argumentation (Section 5.5), and affirmative actions (Section 5.6). Some details on the analytical tool DCUBE, which supports the discrimination discovery process, are provided in Section 5.7. Finally, we summarize the approach and discuss some challenging lines for future research.

Attributes

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

Decision

CLASS, with values GOOD (grant credit) and BAD (deny credit)

Potentially discriminatory (PD) items

PERSONAL_STATUS=FEMALE *(female)*
AGE=GT_52 *(senior people)*
FOREIGN_WORKER=YES *(foreign workers)*

PERS_STATUS	AGE	JOB	PURPOSE	CREDIT_AMNT	HOUSING	...	CLASS
female	gt_52	self_emp	new_car	lt_38_k	rent	...	bad
male married	30_to_41	unemp	used_car	39k_to_75_k	own	...	good
male single	42_to_51	skilled	business	75k_to_111k	for_free	...	good
female	gt_52	unemp	furniture	lt_38_k	own	...	bad
...	...	...	...	...	...	...	...

Table 5.1 The German credit case study: attributes (top) and an excerpt of the dataset (bottom)

5.2 Classification Rules for Discrimination Discovery

As a running example throughout the chapter, we refer to the public domain German credit dataset, publicly available from the UCI repository of machine learning datasets (Newman et al., 1998). The dataset consists of 1000 records over bank account holders. It includes 20 nominal (or discretized) attributes as shown in Table 5.1. The decision attribute takes values representing the good/bad creditor classification of the bank account holder.

5.2.1 Classification Rules

Given a relation with n attributes, we refer to an *item* as an expression $a = v$, where a is an attribute and v one of its possible values. For example PERSONAL_STATUS = MALE SINGLE is an item for the German credit dataset. One of the attributes is taken as the class attribute, i.e., the attribute referring to the decision. In our running example, the class is named CLASS and the two possible items are CLASS = GOOD, that is credit is granted, and CLASS = BAD, that is credit is denied.

A *transaction* T is a set of items, one for each attribute of the relation. Intuitively, a transaction is the set of items corresponding to a row of a table. By an *itemset* $\mathbf{X}$ we mean a set of items, and we say that a transaction T *supports* an itemset $\mathbf{X}$ if every item in $\mathbf{X}$ belongs to T as well, in symbols $\mathbf{X} \subseteq T$. As an example, the transaction corresponding to the first row in Table 5.1 supports the itemset PERSONAL_STATUS = FEMALE, AGE = GT_52 but not PERSONAL_STATUS = MALE SINGLE, AGE =

GT_52. A dataset $\mathscr{D}$ is a set of transactions. Intuitively, it corresponds to the transactions built from a table.

The support of an itemset $\mathbf{X}$ w.r.t. $\mathscr{D}$ is the proportion of transactions in $\mathscr{D}$ supporting $\mathbf{X}$: $supp(\mathbf{X}) = |\{\ T \in \mathscr{D} \mid \mathbf{X} \subseteq T\ \}|/|\mathscr{D}|$, where $|\ |$ is the cardinality operator.

An association rule is an expression $\mathbf{X} \to \mathbf{Y}$, where $\mathbf{X}$ and $\mathbf{Y}$ are disjoint itemsets. $\mathbf{X}$ is called the *premise* and $\mathbf{Y}$ is called the *consequence* of the association rule. We say that $\mathbf{X} \to \mathbf{Y}$ is a *classification rule* if $\mathbf{Y}$ is a class item. As an example, PERSONAL_STATUS = FEMALE, AGE = GT_52 → CLASS = BAD is a classification rule for the German credit dataset.

The support of $\mathbf{X} \to \mathbf{Y}$ is the support of the itemset obtained by the union of $\mathbf{X}$ and $\mathbf{Y}$, in symbols $supp(\mathbf{X}, \mathbf{Y})$, where $\mathbf{X}, \mathbf{Y}$ is the union of $\mathbf{X}$ and $\mathbf{Y}$. Intuitively, the support of a rule states how often the rule is satisfied in the dataset. A support of 0.1 for the rule PERSONAL_STATUS = FEMALE, AGE = GT_52 → CLASS = BAD means that 10% of the transactions support both the premise and the consequence of the rule, i.e., support PERSONAL_STATUS = FEMALE, AGE = GT_52, CLASS = BAD. The confidence of $\mathbf{X} \to \mathbf{Y}$, defined when $supp(\mathbf{X}) > 0$, is:

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

Confidence states the proportion of transactions supporting $\mathbf{Y}$ among those supporting $\mathbf{X}$. A confidence of 0.7 for the rule above means that 70% of the transactions supporting PERSONAL_STATUS = FEMALE, AGE = GT_52 also support CLASS = BAD. Support and confidence range over $[0, 1]$. Since the seminal paper by (Agrawal & Srikant, 1994), many well explored algorithms have been designed for extracting 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 & Hamilton, 2006); a repository of implementations is maintained by (Goethals, 2010).

5.2.2 Measures of Discrimination

A critical 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). 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 against, $\mathbf{B}$ is the itemset that chacterizes the context, and $\mathbf{C}$ is the decision (class) item, then the analysis of discrimination is pursued by studying the rule $\mathbf{A}, \mathbf{B} \to \mathbf{C}$, together with its confidence with respect to the underlying decision dataset - namely, how often such a rule is true in the dataset itself.

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 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 owning a car.

However, discrimination typically occurs for subgroups rather than for the whole group (the US courts coined the term "gender-plus allegations" to describe conducts breaching the law on the ground of sex-plus-something-else), or it may occur for multiple causes (called *multiple discrimination* in ENAR, 2007). 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 generalize the notion of potentially discriminatory *item* to the one of potentially discriminatory (PD) *itemset*, and assume that the downward closure property holds for PD itemsets (Ruggieri et al., 2010a).

Definition 1. If $\mathbf{A}_1$ and $\mathbf{A}_2$ are PD itemsets, then $\mathbf{A}_1, \mathbf{A}_2$ is a PD itemset as well.

On the technical side, the downward closure property is a sufficient condition for separating PD itemsets in the premise of a classification rule, namely, there is only one way $\mathbf{A}, \mathbf{B}$ of splitting the premise of a rule into a PD itemset $\mathbf{A}$ and a PND itemset $\mathbf{B}$.

Definition 2. 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. PD rules cannot be extracted from datasets that do not contain potentially discriminatory items. In such a case, PND rules can still indirectly unveil discriminatory practices (see Section 5.4).

Let us consider now how to quantitatively measure the "burden" imposed on such groups and unveiled by a discovered PD rule. 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, 2011; U.K. Legislation, 2011) than others, or such that "a higher proportion of people without the attribute comply or are able to comply" (see Australian Legislation, 2011) to a qualifying criterium. We recall from (Ruggieri et al., 2010a) the notion of extended lift[1], 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.

[1] The term "extended lift" originates from the fact that it conservatively extends the well-known measure of *lift* (or *interest factor*) of an association rule (Tan et al., 2004), which is obtained, as a special case, when $\mathbf{B}$ empty. Conversely, the extended lift of $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$ corresponds to the lift of $\mathbf{A} \rightarrow \mathbf{C}$ over the set of transactions supporting $\mathbf{B}$.

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

	benefit ($\mathbf{C}$)		
group	denied	granted	
protected ($\mathbf{A}$)	a	b	n_1
unprotected ($\neg\mathbf{A}$)	c	d	n_2
	m_1	m_2	n (total of $\mathbf{B}$)

$$p_1 = a/n_1 \quad p_2 = c/n_2 \quad p = m_1/n$$

$$RD = p_1 - p_2 \quad RR = \frac{p_1}{p_2} \quad RC = \frac{1-p_1}{1-p_2} \quad OR = \frac{RR}{RC} = \frac{a/b}{c/d}$$

$$ED = p_1 - p \quad ER = \frac{p_1}{p} \quad EC = \frac{1-p_1}{1-p}$$

Fig. 5.1 Contingency table and discrimination measures.

Definition 3. 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 → CREDIT=NO with an extended lift of 3 means that being a female increases 3 times the probability of being 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. 5.5 for a discussion). An alternative way, yet equivalent, 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 group under-representation in benefit granting. In addition to extended lift, other measures can be formalized starting from different definitions of discrimination provided by laws. They can be defined over the 2 × 2 contingency table shown in Figure 5.1, showing the absolute number of transactions in the underlying dataset $\mathscr{D}$ satisfying the itemsets in the X-Y coordinates and the context $\mathbf{B}$. Let p_1 (resp., p_2) be the proportion of people in the protected group (resp., not in the protected group) that were not granted a benefit, and let p be the proportion of all people (both protected and not) that were not granted the benefit. The following discrimination measures can be defined:

- *risk difference* (RD $= p_1 - p_2$), also known as *absolute risk reduction*,
- *risk ratio* or *relative risk* (RR $= p_1/p_2$),
- *relative chance* (RC $= (1-p_1)/(1-p_2)$), also known as *selection rate*,

- *odds ratio* ($\mathrm{OR} = p_1(1-p_2)/(p_2(1-p_1))$),

and the versions of RD, RR, and RC when the protected group is compared to the average proportion p, rather than to the proportion of the unprotected group:

- *extended difference* ($\mathrm{ED} = p_1 - p$);
- *extended ratio* or *extended lift* ($\mathrm{ER} = p_1/p$);
- *extended chance* ($\mathrm{EC} = (1-p_1)/(1-p)$).

Since one is interested in contexts of higher benefit denial (resp., lower benefit granting) for the protected group compared to the unprotected group or to the average, the values of interest for RR, OR, and ER are those greater than 1; for RD and ED are those greater than 0; and for RC and EC are those lower than 1. Confidence intervals and tests of statistical significance of the above measures are discussed in (Pedreschi et al., 2009; Ruggieri et al., 2010c). Here, we only mention that statistical tests will rank the rules according to how unlikely it is that they would be observed if there was equal treatment, not according to the severity of discrimination. As an example, a mild discrimination among a large population will be ranked higher than a much more severe discrimination in a small community.

From the legal side, different measures are adopted worldwide. UK law (U.K. Legislation, 2011, (a)) mentions risk difference, EU Court of Justice has given more emphasis to the risk ratio (see Schiek et al., 2007, Section 3.5), and US laws and courts mainly refer to the selection rate[2]. Notice that the risk ratio is the ratio of the proportions of *benefit denial* between the protected and unprotected groups, while selection rate is the ratio of the proportions of *benefit granting*. The EU is more concerned about the ratio of denials, while the US is more concerned about the ratio of grants; unfortunately, they do not lead to the same conclusions in discrimination discovery.

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 (for simplicity, we limit ourselves to the extended lift measure).

Definition 4 (a-protection). We say that a PD classification rule $\mathbf{A},\mathbf{B} \to \mathbf{C}$ is a-protective if $\mathit{elift}(\mathbf{A},\mathbf{B} \to \mathbf{C}) < a$. Otherwise, we say that it is a-discriminatory.

Intuitively, a is a fixed threshold stating an acceptable level of discrimination according to laws, regulations, and jurisprudence. Classification rules denying a benefit 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*[3] evidence of discrimination. While a-protection is defined with reference to *elift*,

[2] (U.S. Federal Legislation, 2011, (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". This is called the *four-fifths rule*. It turns out to fix a minimum threshold value for RC of $4/5 = 0.8$.

[3] *Prima facie* is a Latin term meaning "at first look," or "on its face," and refers to evidence that, unless rebutted, would be sufficient to prove a particular proposition or fact.

its definition clearly applies to any measure from Figure 5.1. An extension of *a*-protection to account for its statistical significance is proposed in (Pedreschi et al., 2009; Ruggieri et al., 2010c). Also, we refer the reader to (Ruggieri et al., 2010a, 2010c) for the presentation and experimentation of data mining algorithms able to efficiently extract *a*-protective classification rules from a large dataset of historical decision records. Finally, (Pedreschi et al., 2012) show that the choice of a reference measure from Figure 5.1 has a critical impact on the ranking imposed over the set of PD classification rules. In other words, selecting a specific discrimination measure is not a neutral choice, in that it implicitly implies a specific moral criterion to evaluate the degree of discrimination in a specific context; i.e., different ways to establish how bad is a discriminatory action. We found it interesting that our quantitative logical framework for discriminatory rules can help understanding the consequences of such choices in law and jurisprudence.

5.3 Direct Discrimination Discovery

From this section on, we formalize various legal concepts in discrimination analysis and discovery as reasonings over the set of extracted classification rules. 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 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:

$$\mathbf{A},\mathbf{B} \rightarrow \text{ BENEFIT=DENIED}$$

i.e., where the consequent consists of denying a benefit (a loan, school admission, a job, etc.). Rules of the form above are then screened by selecting/ranking those with a minimum value of a reference discrimination measure. In terms of Def. 4, we are then looking for “*a*-discrimination of PD classification rules denying benefit”.

As an example, consider our running example dataset and fix the PD items as in Table 5.1. By ranking classification rules of the form $\mathbf{A},\mathbf{B} \rightarrow$ CLASS=BAD accordingly to their extended lift measure, we found near the top positions the following:

PERSONAL_STATUS=FEMALE, FOREIGN_WORKER=YES,

$$\text{PURPOSE=NEW\_CAR} \rightarrow \text{CLASS=BAD}$$

with an extended lift of 1.58. The rule can be interpreted as follows: among those applying for loans to buy a new car, female foreign workers had 1.58 times the average chance of being refused the requested credit. The rule above has a confidence of 0.277, meaning that female foreign workers asking a loan to buy a new car had credit denied in 27.7% of cases (precisely, 13 transactions out of 47). The rule for the generality of applicants:

$$\text{PURPOSE=NEW\_CAR} \rightarrow \text{CLASS=BAD}$$

has a confidence of 0.175, meaning that people asking a loan to buy a new car had credit denied in 17.5% of cases.

5.4 Indirect Discrimination Discovery

The EU Directives (see European Union Legislation, 2011; Tobler, 2008) provide a broad definition of indirect discrimination (also known as systematic discrimination or disparate impact) 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 $\mathbf{A},\mathbf{B} \rightarrow \mathbf{C}$ that is *a*-directly discriminatory, while an "apparently neutral provision" is modelled by a potentially non-discriminatory (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 $\mathbf{A},\mathbf{B} \rightarrow \mathbf{C}$ may be unavailable[4], e.g., because the dataset 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 our approach to indirect discrimination, the problem consists then of inferring some PD rule (with a high discrimination measure value) starting from the set of PND rules, and, possibly, from additional background knowledge. The adjective *potentially non-discriminatory* was chosen exactly to underline that, since the rule does not refer to protected groups, it does not unveil any discriminatory practice in a direct way. Nevertheless, it could do that indirectly.

A remarkable example is *redlining*, a form of indirect discrimination that is explicitly banned in the US (U.S. Federal Legislation, 2011, (b)). As sharply pointed out in Figure 5.2, racial segregation very often emerges in most cities characterized by ethnic diversity: the spatial clustering of a city into racially homogeneous areas is observed in reality much more often than the dispersion of races into an integrated structure. We know from Schelling's segregation model (Schelling, 1971) that a natural tendency to spatial segregation emerges, as a collective phenomenon,

[4] Otherwise, the technique of Section 5.3 can be adopted to unveil the effects of both direct and indirect discrimination.

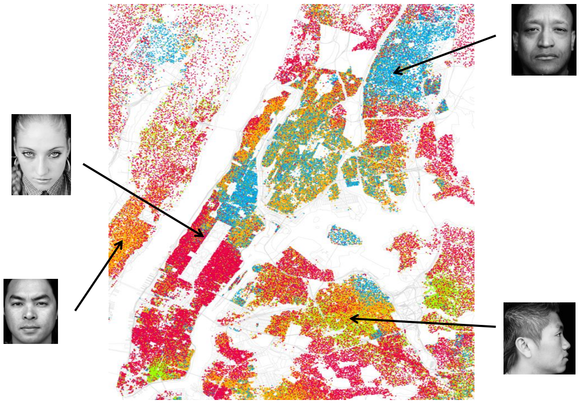

Fig. 5.2 Racial segregation in New York City, based on Census 2000 data (Fischer, 2011). One dot for each 500 residents. Red dots are Whites, blue dots are Blacks, green dots are Asian, orange dots are Hispanic, and yellow dots are other races.

even if each individual person is relatively tolerant and open-minded: in his famous abstract simulation model, Schelling showed how segregation eventually appears even if each person changes his residence only if less than 30% of his neighbors are of his same race. That's why so many urban territories world-wide, in absence of social restrictions or incentives, developed a structure such that depicted in Figure 5.2; in turn, this explains why denying credit or benefits on the basis of residence – drawing a red line on the border of an urban neighborhood – is often an indirect way to discriminate on the basis of race. Let us consider an example of inference in the context of redlining inspired by the *Hussein vs Saints Complete House Furniture* case reported by (Makkonen, 2006), albeit the numbers reported here are fictious. Assume that a Liverpool furniture store refuses to consider 99% of applicants to a job from a particular postal area ZIP=1234 which had a high rate of unemployment. The extracted classification rule ZIP=1234, CITY=LIVERPOOL → APP=NO with confidence $\gamma = 0.99$ is apparently neutral with respect to race discrimination. Assume also that the average refusal rate in the Liverpool area is much lower, say 9%. With our notation, the rule CITY=LIVERPOOL → APP=NO has then confidence $p = 0.09$. Assume now to know, e.g., from census background knowledge, that 80% of the population in the postal area ZIP=1234 is black, i.e., that the area is mainly populated by minorities. In formal terms, the association rule ZIP=1234, CITY=LIVERPOOL → RACE=BLACK has confidence $\beta = 0.8$. It is now legitimate to ask ourselves whether from such rules, one can conclude a form of redlining, namely the use of ZIP=1234 as a proxy for excluding blacks from a benefit (accepting the side effect of possibly excluding some whites from the same neighborhood). Formally, we want to check whether the extended lift of:

$$(\text{ZIP=1234, RACE=BLACK}),\ \text{CITY=LIVERPOOL} \rightarrow \text{APP=NO} \qquad (\star)$$

is particularly high, where the PD itemset **A** is ZIP=1234, RACE=BLACK, denoting blacks living in the area, and the context **B** is CITY=LIVERPOOL, denoting that the comparison is made against the overall population of that city. The extended lift of such a rule can be read as the ratio of the refusal rate of black people in the ZIP over the mean refusal rate of the whole city. A lower bound for the confidence p_1 of the classification rule ($\star$) can be obtained as $p_1 \geq 1-(1-\gamma)/\beta = 1-0.01/0.8 = 0.9875$ (for details, see Ruggieri et al., 2010a). Intuitively, even in the extreme case that the whole 1% of people in the area who were admitted are blacks, the ratio of un-admitted blacks cannot be lower than 98.75%. By knowing that the average admission rate for the generality of people from Liverpool is 9%, the lower bound for the *elift* measure of ($\star$) is $p_1/p \geq 0.9875/0.09 = 10.97$ – and extremely high ratio stating that black people from that area had at least 10.97 times the average chance (of a Liverpool applicant) of seeing their application refused.

We conclude by mentioning that the redlining inference strategy is one possible inference reasoning for deducing unknown discriminatory effects from observed, apparently non-discriminatory, ones. Additional inference strategies are proposed in (Ruggieri et al., 2010a). In general, an inference strategy consists of deriving lower bounds for a discrimination measure of an unavailable PD rule starting from: assumptions on the form of the premise of the rule; and some background knowledge, which in our framework is coded in the form of association rules. The situation resembles here what occurs in privacy-preserving data mining (Agrawal & Srikant, 2000; Sweeney, 2001), where coupling an anonymized dataset with external knowledge might allow for the inference of the identity of individuals through some attack strategy.

5.5 Argumentation

Consider a PD classification rule denying some benefit:

$$\mathbf{A}, \mathbf{B} \rightarrow \text{BENEFIT=DENIED}$$

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 **A**, she satisfies the context conditions **B** and the rule is a-directly discriminatory 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, 2011, (f)). For instance, assume that the complainant claims for discrimination against women among applicants to a job position. A classification rule SEX=FEMALE, CITY=NYC → HIRE=NO with high extended lift supports her position. The respondent might argue that the rule is an instance of a more general rule DRIVE_TRUCK=FALSE, CITY=NYC → 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 $\mathbf{D},\mathbf{B}\rightarrow\mathbf{C}$ when:

- a transaction satisfying $\mathbf{A}$ in context $\mathbf{B}$ satisfies condition $\mathbf{D}$ as well, or, in symbols, $conf(\mathbf{A},\mathbf{B}\rightarrow\mathbf{D})$ is close to 1;
- and, the rule $\mathbf{D},\mathbf{B}\rightarrow\mathbf{C}$ holds at the same or higher confidence, or, in symbols, $conf(\mathbf{D},\mathbf{B}\rightarrow\mathbf{C}) \geq conf(\mathbf{A},\mathbf{B}\rightarrow\mathbf{C})$;

A respondent argumenting against discriminatory allegations supported by a PD rule $\mathbf{A},\mathbf{B}\rightarrow\mathbf{C}$ must show that the rule is an instance of some PND rule $\mathbf{D},\mathbf{B}\rightarrow\mathbf{C}$, and with $\mathbf{D}$ modelling a genuine occupational requirement. On the contrary, a complainant or a control authority can prevent respondent's argumentation by showing that the PD rule $\mathbf{A},\mathbf{B}\rightarrow\mathbf{C}$ is not an instance of any PND rule $\mathbf{D},\mathbf{B}\rightarrow\mathbf{C}$. In (Ruggieri et al., 2010c), the concept of "instance" has been relaxed to the notion of p-instance, requiring $conf(\mathbf{A},\mathbf{B}\rightarrow\mathbf{D}) \geq p$ and $conf(\mathbf{D},\mathbf{B}\rightarrow\mathbf{C}) \geq p \cdot conf(\mathbf{A},\mathbf{B}\rightarrow\mathbf{C})$. On the experimental side, the vast majority of discriminatory PD rules extracted from the German credit dataset result (p-)instances of some PND rule, thus concluding that it is (fortunately) extremely difficult to characterize *prima facie* evidence of discrimination.

Another defence strategy of the respondent is to resort to the well-known Sympson's paradox. (Bickel, Hammel, & O'Connell, 1975) describes a real case of possible discrimination against women in university admission. Let us rephrase it using our notation. Assume that the rule SEX=FEMALE → ADMITTED=NO has an high extended lift, so that a possible discrimination is raised. By examining each individual department A of the university, however, it can happen that each rule SEX=FEMALE, DEPT=A → ADMITTED=NO has a very low extended lift, denoting no discrimination at all. The paradox is that the discrimination observed at university level did not actually occur in any department. If the examination commissions worked at department level, then the department attribute is causal factor, and the standard approach (Pearl, 2009) is to condition probabilities and rules on it. As a consequence, the rules at department level are the correct ones to be looked at, whilst the rule at university level contains confounding factors (the commissions that took decisions).

5.6 Affirmative Actions

Affirmative actions (see ENAR, 2008; Sowell, 2005), sometimes called positive actions or reverse discrimination, are 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 underrepresented groups to quotas in favor of those groups. For instance, US federal contractors are required to identify and set goals for hiring under-utilized minorities and women. Also, universities have voluntarily implemented admission policies that give preferential treatment to women and minority candidates. 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, 2011, (a)). It is therefore important to assess and to monitor the application of affirmative actions. In our approach, affirmative actions can be unveiled 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:

$$\mathbf{A},\mathbf{B} \rightarrow \text{ BENEFIT=GRANTED}$$

i.e., where the consequent consists of granting a benefit (a loan, a school admission, a job, etc.). Rules of this form with a value of the discrimination measure greater than a fixed threshold highlight contexts **B** where the disadvantaged group **A** was actually favored.

Once again, consider our running example dataset. By ranking classification rules of the form $\mathbf{A},\mathbf{B} \rightarrow$ CLASS=GOOD accordingly to their extended lift measure, we found near the top positions the following:

$$\text{AGE = GT\_}52\text{, JOB = UNEMPLOYED } \rightarrow \text{ CLASS=GOOD}$$

with an extended lift of 1.39. The rule can be interpreted as follows: among those unemployed, people older than 52 had 1.39 times the average chance of being granted the requested credit. This could be the case, for instance, of some affirmative actions supporting economic initiatives of unemployed older people.

5.7 The DCUBE Tool

The various concepts and analyses so far discussed, originally implemented as stand-alone programs for achieving the best performances, have been re-designed around an Oracle database, used to store extracted rules, and a collection of functions, procedures and snippets of SQL queries that implement the various legal reasonings for discrimination analysis. The resulting implementation, called DCUBE (Discrimination Discovery in Databases) (Ruggieri et al., 2010b), can be accessed and exploited by a wider audience if compared to a stand-alone monolithic application. In fact, SQL is the dominant query language for relational data, with database administrators already mastering issues such as data storage, query optimization, and import/export towards other formats. Discrimination discovery is an interactive and iterative process, where analyses assume the form of deductive reasoning over extracted rules. An appropriately designed database, with optimized indexes, functions and SQL query snippets, can be welcome by a large audience of users, including owners of socially-sensitive decision data, government anti-discrimination analysts, technical consultants in legal cases, researchers in social sciences, economics and law. Typical discrimination discovery questions that DCUBE is able to answer include:

Direct discrimination discovery: *"How much have women been under-represented in obtaining the loan?"* or *"List under which conditions blacks were suffering an extended lift higher than 1.8 in our recruitment data"*. DCUBE comes with all of the legally-grounded measures from Figure 5.1 predefined. The user can adopt any of them or, even, she can easily define new measures over a 4-fold contingency table by adding methods to an Oracle user defined data type.

Indirect discrimination discovery, such as the following redlining question *"I don't have the race attribute in my data, but have the ZIP of residence. By adding background knowledge on the distribution of race over ZIP codes, infer cases where ZIP actually disguises race discrimination."*

Affirmative actions and favoritism: *"List cases where our university admission policies actually favored blacks"*, and *"Under which conditions white males are given the best mortgage rate in comparison to the average?"*

On-line documentation, demo, and download of the DCUBE system can be accessed from http://kdd.di.unipi.it/dcube.

5.8 Conclusions

We presented a data mining approach for the analysis and discovery of discrimination in a dataset of socially-sensitive decisions. The approach consists first of extracting frequent classification rules, and then screening/ranking them on the basis of quantitative measures of discrimination. The key legal concepts of protected-by-law groups, direct discrimination, indirect discrimination, genuine occupational requirement, and affirmative actions are formalized as reasonings over the set of extracted rules and, possibly, additional background knowledge. The approach has been implemented in the DCUBE tool and made publicly available. Chapter 13 builds on our approach for the purpose of designing data mining classifiers that do not learn to discriminate, an issue known as discrimination prevention.

As future work, we aim to achieve two goals: on one hand, to improve the methods and the technologies for discovering discrimination, especially looking at data mining methods such at classification and clustering, driven by constraints over specific application contexts (racial profiling, labour market, credit scoring, etc.); on the other hand, to further interact with legal experts both to find out new measures and rules that we may support with our tools and to influence their design and interpretation of legislation. Finally, we are looking at other fields of application, other than credit scoring. An interesting one is discovering possible discrimination (with respect to sex, nationality, etc.) in funding research projects.

References

Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules in large databases. In *Proc. of Int. Conf. on Very Large Data Bases (VLDB 1994)* (p. 487-499). Morgan Kaufmann.

Agrawal, R., & Srikant, R. (2000). Privacy-preserving data mining. In *Proc. of the ACM SIGMOD Int. Conf. on Management of Data (SIGMOD 2000)* (p. 439-450). ACM.

Australian Legislation. (2011). *(a) Age Discrimination Act, 2004; (b) Australian Human Rights Commission Act, 1986; (c) Disability Discrimination Act, 1992; (d) Racial Discrimination Act, 1975; (e) Sex Discrimination Act, 1984; (f) Victoria Equal Opportunity Act, 1995, (g) Queensland Anti-Discrimination Act, 1991.* (http://www.hreoc.gov.au)

Bickel, P., Hammel, E., & O'Connell, J. (1975). Sex bias in graduate admissions: Data from berkeley. *Science*, *187*(4175), 398-404.

Chopin, I., & Do, T. U. (2010). *Developing anti-discrimination law in europe*. European Network of Legal Experts in Anti-Discrimination. (http://ec.europa.eu)

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. (2011). *(a) European Convention on Human Rights, 1950; (b) Racial Equality Directive, 2000; (c) Employment Equality Directive, 2000; (d) Gender Goods and Services Directive, 2004; (e) Gender Employment Directive, 2006; (f) Equal Treatment Directive (proposal), 2008.* (http://eur-lex.europa.eu)

Finkelstein, M. O., & Levin, B. (Eds.). (2001). *Statistics for lawyers* (2nd ed.). Springer-Verlag.

Fischer, E. (2011). *Distribution of race and ethnicity in US major cities.* (Published on line at http://www.flickr.com/photos/walkingsf under Creative Commons licence, CC BY-SA 2.0)

Geng, L., & Hamilton, H. J. (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.

Knopff, R. (1986). On proving discrimination: Statistical methods and unfolding policy logics. *Canadian Public Policy*, *12*(4), 573-583.

Makkonen, T. (2006). *Measuring discrimination: Data collection and the EU equality law*. European Network of Legal Experts in Anti-Discrimination. (http://www.migpolgroup.com)

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

Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2nd ed.). New York, USA: Cambridge University Press.

Pedreschi, D., Ruggieri, S., & Turini, F. (2009). Measuring discrimination in socially-sensitive decision records. In *Proc. of the SIAM Int. Conf. on Data Mining (SDM 2009)* (pp. 581–592). SIAM.

Pedreschi, D., Ruggieri, S., & Turini, F. (2012). A study of top-k measures for discrimination discovery. In *Proc. of ACM Int. Symposium On Applied Computing (SAC 2012)* (p. 126-131). ACM.

Ruggieri, S., Pedreschi, D., & Turini, F. (2010a). Data mining for discrimination discovery. *ACM Trans. on Knowledge Discovery from Data*, *4*(2), 1–40.

Ruggieri, S., Pedreschi, D., & Turini, F. (2010b). DCUBE: Discrimination discovery in databases. In *Proc. of the ACM SIGMOD Int. Conf. on Management of Data (SIGMOD 2010)* (pp. 1127–1130). ACM.

Ruggieri, S., Pedreschi, D., & Turini, F. (2010c). Integrating induction and deduction for finding evidence of discrimination. *Artificial Intelligence and Law*, *18*(1), 1–43.

Schelling, T. C. (1971). Dynamic models of segregation. *Journal of Mathematical Sociology*, *1*(May), 143–186.

Schiek, D., Waddington, L., & Bell, M. (Eds.). (2007). *Cases, materials and text on national, supranational and international non-discrimination law*. Hart Publishing.

Sowell, T. (Ed.). (2005). *Affirmative action around the world: An empirical analysis*. Yale University Press.

Sweeney, L. (2001). *Computational disclosure control: A primer on data privacy protection*. Unpublished doctoral dissertation, MIT, Cambridge, MA.

Tan, P.-N., Kumar, V., & Srivastava, J. (2004). Selecting the right objective measure for association analysis. *Information Systems*, *29*(4), 293–313.

Tobler, C. (2008). *Limits and potential of the concept of indirect discrimination*. European Network of Legal Experts in Anti-Discrimination. (http://www.migpolgroup.com)

U.K. Legislation. (2011). *(a) Sex Discrimination Act, 1975, (b) Race Relation Act, 1976.* (http://www.statutelaw.gov.uk)

United Nations Legislation. (2011). *(a) Convention on the Elimination of All forms of Racial Discrimination, 1966, (b) Convention on the Elimination of All forms of Discrimination Against Women, 1979.* (http://www.ohchr.org)

U.S. Federal Legislation. (2011). *(a) Equal Credit Opportunity Act, 1974; (b) Fair Housing Act, 1968; (c) Employment Act, 1967; (d) Equal Pay Act, 1963; (e) Pregnancy Discrimination Act, 1978; (f) Civil Right Act, 1964, 1991.* (http://www.eeoc.gov)