Chapter 5
The Discovery of Discrimination

Dino Pedreschi, 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 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 argue 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: \( \text{PREMISES} \rightarrow \text{DECISION} \), where