Chapter 13
Direct and Indirect Discrimination Prevention Methods

Sara Hajian and Josep Domingo-Ferrer

Universitat Rovira i Virgili

Department of Computer Engineering and Mathematics

UNESCO Chair in Data Privacy

{sara.hajian,josep.domingo}@urv.cat

Abstract Along with privacy, discrimination is a very important issue when considering the legal and ethical aspects of data mining. It is more than obvious that most people do not want to be discriminated because of their gender, religion, nationality, age and so on, especially when those attributes are used for making decisions about them like giving them a job, loan, insurance, etc. Discovering such potential biases and eliminating them from the training data without harming their decision-making utility is therefore highly desirable. For this reason, anti-discrimination techniques including discrimination discovery and prevention have been introduced in data mining. Discrimination prevention consists of inducing patterns that do not lead to discriminatory decisions even if the original training datasets are inherently biased. In this chapter, by focusing on the discrimination prevention, we present a taxonomy for classifying and examining discrimination prevention methods. Then, we introduce a group of pre-processing discrimination prevention methods and specify the different features of each approach and how these approaches deal with direct or indirect discrimination. A presentation of metrics used to evaluate the performance of those approaches is also given. Finally, we conclude our study by enumerating interesting future directions in this research body.

13.1 Introduction

Unfairly treating people on the basis of their belonging to a specific group, namely race, ideology, gender, etc., is known as discrimination. In law, economics and social sciences, discrimination has been studied over the last decades and anti-discrimination laws have been adopted by many democratic governments. Some examples are the US Employment Non-Discrimination Act (United States Congress 1994), the UK Sex Discrimination Act (Parliament of the United Kingdom 1975) and the UK Race Relations Act (Parliament of the United Kingdom 1976).

There are several decision-making tasks which lend themselves to discrimination, e.g. loan granting, education, health insurances and staff selection. In many scenarios, decision-making tasks are supported by information systems. Given a set of information items on a potential customer, an automated system decides whether the customer is to be recommended for a credit or a certain type of life insurance. Automating such decisions reduces the workload of the staff of banks and insurance companies, among other organizations. The use of information systems based on data mining technology for decision making has attracted the attention of many researchers in the field of computer science. In consequence, automated data collection and a plethora of data mining techniques such as association/classification rule mining have been designed and are currently widely used for making automated decisions.

At first sight, automating decisions may give a sense of fairness: classification rules (decision rules) do not guide themselves by personal preferences. However, at a closer look, one realizes that classification rules are actually learned by the system based on training data. If the training data are inherently biased for or against a particular community (for example, foreigners), the learned model may show a discriminatory prejudiced behavior. For example, in a certain loan granting organization, foreign people might systematically have been denied access to loans throughout the years. If this biased historical dataset is used as training data to learn classification rules for an automated loan granting system, the learned rules will also show biased behavior toward foreign people. In other words, the system may infer that just being foreign is a legitimate reason for loan denial. A more detailed analysis of this fact is provided in Chapter 3.

Figure 13.1 illustrates the process of discriminatory and non-discriminatory decision rule extraction. If the original biased dataset *DB* is used for data analysis without any anti-discrimination process (i.e. discrimination discovery and prevention), the discriminatory rules extracted could lead to automated unfair decisions. On the contrary, *DB* can go through an anti-discrimination process so that the learned rules are free of discrimination, given a list of discriminatory attributes (*e.g.* gender, race, age, etc.). As a result, fair and legitimate automated decisions are enabled.

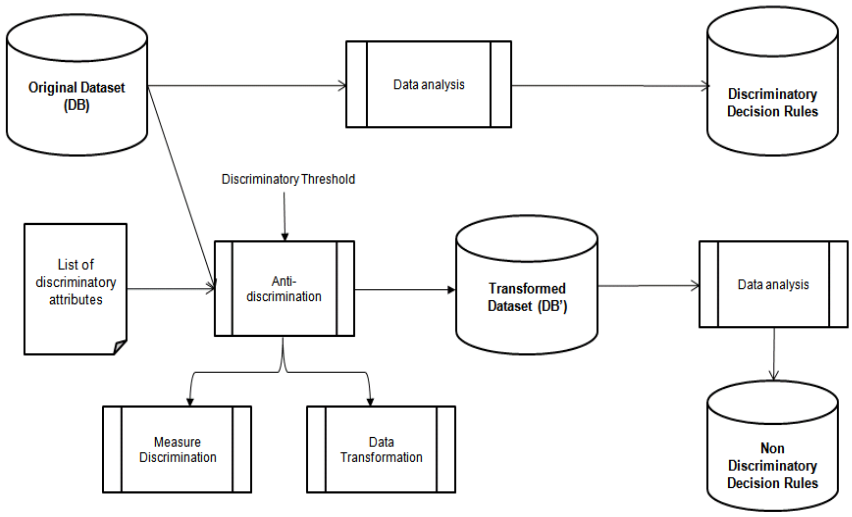

Fig. 13.1. The process of extracting biased and unbiased decision rules.

Despite the wide deployment of information systems based on data mining technology in decision making, the issue of anti-discrimination in data mining did not receive much attention until 2008 (Pedreschi et al. 2008). After that, some proposals have addressed the discovery and measure of discrimination. Others deal with the prevention of discrimination. The discovery of discriminatory decisions was first proposed by Pedreschi *et al.* (2008) and Ruggieri *et al.* (2010). The approach is based on mining classification rules (the inductive part) and reasoning on them (the deductive part) on the basis of quantitative measures of discrimination that formalize legal definitions of discrimination. For instance, the U.S. Equal Pay Act (United States Congress 1963) 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".

Discrimination can be either direct or indirect (also called systematic, see Pedreschi *et al.* (2008)). Direct discriminatory rules indicate biased rules that are directly inferred from discriminatory items (*e.g.* Foreign worker = Yes). Indirect discriminatory rules (redlining rules) indicate biased rules that are indirectly inferred from non-discriminatory items (*e.g.* Zip = 10451) because of their correlation with discriminatory ones. Indirect discrimination could happen because of the availability of some background knowledge (rules), for example, indicating that a certain zipcode corresponds to a deteriorating area or an area with a mostly black population. The background knowledge might be accessible from publicly available data (*e.g.* census data) or might be obtained from the original dataset itself because of the existence of non-discriminatory attributes that are highly correlated with the sensitive ones in the original dataset.

One might conceive that, for direct discrimination prevention, removing discriminatory attributes from the dataset and, for indirect discrimination prevention, removing non-discriminatory attributes that are highly correlated with the sensitive ones could be a basic way to handle discrimination. However, in practice this is not advisable because in this process much useful information would be lost and the quality/utility of the resulting training datasets and data mining models would substantially decrease.

The rest of this chapter is as follows. Section 13.2 contains notation and background on direct and indirect discriminatory rules. Section 13.3 gives a taxonomy of discrimination prevention methods. Section 13.4 describes several preprocessing discrimination prevention methods we have proposed in recent papers. Metrics to measure the success at removing discriminatory rules are given in Section 13.5. Data quality metrics are listed in Section 13.6. Section 13.7 contains experimental results for the direct discrimination prevention methods proposed. Conclusions and suggestions for future work are summarized in Section 13.8.

13.2 Preliminaries

In this section we briefly recall some basic concepts which are useful to better understand the study presented in this chapter.

13.2.1 Basic Notions

- A *dataset* is a collection of data objects (records) and their attributes. Let *DB* be the original dataset.
- An *item* is an attribute along with its value, *e.g.* {Race=black}.
- An *itemset*, *i.e. X*, is a collection of one or more items, *e.g.* {Foreign worker=Yes, City=NYC}.
- A *classification rule* is an expression $X \rightarrow C$, where *C* is a class item (a yes/no decision), and *X* is an itemset containing no class item, e.g. {Foreign worker=Yes, City=NYC} → {hire=no}. *X* is called the premise of the rule.
- The *support* of an itemset, *supp(X)*, is the fraction of records that contain the itemset *X*. We say that a rule $X \rightarrow C$ is *completely supported* by a record if both *X* and *C* appear in the record.
- The *confidence* of a classification rule, *conf*($X \rightarrow C$), measures how often the class item *C* appears in records that contain *X*. Hence, if *supp(X)*> 0

$$conf(X \rightarrow C) = \frac{supp(X, C)}{supp(X)}$$

Support and confidence range over *[0,1]*.

- A *frequent classification rule* is a classification rule with a support or confidence greater than a specified lower bound. Let *FR* be the database of frequent classification rules extracted from *DB*.
- *Discriminatory attributes and itemsets (protected by law):* Attributes are classified as discriminatory according to the applicable anti-discrimination acts (laws). For instance, U.S. federal laws prohibit discrimination on the basis of the following attributes: race, color, religion, nationality, sex, marital status, age and pregnancy (Pedreschi et al. 2008). Hence these attributes are regarded as discriminatory and the itemsets corresponding to them are called discriminatory itemsets. {Gender=Female, Race=Black} is just an example of a discriminatory itemset. Let DA_s be the set of predetermined discriminatory attributes in *DB* and DI_s be the set of predetermined discriminatory itemsets in *DB*.
- *Non-discriminatory attributes and itemsets*: If A_s is the set of all the attributes in *DB* and I_s the set of all the itemsets in *DB*, then nDA_s (*i.e.* set of *non-discriminatory attributes*) is A_s - DA_s and nDI_s (*i.e.* set of *non-discriminatory itemsets*) is I_s - DI_s. An example of non-discriminatory itemset could be {Zip= 10451, City=NYC}.

- The *negated itemset*, *i.e.* $\sim X$ is an itemset with the same attributes as *X*, but such that the attributes in $\sim X$ take any value except those taken by attributes in *X*. In this chapter, we use the ~ notation for itemsets with binary or categorical attributes. For a binary attribute, *e.g.* {Foreign worker=Yes/No}, if *X* is {Foreign worker=Yes}, then $\sim X$ is {Foreign worker=No}. Then, if *X* is binary, it can be converted to $\sim X$ and vice versa. However, for a categorical (non-binary) attribute, *e.g.* {Race=Black/White/Indian}, if *X* is {Race=Black}, then $\sim X$ is {Race=White} or {Race=Indian}. In this case, $\sim X$ can be converted to *X* without ambiguity, but the conversion of *X* into $\sim X$ is not uniquely defined, which we denote by $\sim X \Rightarrow X$. In this chapter, we use only non-ambiguous negations.

13.2.2 Direct and Indirect Discriminatory Rules

As more precisely discussed in Chapter 5, frequent classification rules fall into one of the following two classes: 1) A classification rule (r: $X \rightarrow C$) with negative decision (e.g. denying credit or hiring) is potentially discriminatory (PD) if $X \cap DI_s \neq \emptyset$, otherwise r is potentially non-discriminatory (PND). For example, if DI_s = {Foreign worker=Yes}, a classification rule {Foreign worker=Yes; City=NYC}→Hire=No is PD, whereas {Zip=10451, City=NYC} → Hire=No, or {Experience=Low; City=NYC} → Hire=No are PND.

The word ``potentially" means that a PD rule could probably lead to discriminatory decisions, hence some measures are needed to quantify the direct discrimination potential. Also, a PND rule could lead to discriminatory decisions in combination with some background knowledge; *e.g.*, if the premise of the PND rule contains the zipcode as attribute and one knows that zipcode 10451 is mostly inhabited by foreign people. Hence, measures are needed to quantify the indirect discrimination potential as well.

As mentioned before, Pedreschi *et al.* (2008) and Pedreschi *et al.* (2009a) translated qualitative discrimination statements in existing laws, regulations and legal cases into quantitative formal counterparts over classification rules and they introduced a family of measures over PD rules (for example *elift*) for direct discrimination discovery and over PND rules (for example *elb*) for indirect discrimination discovery. Then, by thresholding *elift* it can be assessed whether the PD rule has direct discrimination potential. Based on this measure (*elift*), a PD rule ($r: X \rightarrow C$) is said to be *discriminatory* if $elift(r) \geq \alpha$[1] or *protective* if $elift(r) < \alpha$. In addition, whether the PND rule has indirect discrimination potential can be assessed by thresholding *elb*. Based on this measure (*elb*), a PND rule ($r': X \rightarrow C$) is said to be *redlining* if $elb(r') \geq \alpha$ or *non-redlining (legitimate)* if $elb(r') < \alpha$. For more detailed information and definitions of these measures, see Chapter 5.

13.3 Taxonomy of Discrimination Prevention Methods

Beyond discrimination discovery, preventing knowledge-based decision support systems from making discriminatory decisions (discrimination prevention) is a more challenging issue. The challenge increases if we want to prevent not only direct discrimination but also indirect discrimination or both at the same time. In this section, we present a taxonomy of discrimination prevention methods after having reviewed a collection of independent works in the area. Figure 13.2 shows this taxonomy. In order to be able to classify the various approaches, we consider two orthogonal dimensions based on which we present the existing approaches. As a

[1] Note that α is a fixed threshold stating an acceptable level of discrimination according to laws and regulations. For example, the four-fifths rule of U.S. Federal Legislation sets α=1.25.

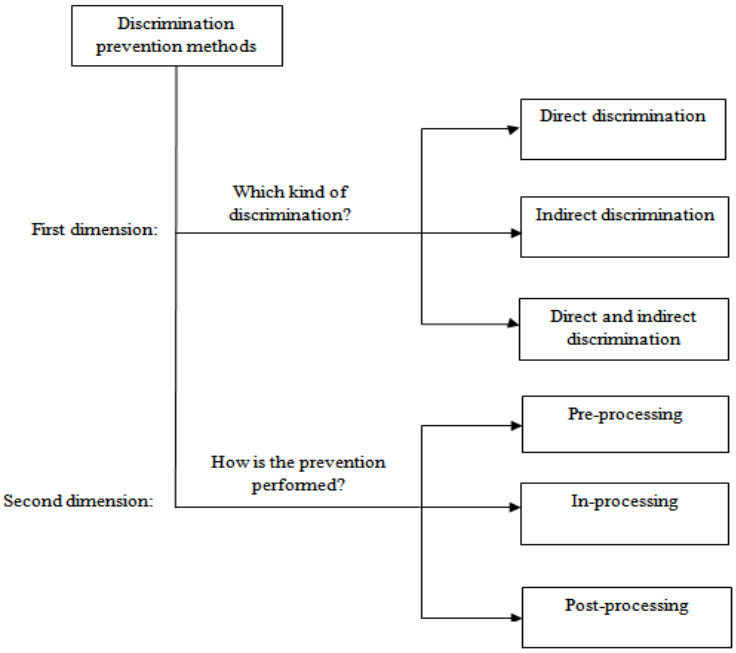

Fig. 13.2. The taxonomy of discrimination prevention methods

first dimension, we consider whether the approach deals with direct discrimination, indirect discrimination, or both at the same time. In this way, we separate the discrimination prevention approaches into three groups: *direct discrimination prevention methods*, *indirect discrimination prevention methods*, and *direct and indirect discrimination prevention methods*. The second dimension in the classification relates to the phase of the data mining process in which discrimination prevention is done. Based on this second dimension, discrimination prevention methods fall into three groups (Ruggieri et al. 2010): *pre-processing*, *in-processing* and *post-processing* approaches. We next describe these groups:

- Pre-processing. Methods in this group transform the source data in such a way that the discriminatory biases contained in the original data are removed so that no unfair decision rule can be mined from the transformed data; any of the standard data mining algorithms can then be applied. The pre-processing approaches of data transformation and hierarchy-based generalization can be adapted from the privacy preservation literature. Along this line, Kamiran and Calders (2009), Kamiran and Calders (2010), Hajian *et al.* (2011a and 2011b) and Hajian and Domingo-Ferrer (2012) perform a controlled distortion of the training data from which a classifier is learned by making minimally intrusive modifications leading to an unbiased dataset.
- In-processing. Methods in this group change the data mining algorithms in such a way that the resulting models do not contain unfair decision rules (Calders and Verwer 2010, Kamiran *et al.* 2010). For example, an alternative ap-

proach to cleaning the discrimination from the original dataset is proposed in Calders and Verwer (2010) whereby the non-discriminatory constraint is embedded into a decision tree learner by changing its splitting criterion and pruning strategy through a novel leaf re-labeling approach. However, it is obvious that in-processing discrimination prevention methods must rely on new special-purpose data mining algorithms; standard data mining algorithms cannot be used because they ought to be adapted to satisfy the non-discrimination requirement.

- Post-processing. These methods modify the resulting data mining models, instead of cleaning the original dataset or changing the data mining algorithms. For example, in Pedreschi *et al.* (2009a), a confidence-altering approach is proposed for classification rules inferred by the rule-based classifier: CPAR (classification based on predictive association rules) algorithm (Yin et al. 2003).

13.4 Types of Pre-processing Discrimination Prevention Methods

Although some methods have already been proposed for each of the above mentioned approaches (pre-processing, in-processing, post-processing), discrimination prevention stays a largely unexplored research avenue. In this section, we concentrate on a group of discrimination prevention methods based on pre-processing (first dimension) that could deal with direct or indirect discrimination (second dimension), because pre-processing has the attractive feature of being independent of the data mining algorithms and models. More details, algorithms and experimental results on these methods are presented in Hajian *et al.* (2011a and 2011b) and Hajian and Domingo-Ferrer (2012). The purpose of all these methods is to transform the original data *DB* in such a way as to remove direct or indirect discriminatory biases, with minimum impact on the data and on legitimate decision rules, so that no unfair decision rule can be mined from the transformed data. As part of this effort, the metrics that specify which records should be changed, how many records should be changed and how those records should be changed during data transformation are developed.

There are some assumptions common to all methods in this section. First, we assume the class attribute in the original dataset *DB* to be binary (*e.g.* denying or granting credit). Second, we obtain the database of *discriminatory* and *redlining* rules as output of a discrimination measurement (discovery) phase based on measures proposed in Pedreschi *et al.* (2008) and Pedreschi *et al.* (2009a); discrimination measurement is performed to identify *discriminatory* and *redlining rules* (based on the work in Chapter 5); then a data transformation phase is needed to transform the data in order to remove all evidence of direct or indirect discriminatory biases associated to *discriminatory* or *redlining* rules. Third, we assume the discriminatory itemsets (*i.e. A*) and the non-discriminatory itemsets (*i.e. D*) to be categorical.

13.4.1 Direct Discrimination Prevention Methods

The proposed solution to prevent direct discrimination is based on the fact that the dataset of decision rules would be free of direct discrimination if it only contained PD rules that are *protective* or PD rules that are instances of at least one *non-redlining (legitimate)* PND rule. Therefore, a suitable data transformation with minimum information loss should be applied in such a way that each *discriminatory* rule either becomes *protective* or an instance of a *non-redlining* PND rule. We call the first procedure direct rule protection and the second one rule generalization.

13.4.1.1 Direct Rule Protection (DRP)

In order to convert each *discriminatory* rule *r': A, B →C*, where *A* is a discriminatory itemset (*A⊆DI$_s$) and B is non-discriminatory itemset (B⊆nDI$_s$)*), into a *protective* rule, two data transformation methods (DTM) could be applied. One method (DTM 1) changes the discriminatory itemset in some records (*e.g.* gender changed from male to female in the records with granted credits) and the other method (DTM 2) changes the class item in some records (*e.g.* from grant credit to deny credit in the records with male gender). Table 13.1 shows the operation of these two methods.

Table 13.1. Data transformation methods for direct rule protection

Direct Rule Protection	
DTM 1	$\sim A, B \rightarrow \sim C \Rightarrow A, B \rightarrow \sim C$
DTM 2	$\sim A, B \rightarrow \sim C \Rightarrow \sim A, B \rightarrow C$

Table 13.1 shows that in DTM 1 some records that support the rule $\sim A, B \rightarrow \sim C$ will be changed by modifying the value of the discriminatory itemset from ~*A* (Sex=Male) to *A* (Sex=Female) until *discriminatory* rule *r': A, B →C becomes protective* (i.e. *elift(r')* < α). In order to score better in terms of the utility measures presented in Section 13.5 and 13.6, the changed records should be those among the ones supporting the above rule that have the lowest impact on the other (protective) rules. Similar records are also chosen in DTM 2 with the difference that, instead of changing discriminatory itemsets, the class item is changed from ~*C* (grant credit) into *C* (deny credit) to make *r'* protective.

13.4.1.2 Rule Generalization

Rule generalization is another data transformation method for direct discrimination prevention. It is based on the fact that if each *discriminatory* rule *r': A, B →C* in the database of decision rules was an instance of at least one *non-redlining (legitimate)* PND rule *r: D, B →C* where *D* is a non-discriminatory itemset ($D \subseteq nDI_s$), the dataset would be free of direct discrimination. To formalize this dependency among rules (i.e. *r'* is an instance of *r*), Pedreschi et al. in (Pedreschi et al. 2009b) say that a PD classification rule *r'* is an instance of a PND rule *r* if rule *r* holds with the same or higher confidence, namely *conf(r: D,B → C) ≥ conf(r': A,B→C)*, and a case (record) satisfying discriminatory itemset *A* in context *B* satisfies legitimate itemset *D* as well, namely *conf(A, B → D) = 1*.

Based on this concept, a data transformation method (i.e. rule generalization) could be applied to transform each *discriminatory* rule *r': A, B →C* into an instance of a legitimate rule. Then, rule generalization can be achieved for discriminatory rules *r'* for which there is at least one *non-redlining* PND rule *r* by changing the class item in some records (e.g. from "Hire no" to "Hire yes" in the records of foreign and low-experienced people in NYC city). Table 13.2 shows the function of this method.

Table 13.2. Data transformation method for rule generalization

Rule Generalization	
DTM	$A, B, \sim D \rightarrow C \Rightarrow A, B, \sim D \rightarrow \sim C$

Table 13.2 shows that in DTM some records that support the rule *A, B, ~D → C* will change by modifying the value of class item from *C* (e.g. deny credit) into $\sim C$ (e.g. grant credit) until *discriminatory* rule *r': A, B →C becomes an instance of a non-redlining (legitimate)* PND rule *r: D, B →C* . Similar to DRP methods, in order to score better in terms of the utility measures presented in Section 13.5 and 13.6, the changed records should the ones among those supporting the above rule that have the lowest impact on the other (protective) rules.

13.4.1.3 Direct Rule Protection and Rule Generalization

Since rule generalization might not be applicable to all *discriminatory* rules, rule generalization cannot be used alone for direct discrimination prevention and must be combined with direct rule protection. When applying both rule generalization and direct rule protection, *discriminatory* rules are divided into two groups:

- *Discriminatory* rules *r'* for which there is at least one non-redlining PND rule *r* such that *r'* could be an instance of *r*. For these rules, rule generalization is performed unless direct rule protection requires less data transformation (in which case direct rule protection is used).

- *Discriminatory* rules *r'* such that there is no such PND rule. For these rules, direct rule protection (DTM 1 or DTM 2) is used.

13.4.2 Indirect Discrimination Prevention Methods

The solution proposed in Hajian *et al.* (2011b) to prevent indirect discrimination is based on the fact that the dataset of decision rules would be free of indirect discrimination if it contained no *redlining* rules. To achieve this, a suitable data transformation with minimum information loss should be applied in such a way that *redlining* rules are converted to *non-redlining* rules. We call this procedure indirect rule protection (IRP).

In order to turn a *redlining* rule $r: D, B \rightarrow C$, where *D* is a non-discriminatory itemset that is highly correlated to the discriminatory itemset *A*, into a *non-redlining* rule based on the indirect discriminatory measure (*elb*), two data transformation methods could be applied, similar to the ones for direct rule protection. One method (DTM 1) changes the discriminatory itemset in some records (*e.g.* from non-foreign worker to foreign worker in the records of hired people in NYC city with Zip≠10451) and the other method (DTM 2) changes the class item in some records (e.g. from "Hire yes" to "Hire no" in the records of non-foreign worker of people in NYC city with Zip≠10451). Table 13.3 shows the operation of these two methods.

Table 13.3. Data transformation methods for indirect rule protection

Indirect Rule Protection	
DTM 1	$\sim A, B, \sim D \rightarrow \sim C \Rightarrow A, B, \sim D \rightarrow \sim C$
DTM 2	$\sim A, B, \sim D \rightarrow \sim C \Rightarrow \sim A, B, \sim D \rightarrow C$

Table 13.3 shows that in DTM 1 some records in the original data that support the rule *~A, B, ~D → ~C* will be changed by modifying the value of the discriminatory itemset from *~A* (Sex=Male) into *A* (Sex=Female) in these records until the *redlining* rule *r: D, B →C* becomes *non-redlining* (i.e. *elb(r)* < α). With the aim of scoring better in terms of the utility measures presented in Section 13.5 and 13.6, among the records supporting the above rule, one should change those with lowest impact on the other (*non-redlining*) rules. Similar records are also chosen in DTM 2 with the difference that, instead of changing discriminatory itemsets, the class item is changed from *~C* (e.g. grant credit) into *C* (e.g. deny credit) in these records to make *r non-redlining*.

The difference between the DRP and IRP methods shown in Tables 1 and 3 is about the set of records chosen for transformation. As shown in Table 3, in IRP

the chosen records should not satisfy the D itemset (chosen records are those with $\sim A, B \ \sim D \rightarrow \sim C$), whereas DRP does not care about D at all (chosen records are those with $\sim A, B \ \rightarrow \sim C$).

13.5 Measuring Discrimination Removal

Discrimination prevention methods should be evaluated based on two aspects: discrimination removal and data quality. We deal with the first aspect in this section: how successful the method is at removing all evidence of direct and/or indirect discrimination from the original dataset. To measure discrimination removal, four metrics were proposed in Hajian *et al.* (2011a and 2011b) and Hajian and Domingo-Ferrer (2012):

- **Direct Discrimination Prevention Degree (DDPD)**. This measure quantifies the percentage of *discriminatory* rules that are no longer *discriminatory* in the transformed dataset.
- **Direct Discrimination Protection Preservation (DDPP)**. This measure quantifies the percentage of the *protective* rules in the original dataset that remain *protective* in the transformed dataset.
- **Indirect Discrimination Prevention Degree (IDPD).** This measure quantifies the percentage of *redlining rules* that are no longer *redlining* in the transformed dataset.
- **Indirect Discrimination Protection Preservation (IDPP).** This measure quantifies the percentage of *non-redlining* rules in the original dataset that remain *non-redlining* in the transformed dataset.

Since the above measures are used to evaluate the success of the proposed methods in direct and indirect discrimination prevention, ideally their value should be 100%.

13.6 Measuring Data Quality

The second aspect to evaluate discrimination prevention methods is how much information loss (*i.e.* data quality loss) they cause. To measure data quality, two metrics are proposed in Verykios and Gkoulalas-Divanis (2008):

- **Misses Cost** (MC). This measure quantifies the percentage of rules among those extractable from the original dataset that cannot be extracted from the transformed dataset (side-effect of the transformation process).

- **Ghost Cost** (GC). This measure quantifies the percentage of the rules among those extractable from the transformed dataset that were not extractable from the original dataset (side-effect of the transformation process).

MC and GC should ideally be 0%. However, MC and GC may not be 0% as a side-effect of the transformation process.

13.7 Experimental Results

This section presents the experimental evaluation of the proposed direct discrimination prevention approaches. We use the German Credit Dataset (Newman *et al.* 1998) in our experiments, since it is a well-known and frequently used dataset in the context of anti-discrimination. This dataset consists of 1,000 records and 20 attributes (without class attribute) of bank account holders. For our experiments with this dataset, we set DI_s= {Foreign worker=Yes, Personal Status=Female and not Single, Age=Old} (cut-off for Age=Old: 50 years old).

Figure 13.3 shows at the left the degree of information loss (as average of MC and GC) and it shows at the right the degree of discrimination removal (as average of DDPD and DDPP) of direct discrimination prevention methods for the German Credit dataset when the value of the discriminatory threshold α varies from 1.2 to 1.7, the minimum support is 5% and the minimum confidence is 10%. The number of direct *discriminatory* rules extracted from the dataset is 991 for α =1.2, 415 for α =1.3, 207 for α =1.4, 120 for α =1.5, 63 for α =1.6 and 30 for α =1.7, respectively.

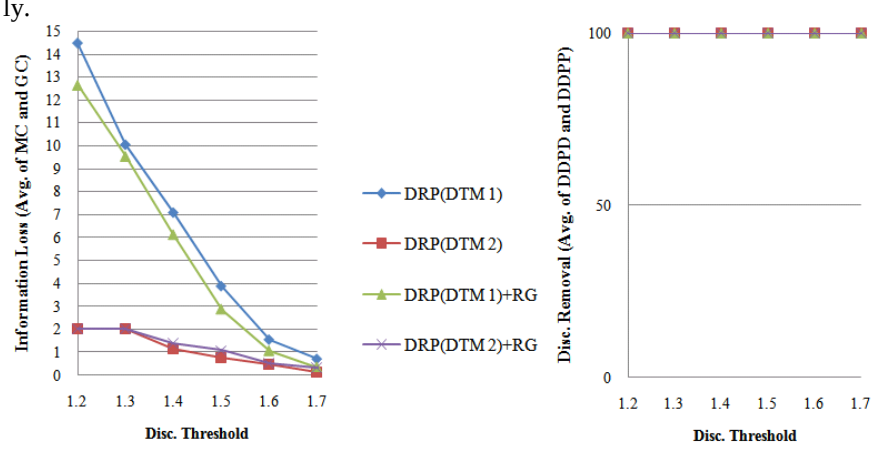

Fig.13.3. Left: Information loss, Right: Discrimination removal degree for direct discrimination prevention methods for α in [1.2, 1.7]. DRP(DTM *i*): Data transformation method *i* for DRP; RG: Rule Generalization.

As shown in Figure 3, the degree of discrimination removal provided by all methods for different values of α is also 100%. However, the degree of information loss decreases substantially as α increases; the reason is that, as α increases, the number of *discriminatory* rules to be dealt with decreases. In addition, as shown in Figure 2, the lowest information loss for most values of α is obtained by DTM 2 for DRP.

Empirical results on indirect discrimination prevention methods can be found in Hajian *et al.* (2011b).

13.8 Conclusions and Future Work

In sociology, discrimination is the prejudicial treatment of an individual based on their membership in a certain group or category. It involves denying to members of one group opportunities that are available to other groups. Like privacy, discrimination could have negative social impact on acceptance and dissemination of data mining technology. Discrimination prevention in data mining is a new body of research focusing on this issue. One of the research questions here is whether we can adapt and use the pre-processing approaches of data transformation and hierarchy-based generalization from the privacy preservation literature for discrimination prevention. In response to this question, we try to inspire on the data transformation methods for knowledge (rule) hiding in privacy preserving data mining (more discussed in Chapter 11) and we devise new data transformation methods (*i.e.* direct and indirect rule protection, rule generalization) for converting direct and/or indirect discriminatory decision rules to legitimate (non-discriminatory) classification rules; our current results are convincing in terms of discrimination removal and information loss. However, there are many other challenges regarding discrimination prevention that could be considered in the rest of this research. For example, the perception of discrimination, just like the perception of privacy, strongly depends on the legal and cultural conventions of a society. Although we argued that discrimination measures based on *elift* and *elb* are reasonable, if substantially different discrimination definitions and/or measures were to be found, new data transformation methods would need to be designed.

Another challenge is the relationship between discrimination prevention and privacy preservation in data mining. It would be extremely interesting to find synergies between rule hiding for privacy-preserving data mining and rule hiding for discrimination removal. Just as we were able to show that indirect discrimination removal can help direct discrimination removal, it remains to see whether privacy protection can help anti-discrimination or viceversa.

Disclaimer and Acknowledgments

The authors are with the UNESCO Chair in Data Privacy, but the views expressed in this article do not necessarily reflect the position of UNESCO nor commit that organization. This work was partly supported by the Spanish Government through projects TSI2007-65406-C03-01 "E-AEGIS", TIN2011-27076-C03-01 "CO-PRIVACY" and CONSOLIDER INGENIO 2010 CSD2007-00004 "ARES", by the Government of Catalonia under grant 2009 SGR 1135 and by the European Commission under FP7 project "DwB". The second author is partly supported as an ICREA Acadèmia Researcher by the Government of Catalonia.

References

Calders, T., & Verwer, S. (2010). Three naive Bayes approaches for discrimination-free classification. Data Mining and Knowledge Discovery, 21(2):277-292.

Hajian, S., Domingo-Ferrer, J. & Martinez-Ballesté, A. (2011a). Discrimination prevention in data mining for intrusion and crime detection. Proc. of the IEEE Symposium on Computational Intelligence in Cyber Security (CICS 2011), pp. 47-54. IEEE.

Hajian, S., Domingo-Ferrer, J. & Martinez-Ballesté, A. (2011b). Rule protection for indirect discrimination prevention in data mining. Modeling Decisions for Artificial Intelligence-MDAI 2011, Lecture Notes in Computer Science 6820, pp. 211-222.

Hajian, S. & Domingo-Ferrer, J. (2012). A methodology for direct and indirect discrimination prevention in data mining. Manuscript.

Kamiran, F. & Calders, T. (2009). Classification without discrimination. Proc. of the 2nd IEEE International Conference on Computer, Control and Communication (IC4 2009). IEEE.

Kamiran, F. & Calders, T. (2010). Classification with no discrimination by preferential sampling. Proc. of the 19th Machine Learning conference of Belgium and The Netherlands.

Kamiran, F., Calders, T. & Pechenizkiy, M. (2010). Discrimination aware decision tree learning. Proc. of the IEEE International Conference on Data Mining (ICDM 2010), pp. 869-874. ICDM.

Newman, D. J., Hettich, S., Blake, S. L. & Merz, C.J. (1998). UCI Repository of Machine Learning Databases. http://archive.ics.uci.edu/ml.

Parliament of the United Kingdom. (1975). Sex Discrimination Act. http://www.opsi.gov.uk/acts/acts1975/ PDF/ukpga 19750065 en.pdf.

Parliament of the United Kingdom. (1976). Race Relations Act. http://www.statutelaw.gov.uk/content.aspx? activeTextDocId=2059995.

Pedreschi, D., Ruggieri, S. & Turini, F. (2008). Discrimination-aware data mining. Proc. of the 14th ACM International Conference on Knowledge Discovery and Data Mining (KDD 2008), pp. 560-568. ACM.

Pedreschi, D., Ruggieri, S. & Turini, F. (2009a). Measuring discrimination in socially-sensitive decision records. Proc. of the 9th SIAM Data Mining Conference (SDM 2009), pp. 581-592. SIAM.

Pedreschi, D., Ruggieri, S. & Turini, F. (2009b). Integrating induction and deduction for finding evidence of discrimination. Proc. of the 12th ACM International Conference on Artificial Intelligence and Law (ICAIL 2009), pp. 157-166. ACM.

Ruggieri, S., Pedreschi, D. & Turini, F. (2010). Data mining for discrimination discovery. ACM Transactions on Knowledge Discovery from Data, 4(2) Article 9.

United States Congress. (1994). Employment Non-Discrimination Act. http://www.govtrack.us/congress/bill.xpd? bill=h111-3017.

United States Congres. (1963). US Equal Pay Act. http://archive.eeoc.gov/epa/anniversary/epa-40.html.

Verykios, V. & Gkoulalas-Divanis, A. (2008). A survey of association rule hiding methods for privacy. In C. C. Aggarwal and P. S. Yu (Eds.), Privacy- Preserving Data Mining: Models and Algorithms. Springer.
Yin, X. & Han, J. (2003). CPAR: Classification based on Predictive Association Rules. In Proc. of SIAM ICDM 2003. SIAM.