Knowl Inf Syst (2013) 35:613–644
DOI 10.1007/s10115-012-0584-8

REGULAR PAPER

Quantifying explainable discrimination and removing illegal discrimination in automated decision making

Faisal Kamiran · Indrė Žliobaitė · Toon Calders

Received: 29 February 2012 / Revised: 18 April 2012 / Accepted: 20 October 2012 /
Published online: 18 November 2012

Abstract Recently, the following discrimination-aware classification problem was introduced. Historical data used for supervised learning may contain discrimination, for instance, with respect to gender. The question addressed by discrimination-aware techniques is, given sensitive attribute, how to train discrimination-free classifiers on such historical data that are discriminative, with respect to the given sensitive attribute. Existing techniques that deal with this problem aim at removing all discrimination and do not take into account that part of the discrimination may be explainable by other attributes. For example, in a job application, the education level of a job candidate could be such an explainable attribute. If the data contain many highly educated male candidates and only few highly educated women, a difference in acceptance rates between woman and man does not necessarily reflect gender discrimination, as it could be explained by the different levels of education. Even though selecting on education level would result in more males being accepted, a difference with respect to such a criterion would not be considered to be undesirable, nor illegal. Current state-of-the-art techniques, however, do not take such gender-neutral explanations into account and tend to overreact and actually start reverse discriminating, as we will show in this paper. Therefore, we introduce and analyze the refined notion of conditional non-discrimination in classifier design. We show that some of the differences in decisions across the sensitive groups can be explainable and are hence tolerable. Therefore, we develop methodology for quantifying

A short version of this paper appeared in ICDM'12 [47].

F. Kamiran (✉)
Mathematical and Computer Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia
e-mail: faisal.kamiran@gmail.com

I. Žliobaitė
Bournemouth University, Poole, UK
e-mail: izliobaite@bournemouth.ac.uk

T. Calders
Eindhoven University of Technology, Eindhoven, The Netherlands
e-mail: t.calders@tue.nl

the explainable discrimination and algorithmic techniques for removing the illegal discrimination when one or more attributes are considered as explanatory. Experimental evaluation on synthetic and real-world classification datasets demonstrates that the new techniques are superior to the old ones in this new context, as they succeed in removing almost exclusively the undesirable discrimination, while leaving the explainable differences unchanged, allowing for differences in decisions as long as they are explainable.

Keywords Classification · Independence · Discrimination-aware data mining

1 Introduction

Decision making is a cognitive process which leads to a final choice from a range of alternative options. Decision making is often done by human beings which may lead to highly rational and productive but socially and legally unacceptable outcome as humans have a limited capacity to explore every perspective and consequence of a certain decision. In particular, when humans make subjective decisions, discrimination toward individuals belonging to certain groups may occur. For instance, a job screening committee may subjectively prefer and thus select Caucasian candidates more generously than Afro-American candidates. Such cases can be brought to court for an in-depth analysis of the circumstances.

But not only humans can discriminate. Nowadays, more and more decisions in lending, recruitment, grant or study applications are partially being automated based on models premised on historical data. Classification is an important data mining technique that is widely used to automate the future decision-making process. In classification we build models to predict class of future data objects based on already labeled examples available in the historical data. That historical data may contain legally and socially unacceptable discrimination, for instance, racial discrimination in the recruitment of job candidates. In such a case classifiers are likely to learn the discriminatory relation present in the historical data and apply it when making predictions. Inappropriately trained models may hence discriminate systematically, which is a lot more harmful than individual cases.

It is in the best interest of the decision makers (e.g., banks, consultancies, universities) to ensure that the classifiers they build are discrimination-free even if the historical data are discriminatory. The following case illustrates the legal context and the difficulty of the task. Recently, one of the world's largest consultancy firms was accused of discrimination against ethnic minorities in a law suit [1]. The firm used existing criminal records to turn down candidates in pre-employment screening. Not the use of criminal records itself was considered problematic. In these data race and criminality were correlated, and the use of criminal records indirectly leads to racial discrimination. Thus, even though the company did not intend to discriminate, the decisions were deemed discriminatory by the court, while having been convicted was deemed to be not relevant for prescreening purposes. This example shows that discrimination may occur even if the sensitive information is not directly used in the model and that such indirect discrimination is as well forbidden. Many attributes can be used only to the extent that they do not lead to indirect discrimination.

The current solutions to make classifiers discrimination-free [6,7,24–26] aim at removing all discrimination present in the data; the probability of a positive decision by the learned classifier must be equal for all subgroups defined by the sensitive attribute (e.g., male and female).

Table 1 Summary statistics of the Adult dataset [2]

	Hours per week	Annual income (K$)
Female	36.4	10.9
Male	42.4	30.4
All data	40.4	23.9

The authors of [6,24,25] propose discrimination-aware preprocessing techniques to remove all the discrimination from training data before learning a classifier, and discrimination-aware methods proposed in [7,26] adapt the classifier learning process itself to make the learnt classifier impartial. As we observe in this paper, however, such approaches have a significant limitation, as they do not take into account the fact that some part of the differences in the probability of acceptance for the two groups may be objectively explainable by other attributes.

For instance, in the Adult dataset [2], females on average have a lower annual income than males. However, from Table 1 one can observe that females work less hours per week on average. If we assume that job requires the attendance of employee for full working hours (e.g., job at information desk), work hours per week give a good justification for low income. Suppose a human resource consultancy company wants to build a classifier to automatically suggest a salary, given an applicant. Suppose also that the company is aiming to prevent gender discrimination in the classification decisions. The existing discrimination-free classifiers would correct the decision making in such a way that males and females would get on average the same income, say 20 K$, leading to a reverse discrimination as it would result in male employees being assigned a lower salary than female for the same amount of working hours. This example suggests that making the probabilities of acceptance equal for both would lead to favoring the group which is being deprived. In reality, if the difference in the decisions can be justified, it is not considered as illegal discrimination.

This paper takes a step forward in designing discrimination-free classifiers as well as extends the discrimination problem setting and makes the following contributions:

1. The paper introduces a methodology for analytically quantifying explainable and illegal discrimination in automated decision making, considering one or more attributes as explanatory. We argue that only the discrimination, which is conditioned on an explanatory attribute should be removed. We refer to this methodology as *conditional discrimination-aware classification*.
2. Using these analytical results, the paper introduces three algorithmic techniques for removing only unexplainable (illegal) discrimination in classification. The techniques can be used as wrappers to classifiers of user's choice.
   - **Local massaging** builds upon data preprocessing method due [6,25], where labels of some instances in the dataset are modified to make the input dataset discrimination-free. The proposed local massaging technique partitions the dataset on the basis of explanatory attribute, quantifies the explainable and the illegal discrimination for each partition and then applies *Massaging* [25] to the partitions.
   - **Local preferential sampling** builds upon data preprocessing method due [24,25], where data are resampled with replacement in such a way that the input dataset becomes discrimination-free. As in the local massaging, the partitioning and the analytical quantification of discrimination is conditioned on an explanatory attribute using the new analytical results, and then the *Preferemncial sampling* procedure [24,25] is applied.

- **Local direct classification** is a baseline technique that uses our new analytical results to quantify discrimination, but instead of learning a new classifier on the preprocessed data, this technique directly adjusts the decision boundaries of trained classifiers.

3. For the tasks where more than one attribute needs to be considered explanatory, we present a framework for aggregating explanatory attributes and demonstrate how to apply the proposed theory and techniques in such situations.

Our experimental evaluation in the controlled settings and on the real-world classification problems demonstrates that the new techniques effectively remove the illegal discrimination, allowing the differences in decisions to be present as long as they are explainable.

The remainder of the paper is organized as follows. We motivate the conditional discrimination-aware problem with legal and social evidences in Sect. 2. In Sect. 3 we define a formal discrimination model, and in Sect. 4 we analytically quantify how much of the discrimination is explainable. In Sect. 5 we present two techniques to remove illegal discrimination from the training data. Section 6 presents experimental evaluation. In Sect. 7 we extend our techniques to handle multiple explanatory attributes. Section 8 discusses the related work. Section 9 concludes the study.

2 Background and motivation

The word discrimination originates from the Latin word *discriminate*, which means to *distinguish between*. Discrimination is widely studied in social sciences [22] where it refers to the unfair treatment of individuals of a certain group solely based on their affiliation with a particular group, category or class. Such discriminatory practices suppress the opportunities for the members of deprived groups in employment, income, education, finance and many other social activities on the basis of age, gender, skin color, religion, race, language, culture, marital status and economic condition. Discrimination is increasingly often considered unacceptable from social, ethical and legal perspectives.

In this paper we consider two types of discrimination: explainable discrimination and illegal discrimination. We consider that only the illegal discrimination should be avoided in the future decision making. In this section we discuss this setting in the context of evidence from legal domain and the historical perspective and demonstrate that intuitively trivial solutions would not solve this problem.

2.1 Legal evidence

To motivate the discrimination-aware classification setting, let us consider legal environment of automated decision making. There are many anti-discrimination laws that prohibit discrimination in housing, employment, financing, insurance, wages, etc., on the basis of race, color, national origin, religion, sex, familial status and disability. If we observe these laws in detail, it becomes obvious that these laws often prohibit illegal part of discrimination. If the discriminatory treatment can be justified with some other explanatory attributes, it is not considered as illegal practice. It means that proving a case as discriminatory in court requires proofs that there were no genuine reasons for the biased treatment. As an example, employment practices may be considered discriminatory if they have a disproportionate adverse impact on members of a minority group. We discuss some of laws that prohibit illegal discrimination and show how they relate to our problem statement:

The Australian Sex Discrimination Act 1984 [3]: This act prohibits discrimination in work, education, services, accommodation, land, clubs on the grounds of marital status, pregnancy or potential pregnancy, and family responsibilities. This act defines sexual harassment and other discriminatory practices on different grounds and declares them unlawful. The main objectives of this act are as follows:

(a) to give effect to certain provisions of the Convention on the Elimination of All Forms of Discrimination Against Women; and
(b) to eliminate, so far as possible, discrimination against persons on the ground of sex, marital status, pregnancy or potential pregnancy in the areas of work, accommodation, education, the provision of goods, facilities and services, the disposal of land, the activities of clubs and the administration of Commonwealth laws and programs; and
(ba) to eliminate, so far as possible, discrimination involving dismissal of employees on the ground of family responsibilities; and
(c) to eliminate, so far as possible, discrimination involving sexual harassment in the workplace, in educational institutions and in other areas of public activity; and
(d) to promote recognition and acceptance within the community of the principle of the equality of men and women.

However, section 7B of this law clearly states that if the discriminatory practice is reasonable in the certain scenario and can be justified with the circumstances, it will no longer be considered as discriminatory. Section 7B of this act is as follows: *a person does not discriminate against another person by imposing, or proposing to impose, a condition, requirement or practice that has, or is likely to have, the disadvantaging effect mentioned in subsection 5(2), 6(2) or 7(2) if the condition, requirement or practice is reasonable in the circumstances.*

The US Equal Pay Act 1963 [44]: This act requires that men and women working at the same place should be paid equally for their works. The jobs need not to be identical, but they must be substantially equal. This law covers all forms of pay including salary, overtime pay, bonuses, stock options, profit sharing and bonus plans, life insurance, vacation and holiday pay, cleaning or gasoline allowances, hotel accommodations, reimbursement for travel expenses and benefits. The act describes it as follows: *No employer having employees subject to any provisions of this section shall discriminate, within any establishment in which such employees are employed, between employees on the basis of sex by paying wages to employees in such establishment at a rate less than the rate at which he pays wages to employees of the opposite sex in such establishment for equal work on jobs the performance of which requires equal skill, effort and responsibility, and which are performed under similar working conditions, except where such payment is made pursuant to (i) a seniority system; (ii) a merit system; (iii) a system which measures earnings by quantity or quality of production; or (iv) a differential based on any other factor other than sex: Provided that an employer who is paying a wage rate differential in violation of this subsection shall not, in order to comply with the provisions of this subsection, reduce the wage rate of any employee.* This act clearly states that if the employees of one gender are more experienced and more productive, it is perfectly valid to pay differently. This is exactly what we argue in this paper as a next step of the previously done discrimination-aware works.

2.2 Redlining

The discrimination-free classification problem is non-trivial needs advanced solutions. One can consider a straightforward solution to make a classifier discrimination-free by removing the sensitive attribute (e.g., race) from the input space. Unfortunately, that would not help

if some of the input attributes are not independent from the sensitive attribute. For instance, a postal code may be strongly related with the race. If it is not allowed to use race in the decision making, discriminatory decisions still can be made by using postal code. That would be an indirect discrimination.

Consider the German Credit Dataset from the UCI repository [2] as an example of decisions to grant loans based on demographic information of applicants. Loan decisions correlate with the age of an applicant, and the correlation is 0.09. Suppose using age in deciding upon loans is forbidden by law. If we remove the *age* attribute from the data, it will not remove the age discrimination, as other attributes, such as *own_house*, indicating if the applicant is a home-owner, give information about the *age* of a loan applicant. In fact, eight attributes are correlated with *age* by more than 0.1.

A parallel can be drawn with the practice of *redlining*: denying inhabitants of certain racially determined areas from services such as loans. It describes the practice of marking a red line on a map to delineate the area where banks would not invest; later, the term was applied to discrimination against a particular group of people (usually by race or sex) no matter the geography. During the heyday of *redlining*, the areas most frequently discriminated against were black inner city neighborhoods. Through at least the 1990s, this practice meant that banks would often lend to lower-income whites but not to middle- or upper-income blacks [12]. The concept of redlining is really important because it illustrates the situations when the direct use of sensitive attribute in the decision making is not allowed by law. In such a situation a decision maker could be tempted to use the related attribute of sensitive attribute as a proxy. Such profiling will lead to higher gains for the decision makers; nevertheless, it is ethically and legally unacceptable.

To get rid of such discriminatory relations among attributes, one would also need to remove the attributes that are correlated with the sensitive attribute. It is not a good solution if these attributes carry the objective information about the class label, as in such case the predictions will become less accurate. For instance, a postal code in addition to the racial information may carry information about real estate prices in the neighborhood, which is objectively informative for loan decisions. Thus, our goal is to use the objective information, but not the sensitive information of such attributes.

3 Formal model of discrimination in decision making

The setting of conditional discrimination-aware classification is formally defined as follows. Let X be an instance in p-dimensional space, let $y \in \{+, -\}$ be its label. The task is to learn a classifier $\mathcal{L} : X \rightarrow y$. In addition to X, let $s \in \{f, m\}$ be a sensitive attribute. In this paper we will consider gender as sensitive attribute with values female (f) and male (m). In reality, many other attributes, for example, ethnicity, religion, age, citizenship, etc., can be considered as sensitive attributes. We assume that we have background knowledge that which attribute is a sensitive attribute and it is forbidden by law to make decisions based on such attribute.

3.1 Discrimination model

To analyze the effects of discrimination and design discrimination-free learning techniques, a model describing how discrimination happens needs to be assumed. We consider that discrimination happens in the following way in relation to experimental findings reported in [22]. The historical data originate from decision making by human experts. First, the qualifications of a

candidate are evaluated, and a preliminary score is obtained. The qualifications are evaluated objectively. Then the score is corrected with a discrimination bias by looking at, for example, the gender of a candidate and either adding or subtracting a fixed (the same) bias from the qualification score. We can consider the historical data originated from human decision making as a classifier $\mathcal{L}$. That classifier consists of three main parts:

1. a function from attributes to a qualification score $r = G(X)$, where X does not include the sensitive attribute;
2. a discrimination bias function
$$B(s) = \begin{cases} b & \text{if } s = m \\ -b & \text{if } s = f \end{cases};$$
3. the final decision function $y = \mathcal{L}\left(G(X) + B(s)\right)$.

According to this model a decision is made in the following way. First, the qualifications of a candidate are evaluated based on the attributes in X, and a preliminary score is obtained $r = G(X)$. The qualifications are evaluated objectively. Then the discrimination bias is introduced by looking at the gender of a candidate and either adding or subtracting a fixed bias from the qualification score to obtain $r^* = G(X) + B(s) = r \pm b$. The final decision is made by $\mathcal{L}(r^*)$. Decision making can have two major forms: *online* and *offline*. With the off-line decision the candidates are ranked based on their scores r^*, and n candidates that have the highest scores are accepted. With the online decision an acceptance threshold θ is set, and the incoming candidates that have the score $r^* > \theta$ are accepted.

This discrimination model has two important implications. First, the decision bias is more likely to affect the individuals that are close to the decision boundary according to their score r. If an individual is far from the decision boundary, adding or subtracting the discriminatory bias b does not change the final decision. This observation is consistent with experimental findings how discrimination happens in practice [22].

Second, traditional classifiers try to learn r^*, whereas discrimination-aware classification also involves decomposing r^* into $G(X)$ and $B(s)$ and reverting the influence of $B(s)$. There may be attributes within X, however, that contribute to $G(X)$, but at the same time are correlated with the sensitive attribute s, and through s, with $B(s)$. When observing the decisions, it would seem due to correlation that the decision is using s. Previous works have been very conservative in assuming that all the correlation between r^* and s is due to the discrimination bias $B(s)$. In this paper we refine this viewpoint.

It is important to mention here that this discrimination model does not guarantee to cover the all possible scenarios that lead to discrimination; however, it covers the most important and typical scenario.

3.2 Explanatory attribute

The explanatory attribute is the attribute e (among X) that is (cor)related with the sensitive attribute s and at the same time gives some objective information about the label y. Both relations can be measured in data, for instance, as the information gain about s given e, and about y given e. Our reasoning is built upon only one explanatory attribute. Nevertheless, this setting does not delimit taking into account multiple explanatory attributes if they are grouped into a singe representation, as we will demonstrate in Sect. 7.

In general there is no objective truth which attribute is more reasonable to use as the explanation for discrimination. For instance, when gender is the sensitive attribute, some attributes, such as relationships (*wife* or *husband*), may not be a good explanation, as semantically they are closely related to gender, while different working hours may be an appropriate

reason to have different monthly salaries. What is discriminatory and what is legal to use as an explanation depends on the law and goals of the anti-discrimination policies. Thus, the interpretation of the attributes needs to be fixed externally by law or domain experts. When non-discrimination needs to be enforced, the law sets the constraints, while we build the techniques to incorporate those constraints into classification. Otherwise, the selection of explanatory attribute becomes very confusing and debatable because one reasonable explanation could be highly unreasonable for the other one.

This study is built upon and valid with the following assumptions:

1. the sensitive and explanatory attributes are nominated externally by law or a domain expert, for example, lawyer, legal experts, etc.
2. the explanatory attribute is *not independent* from the sensitive attribute and at the same time gives objective information about the class label;
3. the illegal discrimination contained in the historical data is due to direct discrimination based on the sensitive attribute. It means no *redlining* (hidden discrimination) in the historical data; however, *redlining* may be introduced as a result of training a classifier on these data.

This study is *not* restricted to one *explanatory* attribute, while it is restricted to one binary *sensitive* attribute.

3.3 Measuring discrimination in classification

In the existing discrimination-aware classification, the discrimination is considered to be present if the probabilities of acceptance for the favored community (denote m) and the deprived community (denote f) were not equal, that is, $P(y = +|s = m) \neq P(y = +|s = f)$. Discrimination is measured as the difference between the two probabilities:

$$D_{\text{all}} = P(y = +|s = m) - P(y = +|s = f). \tag{1}$$

In the previous works all the difference in acceptance between the two groups was considered undesirable. In this study, however, we argue that some of the difference may be objectively explainable by the explanatory attribute. Thus, we can describe the difference in the probabilities as a sum of the explainable and illegal discrimination:

$$D_{\text{all}} = D_{\text{expl}} + D_{\text{illegal}}. \tag{2}$$

In this study we are interested to remove and thus measure D_{illegal}, which from Eq. (2) is

$$D_{\text{illegal}} = D_{\text{all}} - D_{\text{expl}}. \tag{3}$$

For that we need to find an expression for D_{expl}.

4 Explainable and illegal discrimination

We use a toy model about admission to a fictitious university[1] to explain the difference between the explainable and illegal discrimination considered. Note that the model presents a simplified version of reality and is intended to cover the key mechanisms of decision making and does not cover a full application process. In our admission example we take

[1] This model does not express our belief how admission procedures happen. We use it for the purpose of illustration only.

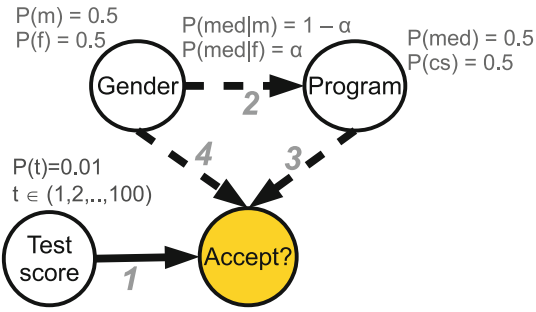


Fig. 1 University admission example

gender as the sensitive attribute; male (m) and female (f) are the sensitive groups, against which discrimination may occur. There are two programs: medicine (med) and computer science (cs) with potentially different acceptance standards. We consider the program as the explanatory attribute; thus, the differences in acceptance rates that can be attributed to different application rates into the programs between male and female are acceptable. All applicants take a test for which their score is recorded (T). The acceptance (+) decision is made personally for each candidate during the final interview. Figure 1 shows the setting.

There are four relations between variables in this example. Relation (1) shows that the final decision whether to accept partially depends on the test score. Notice that the test scores are assumed to be independent from gender or program. Relation (3) shows that the probability of acceptance depends on the program. For example, the competition to medicine may be higher; thus, less applicants are accepted in total. Relation (2) shows that the choice of program depends on gender. For instance, the larger part of the female candidates may apply to medicine, while more males apply to computer science. Relation (4) shows that acceptance also depends on gender, which is a bias in the decision making that is clearly a case of illegal discrimination. The presence illegal, explanatory or both discriminations in the data will depend on the relations (2), (3) and (4), as we will see in the following two examples.

4.1 Quantifying explainable discrimination

We present different scenarios to investigate different combinations of illegal and explainable discrimination by using the university admission model presented in Fig. 1.

Example 1 demonstrates that all the discrimination may be explainable. Suppose there are 2,000 applicants, 1,000 males and 1,000 females. Each program receives the same number of applicants, but medicine is more popular among females, $P(\text{med}|f) = 0.8$. Assume that medicine is more competitive, $P(+|\text{med}) < P(+|\text{cs})$. Within each program, males and females are treated equally, as described in Table 2. However, the aggregated scores indicate

Table 2 No illegal discrimination (Example 1)

	Medicine		Computer	
	Female	Male	Female	Male
Number of applicants	800	200	200	800
Acceptance rate (%)	20	20	40	40
Accepted (+)	160	40	80	320

Table 3 Illegal discrimination is present (Example 2)

	Medicine		Computer	
	Female	Male	Female	Male
Number of applicants	800	200	200	800
Acceptance rate (%)	15	25	35	45
Accepted (+)	120	50	70	360

that 36 % of males were accepted, but only 24 % of females. The difference is explained by the fact that more females applied to the more competitive program. Thus, there is no illegal discrimination.

We can also report a similar case from the Berkely study [5] where the examination of aggregate data on graduate admissions to the University of California, Berkeley, for fall 1973 shows a clear but misleading pattern of bias against female applicants. It shows that overall 44 % of males and 35 % of female applicants are admitted; thus, it seems that there is 9 % discrimination (D_{all}) toward female applicants. However, the examination of pooled data w.r.t. different departments shows that there is a small but statistically significant bias in favor of females. It means that the overall low admission rate for females is explainable by their tendency to apply to graduate departments that were more competitive for the applicants of either gender to enter. This case concludes that in-depth analysis of discrimination cases is really important to prove that whether some discriminatory practice was exercised or it was just a misconceptions.

Example 2 presents a case in which both explainable and illegal discrimination happened. Suppose a similar situation to Example 1 occurs, but the decision making is biased in favor of males, $P(+|m, e_i) > P(+|f, e_i)$, where e_i is a program, as presented in Table 3. The decisions result in different aggregated acceptance rates for the programs: medicine 17 % and computer science 43 %. It appears that in total 19 % of females and 41 % of males are accepted. Our goal is to determine which part of this difference is explainable by program, and which part is due to illegal discrimination.

First, we need to settle what would have been the correct acceptance rates $P^\star(+|\text{med})$ and $P^\star(+|\text{cs})$ within each program, if males and females would have been treated equally. Then we can find which part of the difference between the genders is explainable and treat the remaining part as illegal discrimination that needs to be removed. Finding the correct acceptance rates, however, is challenging, as there is no unique way to do it. Would all the acceptance rate have been as for males now, all as for females, or some average of the two?

To find the correct acceptance rates, we refer to the discrimination model given in Sect. 3.1. Under this model, it is reasonable to assume that roughly the same fraction of males benefit from the bias (those that are at most d below the acceptance threshold), as there are females that have a disadvantage due to the bias (those that are at most d above the threshold), as within the programs males and females are assumed to be equally capable. Under this assumption we need to take the average of the acceptance probability of males and females, resulting in $P^\star(+|\text{med}) = 20\,\%$ for medicine and $P^\star(+|\text{med}) = 40\,\%$ for computer science. Alternatively, if we fix the number of positive labels in the groups to the number observed in the discriminatory data, we would get $170/1000 = 17\,\%$ acceptance for medicine and $440/1000 = 44\,\%$ for computer science. Following the rationale of the discrimination model, however, these numbers are skewed and would result in programs more popular among females to be perceived as being more selective, leading to *redlining*.

Table 4 Calculating the explainable difference

	Medicine		Computer	
	Female	Male	Female	Male
Number of applicants	800	200	200	800
Acceptance rate (Example 2) (%)	15	25	35	45
Corrected acceptance rate (%)	20		40	
Accepted explainable	160	40	80	320

This way, when decisions are automated, the discrimination would transfer from gender to program; a program with lots of females would receive an overall lower acceptance.

Thus, we assume that the acceptance thresholds would have been fixed as the average of the historical acceptance thresholds for males and females. This choice is motivated by the scenario where the candidates come continuously and that any candidate that is sufficiently qualified would get a position, or salary level, or a loan. Hence, there is no resource constraint, and the number of positive outputs only depends upon the number of instances meet a certain threshold. An alternative scenario would be to assume that all the applications are collected together at a deadline. Then the candidates are ranked, and a fixed number of the best candidates are offered a position. Whether to keep the number of accepted individuals fixed or to keep the acceptance threshold fixed depends on the application domain. For instance, in case of scholarships and job application, university acceptance fixing the number of persons may be more reasonable, since the applicants come in batch at the deadline. In case of deciding to grant a credit or what salary level to apply, fixing the threshold makes more sense (accept all individuals that pass qualification requirements), since the individuals come one by one. We argue that the choice of acceptance scenario is situation dependent and hence not part of the design of non-discrimination techniques.

Table 4 illustrates calculation of the explainable part for the discrimination toward females, as presented in Example 2. We find the correct acceptance rate within each program as the average of male and female acceptance. Thus, $D_{\text{expl}} = 36\,\% - 24\,\% = 12\,\%$. From the original data $D_{\text{all}} = 41\,\% - 19\,\% = 22\,\%$. Thus, from Eq. (3) we get $D_{\text{illegal}} = D_{\text{all}} - D_{\text{expl}} = 22\,\% - 12\,\% = 10\,\%$; the data have 10 % of illegal discrimination.

Formally, the explainable discrimination is the difference between acceptance of males and females

$$P^{\star}(+|e_i) := \frac{P(+|e_i, m) + P(+|e_i, f)}{2}, \tag{4}$$

if every individual with a fixed value of the explanatory attribute value e_i would have the same chance to be accepted,[2] independently of the gender:

$$\begin{aligned} D_{\text{expl}} &= \sum_{i=1}^{k} P(e_i|m) P^{\star}(+|e_i) - \sum_{i=1}^{k} P(e_i|f) P^{\star}(+|e_i) \\ &= \sum_{i=1}^{k} \left(P(e_i|m) - P(e_i|f)\right) P^{\star}(+|e_i), \end{aligned}$$

[2] Short notation of probabilities: $P(+|e_i)$ means $P(y = +|e = e_i)$.

where $e \in \{e_1, \ldots, e_k\}$, $P(e_i|m)$ and $P(e_i|f)$ are observed from data, and $P_c^\star(+|e_i)$ is calculated as in Eq. (4). The illegal discrimination can thus be computed as the difference between D_{all} (Eq. (1)) and D_{expl}:

$$D_{\text{illegal}} = P(+|m) - P(+|f) - \sum_{i=1}^{k} (P(e_i|m) - P(e_i|f))\, P^\star(+|e_i). \tag{5}$$

4.2 Effects of redlining

Till now we have formalized the difference between illegal and explainable discrimination, our next step is to analyze under what circumstances a trained classifier risks to capture illegal discrimination. We discuss a scenario where it is no longer allowed to discriminate females directly, the gender information is kept hidden from the admission committee (or not used by the classier for future decision making) to avoid the gender discrimination. The committee will treat male and female applicants within medicine and within computer science equally. However, knowing the fact that females prefer to apply to medicine, it is still possible to discriminate indirectly (without knowing the gender of an applicant). A decision maker who wants to discriminate may reduce the overall acceptance rates to medicine and increase the acceptance rate to computer science.

For our analysis we use synthetic data that are generated based on our toy model introduced in Fig. 1. We generate 10,000 male and 10,000 female instances. The (integer) test scores $T \in [1, 100]$ are assigned uniformly for any individual. In every experiment all probabilities in the Belief network (given in Fig. 1) are fixed, except for the probabilities $P(e_i|s)$: for $\alpha \in [0, 1]$, we generate data with: $P(\text{med}|f) = \alpha$, $P(\text{cs}|f) = 1 - \alpha$, $P(\text{med}|m) = 1 - \alpha$, and $P(\text{cs}|m) = \alpha$. In this way we can study the influence of the strength of the relationship between gender and program on the discrimination, while the total number of people applying for medicine (and computer science respectively) remains the same. For interpretation reasons denote $\beta = P(\text{med}|f) - P(\text{cs}|f) = \alpha - (1-\alpha) = 2\alpha - 1$, then $\beta \in [-1, 1]$ can be interpreted as correlation between the gender and the program. The closer $|\beta|$ is to 1, the stronger the dependency between the explainable and sensitive attribute becomes; $\beta = 0$ means that the gender and the program are independent. Hence, the closer β will be to 0, the less explainable discrimination there will be.

Following the discrimination model introduced in Sect. 3.1 we assign the label to an individual in the toy dataset as

$$y = \delta\left[\left(t + a(-1)^{\delta[\text{med}]} + b(-1)^{\delta[f]}\right) > 70\right], \tag{6}$$

where $\delta[.]$ is a function that outputs 1 if its argument is true and 0 otherwise, t is the test score assigned to an individual, a is the effect to acceptance decisions due to program, and b is the effect to the acceptance due to gender discrimination bias.

We report three cases with different acceptance decisions determined from Eq. (6) under discrimination scenarios. The scenarios are summarized in Table 5. In Case I acceptance depends only on the program choice and the test; thus, all the discrimination is explainable. In Case II both programs have the same acceptance thresholds, but the acceptance decision depends on gender; thus, all the discrimination is illegal. Case III is a combination of illegal and explainable discrimination, and the acceptance depends on the test, the program and the gender.

Table 5 Three discrimination scenarios for analysis

	$P(t)$	a	b	$P(\text{med}\|f)$
Case I, only explainable	0.01	10	0	α
Case II, only illegal	0.01	0	5	α
Case III, explainable and illegal	0.01	10	5	α

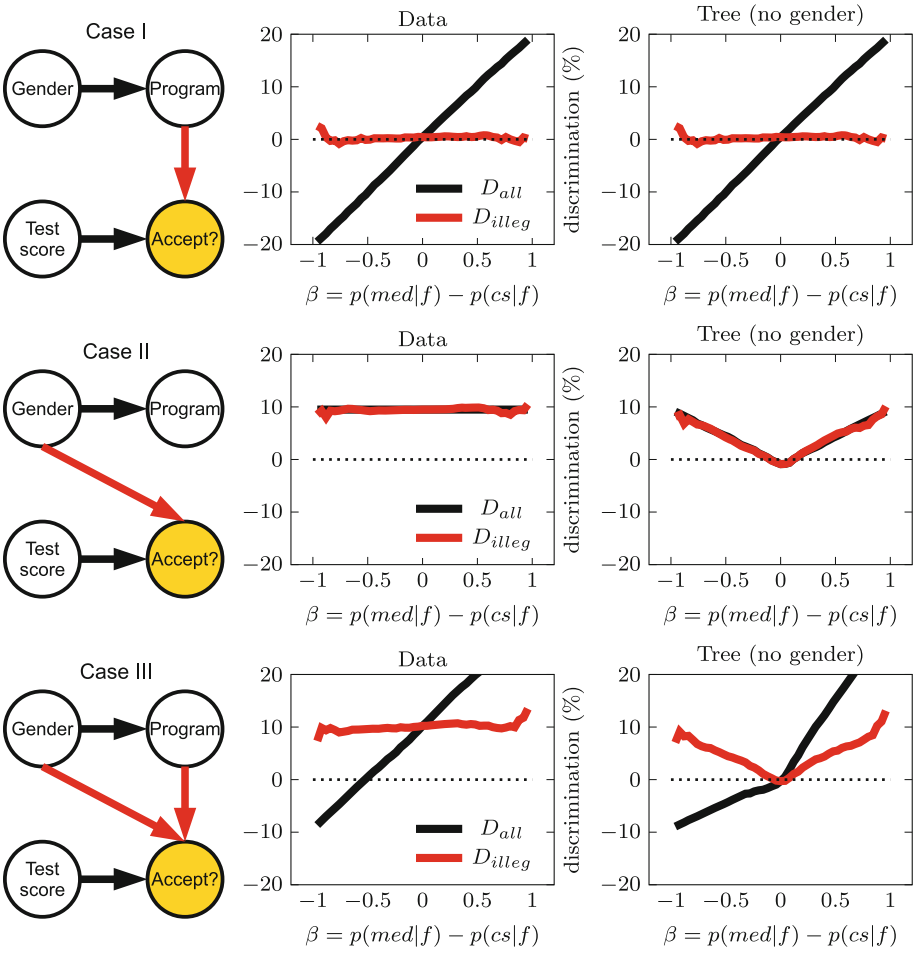


Fig. 2 Interactions between explainable and illegal discrimination

Figure 2 presents the discrimination in function of $\beta = P(\text{med}|f) - P(\text{cs}|f)$. The left plots show the discriminations D_{all} and D_{illegal} in the testing data with the original labels. The right plots show the resulting discriminations with the predicted labels by a decision tree. A decision tree is trained on the data from which gender has been removed, and the training data include only the program and the test score. We analyze the interaction between D_{all} and D_{illegal}.

Case I illustrates the situation from Example 1, where all the difference in acceptance is explainable by program. The difference in acceptance, which we observe as D_{all}, depends on the relation between gender and program, and it is all explainable and thus can be tolerable.

Case II illustrates an opposite situation, where all the discrimination is illegal. Therefore, we observe that D_{all} and D_{illegal} in the plots overlap. In this case the program and the label are not directly related. When the gender attribute is removed, the learned decision tree captures the discriminatory decisions indirectly through program. This way the *redlining effect* appears, which is strong when gender and program are strongly dependent. If program and gender are independent ($\beta = 0 : P(\text{med}|f) = P(\text{med}|m) = P(\text{med}) = 0.5$), then no *redlining* is observed ($D_{\text{illegal}} = 0$). Notice that in this extreme case the classifier can be easily made discrimination-free by removing both gender and program from the input space, without losing any useful information.

In Case III, which corresponds to Example 2, the explainable and the illegal discrimination act together. Some of the difference in acceptance appears due to illegal discrimination, while some is explainable by the program choice and thus can be tolerated. The learned decision tree shows the same illegal discrimination (D_{illegal}) as in Case II. However, the probabilities of acceptance for males and females are different in Case II and Case III. D_{all} in Case III becomes negative for $\beta < 0$. We can see that if very few females apply to medicine ($P(\text{med}|f)$ is close to zero), which is more competitive program, then $D_{\text{all}} < 0$ indicates that females are favored, while in fact they are deprived, as 10 % of illegal discrimination is present ($D_{\text{illegal}} \neq 0$). This case illustrates the Simpson's paradox [38], in which a relation present in different groups is reversed when the groups are combined. Thus, to assess the true illegal discrimination, we need to be able to measure D_{illegal}, and we propose the methodology to measure it in this study.

To sum, the experiments demonstrate the following effects:

- removing the sensitive attribute does not remove discrimination if the sensitive attribute is (cor)related with other attributes (Cases II and III);
- if an input attribute is (cor)related with the sensitive attribute *and* the label and is nominated as explanatory, not all the difference in acceptance is illegal and removing all the difference would result in the reverse discrimination;
- Case III demonstrates that there is a need for advanced training strategies to remove discrimination, and at the same time to preserve the objective information that could be captured by one and the same variable.

5 Removing the illegal discrimination when training a classifier

As we observed in the synthetic examples, a naive approach to remove the sensitive attribute before training will not work if any other attribute is (cor)related with the sensitive attribute. Removing the explanatory attribute would help to remove illegal discrimination, but the accuracy will suffer, as the explanatory attribute at the same time bears the objective information about the label. For instance, in our example the program objectively explains the difference in decisions as acceptance rates differ for different programs. Thus, in real-life scenarios more involved strategies to remove discrimination are required.

In order to ensure that the built classifier is discrimination-free, one needs to control both

1. $P_c(+|e_i, m) = P_c(+|e_i, f)$, where P_c is the probability assigned by the classifier, and

2. $P_c(+|e_i) = P^\star(+|e_i)$, where $P^\star(+|e_i)$ is defined in Eq. (4). This means that the prediction is consistent with the original distribution of the data.

As discussed before, the first condition in isolation is insufficient due to the *redlining effect*. A classifier that only takes this condition into account would underestimate the positive class probability of a group in which females are overrepresented.

We distinguish two main strategies that could make classifiers free from illegal discrimination. The first strategy is to remove the relation between the sensitive attribute and the class label from the training data, which is the source of the illegal discrimination (relation (4) in Fig. 1). Note that removing the relation is not the same as removing the sensitive attribute itself, it means making $P(+|\text{med}, f) = P(+|\text{med}, m) = P^\star(+|\text{med})$. We can achieve that, for instance, by modifying the original labels of the training data.

The alternative strategy is to split the data into smaller groups based on the explanatory attribute. That would remove the relation between the sensitive and the explanatory attributes (relation (2) in Fig. 1). Then individual classifiers can be trained for each group. This strategy would also require to correct the training labels in each groups, otherwise the *redlining effect* will manifest. In addition, it would significantly reduce the data available for training a classifier, which may result in much lower accuracy than the global model. Thus, in this study we adopt the first type of strategy.

In this work we propose three algorithmic techniques for removing the illegal discrimination. The techniques first preprocess historical training data to satisfy the conditional non-discrimination constraints: $P'(+|e_i, f) = P'(+|e_i, m) = P'^\star(+|e_i)$ and $P^\star(+|e_i)$ is fixed so that no *redlining* is introduced (P' denotes the probability in the modified data). First we need to fix the desired probabilities of acceptance $P^\star(+|e_i)$, which would have been correct. We set $P^\star(+|e_i)$ to be the average of male and female acceptance rates, Eq. (4), as motivated in Sect. 4.1. After finding $P^\star(+|e_i)$ for all $e_i \in dom(e)$, the remaining part is to change the labels of the training data so that $P'(+|e_i, f) = P'(+|e_i, m) = P^\star(+|e_i)$. The local massaging and the local preferential sampling techniques anticipate that the classifiers trained on the modified data, which do not contain illegal discrimination, will produce outputs that would satisfy $P_c(+|e_i, f) = P_c(+|e_i, m) = P^\star(+|e_i)$ (P_c denotes the probability in the outputs of a classifier). The third technique, introduced as a baseline, uses the preprocessed data to correct the decision boundaries of the existing discriminatory classifiers directly; it does not train a new classifier on the preprocessed data. The role of the proposed techniques is using our theory on conditional non-discrimination (Sect. 4) to decide which instances in the historical data need to be modified and in what way.

5.1 Local massaging

The local massaging for every partition in the training data induced by the explanatory attribute will modify the values of labels until both $P'(+|m, e_i)$ and $P'(+|f, e_i)$ become equal to $P^\star(+|e_i)$. The discrimination model in Sect. 3.1 implies that discrimination is more likely to affect the objects that are closer to the decision boundary. To this end, massaging identifies the instances that are close to the decision boundary and changes the values of their labels to the opposite. For that purpose individuals need to be ordered according to their probability of acceptance. To be able to order, we need to convert the original binary labels (accept or reject) to real-valued probabilities of acceptance. For that we learn an internal ranker (a classifier that outputs the posterior probabilities).

Suppose females have been discriminated as in our university admission model and the discrimination is reflected in the historical data. The local massaging will identify a number of females that were almost accepted and make their labels positive and identify a number of males that were very likely, but have not been rejected, and make their labels negative.

This technique is related to the massaging proposed in [26], while, given the new theory, now it can handle the illegal discrimination. Algorithm 1 gives the pseudo-code. The procedure for local massaging is illustrated in Fig. 3.

Algorithm 1: Local massaging

input : dataset $(\mathbf{X}, \mathbf{s}, \mathbf{e}, \mathbf{y})$
output: modified labels $\hat{\mathbf{y}}$

`PARTITION` $(\mathbf{X}, \mathbf{e})$ (Algorithm 3);
for *each partition* $X^{(i)}$ **do**
  learn a ranker $p(+|X^{(i)}, e_i) = \mathcal{H}_i(X^{(i)})$;
  rank `males` using $\mathcal{H}_i$ according to $p(+|X^{(i)}, e_i)$;
  `relabel DELTA` (`male`) `males` that are the closest to the decision boundary from $+$ to $-$ (Algorithm 4);
  rank `females` using $\mathcal{H}_i$ according to $p(+|X^{(i)})$;
  `relabel DELTA` (`female`) `females` that are the closest to the decision boundary from $-$ to $+$
end

5.2 Local preferential sampling

The preferential sampling technique does not modify the training instances or labels; instead, it modifies the composition of the training set. It deletes and duplicates training instances so that the labels of new training set contain no discrimination and satisfy the criteria $P'(+|m, e_i) = P'(+|f, e_i) = P^{\star}(+|e_i)$. Following the discrimination model where the discrimination is more likely to affect the objects that are closer to the decision boundary, the preferential sampling deletes the "wrong" instances that are close to the decision boundary and duplicates the instances that are "right" and close to the boundary. To select the instances, they are ordered according to their probability of acceptance using a ranker learned on each group in the same way as in the local massaging.

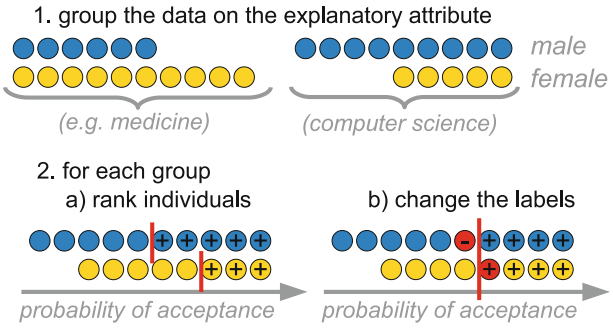


Fig. 3 Local massaging

Algorithm 2: Local preferential sampling

input : dataset $(\mathbf{X}, \mathbf{s}, \mathbf{e}, \mathbf{y})$
output: resampled dataset (a list of instances)

PARTITION $(\mathbf{X}, \mathbf{e})$ (see Algorithm 3);
for *each partition* $X^{(i)}$ **do**
  learn a ranker $p(+|X^{(i)}, e_i) = \mathcal{H}_i(X^{(i)})$;
  rank males using $\mathcal{H}_i$ according to $p(+|X^{(i)}, e_i)$;
  delete $\frac{1}{2}$DELTA (male) (see Algorithm 4) males $+$ that are the closest to the decision boundary from $-$ to $+$;
  duplicate $\frac{1}{2}$DELTA (male) males $-$ that are the closest to the decision boundary from $-$ to $+$;
  rank females using $\mathcal{H}_i$ according to $p(y^{(i)} = +|X^{(i)})$;
  delete $\frac{1}{2}$DELTA (female) females $-$ that are the closest to the decision boundary from $-$ to $+$;
  duplicate $\frac{1}{2}$DELTA (female) females $+$ that are the closest to the decision boundary from $-$ to $+$;
end

Algorithm 3: subroutine PARTITION$(\mathbf{X}, \mathbf{e})$

find all unique values of e: $\{e_1, e_2, \ldots, e_k\}$;
for $i = 1$ **to** k **do**
  make a group $X^{(i)} = \{X : e = e_i\}$;
end

In the university example the local preferential sampling will delete a number of males that were almost rejected and duplicate the males that were almost accepted. It will also delete a number of females that were almost accepted and duplicate the females that were almost rejected.

Algorithm 4: subroutine DELTA(gender)

return $G_i|p(+|e_i, \text{gender}) - p^\star(+|e_i)|$,
where $p^\star(+|e_i)$ comes from (Eq. (4)),
G_i is the number of gender people in $X^{(i)}$;

This technique is related to the preferential sampling [24], while, given the new theory, now it can handle the explainable discrimination. Algorithm 2 gives the pseudo-code. The procedure for local preferential sampling is presented in Fig. 4.

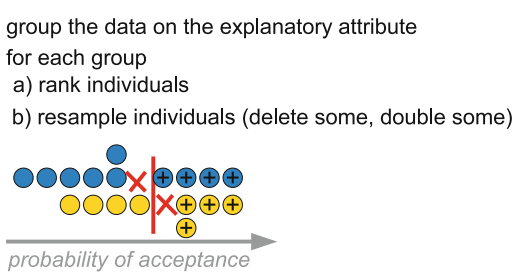


Fig. 4 Local preferential sampling

5.3 Local direct classifier

The local direct classifier technique can be considered as a baseline. It does not train a new discrimination-free classifier; instead, it modifies the decision boundary of the existing discriminatory classifier directly. Firstly, a separate classifier is built for each intersection of the explanatory group and the sensitive group. The instances within each intersection are ranked from the highest probability of acceptance to the lowest. Then the conditional non-discrimination criteria $P'(+|m, e_i) = P'(+|f, e_i) = P^{\star}(+|e_i)$ are computed for each explanatory group. Finally, the decision boundaries of each existing classifier are adjusted to satisfy the conditional non-discrimination criteria in the training data.

In the university example the local direct classifier will rank males and females within medicine and computer science separately. It will compute how many males and females should be accepted to medicine and to computer science to satisfy the criteria. It will use the ranker classifier directly for decision making for new applicants. Algorithm 5 describes the training procedure, and Algorithm 6 describes the classification procedure.

Algorithm 5: Local direct training

input : dataset $(\mathbf{X}, \mathbf{s}, \mathbf{e}, \mathbf{y})$
output: classifiers $\mathcal{H}^{(s)_e}$ with decision thresholds $\Theta^{(s)_e}$
PARTITION $(\mathbf{X}, \mathbf{e})$ (Algorithm 3);
for *each partition* $X^{(i)}$ **do**
  learn a ranker for males $p(+|X^{(i)}, e_i, m) = \mathcal{H}_i^{(m,i)}(X^{(i)})$;
  learn a ranker for females $p(+|X^{(i)}, e_i, f) = \mathcal{H}_i^{(f,i)}(X^{(i)})$;
  set the decision boundary from $+$ to $-$ for males according to the j^{th} ranked male: $\Theta^{(m)_{e_i}} = \mathcal{H}_{e_i}^{(m)}(X_j^{(i)})$, where $j = p^{\star}(+|e_i)$ from Eq. (4);
  set the decision boundary for females as $\Theta^{(f)_{e_i}} = \mathcal{H}_{e_i}^{(f)}(X_j^{(i)})$
end

Algorithm 6: Local direct classification

input : new data instance (X, s, e)
output: decision $\hat{y}$
if $p(+|X, e, s) \geq \Theta_e^{(s)}$ **then**
  $\hat{y} = +$;
else
  $\hat{y} = -$
end

The local direct classifier uses the same model for internal ranking and for decision making. We refer to this technique as a baseline since in a general setting we expect that different classification models may be needed to produce good rankings and good classification decisions.

6 Experimental evaluation

We evaluate the performance of the proposed local discrimination handling techniques in line with their global counterparts. The objective is to minimize the absolute value of the *illegal* discrimination while keeping the accuracy as high as possible. It is important not to overshoot and end up with a reverse discrimination. The goals of our experiments are:

1. to present a motivation for conditional discrimination-aware classification research,

2. to explore how well the proposed techniques remove illegal discrimination as compared to the existing techniques for global non-discrimination and
3. to analyze the effects of removing discrimination on the final classification accuracy.

We explore the performance of the methods that aim to remove the relation between the sensitive attribute and the label. We test the local massaging and the local preferential sampling.

6.1 Data

We use three real-world datasets. In the **Adult** dataset [2], the task is to classify individuals into *high-* and *low-*income classes. We use a uniform sample of 15,696 instances, which are described by 13 attributes (we discretize the 6 numeric attributes) and a class label. Gender is the sensitive attribute, and income is the label. We repeat our experiments several times, where any of the other attributes in turn is selected as explanatory.

The second dataset is the **Dutch Census of 2001** [15] (further referred as Dutch) that represents aggregated groups of inhabitants of the Netherlands. We formulate a binary classification task to classify the individuals into *high-income* and *low-income* professions, using occupation as the class label. Individuals are described by 11 categorical attributes. After removing the records of underaged people, several professions in the middle level and people with unknown professions, our dataset consists of 60,420 instances. Gender is treated as the sensitive attribute.

The third one is the **Communities and Crimes** dataset [2]. This dataset has 1 994 instances which give information about different communities and crimes within the United States. Each instance is described by 122 predictive attributes which are used to predict the total number of violent crimes per 100 K population. In our experiments we discretize some numerical attributes to use them as explanatory attributes. We add a sensitive attribute *Black* to divide the communities by thresholding the numerical attribute *racepctblack* at 0.06. We use *kid-2-parents, pct-illegal, pct-div, under-poverty and population* attributes as explanatory attributes as they are correlated with both the sensitive attribute and the class attribute. We discretize the class attribute to divide the data objects into major and minor violent communities.

Figure 5 shows the discrimination in the datasets. Here and in the next plots, the attributes on the horizontal axis are ordered from the largest correlation with the sensitive attribute to the lowest.

In the Adult dataset a number of attributes are weakly related with gender (such as workclass, education, occupation, race, capital loss and native country). Therefore, nominating any of those attributes as explanatory would not explain much of the discrimination. For instance, we know from biology that race and gender are independent. Thus, race cannot explain the discrimination on gender; that discrimination is either illegal or it is due to some other attributes. Indeed, we observe from the plot that all the discrimination is illegal, when treating race attribute as explanatory.

On the other hand, we observe that the relationship attribute explains a lot of D_{all}. Whether relationship is an acceptable argument to justify differences in income is for lawyers to determine. Judging subjectively, the values of this attribute "*wife* and *husband*" clearly capture the gender information. From a data mining perspective, if we treat it as acceptable, a large part of the discrimination gets explained.

Age and working hours per week are other examples of explanatory attributes. They justify some of the discrimination. Intuitively, these reasons are perfectly valid for having different income, so it makes sense to treat them as explanatory.

In the Communities and crimes dataset, overall discrimination D_{all} is very high, and the attributes *kid-2-parents, pct-illegal and pct-div* can explain nearly half of D_{all}.

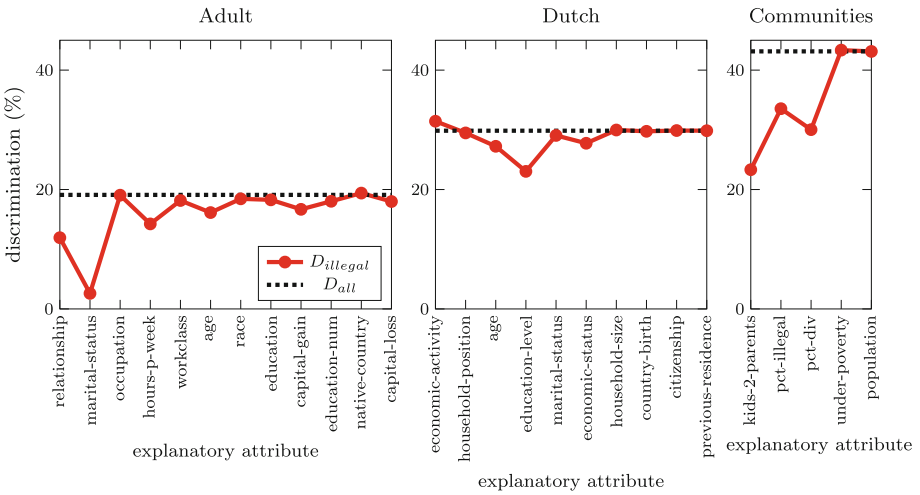


Fig. 5 Discrimination contained in the datasets

In the Dutch dataset the difference between the all and the illegal discrimination is much smaller than in the Adult data. Here many attributes are not that strongly correlated with gender. Simply removing the sensitive attribute should therefore perform reasonably well. Nevertheless, education level, age and economic activity present cases for conditional non-discrimination; thus, we explore this dataset in our experiments.

6.2 Motivation experiments

To give a motivation for our new approach, we demonstrate that the existing techniques do not solve the conditional non-discrimination problem.

6.2.1 Removing the sensitive attribute

First we test a naive approach, which removes the sensitive attribute from the training data. We learn a decision tree with the J48 classifier (Weka implementation) on all the data except the gender attribute, treated as sensitive. Figure 6 shows the resulting discriminations, when the learned tree (no-Sen) is evaluated using 10-fold cross-validation. We can clearly observe the *redlining effect*, especially in the Adult data; even though the sensitive attribute is removed, the illegal discrimination still manifests.

6.2.2 Global techniques

Next we investigate to what extent the two existing global techniques [6,24] remove illegal discrimination. Global massaging (G-Mas) modifies the labels of the training data to make the probabilities of acceptance equal for the two sensitive groups. Global preferential sampling (G-Pre) resamples the training data so that non-discrimination constraints for the label distribution are satisfied. Both methods aim at making D_{all} equal to 0, which is not the same as removing D_{illegal} and will actually reverse the discrimination, as can be seen from Fig. 7. The global techniques do not take into account that the distributions of the sensitive groups

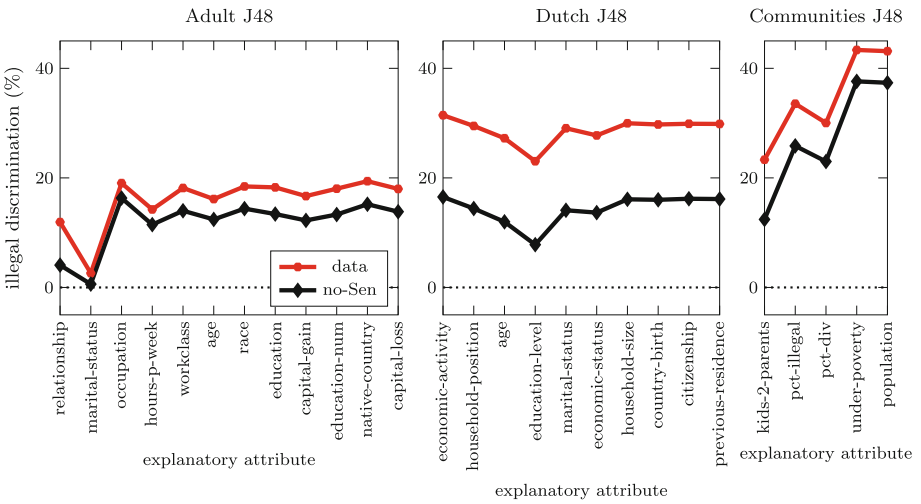


Fig. 6 Discrimination after removing the sensitive attribute (no-Sen)

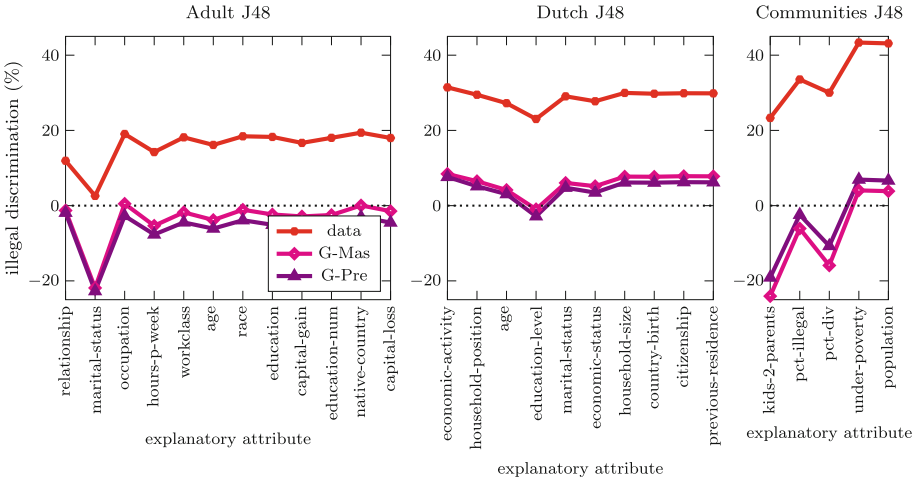


Fig. 7 Discrimination with the *global* techniques (G-Mas and G-Pre)

may differ and thus some of the differences in probabilities are explainable. Hence, the global methods overshoot and a reverse discrimination is introduced, as illustrated in Fig. 7.

As expected, the massaging and the preferential sampling techniques work well for removing all discrimination, for example, for the Adult data after massaging $D_{\text{all}} = 0$. But, if we treat *marital status* as the explanatory attribute, these results introduce a reverse illegal discrimination. The same, but on a smaller scale, holds for several other explanatory attributes, for example, *hours per week* and *age*. For the Dutch Census data, both techniques overshoot if conditioned on *education level*.

These results confirm that a reverse illegal discrimination is introduced when global discrimination handling techniques are applied raising the necessity for local methods.

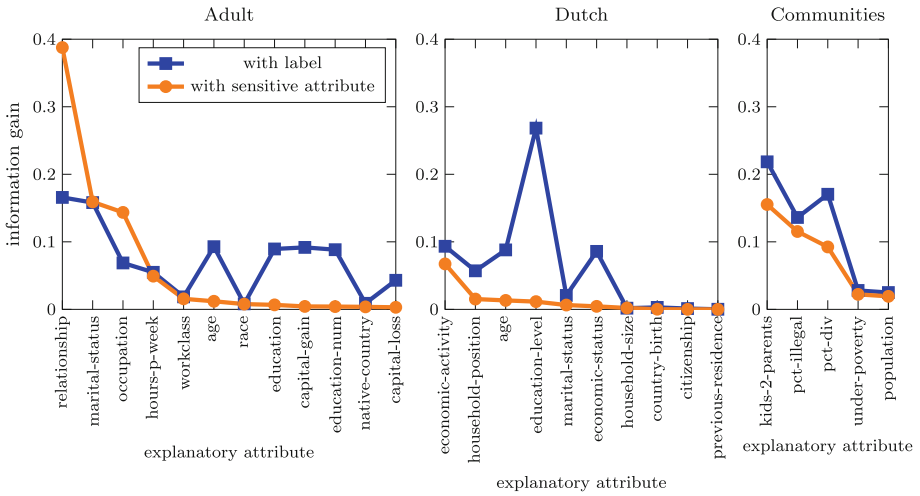


Fig. 8 Relations between sensitive, explanatory attributes and labels

6.2.3 Applicability of the local techniques

The existing techniques fail the most when the difference between D_{all} and D_{illegal} in the data is large. For instance, Fig. 5 shows sharp negative peaks when *marital status* or *relationship* acts as the explanatory attributes in the Adult data. In such cases, the need for the special techniques that can handle conditional discrimination is essential.

A large difference between D_{all} and D_{illegal} implies that a large part of the difference in the decisions is due to the explanatory attribute. We quantify the dependencies between class on the one hand, and sensitive and explanatory attributes on the other hand by the following information gains:

$$G(y, e_i) = H(y) - H(y|e_i), \text{ and}$$
$$G(s, e_i) = H(s) - H(s|e_i).$$

$H(.)$ denotes entropy, s the sensitive attribute, y the label and e_i the explanatory attribute. The information gains for the Adult and the Dutch census datasets are plotted in Fig. 8. The figure confirms the intuition that the stronger the relation with the explanatory attribute (higher information gain), the larger the share of the total discrimination that is explainable. Recall Fig. 5 for the discriminations.

6.3 Removing the illegal discrimination using local techniques

Let us analyze how the proposed local techniques handle discrimination. We expect them to remove exactly the illegal discrimination and nothing more. We test the performance with the decision trees (J48) and the Naive Bayes classifier (NBS) via 10-fold cross-validation.

Figure 9 shows the resulting discrimination after applying the local massaging (L-Mas) and the local preferential sampling (L-Pre). The local direct classifier (L-Dir) is used as a baseline.

The intelligent local techniques L-Mas and L-Pre perform well with J48 on the Adult data. Illegal discrimination is reduced to nearly zero, except for relationship as explanatory

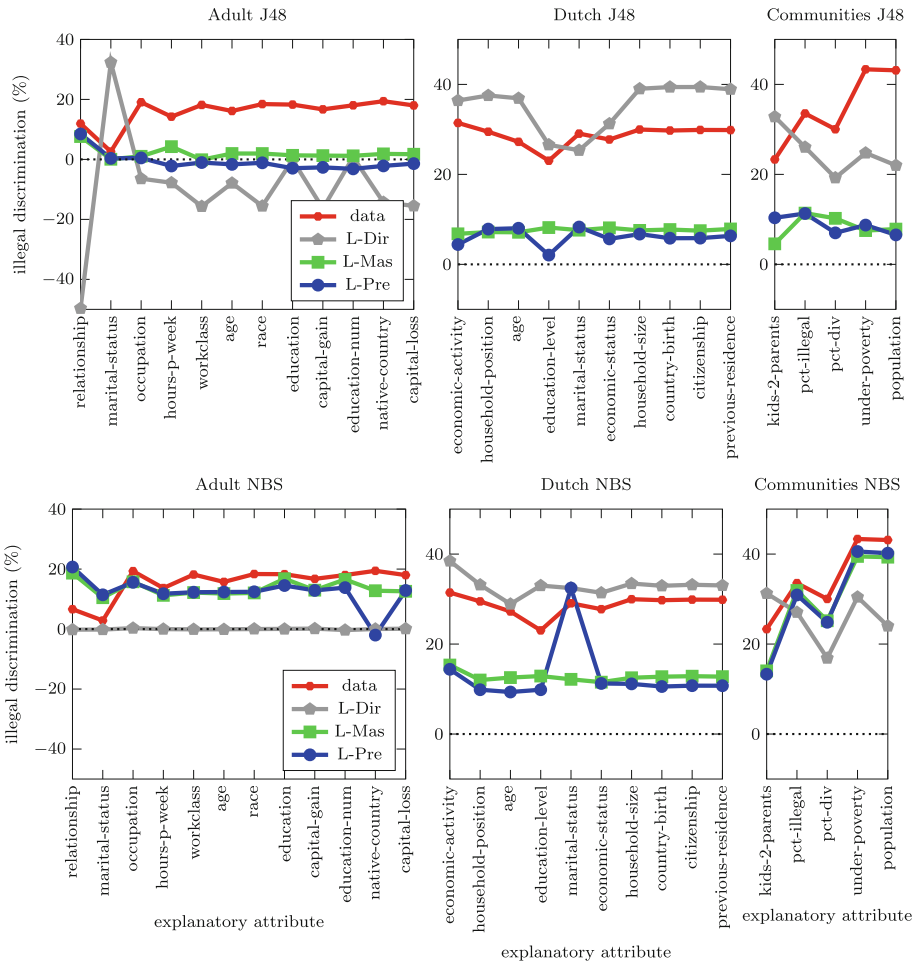


Fig. 9 Discrimination with *the local* techniques (L-Mas, L-Pre and L-Dir)

attribute when massaging is applied to the Adult dataset. Our techniques do not produce the reverse discrimination as, for example, global massaging does.

The proposed solutions do not perform well with J48 on the Dutch census data and the Communities and Crimes data, as the sensitive attribute is not very strongly correlated with any other attribute in the dataset (as we see in Fig. 8). Our proposed local techniques are primarily designed to handle high correlations with the sensitive attribute that induce *redlining*.

Removing discrimination with NBS as the base classifier is not that efficient, as we see from the figure. One explanation for that performance relates to the nature of the Naive Bayes classifier, which treats attributes as independent and effectively prevents pushing toward an opposite discrimination at a microlevel within the explanatory groups. Other explanation is that Naive Bayes is more stable classifier and does not readily pick the changes in the data. It tends to perform consistently even the training data are modified to some extent.

We observe in our experiments that the baseline L-Dir does not perform that well, as shown in Fig. 9. One reason for the poor performance of L-Dir could be that we need sufficient data

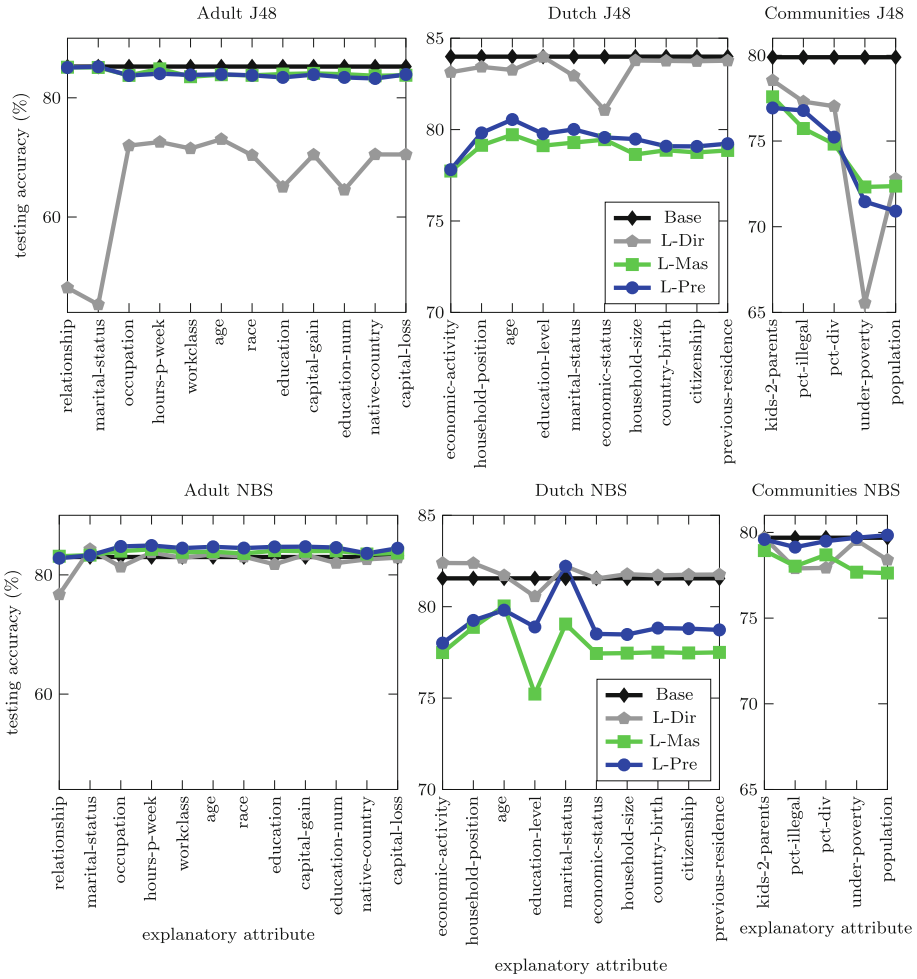


Fig. 10 Accuracy with *the local* techniques (L-Mas, L-Pre and L-Dir)

to select the accurate thresholds for both favored and deprived communities. We observe in our experiments that when we split the data w.r.t. explanatory attribute values, the number of instances in each bin becomes insufficient to determine the accurate decision boundaries. Moreover, the ratio between the instance of favored community (e.g., male) and deprived community (e.g., female) is often not that balanced for thresholding. The poor performance of L-Dir with decision tree can be attributed to the fact that we use the decision tree both as a base classifier and as a ranker; a decision tree is not designed to output smooth posterior probabilities and is not that good as a ranker and consequently shows to be not that suitable for using within L-Dir.

6.4 Accuracy with the local techniques

When classifiers become discrimination-free, they may lose some accuracy, as measured in the historical data. Figure 10 presents the testing accuracy of a decision tree (Base) when

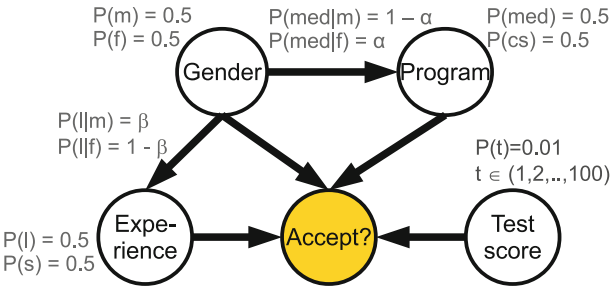


Fig. 11 University admission example with two explanatory attributes

the original historical data with all the attributes are used for training, and the accuracy after our local techniques have been applied. The accuracy of the local techniques L-Mas and L-Pre decreases as the evaluation is carried out on the original data that contain discrimination. Nevertheless, the absolute accuracy remains high; it drops by 5 % at most. L-Dir, on the other hand, shows poor accuracy as expected, particularly when using J48 as a ranker.

Overall, our experiments demonstrate that the local massaging and the local preferential sampling classify future data with reasonable accuracy and maintain low discrimination.

7 Handling multiple explanatory attributes

Our theory and techniques for computing the explainable discrimination are built upon the assumption that there is only one explanatory attribute; however, in reality there may be more than one explanatory attribute that need to be taken into account together (e.g., working hours and experience in determining a salary). This section presents an extension for handling multiple explanatory attributes that may need to be considered together (e.g., working hours and experience in determining a salary).

Let us first consider the following modification of the university admission example. Suppose there are again 2,000 applicants, 1,000 males and 1,000 females. Each program receives the same number of applicants, but medicine is more popular among females, $P(\text{med}|f) = 0.8$. In addition, applicants can have *long* or *short* previous work experience, and females have shorter work experience on average, $P(\text{sh}|f) = 0.6$. The belief network with the assigned probabilities is provided in Fig. 11.

Assume that medicine is more competitive, $P(+|\text{med}) < P(+|\text{cs})$, and that probability of acceptance is higher for the applicants that have long work experience, $P(+|\text{lo}) > P(+|\text{sh})$. Within each program males and females are treated equally, as described in Table 6. The aggregated scores indicate that 37 % of males were accepted, but only 23 % of females. If we try to explain the difference only by the program, we get that within medicine 19 % of females were accepted, while 33 % of males were accepted, and within computer science 39 % of females and 41 % of males were accepted. If, however, we take into account program and experience, we see (Table 6) that given the same experience and the same application program, male and female candidates have been treated equally. Thus, there is no illegal discrimination. However, note that we need to analyze all the combinations of all the values of the explanatory attributes separately.

Table 6 Example 4: no illegal discrimination

	Medicine, long		Medicine, short		Computer, long		Computer, short	
	Female	Male	Female	Male	Female	Male	Female	Male
Number of applicants	320	120	480	80	80	480	120	320
Acceptance rate (%)	25	25	15	15	45	45	35	35
Accepted (+)	80	30	72	12	36	216	42	112

In reality, however, this approach is not applicable. Firstly, if we have more explanatory attributes that can take large sets or wide ranges of values, the number of groups to be considered will explode. In such a case it will be impractical and infeasible to consider all the groups separately. Moreover, some groups can have as little as one or two members, that situation would introduce a lot of noise and inaccuracy in estimating discrimination. More importantly, in such a case it becomes increasingly less likely that two instances will agree on all the attributes. That is a problem, since if we treat every instance as unique, then there we observe no discrimination, as there is nothing to compare an instance with. Thus, we need to form large enough groups to have a pool for comparison within each group.

Therefore, we propose a more practical and meaningful solution for handling multiple explanatory attributes. The idea is to create a synthetic explanatory attribute $\tilde{e}$ that would integrate all the explanatory attributes that we need to consider $\tilde{e} = f(e^{(1)}, e^{(2)}, \ldots, e^{(k)})$, where k is the number of explanatory attributes. Then the new attribute $\tilde{e}$ that describes a group to which a person belongs can be treated as explanatory when applying the theory and techniques proposed in this study. The main intuition behind grouping is to monitor that individuals that are similar to each other in terms of explanatory attributes (fall into one group) are treated in a similar way in decision making regardless of the gender. The resulting groups themselves are expected to be correlated with the sensitive attribute and the label, as the explanatory attributes are.

The major challenge in this approach is how to define the grouping procedure $f()$. In order not to introduce the *redlining*, the grouping *procedure* $f()$ needs to be independent from the sensitive attribute and the label.

In this study we provide an illustration of the proposed approach using clustering of the explanatory attributes as a simple grouping approach. In order to minimize the risk of capturing sensitive information into the grouping procedure, we omit from the clustering input space the attributes that are exceptionally highly correlated with the sensitive attribute.

We report the results of the following experiment on the Adult dataset. In order to form the groups, we run the k-means clustering on the input data. We omit from the clustering input space gender itself, relationship, marital status, occupation and income. None of the attributes is exceptionally correlated with the label; thus, we did not omit due to that.

We compare the illegal discrimination D_{illegal} in the outputs of a decision tree (J48) trained on the original data and on the data that have been preprocessed using the global and our local techniques (massaging and preferential sampling), discussed in Sect. 5. We test the performance via 10-fold cross-validation. We use the same experimental protocol as with one explanatory attribute. Figure 12 presents the resulting illegal discrimination and accuracies. We observe, as in the case with one explanatory attribute, that the global techniques overshoot and introduce the reverse discrimination, while our local techniques remove exactly the illegal discrimination and they preserve reasonable prediction accuracy.

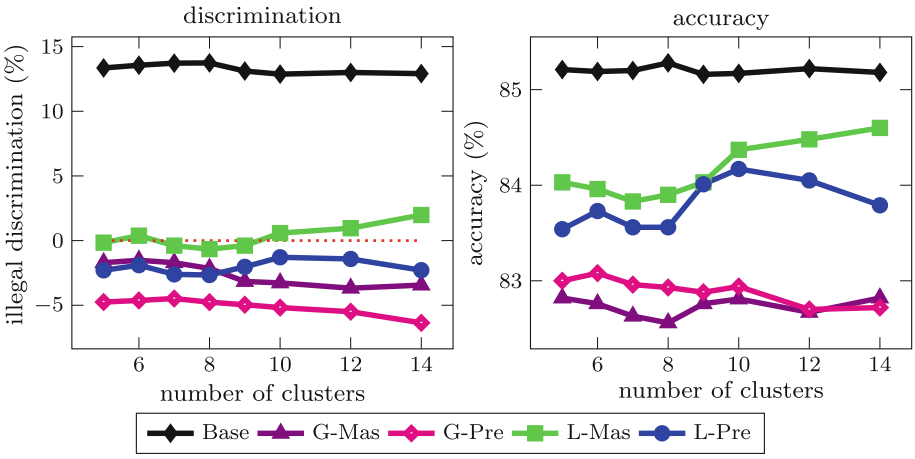


Fig. 12 Discrimination and accuracy with multiple explanatory attributes

8 Related work

The concept of discrimination is relatively new in data mining, but it has been studied in social sciences for a long time. We broadly categorize the related work of the discrimination-aware classification problem into the related works in social sciences and the related works in data mining.

8.1 Social sciences

Social sciences, for example, economics, law, sociology and education, deal with the inculcation of different aspects of society. The general forms of discrimination such as racism, sexism, ableism, ageism, casteism, classism, colorism, linguicism and rankism are referred to the social discrimination on the basis of race, gender, disabilities, age, caste, social class, skin or eye color, language and rank of a person, respectively. In social sciences research many aspects of discrimination have been studied. In this study we only overview the most relevant works concerned with anti-discrimination in the legal and economic domains.

In the legal domain, there are many civil right laws to prohibit the practice of discrimination. In the United States there are many anti-discrimination laws [39] to prevent the discriminatory practices from the society, for example, the equal credit opportunity act [43], equal pay act [44], the civil rights act [42] and the fair housing act [45]. Similarly in the European Union [17,19] and the UK [41], there are many laws which prohibit discrimination and ensure the equal treatment to the people. In addition to the anti-discrimination laws, there are many organizations which are working to protect the civil rights of citizens. For instance, the European Network Against Racism (ENAR) [18] is a network of European NGOs working to combat racism in all EU member states and represents more than 700 NGOs throughout the European Union.

Gary S. Becker [4] analyzes the factors that lead to economic discrimination in the market place, employer and employee discrimination, consumer discrimination and changes in the discrimination over time in his book *The Economics of Discrimination* [4]. He develops a useful model for analyzing the economic effects of discrimination. He treats Negro and white sectors of the United States as if they were separate countries in an international trade model,

and he assumes that the white sector owns a higher ratio of capital to labor than the Negro sector does. The discrimination affects the dealings of Negroes and whites in a similar way as tariff barriers impede trade between two countries. Gary S. Becher's work got a lot of attention from the research community and resulted in many critical reviews [11,13,29,34,37] on this book which proposed new directions to study the economic discrimination.

8.2 Data mining

In data mining the discrimination-aware decision-making problem got the attention from the data mining research community recently; however, we can trace out some research works of similar nature from the data mining and machine learning literature. We give a brief overview of related work in data mining by bucketing different works in discrimination-aware data mining itself, cost-sensitive classification and sampling techniques for unbalanced datasets.

In *discrimination-aware data mining* there are two important directions to work on: discrimination discovery from the given datasets [20,21,32,33,35,36] and the discrimination prevention from the future decision making [7,23–26,47]. The works on discrimination discovery find out the discriminatory practices from the given datasets. A central notion in their works on identifying discriminatory rules is that of the *context* of the discrimination. That is, specific regions in the data are identified in which the discrimination is particularly high. These works assume that the discriminatory attribute is not present in the dataset and background knowledge for the identification of discriminatory guidelines has to be used.

A recent paper [30] proposes a variant of k-NN classification for the discovery of discriminated objects. The authors consider a data object as discriminated if there exist a significant difference of treatment among its neighbors belonging to a protected-by-law group (i.e., the deprived community) and its neighbors not belonging to it (i.e., the favored community). They also propose a discrimination prevention method by changing the class labels of these discriminated objects. This discrimination prevention method is very close to our local massaging technique, especially when the ranker being used is based upon a nearest neighbor classifier. There is, however, one big difference: Whereas in massaging only the minimal number of objects is changed to remove all discrimination from the dataset, the authors of [30] propose to continue relabeling until all labels are consistent. From a legal point of view, the cleaned dataset obtained by [30] is probably more desirable as it contains less "illegal inconsistencies." For the task of discrimination-aware classification, however, it is unclear whether the obtained dataset is suitable for learning a discrimination-free classifier. The exploration of this option could be a promising direction for further research. The authors of [20,21] also propose methods similar to local massaging to preprocess the training data in such a way that only potentially non-discriminatory rules can be extracted. For this purpose they modify all the items in a given dataset that lead to the discriminatory classification rules by applying rule hiding techniques on either given or discovered discriminative rules.

Our current work lies in the category of works on discrimination prevention in the future decision making. However, we differ from the previous discrimination prevention works [7,23–26] in defining of what is considered to be non-discriminatory. The previous works require the acceptance probabilities to be equal across the sensitive groups. It means that if 10 % of male applicants is accepted, also 10 % of female applicants should be accepted. The previous works solve the problem by introducing a reverse discrimination either in the training data [6,24] or pushing constraints into the trained classifiers [7,26]. These works do not consider any difference in the decisions to be explainable and thus tend to overshoot in

removing discrimination so that males become discriminated in future. We are not aware of any study formulating or addressing this problem of conditional non-discrimination from a data mining perspective other than [47].

In *Cost-Sensitive and Utility-Based learning* [8,16,31,40,46], it is assumed that not all types of prediction errors are equal and not all examples are as important. In cost-sensitive learning the goal is no longer to optimize the accuracy of the prediction, but rather the total cost. Domingos proposes a method named MetaCost [14] for making classifiers cost-sensitive by wrapping a cost-minimizing procedure around them. MetaCost assumes that costs of misclassifying the examples are known in advance and are the same for all the examples. It is based on relabeling the training examples with their estimated minimal-cost classes, and applying the error-based learner to the new training set. As such, MetaCost has some similarity with *Local Massaging* with respect to relabeling the training data, but *Local Massaging* relabels only the training examples, which may be potentially misclassified due to the impact of discrimination, while MetaCost changes the labels of all the training examples.

In *Sampling Techniques for Unbalanced Datasets*. [10], a synthetic minority over-sampling technique (SMOTE) for two-class problems that over-sampled the minority class by creating synthetic examples rather than replicating examples is proposed. Chawla et al. [9] also utilize a wrapper [27] approach to determine the percentage of minority class examples to be added to the training set and the percentage to under-sample the majority class examples. Koknar-Tezel and Latecki [28] present an innovative approach that augments the minority class by adding synthetic points in distance spaces and then use Support Vector Machines for classification. These sampling methods show some similarity with our local preferential sampling technique; by increasing the number of samples in one group (the deprived community members with a positive label), we increase the importance of this group such that the classifier learned on the re-sampled dataset is forced to give more attention to this group. Making an error on this group will hence be reflected in more severe penalties than in the original dataset.

9 Conclusion

We have presented the discrimination-aware decision-making problem in a broader and more practical perspective. We have motivated the discrimination problem in automated decision making by establishing its connection with anti-discrimination laws. We have discussed the discrimination-aware classification paradigm in the presence of explanatory attributes that are correlated with the sensitive attribute.

In such a case, as we demonstrated, not all discrimination can be considered illegal, and the existing techniques tend to overshoot and start a reverse discrimination. Therefore, we introduced a new way of measuring discrimination, by explicitly splitting it up into explainable and illegal discrimination. In addition, we have introduced two discrimination prevention techniques that preprocess the training data before learning a classifier in order to remove only illegal discrimination. We have also introduced a third discrimination prevention technique that prevents illegal discrimination by adjusting decision boundaries of a trained discriminatory classifier directly based on the values of the explanatory attribute.

We have presented an extensive experimental evaluation on the multiple real-world datasets to analyze the performance of our proposed methods as compared to the current state-of-the-art methods. The experiments demonstrated the effectiveness of the new local techniques, especially in cases when the sensitive attribute is highly correlated with the explanatory attribute.

Our theory and techniques for computing the explainable discrimination work with one explanatory attribute. In reality more than one explanatory attribute may need to be taken into account. To address that we have developed a framework that allows to aggregate multiple explanatory attributes into a single synthetic attribute and apply our theory and algorithmic techniques for discrimination-aware classification with multiple explanatory attributes.

In this paper, we assumed that the sensitive attribute and the explanatory attributes are nominated by the domain experts, for example, legal experts. Otherwise, the selection of a reasonable combination of the sensitive attribute and the explanatory attribute becomes very confusing and debatable because one reasonable combination could be highly unreasonable for the other one. We have discussed several works in data mining [20,21,32,33,35,36] in Sect. 8.2 which mainly focused on detection of discriminatory patterns within a given dataset. Combining our discrimination-aware classification techniques with their discrimination detection methods is one direction for our future research.

While considering one or more explanatory attributes, we restricted ourselves to a binary classification problem and one binary sensitive attribute. Our current settings may be extended to a multiple class problem by converting a multiclass classification problem to a number of one-against-all binary classification problems. Nevertheless, often there will be a more subtle gradation in desirability between the classes that needs to be taken into account as well. We can handle a sensitive attribute with multiple values in a similar way by choosing some of the values as defining the deprived community, yet again similar objections apply. It becomes even more difficult when the discrimination problem has multiple sensitive attributes that can be combined, for example, if we consider both gender and ethnicity as sensitive attributes at the same time, for example, *black females*. In this case black females may be deprived, while white females may be favored, but overall there is discrimination toward females which makes the problem more challenging to solve. A promising direction could be to extend the work [30] where discriminated instances are identified by finding discrepancies in labeling with its k nearest neighbors in the other community. For the definition of the distance function, we could incorporate the neutrality of certain attributes such as "Number of car crashes in the past" by, for example, giving them a higher weight.

We can conclude that this paper only touches the tip of the iceberg. Much remains to be done to extend the solutions to include a large number of sensitive attributes, deal with numerical sensitive attributes and regression problems. We believe discrimination-aware classification is a practically relevant and interesting research area with many open problems.

References

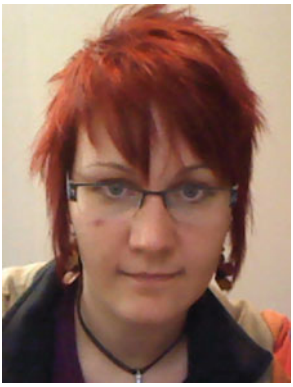
1. Ahearn T (2010) Discrimination lawsuit shows importance of employer policy on the use of criminal records during background checks. via: http://www.esrcheck.com/wordpress/2010/04/12/
2. Asuncion A, Newman D (2007) UCI machine learning repository. Online http://archive.ics.uci.edu/ml/
3. Attorney-General's Dept C (1984) Australian sex discrimination act 1984. via: http://www.comlaw.gov.au/Details/C2010C00056
4. Becker G (1971) The economics of discrimination. University of Chicago Press, Chicago
5. Bickel P, Hammel E, O'Connell J (1975) Sex bias in graduate admissions: data from Berkeley. Science 187(4175):398–404
6. Calders T, Kamiran F, Pechenizkiy M (2009) Building classifiers with independency constraints. In: IEEE ICDM workshop on domain driven data mining (DDDM'09), pp 13–18
7. Calders T, Verwer S (2010) Three naive bayes approaches for discrimination-free classification. Data Mining Knowl Discov 21(2):277–292
8. Chan PK, Stolfo SJ (1998) Toward scalable learning with non-uniform class and cost distributions: a case study in credit card fraud detection. In: Proceedings of ACM SIGKDD conference on knowledge discovery and data mining (KDD'98), pp 164–168

9. Chawla N, Hall L, Joshi A (2005) Wrapper-based computation and evaluation of sampling methods for imbalanced datasets. In: Proceedings of the 1st international workshop on Utility-based data mining, pp 24–33
10. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) Smote: synthetic minority over-sampling technique. J Artif Intell Res 16:321–357
11. Collard D (1972) The economics of discrimination. Econ J 82(326):788–790
12. Dedman B (1988) The color of money: atlanta blacks losing in home loans scramble: banks favor white areas by 5–1 margin. Atlanta J Const
13. Dewey D (1958) The economics of discrimination. South Econ J 24(4):494–496
14. Domingos P (1999) Metacost: a general method for making classifiers cost-sensitive. In: Proceedings of ACM SIGKDD conference on knowledge discovery and data mining (KDD)), pp 155–164
15. Dutch Central Bureau for Statistics (2001) Volkstelling
16. Elkan C (2001) The foundations of cost-sensitive learning. In: Proceedings of the 17th international joint conference on, artificial intelligence (IJCAI'01), pp 973–978
17. Ellis E (2005) EU anti-discrimination law. Oxford University Press, Oxford
18. European Network Against Racism (1998). via: http://www.enar-eu.org/
19. European Union Legislation (2012) via: http://europa.eu/legislation_summaries/index_en.htm
20. Hajian S, Domingo-Ferrer J, Martinez-Balleste A (2011) Discrimination prevention in data mining for intrusion and crime detection. In: IEEE symposium on computational intelligence in cyber security (CICS). IEEE, pp 47–54
21. Hajian S, Domingo-Ferrer J, Martínez-Ballesté A (2011) Rule protection for indirect discrimination prevention in data mining. Model Dec Artif Intell 6820:211–222
22. Hart M (2005) Subjective decisionmaking and unconscious discrimination. Alabama Law Rev 56:741
23. Kamiran F, Calders T (2009) Classifying without discriminating. In: Proceedings of the 2nd international conference on computer, control and, communication (IC4), pp 1–6
24. Kamiran F, Calders T (2010) Classification with no discrimination by preferential sampling. In: Proceedings of the 19th annual machine learning conference of Belgium and the Netherlands (BENELEARN'10), pp 1–6
25. Kamiran F, Calders T (2012) Data preprocessing techniques for classification without discrimination. Knowl Inf Syst 33:1–33
26. Kamiran F, Calders T, Pechenizkiy M (2010) Discrimination aware decision tree learning. In: Proceedings of IEEE international conference on data mining (ICDM), pp 869–874
27. Kohavi R, John GH (1997) Wrappers for feature subset selection. Artif Intell 97(1–2):273–324
28. Koknar-Tezel S, Latecki L (2010) Improving SVM classification on imbalanced time series data sets with ghost points. Knowl Inf Syst 24(2):1–23
29. Krueger A (1963) The economics of discrimination. J Polit Econ 71(5):481–486
30. Luong B, Ruggieri S, Turini F (2011) k-nn as an implementation of situation testing for discrimination discovery and prevention. Technical Report TR-11-04, Dipartimento di Informatica, Universita di Pisa
31. Margineantu D, Dietterich T (1999) Learning decision trees for loss minimization In: Multi-class problems. Technical report, Department of Computer Science, Oregon State University
32. Pedreschi D, Ruggieri S, Turini F (2008) Discrimination-aware data mining. In: Proceedings of ACM SIGKDD conference on knowledge discovery and data mining (KDD'08)
33. Pedreschi D, Ruggieri S, Turini F (2009) Measuring discrimination in socially-sensitive decision records. In: Proceedings of the SIAM international conference on data mining (SDM'09), pp 581–592
34. Reder M (1958) The economics of discrimination. Am Econ Rev 48(3):495–500
35. Ruggieri S, Pedreschi D, Turini F (2010) DCUBE: discrimination discovery in databases. In: Proceedings of the ACM SIGMOD international conference on management of data (SIGMOD'10). ACM, pp 1127–1130
36. Ruggieri S, Pedreschi D, Turini F (2010) Integrating induction and deduction for finding evidence of discrimination. Artif Intell Law 18:1–43
37. Sawhill I (1973) The economics of discrimination against women: some new findings. J Human Res 8(3):383–396
38. Simpson EH (1951) The interpretation of interaction in contingency tables. J R Stat Soc 13:238–241
39. U. The US department of Justice (2011) The US federal legislation, via: http://www.justice.gov/crt
40. Turney P (2000) Cost-sensitive learning bibliography. In: Institute for Information Technology, National Research Council, Ottawa, Canada
41. United Kingdom Legislation, 2012. via:http://www.legislation.gov.uk/
42. The US Civil Rights Act, 2006. via: http://finduslaw.com/
43. U. Us Dept. of Justice. Us equal credit opportunity act, 1974. via: http://www.fdic.gov/regulations/laws/rules/6500-1200.html

44. E. Us Empl. Opp. Comm. Us equal pay act, 1963. via: http://www.eeoc.gov/laws/statutes/epa.cfm
45. US Fair Housing Act (1968). via: http://www.justice.gov/crt/about/hce/
46. Wang B, Japkowicz N (2009) Boosting support vector machines for imbalanced data Sets. Knowl Inf Syst, pp 1–20
47. Zliobaite I, Kamiran F, Calders T (2011) Handling conditional discrimination. In: Proceedings of IEEE international conference on data mining (ICDM'11), pp 992–1001

Author Biographies

Faisal Kamiran got his MSCS (Master in Science and Computer Science) degree from University of the Central Punjab (UCP), Lahore, in 2006. He got the top position in UCP during his MSCS. He received his PhD degree from the Eindhoven University of Technology The Netherlands in 2011. He has done his doctoral research in the Databases and Hypermedia (DH) group under the supervision of Prof. Dr. Toon Calders and Prof. Dr. Paul De Bra. Currently, he is working as a postdoc fellow in King Abdullah University of Science and Technology (KAUST), KSA. His research interests include constraints-based classification, privacy preserving and graph mining.



Indrė Žliobaitė is a Lecturer in Computational Intelligence at Bournemouth University, UK. She received her PhD from Vilnius University, Lithuania. I. Žliobaitė has six years of experience in credit analysis in banking industry. Her research interests and expertise concentrate around adaptive and context-aware machine learning, learning from evolving streaming data, change detection and predictive analytics applications. Recently, she has co-chaired workshops at ECMLPKDD 2010 and ICDM 2011, co-organized tutorials at CBMS 2010 and PAKDD 2011 on adaptive learning. She is a Research Task Leader within the INFER.eu project that is developing robust adaptive predictive systems. For further information see http://zliobaite.googlepages.com.

Toon Calders graduated in 1999 from the University of Antwerp with a diploma in Mathematics. He received his PhD in Computer Science from the same university in May 2003, in the database research group ADReM. From May 2003 until September 2006, he continued working in the ADReM group as a post-doctoral researcher. Since October 2006, he is an assistant professor in the Information Systems group at the Eindhoven Technical University. Toon Calders published over 50 papers on data mining in conference proceedings and journals, was conference chair of the BNAIC 2009 and EDM 2011 conferences and is a member of the editorial board of the Springer Data Mining journal and Area Editor for the Information Systems journal.