Knowl Inf Syst (2012) 33:1–33
DOI 10.1007/s10115-011-0463-8

REGULAR PAPER

Data preprocessing techniques for classification without discrimination

Faisal Kamiran · Toon Calders

Received: 23 November 2010 / Revised: 23 August 2011 / Accepted: 16 November 2011 /
Published online: 3 December 2011

Abstract Recently, the following *Discrimination-Aware Classification Problem* was introduced: Suppose we are given training data that exhibit unlawful discrimination; e.g., toward *sensitive attributes* such as gender or ethnicity. The task is to learn a classifier that optimizes accuracy, but does not have this discrimination in its predictions on test data. This problem is relevant in many settings, such as when the data are generated by a biased decision process or when the sensitive attribute serves as a proxy for unobserved features. In this paper, we concentrate on the case with only one binary sensitive attribute and a two-class classification problem. We first study the theoretically optimal trade-off between accuracy and non-discrimination for pure classifiers. Then, we look at algorithmic solutions that preprocess the data to remove discrimination before a classifier is learned. We survey and extend our existing data preprocessing techniques, being *suppression* of the sensitive attribute, *massaging* the dataset by changing class labels, and *reweighing* or *resampling* the data to remove discrimination without relabeling instances. These preprocessing techniques have been implemented in a modified version of Weka and we present the results of experiments on real-life data.

Keywords Classification · Preprocessing · Discrimination-aware data mining

1 Introduction

Classifier construction is one of the most researched topics within the data mining and machine learning communities. Literally thousands of algorithms have been proposed.

This paper is an extended version of the papers [3,13,14].

F. Kamiran (✉)
HG 7.46, P.O. Box 513, 5600 MB, Eindhoven, The Netherlands
e-mail: faisal.kamiran@gmail.com

T. Calders
HG 7.82a, P.O. Box 513, 5600 MB, Eindhoven, The Netherlands
e-mail: t.calders@tue.nl

The quality of the learned models, however, depends critically on the quality of the training data. No matter which classifier inducer is applied, if the training data are incorrect, poor models will result. In this paper we study cases in which the input data are discriminatory and we want to learn a discrimination-free classifier for future classification. In Sect. 2, we sketch three realistic scenarios in which the discrimination-aware classification problem occurs naturally and we link to relevant anti-discrimination legislation.

Discrimination-aware classification originally stems from [13,14] and was further explored in [3]. The input of the discrimination-aware classification problem is a labeled dataset and one or more sensitive attributes. The output is a classifier to predict the label that should not correlate with the sensitive attribute. The quality of the classifier is measured by its accuracy and discrimination; the more accurate, the better, and the less discriminatory, the better. In the previous papers, and in this one, we restrict ourselves to one binary sensitive attribute S with domain $\{b, w\}$ and a binary classification problem with target attribute *Class* with domain $\{-, +\}$. "+" is the desirable class for the data subjects and the objects satisfying $S = b$ and $S = w$ represent, respectively, the deprived and the favored community. The discrimination of a classifier C is defined as

$$disc_{S=b} := P\big(C(X) = + \mid X(S) = w\big) - P\big(C(X) = + \mid X(S) = b\big) ,$$

where X is a random unlabeled data object. A discrimination larger than 0 reflects that a tuple for which S is w has a higher chance of being assigned the positive label by the classifier C than one where S is b. Our choice for this discrimination measure is further motivated in Sect. 3. Similar as for accuracy, the discrimination of a classifier can be estimated using an independent test-set.

The problem of classification without discrimination w.r.t. a sensitive attribute S is in fact a multi-objective optimization problem; on the one hand, the more discrimination we allow for, the higher accuracy we can obtain while on the other hand, in general we can trade in accuracy in order to reduce the discrimination. Before going into algorithmic solutions, we first present a theoretical study of this trade-off in Sect. 4. After that, we discuss algorithmic solutions.

The following four methods for incorporating non-discrimination constraints into the classifier construction process are discussed in Sect. 5. All four methods are based on preprocessing the dataset after which the normal classification tools can be used to learn a classifier.

1. **Suppression.** We find the attributes that correlate most with the sensitive attribute S. To reduce the discrimination between the class labels and the attribute S, we remove S and these most correlated attributes. This simple and straightforward approach will serve as the baseline in our experiments.
2. **Massaging the dataset.** We change the labels of some objects in the dataset in order to remove the discrimination from the input data. A good selection of which labels to change is essential. To select the best candidates for relabeling, a ranker is used. This method is an extension of the method proposed in [13] where a Naïve Bayesian classifier was used for both the ranking and learning. In this paper we will consider arbitrary combinations of ranker and learner.
3. **Reweighing.** Instead of changing the labels, the tuples in the training dataset are assigned weights. As we will show, by carefully choosing the weights, the training dataset can be made discrimination-free w.r.t. S without having to change any of the labels. The weights on the tuples can be used directly in any method based on frequency counts. This method was first proposed in [3].

4. **Sampling.** For those methods that cannot directly work with weights, the related *Sampling* method can be used instead. We calculate sample sizes for the 4 combinations of *S*- and *Class*-values that would make the dataset discrimination-free. Then, we apply stratified sampling on the four groups; two of the groups will be under-sampled and two over-sampled. We introduce two techniques to select which objects to duplicate, and which to remove. In the first scheme, *Uniform Sampling (US)*, we apply uniform sampling with replacement. In this scheme, every object has a uniform probability to be duplicated to increase the size or to be skipped to decrease the size of a group. In the second scheme, *Preferential Sampling (PS)*, borderline objects get high priority for being duplicated or being skipped. A ranker is used to decide which objects are at the border.

Section 6 contains the results of an extensive empirical study. We present the results supporting the following claims:

(i) Removing the attribute *S* from the dataset does not always result in the removal of the discrimination. This we call the redlining effect [13].
(ii) Especially the *Massaging* and *PS* techniques lead to an effective decrease in discrimination with a minimal loss in accuracy.

We stress that in all experiments, the preprocessing techniques have only been applied to the training data. When evaluating the classifiers always, unmodified, independent test data were used.

2 Motivation

Discrimination refers to the unfair and unequal treatment of individuals of a certain group based solely on their affiliation to that particular group, category, or class. Such discriminatory attitude deprives the members of one group from the benefits and opportunities which are accessible to other groups. Different forms of discrimination in employment, income, education, finance, and in many other social activities may be based on age, gender, skin color, religion, race, language, culture, marital status, economic condition, etc. Such discriminatory practices are usually fueled by stereotypes, an exaggerated or distorted belief about a group. Discrimination is often socially, ethically, and legally unacceptable and may lead to conflicts among different groups.

2.1 Scenarios for discrimination-aware classification

We illustrate the need of discrimination-aware classification with three potential scenarios, each one outlining a different situation in which we need to learn a non-discriminatory classifier on biased data.

Scenario 1: historical discrimination [3]. *Throughout the years, an employment bureau recorded various parameters of job candidates. Based on these parameters, the company wants to learn a model for partially automating the match making between a job and a job candidate. A match is labeled as successful if the company hires the applicant. It turns out, however, that the historical data are biased; for higher board functions, Caucasian males are systematically being favored. A model learned directly on this data will learn this discriminatory behavior and apply it in future predictions. From an ethical and legal point of view it is of course unacceptable that a model discriminating in this way is deployed.*

The historical labels only partially represent the task we want to learn; we only want to model the part of the interest of companies in the applicants' profiles that is not related to gender or race. The discrimination-aware classification problem tackles this problem by imposing an additional constraint on the models to be learned.

Scenario 2: multiple data sources. *A survey is being conducted by a team of researchers; each researcher visits a number of regionally co-located hospitals and enquires some patients. The survey contains ambiguous questions (e.g., "Is the patient anxious?", "Is the patient suffering from delusions?"). Different enquirers will answer to these questions in different ways. Generalizing directly from the training set consisting of all surveys without taking into account these differences among the enquirers may easily result in misleading findings. For example, if many surveys from hospitals in the Eindhoven area are supplied by an enquirer who more quickly than the other enquirers diagnoses anxiety, faulty conclusions such as "Patients in Eindhoven suffer from anxiety symptoms more often than other patients" may emerge.* In this case the discrimination-aware classification paradigm is used to avoid overfitting due to different data sources. Actually, here discrimination-aware classification can be seen as a form of incorporating domain knowledge by making explicit the assumption that we consider it to be more likely that differences between the data sources can be explained by different labeling procedures rather than by differences in the underlying distributions. Notice that similar situations exist when comparing scores for research papers among different reviewers, or movie ratings of different people.

Scenario 3: sensitive attribute as a proxy. In some cases the discrimination in the input data appears when the sensitive attribute serves as a proxy of features that are not present in the dataset. Consider, e.g., the support someone may get financial support from his or her family for repaying a mortgage loan when he or she looses his or her job. Such possibility of support of the family or the absence thereof can critically influence the risk a prospective client represents for a bank. This highly useful parameter, however, is very difficult to observe and quantify. Suppose now that, due to socio-cultural or economical reasons, the possibility of family support correlates to the ethnicity of a person. In such a situation a bank could be tempted to use the ethnicity attribute as a proxy for family support. Such ethnic profiling makes perfect economical sense; it will lead to more accurate models for risk, and thus, indirectly, higher gains for the bank. Nevertheless, it is ethically and legally unacceptable. We quote *Turner and Skidmore* [27] on such cases: *"If lenders think that race is a reliable proxy for factors they cannot easily observe that affect credit risk, they may have an economic incentive to discriminate against minorities. Thus, denying mortgage credit to a minority applicant on the basis of minorities on average—but not for the individual in question—may be economically rational. But it is still discrimination, and it is illegal."*

2.2 Anti-discrimination legislation

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. We discuss some of these laws here and show how they relate to our problem statement:

The Australian Sex Discrimination Act 1984 [2] 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. This law also prohibits indirect and unintentional discrimination: *[…] a person […] discriminates against another person […] on the ground of the sex of the aggrieved person if, by reason of: (a) the sex of the aggrieved person; (b) a characteristic that appertains generally to persons of the sex of the aggrieved person; or (c) a characteristic that is generally imputed to persons of the sex of the aggrieved person; the discriminator treats the aggrieved person less favorably than, in circumstances that are the same or are not materially different, the discriminator treats or would treat a person of the opposite sex.* It is the responsibility of the accused party to prove that his/her intention was not to discriminate the aggrieved party: *the burden of proving that an act does not constitute discrimination […] lies on the person who did the act.* Notice that under this law it is insufficient to remove the sex attribute from a dataset before learning; also indirect discrimination on the basis of a "characteristic that appertains generally to persons of the sex of the aggrieved person" is disallowed.

The US Equal Pay Act 1963 [30] requires that men and women in the same workplace be given equal pay for equal work. The jobs need not 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. This act aimed at abolishing wage disparity based on sex. According to the US Bureau of Labor Statistics, women's salaries vis-à-vis men's have risen dramatically since the enactment of this equal pay act, from 62% of men's earnings in 1970 to 80% in 2004 [6]. This real-world case illustrates a scenario where our historical data are discriminatory due to a biased data generation process, but where classifiers learned on the data are forced to be discrimination-free by law.

The US Equal Credit Opportunity Act 1974 [29] declares unlawful for any creditor to discriminate against any applicant, with respect to any aspect of a credit transaction, on the basis of race, color, religion, national origin, sex or marital status, or age [26]. This law contains similar provisions as the previous two.

Another good example of the need for non-discriminative classifiers despite their potentially lower accuracy is given by the **European Council Directive 2004**. Even though there is clear historical evidence showing higher accident rates for male drivers in traffic, insurance companies are no longer allowed to discriminate based on gender in many countries, as explicated by the following recent ruling of the European Court of Justice [25]: *The European Court of Justice decided on March 1, 2011 that, from December 21, 2012, it will no longer be legal under EU law to charge women less for insurance than men. The verdict means that different priced premiums for men and women drivers will now be considered to be in breach of the EU's anti-discrimination rules.* This ruling is the implementation of the European Council Directive 2004/113/EC of December 13, 2004.

All of the anti-discriminatory laws prohibit discriminatory practices in future. It means that our discrimination-aware classification paradigm clearly applies to these situations. If we are interested to apply classification techniques, and our available *historical* data contain discrimination, it is simply illegal to use traditional classifiers without taking the discrimination aspect into account due to these anti-discrimination laws. Because of the above mentioned laws or due to ethical concerns, such use of existing classification techniques is unacceptable. In such a situation our "classification without discrimination" paradigm applies: we want to learn non-discriminatory classification models from biased historical data such that they generate accurate predictions for future decision making, yet do not discriminate with respect to a given sensitive attribute.

3 Problem statement: discrimination-aware classification

In this section we formally define and motivate a measure for the discrimination of a classifier, and we introduce the *Discrimination-Aware Classification Problem*.

3.1 A measure for discrimination

We assume a set of attributes $A = \{A_1, \ldots, A_n\}$ and their respective domains $dom(A_i)$, $i = 1, \ldots, n$ have been given. A *tuple* X over the schema $(A_1, \ldots, A_n)$ is an element of $dom(A_1) \times \ldots \times dom(A_n)$. We denote the value of X for attribute A_i by $X(A_i)$. A dataset over the schema $(A_1, \ldots, A_n)$ is a finite set of such tuples and a labeled dataset is a finite set of tuples over the schema $(A_1, \ldots, A_n, Class)$. Throughout the paper we will assume $dom(Class) = \{-, +\}$.

We assume that a special attribute $S \in A$, called the *sensitive attribute*, and a special value $b \in dom(S)$, called the *deprived community* have been given. The semantics of the pair S, b is that it defines the discriminated community; for example, S could be "ethnicity" and b "Black." For reasons of simplicity we will assume that the domain of S is binary; i.e., $dom(S) = \{b, w\}$. Obviously, we can easily transform a dataset with multiple attribute values for S into a binary one by replacing all values $v \in dom(S) \setminus \{b\}$ with a new dedicated value w.

We define the discrimination in the following way:

Definition 1 (Discrimination in labeled dataset): Given a labeled dataset D, an attribute S and a value $b \in dom(S)$. The discrimination in D w.r.t. the group $S = b$, denoted $disc_{S=b}(D)$, is defined as:

$$disc_{S=b}(D) := \frac{|\{X \in D \mid X(S) = w, X(Class) = +\}|}{|\{X \in D \mid X(S) = w\}|} - \frac{|\{X \in D \mid X(S) = b, X(Class) = +\}|}{|\{X \in D \mid X(S) = b\}|}.$$

That is, the difference of the probability of being in the positive class between the tuples X in D having $X(S) = w$ in D and those having $X(S) = b$. When clear from the context we will omit $S = b$ from the subscript.

Definition 2 (Discrimination in a classifier's predictions): Given an unlabeled dataset D, an attribute S and a value $b \in dom(S)$. The discrimination of the classifier C w.r.t. the group $S = b$ in dataset D, denoted $disc_{S=b}(D)$, is defined as:

$$disc_{S=b}(C, D) := \frac{|\{X \in D \mid X(S) = w, C(X) = +\}|}{|\{X \in D \mid X(S) = w\}|} - \frac{|\{X \in D \mid X(S) = b, C(X) = +\}|}{|\{X \in D \mid X(S) = b\}|}.$$

That is, it is the difference in probability of being assigned the positive class by the classifier between the tuples of D having $X(S) = w$ and those having $X(S) = b$. When clear from the context we will omit $S = b$ from the subscript, and D from the list of arguments.

Example 1 In Table 1, an example dataset is given. This dataset contains the Sex, Ethnicity, and Highest Degree for 10 job applicants, the Job Type they applied for, and the outcome of the selection procedure, *Class*. In this dataset, the discrimination w.r.t. the attribute *Sex* and *Class* is: $disc_{Sex=f}(D) := \frac{4}{5} - \frac{2}{5} = 40\%$. It means that in the dataset, a female is, in absolute numbers, 40% less likely to be accepted for a job than a male.

Table 1 Sample relation for the job-application example

Sex	Ethnicity	Highest degree	Job type	Class
M	Native	H. school	Board	+
M	Native	Univ.	Board	+
M	Native	H. school	Board	+
M	Non-nat.	H. school	Healthcare	+
M	Non-nat.	Univ.	Healthcare	−
F	Non-nat.	Univ.	Education	−
F	Native	H. school	Education	−
F	Native	None	Healthcare	+
F	Non-nat.	Univ.	Education	−
F	Native	H. school	Board	+

3.2 Motivation for the discrimination measure

Our way of measuring discrimination as the difference in positive class probability between the two groups is based upon the following observation. Suppose we have data on employees that applied for jobs and whether or not they got the job, and we want to test if there is gender discrimination. Therefore, we consider the proportion of men that were hired versus the proportion of women that were hired. A statistically significant difference in these proportions would indicate discrimination. Let us indicate the true (resp. observed) proportion of males that were hired as m_1 ($\overline{x}_1$), and the proportion for the females as m_2 ($\overline{x}_2$). Notice that our discrimination measure equals $\overline{x}_1 - \overline{x}_2$. The standard statistical approach for testing if females are discriminated would be to test if a one-sided null hypothesis $h_0 : m_2 \geq m_1$ can be rejected. If the hypothesis gets rejected, the probability is high that there is discrimination. Many different statistical tests could be used in this example; popular tests that apply are the *two-sample t test* or the *two-proportion Z test*. Besides trying to refute the null hypothesis h_0, we could also go for a test of independence between the attributes gender and class with, e.g., the χ^2 *test* or the *G test*. Unfortunately there is no single best test; depending on the situation (usually the absence or presence of abundant data or proportions taking extreme values) one test may be preferable over another. Here we can reasonably assume, since we are working in a data mining context, that sufficient data are available. We also assume that none of the proportions takes extreme values. As such, the choice of test is not that important, as long as we restrict ourselves to one test. The test statistic that would be used for a two-sample t test (assuming unknown and potentially different variances) is:

$$\frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{disc_{gender=f}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} ,$$

where s_1 and s_2 denote the empirical standard deviations of the two groups and n_1 and n_2 their respective sizes. The statistical test, however, only tells us if there is discrimination, but does not indicate the severity of discrimination. In this respect notice that the test statistic for the hypothesis $h_0 : m_1 - m_2 = d_0$ is:

$$\frac{\overline{x}_1 - \overline{x}_2 - d_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} .$$

As this example shows, it is not unreasonable to take the difference between proportions as a measure for the severity of discrimination. Nevertheless, we want to emphasize that similar arguments can be found for defining the discrimination as a ratio, or for using measures based on mutual information gain between sensitive attribute and class or entropy-based measures (such as the G test). In our work we made the choice for the difference in proportions because, statistically speaking, it makes sense, and it has the advantage of having a clear and intuitive meaning of expressing the magnitude of the observed discrimination.

3.3 Definition of the discrimination-aware classification problem

The problem we study in the paper can now be stated as follows:

Definition 3 Discrimination-Aware Classification. Given a labeled dataset D, an attribute S, and a value $b \in dom(S)$, learn a classifier C such that:

(a) the accuracy of C for future predictions is high; and
(b) the discrimination of new examples classified by C is low.

Clearly there will be a trade-off between the accuracy and the discrimination of the classifier. In general, lowering the discrimination will result in lowering the accuracy and vice versa. This trade-off is further elaborated upon in the next section. In this paper we are making three strong assumptions:

A1 We are implicitly assuming that the primary intention is learning the most accurate classifier for which the discrimination is 0. When we assume the labels result from a biased process, insisting on high accuracy may be debatable. Nevertheless, any alternative would imply making assumptions on which objects are more likely to have been mislabeled. Such assumptions would introduce an unacceptable bias in the evaluation of the algorithms toward favoring those that are based on these assumptions. In the case where the labels are correct, yet the discrimination comes from the sensitive attribute being a proxy for absent features, optimizing accuracy is clearly the right thing to do.

A2 Ideally the learned classifier should not use the attribute S to make its predictions. Knowing the attribute S at prediction time may lead to a so-called "reverse discrimination" to cancel out the bad discrimination, which is not always desirable when one can be held legally accountable for decisions based on the classifier's predictions. Besides, it is contradictory to explicitly use the sensitive attribute in decision making while the goal is to ensure that decisions do not depend on the sensitive attribute.

A3 The total ratio of positive predictions of the learned classifier should be similar to the ratio of positive labels in the dataset D. This assumption would hold when assigning a positive label to a person implies an action for which resources are limited; e.g., a bank that can assign only a limited number of loans or a university having bounded capacity for admitting students.

4 Theoretical analysis of the accuracy: discrimination trade-off

Before going into the proposed solutions, we first theoretically study the trade-off between discrimination and accuracy in a general setting.

Definition 4 Let C and C' be two classifiers. We say that C *dominates* C' if the accuracy of C is larger than or equal to the accuracy of C', and the discrimination of C is at most as

high as the discrimination of C'. *C strictly dominates C'* if at least one of these inequalities is strict.

Given a set of classifiers $\mathcal{C}$, we call a classifier $C \in \mathcal{C}$ *optimal w.r.t. discrimination and accuracy (DA-optimal) in* $\mathcal{C}$ if there is no other classifier in $\mathcal{C}$ that strictly dominates C.

Notice that according to this definition a classifier with a high negative discrimination could be DA-optimal. Although this may seem counter-intuitive, it makes sense if we consider a DA-optimal classifier as a classifier with maximal accuracy among all classifiers with the same or lower discrimination. If a user wants a classifier that goes beyond just removing discrimination, and even reverts the discrimination (makes it negative), he or she will have to trade in even more accuracy. Furthermore, as we will show further on, the relation between discrimination and accuracy will be monotone for DA-optimal classifiers; i.e., a DA-optimal classifier with lower discrimination will imply a lower accuracy. If we would have defined the domination relation w.r.t. absolute value of discrimination, this would result in all classifiers with negative discrimination to be dominated by the one that corresponds to a discrimination of 0.

For reasons of simplicity, in our theoretical exposition we assume that a dataset D is given against which discrimination and accuracy of all classifiers is measured. This assumption is not limiting our theoretical results since all our results still hold when the cardinality of D is infinite; i.e., we can think of D as a perfect description of the true underlying probability distribution. Furthermore, we assume a sensitive attribute S and value b have been given. We will use $acc(C)$ and $disc(C)$ to denote, respectively, the accuracy of the classifier C in D and the discrimination of C in D w.r.t. $S = b$. We will use $\mathcal{C}_{all}$ to denote the set of all classifiers and $\mathcal{C}^*_{all}$ to denote the set of all classifiers C such that $P(C(X) = +|X \in D) = P(X(Class) = +|X \in D)$; i.e., all classifiers that have the same overall probability of assigning the positive label as observed in D.

4.1 Perfect classifiers

We first study the trade-off between accuracy and discrimination if we have perfect knowledge about the probability distribution; i.e., we have a perfect classifier C^{Perf} for D; that is, $C^{Perf}(X) = X(Class)$ for all $X \in D$. This perfect classifier is clearly DA-optimal in $\mathcal{C}_{all}$ and $\mathcal{C}^*_{all}$ as no other classifier has the same accuracy of 100%. Our first theorem will explain what is the most optimal way to change this classifier to get other classifiers that are no longer as accurate, but that are DA-optimal because of their decreased discrimination. The rate at which these DA-optimal classifiers have to trade in accuracy to reduce discrimination is what we understand as the *accuracy-discrimination trade-off*.

Let D_b and D_w be defined as follows:

$$D_b := \{X \in D \mid X(S) = b\}$$
$$D_w := \{X \in D \mid X(S) = w\}$$

and let d_b and d_w be, respectively, $|D_b|$ and $|D_w|$. d denotes $|D|$.

Theorem 1 *A classifier C is DA-optimal in $\mathcal{C}_{all}$ iff*

$$acc(C^{Perf}) - acc(C) \;=\; \frac{\min(d_b, d_w)}{d}\left(disc(C^{Perf}) - disc(C)\right)$$

A classifier C is DA-optimal in $\mathcal{C}^*_{all}$ iff

$$acc(C^{Perf}) - acc(C) \;=\; 2\frac{d_b}{d}\frac{d_w}{d}\left(disc(C^{Perf}) - disc(C)\right)$$

Proof Let C be a DA-optimal classifier. We denote the number of true negatives, true positives, false positives, and false negatives of C by, respectively, tn, tp, fp, and fn; e.g., $tp = |\{X \in D \mid X(Class) = C(X) = +\}|$. tp_b denotes the number of true positives that have $S = b$. tp_b, fp_b, ..., and fn_w are defined similarly. With these conventions, we can express the accuracy and discrimination of C as follows:

$$acc(C) = \frac{tp + tn}{d} = \frac{tp_b + tn_b + tp_w + tn_w}{d}$$
$$disc(C) = \frac{tp_w + fp_w}{d_w} - \frac{tp_b + fp_b}{d_b}$$

Let n_b denote the number of objects X in D with $X(Class) = -$ and $X(S) = b$. Similarly we define p_b, n_w, and p_w Notice that $acc(C)$ and $disc(C)$ only depend on tp_b, fp_b, tp_w, fp_w. The other quantities are determined by these four; e.g., $tn_b = n_b - fp_b$. Furthermore, for every choice of $tp_b \in [0, p_b], fp_b \in [0, n_b], tp_w \in [0, p_w], fp_w \in [0, n_w]$, there is a classifier in $\mathcal{C}$ that corresponds to this choice. Therefore, if C is DA-optimal in $\mathcal{C}$, $disc(C)$ must be equal to the solution of the following integer optimization problem:

Minimize

$$\frac{tp_w + fp_w}{d_w} - \frac{tp_b + fp_b}{d_b}$$

in function of the integer variables tp_b, fp_b, tp_w, fp_w, subject to the following constraints:

$$\begin{cases} \dfrac{tp_b + (n_b - fp_b) + tp_w + (n_w - fp_w)}{d} = acc(C) \\ 0 \le tp_b \le p_b \\ 0 \le fp_b \le n_b \\ 0 \le tp_w \le p_w \\ 0 \le fp_w \le n_w \end{cases}$$

In the case of $\mathcal{C}^*$, the constraint

$$tp_b + fp_b + tp_w + fp_w = p$$

needs to be added, where p denotes $|\{X \in D \mid X(Class) = +\}|$.

In both cases, i.e., $\mathcal{C}$ and $\mathcal{C}^*$, any DA-optimal classifier will have $fp_w = 0$ and $tp_b = p_b$. For the case $\mathcal{C}$, this is clear as decreasing fp_w and increasing tp_b both decrease $disc(C)$ and increase $acc(C)$. For $\mathcal{C}^*$, we split into two cases:

$p_b - tp_b > fp_w$ The following solution strictly dominates C, unless $fp_w = 0$ and $tp_b = p_b$:

$$\begin{cases} tp'_b = p_b & tp'_w = tp_w \\ fp'_b = fp_b + tp_b + fp_w - p_b & fp'_w = 0 \end{cases}$$

This solution satisfies all inequalities and has a lower discrimination and higher accuracy.

$p_b - tp_b \le fp_w$ The following solution strictly dominates C, unless $fp_w = 0$ and $tp_b = p_b$:

$$\begin{cases} tp'_b = p_b & tp'_w = tp_b + tp_w + fp_w - p_b \\ fp'_b = fp_b & fp'_w = 0 \end{cases}$$

Again, this solution satisfies all inequalities and has a lower discrimination and higher accuracy.

Hence, we get the following formulas for the difference in accuracy and discrimination between C and C^{Perf}:

$$1 - acc(C) = \frac{fp_b + fn_w}{d}$$
$$disc(C^{Perf}) - disc(C) = \frac{fn_w}{d_w} + \frac{fp_b}{d_b}$$

The extra condition for $\mathcal{C}^*$ becomes:

$$fp_b = fn_w \ .$$

From these equalities the theorem now easily follows. □

It is interesting to see that the discrimination-accuracy trade-off is linear; lowering the discrimination level by 1% results in an accuracy decrease of min $\left(\frac{d_b}{d}, \frac{d_w}{d}\right)$ % and an accuracy decrease of $2\frac{d_b}{d}\frac{d_w}{d}$% if the class distribution needs to be maintained. These DA-optimal classifiers can be constructed from the perfect classifier.

4.2 Imperfect classifiers

In the last theorem we assumed a perfect classifier. In most cases, however, we will only have an imperfect classifier at our disposal. We will now assume that we have such an imperfect classifier C of which we want to reduce its discrimination by randomly changing some of its predictions. The probability with which we will change a prediction of an instance X will depend on $X(S)$ and $X(Class)$ only. We will denote these four probabilities by p_{b+}, p_{b-}, p_{w+}, and p_{w-}. The resulting classifier is denoted $C[p_{b+}, p_{b-}, p_{w+}, p_{w-}]$; i.e., $C[p_{b+}, p_{b-}, p_{w+}, p_{w-}](X)$ equals $C(X)$ with probability $p_{X(S)C(X)}$. Notice that the accuracy and discrimination of this random classifier in fact represents the expected accuracy and discrimination of all deterministic classifiers with p_{b+}, p_{b-}, p_{w+}, p_{w-} correspondence with C. We will denote the class of all classifiers that can be derived from C in this way by $\mathcal{C}_C$. $\mathcal{C}_C^*$ will denote all classifiers C' in $\mathcal{C}_C$ for which it holds that $P(C'(X) = +) = P(C(X) = +)$. The following theorem characterizes the DA-optimal classifiers of $\mathcal{C}_C$ and of $\mathcal{C}_C^*$.

Theorem 2 *If classifier C' is DA-optimal in $\mathcal{C}_C$, then*

$$E[acc(C) - acc(C')] \ = \ (2acc(C) - 1)\frac{\min(d_b, d_w)}{d}(disc(C) - disc(C'))$$

If classifier C' is DA-optimal in $\mathcal{C}_C^*$, then

$$E[acc(C) - acc(C')] \ = \ 2(2acc(C) - 1)\frac{d_b}{d}\frac{d_w}{d}(disc(C) - disc(C'))$$

$E[.]$ denotes here the expected value over all databases D on which C has accuracy $acc(C)$ and discrimination $disc(C)$.

Proof We assume, without loss of generality, that $acc(C) \geq 0.5$; if this is not the case, we switch all predictions of C to obtain a new classifier with an accuracy of $1 - acc(C)$. Let now C' be any classifier with $corr(C, C') = \gamma$; i.e., C and C' agree (correspond) on a fraction γ of the dataset D. Then, the expected value for the accuracy of C' can be computed as follows:

$$\begin{aligned} E[acc(C')] &= [P(C(X) = C'(X)) \times P(C(X) = X(Class)) \\ &\quad + P(C(X) \neq C'(X)) \times P(C(X) \neq X(Class))] \\ &= corr(C, C')acc(C) + (1 - corr(C, C'))(1 - acc(C)) \\ &= corr(C, C')(2acc(C) - 1) + (1 - acc(C)) \end{aligned}$$

Notice that in the given derivation we assume that agreement of C and C' on an instance X is independent from correctness of the prediction of C for X. $C[p_{b+}, p_{b-}, p_{w+}, p_{w-}]$ satisfies this condition. As such, the expected accuracy of the classifiers in $\mathcal{C}_C$ and $\mathcal{C}_C^*$ only depend on their correspondence with C, and the higher the correspondence, the higher the accuracy. Furthermore,

$$E[acc(C) - acc(C')] \;=\; (2acc(C) - 1)(1 - corr(C, C'))$$

On the other hand, we can use Theorem 1 to find the relation between the maximal correspondence with C and the discrimination of the classifier C'; the maximal reduction in discrimination linked to the minimal reduction in correspondence is as follows:

$$\min_{C' \in \mathcal{C}_C, disc(C') = \delta} (1 - corr(C, C')) \;=\; \frac{\min(d_b, d_w)}{d}(disc(C) - \delta)$$

and for $\mathcal{C}_C^*$,

$$\min_{C' \in \mathcal{C}_C^*, disc(C') = \delta} (1 - corr(C, C')) \;=\; 2\frac{d_b}{d}\frac{d_w}{d}(disc(C) - \delta)$$

Combining these two facts leads directly to the theorem. □

Again we see a linear trade-off. This linear trade-off could be interpreted as bad news: no matter what we do, we will always have to trade in accuracy proportional to the decrease in discrimination we want to achieve. Especially when the classes are balanced, this is a high price to pay.

4.3 Classifiers based on rankers

On the bright side, however, most classification models actually provide a score or probability for each tuple for being in the positive class instead of only giving the class label. Such a scoring classifier, called a ranker, actually ranks the objects according to its assessment of the probability that the object is in the positive class. The score allows us for a more careful choice of objects of which to change the prediction: instead of using a uniform chance for all tuples with the same predicted class and S-value, the score can be used as follows. Assume a scoring classifier R that assigns to all objects a score. We can dynamically set different cut-off c_b and c_w for, respectively, tuples with $S = b$ and $S = w$ to obtain the classifier $R(c_b, c_w)$ that will predict $+$ for a tuple X if $X(S) = b$ and $R(X) \geq c_b$ and if $X(S) = w$ and $R(X) \geq c_w$. Otherwise $-$ is predicted. We denote the class of all classifiers $R(c_b, c_w)$ by $\mathcal{C}_R$. Intuitively one expects that slight changes to the discrimination will only incur minimal changes to the accuracy, as the tuples that are being changed are the least certain ones and hence sometimes a change will actually result in a better accuracy. The decrease in accuracy will thus no longer be linear in the change in discrimination, but its rate will increase as the change in discrimination increases, until in the end it becomes linear again, because the tuples we change will become increasingly more certain leading to a case similar to that of the perfect classifier. A full analytical exposition of this case is far beyond the scope of this paper. Instead, we tested this trade-off empirically. The results of this study are shown in

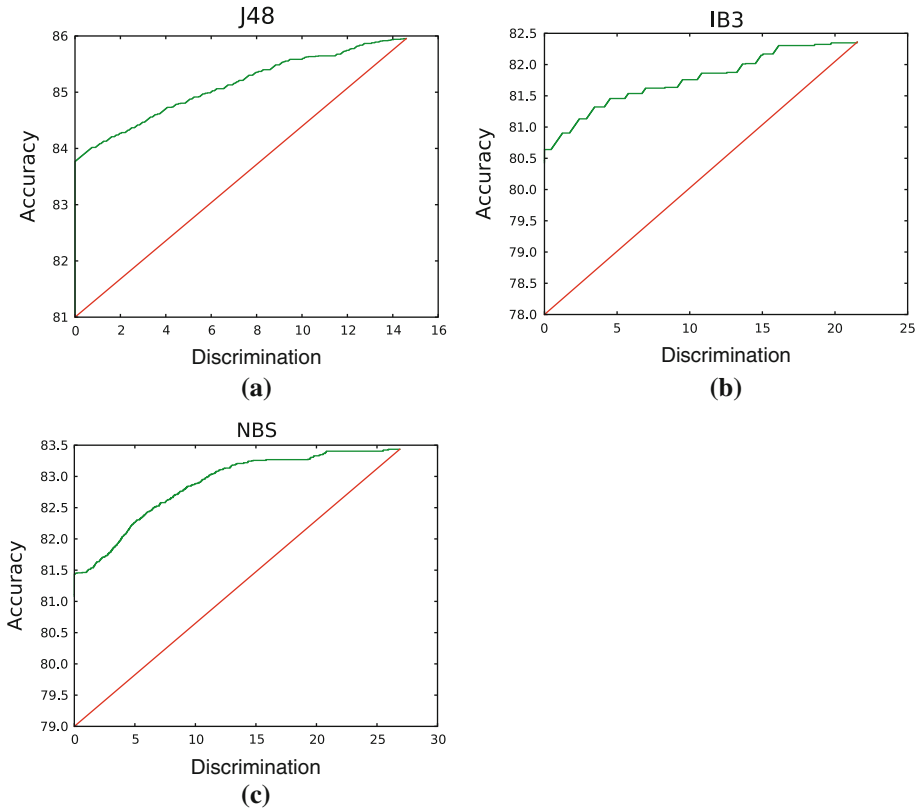


Fig. 1 Trade-off between accuracy and discrimination (dependence) for the DA-optimal classifiers in $\mathcal{C}_R$ and $\mathcal{C}_C$. **a** Decision tree. **b** IBk. **c** Naïve Bayes

Fig. 1. In this figure the DA-optimal classifiers in the classes $\mathcal{C}_R$ (curves) and $\mathcal{C}$ (straight line) are shown for the Census Income dataset [1]. The three classifiers are a Decision Tree (J48), an instance-based classification model with three neighbors (IBk), and a Naïve Bayesian Classifier (NBS). The ranking versions are obtained from, respectively, the (training) class distribution in the leaves, a distance-weighted average of the labels of the 3 nearest neighbors, and the posterior probability score. The classifiers based on the scores perform considerably better than those based on the classifier only.

4.4 Conclusion

In this section the accuracy-discrimination trade-off is clearly illustrated. It is theoretically shown that if we rely on classifiers, and not on rankers, the best we can hope for is a linear trade-off between accuracy and discrimination. For important classes of classifiers the DA-optimal classifiers were explicitly constructed. Notice, however, that the theoretical solutions proposed in this section violate our assumption A2; the classifiers $C[p_{b+}, p_{b-}, p_{w+}, p_{w-}]$ and $R(c_b, c_w)$ heavily use the attribute S to make their predictions. Therefore, these optimal solutions are not suitable for our purposes. In the next section, three solutions will be proposed that do not make use of the attribute S *at prediction time*, but only in the learning phase. The theoretically optimal solutions proposed in this section can be seen as "top-lines" which

in theory we cannot outperform (without S we have strictly less information and hence, if our learning methods would be perfect, our model of the distribution that generated the data deteriorates). The theoretical results represent the goal that we want to approach as close as possible.

5 Solutions: data preprocessing techniques

In this section we propose three solutions to learn a non-discriminating classifier that uses the attribute S only during learning and not at prediction time. All solutions are based on removing the discrimination from the training dataset. Subsequently, a classifier is learned on this cleaned dataset. Our rationale for this approach is that, since the classifier is trained on discrimination-free data, it is likely that its predictions will be (more) discrimination-free as well. The empirical evaluation in Sect. 6 will confirm this statement. The first approach we present, called *Massaging the data*, is based on changing the class labels in order to remove the discrimination from the training data. A preliminary version of this approach was presented in [13]. The second approach is less intrusive as it does not change the class labels. Instead, weights are assigned to the data objects to make the dataset discrimination-free. This approach will be called *Reweighing*. Since reweighing requires the learner to be able to work with weighted tuples, we propose another solution without this requirement, in which we re-sample the dataset in such a way that the discrimination is removed. We will refer to this approach as *Sampling*. Two ways of sampling will be presented and tested.

5.1 Massaging

Algorithm 1: *Learn Classifier on Massaged Data*

Input: Labeled dataset D, sensitive attribute S and value b, desired class $+$
Output: Classifier C, learned on massaged D
1: $(pr, dem) := Rank(D, S, b, +)$
2: $M := \dfrac{disc_{S=b}(D) \times |\{X \in D \mid X(S) = b\}| \times |\{X \in D \mid X(S) = w\}|}{|D|}$
3: Select the top-M of pr
4: Change the class label of the M selected objects to $+$
5: Select the top-M objects of dem
6: Change the class label of the M selected objects to $-$
7: Train a classifier C on the modified D
8: **return** C

In *Massaging*, we will change the labels of some objects X with $X(S) = b$ from $-$ to $+$, and the same number of objects with $X(S) = w$ from $+$ to $-$. In this way the discrimination decreases, yet the overall class distribution is maintained. From the proof of Theorem 1 we know that this strategy reduces the discrimination to the desirable level with the least number of changes to the dataset while keeping the overall class distribution fixed. The set pr of objects X with $X(S) = b$ and $X(Class) = -$ will be called the *promotion candidates* and the set dem of objects X with $X(S) = w$ and $X(Class) = +$ will be called the *demotion candidates*.

We will not randomly pick promotion and demotion candidates to relabel. On the training data, a ranker R for ranking the objects according to their positive class probability is

Algorithm 2: *Rank*

Input: Labeled dataset D, Sensitive attribute and value S, b, desired class $+$
Output: Ordered promotion list pr and demotion list dem
1: Learn a ranker R for prediction $+$ using D as training data
2: $pr := \{X \in D \mid X(S) = b, X(Class) = -\}$
3: $dem := \{X \in D \mid X(S) = w, X(Class) = +\}$
4: Order pr descending w.r.t. the scores by R
5: Order dem ascending w.r.t. the scores by R
6: **return** (pr, dem)

learned. We assume that higher scores indicate a higher chance to be in the positive class. With this ranker, the promotion candidates are sorted according to descending score by R and the demotion candidates according to ascending score. When selecting promotion and demotion candidates, first the top elements will be chosen. In this way, the objects closest to the decision border are selected first to be relabeled, leading to a minimal effect on the accuracy. This modification of the training data is continued until the discrimination becomes zero. The number M of pairs needed to be modified to make a dataset D discrimination-free can be calculated as follows. If we modify M pairs, the resulting discrimination will be:

$$\frac{p_w - M}{|D_w|} - \frac{p_b + M}{|D_b|} = disc(D) - M\left(\frac{1}{|D_b|} + \frac{1}{|D_w|}\right) = disc(D) - \left(M\frac{|D|}{|D_w||D_b|}\right)$$

To reach zero discrimination, we hence have to make:

$$M = \frac{disc(D) \times |D_b| \times |D_w|}{|D|}$$

modifications. Recall that D_b and D_w denote the objects in D with $S = b$ and $S = w$, respectively, and p_b and p_w are the number of positive objects with, respectively, $S = b$ and $S = w$. If the resultant number M is not a whole number, we round it up, which will result a slight negative discrimination. We relabel the M top elements from both the promotion and demotion lists.

Example 2 Consider again the dataset D given in Table 1. We want to learn a classifier to predict the class of objects for which the predictions are non-discriminatory toward $Sex = f$. In this example we rank the objects by their positive class probability given by a *Naïve Bayes* classification model. In Table 2 the positive class probabilities as given by this ranker are added to the table for reference (calculated using the "NBS" classifier of Weka). In the second step, we arrange the data separately for *female* applicant with class $-$ in descending order and for *male* applicants with class $+$ in ascending order with respect to their positive class probability. The ordered promotion and demotion candidates are given in Table 3.

The number M of labels of promotion and demotion candidates we need to change equals:

$$M = \frac{disc(D) \times |D_{female}| \times |D_{male}|}{|D|} = \frac{40\% \times 5 \times 5}{10} = 1$$

So, relabeling one promotion candidate and one demotion candidate makes the data discrimination-free. We hence relabel the highest scoring female with a negative label and the lowest scoring male with a positive label. After the labels for these instances have been changed, the discrimination becomes 0%. The resulting dataset will be used as a training set for classifier induction. □

Table 2 Sample job-application relation with positive class probability

Sex	Ethnicity	Highest degree	Job type	Cl.	Prob (%)
M	Native	H. school	Board	+	98
M	Native	Univ.	Board	+	89
M	Native	H. school	Board	+	98
M	Non-nat.	H. school	Healthcare	+	69
M	Non-nat.	Univ.	Healthcare	−	30
F	Non-nat.	Univ.	Education	−	2
F	Native	H. school	Education	−	40
F	Native	None	Healthcare	+	76
F	Non-nat.	Univ.	Education	−	2
F	Native	H. school	Board	+	93

Table 3 Promotion candidates (negative objects with $Sex = f$ in descending order) and demotion candidates (positive objects with $Sex = m$ in ascending order)

Sex	Ethnicity	Highest degree	Job type	Cl.	Prob (%)
F	Native	H. school	Education	−	40
F	Non-nat.	Univ.	Education	−	2
F	Non-nat.	Univ.	Education	−	2
M	Non-nat.	H. school	Healthcare	+	69
M	Native	Univ.	Board	+	89
M	Native	H. school	Board	+	98
M	Native	H. school	Board	+	98

Algorithm The pseudocode of our algorithm is given in Algorithms 1 and 2. Algorithm 1 describes changing the class labels and classifier learning, and Algorithm 2 the sorting of the promotion and demotion lists.

5.2 Reweighing

The *Massaging* approach is rather intrusive as it changes the labels of the objects. Our second approach does not have this disadvantage. Instead of relabeling the objects, different weights will be attached to them. For example, objects with $X(S) = b$ and $X(Class) = +$ will get higher weights than objects with $X(S) = b$ and $X(Class) = -$ and objects with $X(S) = w$ and $X(Class) = +$ will get lower weights than objects with $X(S) = w$ and $X(Class) = -$. We will refer to this method as *Reweighing*. Again we assume that we want to reduce the discrimination to 0 while maintaining the overall positive class probability. We now discuss the idea behind the weight calculation.

If the dataset D is unbiased, i.e., S and $Class$ are statistically independent, the expected probability $P_{exp}(S = b \wedge Class = +)$ would be:

$$P_{exp}(S = b \wedge Class = +) := \frac{|\{X \in D \mid X(S) = b\}|}{|D|} \times \frac{|\{X \in D \mid X(Class) = +\}|}{|D|} .$$

Algorithm 3: *Reweighing*

Input: $(D, S, Class)$
Output: Classifier learned on reweighed D
1: **for** $s \in \{b, w\}$ **do**
2: **for** $c \in \{-, +\}$ **do**
3: Let $W(s, c) := \frac{|\{X \in D \mid X(S) = s\}| \times |\{X \in D \mid X(Class) = c\}|}{|D| \times |\{X \in D \mid X(Class) = c \text{ and } X(S) = s\}|}$
4: **end for**
5: **end for**
6: $D_W := \{\}$
7: **for** X in D **do**
8: Add $(X, W(X(S), X(Class)))$ to D_W
9: **end for**
10: Train a classifier C on training set D_W, taking onto account the weights
11: **return** Classifier C

In reality, however, the observed probability in D,

$$P_{obs}(S = b \wedge Class = +) := \frac{|\{X \in D \mid X(S) = b \wedge X(Class) = +\}|}{|D|}$$

might be different. If the expected probability is higher than the observed probability value, it shows the bias toward class $-$ for those objects X with $X(S) = b$.

To compensate for the bias, we will assign lower weights to objects that have been deprived or favored. Every object X will be assigned weight:

$$W(X) := \frac{P_{exp}(S = X(S) \wedge Class = X(Class))}{P_{obs}(S = X(S) \wedge Class = X(Class))} ;$$

i.e., the weight of an object will be the expected probability to see an instance with its sensitive attribute value and class given independence, divided by its observed probability.

In this way we assign a weight to every tuple according to its S and $Class$-values. We will call the dataset D with the added weights, D_W. It is easy to see that D_W is unbiased; i.e., if we multiply the frequency of every object by its weight, the discrimination would be 0. On this balanced dataset the discrimination-free classifier is learned.

Example 3 Consider again the dataset in Table 1. The weight for each data object is computed according to its S- and $Class$-value. We calculate the weight of a data object with $X(S) = f$ and $X(Class) +$ as follows. We know that 50% objects have $X(S) = f$ and 60% objects have $Class$-value $+$, so the expected probability of the object should be:

$$P_{exp}(Sex = f \ \wedge \ X(Class) \ = \ +) = 0.5 \times 0.6 = 30\%$$

but its actually observed probability is 20%. So the weight $W(X)$ will be:

$$W(X) = \frac{0.5 \times 0.6}{0.2} = 1.5 \ .$$

Similarly, the weights of all other combinations are as follows:

$$W(X) := \begin{cases} 1.5 & \text{if } X(Sex) = f \text{ and } X(Class) \ = \ + \\ 0.67 & \text{if } X(Sex) = f \text{ and } X(Class) \ = \ - \\ 0.75 & \text{if } X(Sex) = m \text{ and } X(Class) \ = \ + \\ 2 & \text{if } X(Sex) = m \text{ and } X(Class) \ = \ - \end{cases}$$

The weight of each data object of the Table 1 is given in Table 4.

Table 4 Sample job-application relation with weights

Sex	Ethnicity	Highest degree	Job type	Cl.	Weight
M	Native	H. school	Board	+	0.75
M	Native	Univ.	Board	+	0.75
M	Native	H. school	Board	+	0.75
M	Non-nat.	H. school	Healthcare	+	0.75
M	Non-nat.	Univ.	Healthcare	−	2
F	Non-nat.	Univ.	Education	−	0.67
F	Native	H. school	Education	−	0.67
F	Native	None	Healthcare	+	1.5
F	Non-nat.	Univ.	Education	−	0.67
F	Native	H. school	Board	+	1.5

The pseudocode of the algorithm describing our *Reweighing* approach is given in Algorithm 3.

5.3 Sampling

Since not all classifier learners can directly incorporate weights in their learning process, we also propose a *Sampling* approach. The dataset with weights is transformed by sampling the objects with replacement according to their weights.

We partition the dataset into four groups: DP (*D*eprived community with *P*ositive class labels), DN (*D*eprived community with *N*egative class labels), FP (*F*avored community with *P*ositive class labels), and FN (*F*avored community with *N*egative class labels):

$$DP := \{X \in D \mid X(S) = b \wedge X(Class) = +\}$$
$$DN := \{X \in D \mid X(S) = b \wedge X(Class) = -\}$$
$$FP := \{X \in D \mid X(S) = w \wedge X(Class) = +\}$$
$$FN := \{X \in D \mid X(S) = w \wedge X(Class) = -\}.$$

Consider Fig. 2, representing a dataset with 40 data points. The data points in the positive class are represented by $+$, the data points of the negative class by $-$. The projection on the horizontal axis represents the probability of each data object to be in the positive class: the more to the right is the point, the higher its positive class probability. This probability comes, e.g., from a ranker we learned on the training data. This probability will only be of interest for our second sampling method, the preferential sampling, and can for the moment be ignored. The data points plotted in the upper half of the graph, respectively, the lower half, represent the deprived, respectively, the favored community. In the case of discrimination, the relative size of DN versus DP will be larger than the relative size of FN versus FP.

Similar as in *Reweighing*, we compute for each of the groups FN, FP, DP, and DN their expected sizes if the given dataset would have been non-discriminatory, as shown in the following table:

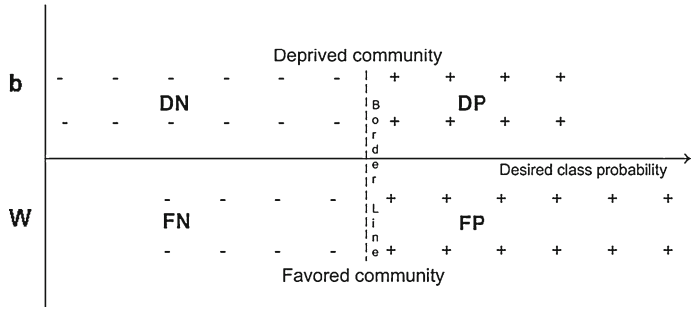


Fig. 2 Pictorial representation of a dataset with 40 datapoints. The points are split into the deprived community (*top*) and the favored community (*bottom*). Their relative position on the *horizontal axis* represents the probability that they belong to the positive class according to the distribution of the data. Since the dataset contains discrimination, the favored objects tend to be more on the right than the deprived objects

Algorithm 4: *Uniform Sampling*

Input: $(D, S, Class)$
Output: Classifier C learned on resampled D
1: **for** $s \in \{b, w\}$ **do**
2: **for** $c \in \{-, +\}$ **do**
3: Let $W(s, c) := \frac{|\{X \in D \mid X(S) = s\}| \times |\{X \in D \mid X(Class) = c\}|}{|D| \times |\{X \in D \mid X(Class) = c \text{ and } X(S) = s\}|}$
4: **end for**
5: **end for**
6: Sample uniformly $W(b, +) \times |DP|$ objects from DP;
7: Sample uniformly $W(w, +) \times |FP|$ objects from FP;
8: Sample uniformly $W(b, -) \times |DN|$ objects from DN;
9: Sample uniformly $W(w, -) \times |FN|$ objects from FN;
10: Let D_{US} be the bag of all samples generated in steps 6 to 9
11: **return** Classifier C learned on D_{US}

Sample size	DP	DN	FP	FN
Actual	8	12	12	8
Expected	10	10	10	10

This time, however, the ratio between the expected group size and the observed group size will not be used as a weight to be added to the individual objects, but instead we will sample each of the groups separately, until its expected group size is reached. For the groups FP and DN this means that they will be under-sampled (the objects in those groups have a weight of less than 1), whereas the other groups FN and DP will be over-sampled.

5.3.1 Uniform sampling

As the name already suggests, in *US* all the data objects of the same group have the same chance of being duplicated or skipped; if we need to sample n objects from a group P, *US* will apply uniform sampling with replacement. In Fig. 3 a possible re-sampling of the dataset is given; the bold elements are duplicated while the encircled objects are removed. Algorithm 4 gives a formal description of the *US* method.

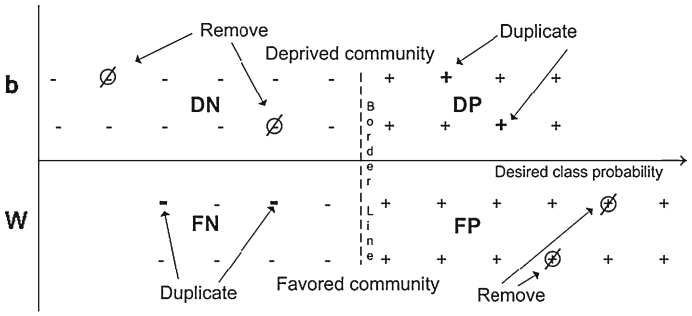


Fig. 3 Pictorial representation of the *Uniform Sampling* scheme. The re-substituted data points are in *bold* while the *encircled* ones are skipped

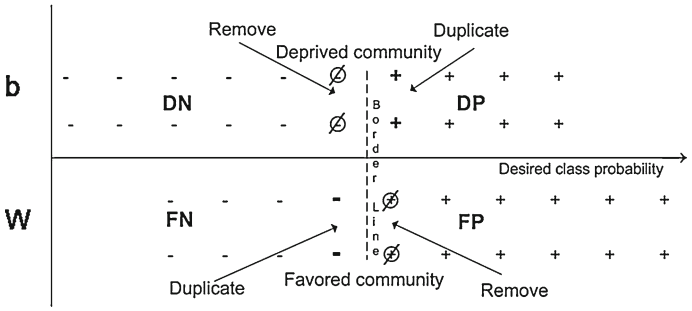


Fig. 4 Pictorial representation of *Preferential Sampling* scheme. The re-substituted data points are in *bold* while the *encircled* ones are skipped

5.3.2 Preferential sampling

In *Preferential Sampling (PS)* we use the idea that data objects close to the decision boundary are more likely to be discriminated or favored due to discrimination in the dataset. To identify the borderline objects, *PS* starts by learning a ranker on the training data. *PS* uses this ranker to sort the data objects of DP and FP in ascending order, and the objects of DN and FN in descending order w.r.t. the positive class probability. Such arrangement of data objects makes sure that the higher up in the ranking an element occurs, the closer it is to the boundary.

PS starts from the original training dataset and iteratively duplicates (for the groups DP and FN) and removes objects (for the groups DN and FP) in the following way:

– Decreasing the size of a group is always done by removing the data objects closest to the boundary; i.e., the top elements in the ranked list.
– Increasing the sample size is done by duplication of the data object closest to the boundary. When an object has been duplicated, it is moved, together with its duplicate, to the bottom of the ranking. We repeat this procedure until the desired number of objects is obtained.

Figure 4 gives the resampling of dataset shown in Fig. 2 to make it discrimination-free by using the preferential sampling method. In most cases, only a few data objects have to be duplicated or removed. The exact algorithm is given in Algorithm 5.

Algorithm 5: *Preferential Sampling*

Input: $(D, S, Class)$
Output: Classifier C learned on resampled D
1: **for** $s \in \{b, w\}$ **do**
2: **for** $c \in \{-, +\}$ **do**
3: Let $W(s, c) := \dfrac{|\{X \in D \mid X(S) = s\}| \times |\{X \in D \mid X(Class) = c\}|}{|D| \times |\{X \in D \mid X(Class) = c \text{ and } X(S) = s\}|}$
4: **end for**
5: **end for**
6: Learn a ranker R for predicting $+$ using D as training set
7: $D_{PS} := \{\}$
8: Add $\lfloor W(b, +) \rfloor$ copies of DP to D_{PS}
9: Add $\lfloor W(b, +) - \lfloor W(b, +) \rfloor \times |DP| \rfloor$ lowest ranked elements of DP to D_{PS}
10: Add $\lfloor W(b, -) \times |DN| \rfloor$ lowest ranked elements of DN to D_{PS}
11: Add $\lfloor W(w, +) \times |FP| \rfloor$ highest ranked elements of FP to D_{PS}
12: Add $\lfloor W(w, -) \rfloor$ copies of FN to D_{PS}
13: Add $\lfloor W(w, -) - \lfloor W(b, -) \rfloor \times |FN| \rfloor$ highest ranked elements of FN to D_{PS}
14: **return** Classifier C learned on D_{PS}

6 Experiments

All preprocessing methods introduced in the paper have been implemented and tested. We compare the following algorithms:

1. The preprocessing techniques introduced in the paper:
   (a) The **Massaging** approach with different rankers. We consider two different rankers: one based on a Naïve Bayes classifier (M_NBS) and one based on a nearest neighbor classifier with 7 neighbors (M_IBk7). These rankers are used to relabel the dataset to make it discrimination-free.
   (b) **Reweighing** (RW) and **Uniform Sampling** (US); these methods are parameter-free as they do not rely on a ranker.
   (c) **Preferential Sampling** (PS) with a Naïve Bayes classifier as ranker.

This gives a total of 5 preprocessing methods to clean away the discrimination of the input data. On the cleaned data, different base classifiers were trained: a Naïve Bayes Classifier (NBS), three nearest neighbor classifiers with, respectively, 1, 3, and 7 neighbors (IBk1, IBk3, and IBk7), and a decision tree learner: the Weka implementation of the C4.5 classifier (J48). This gives a total of $5 \times 5 = 25$ combinations. Many more combinations have been tested (including, e.g., Adaboost and a large variety of rankers for the Massaging approach) but we restricted ourselves to the choices above as they present a good summary of the obtained results; for the other classifiers, similar results were obtained.

2. Two **baseline** approaches:
   (a) An **out-of-the-box classifier** not taking any anti-discrimination measures into account in any way (labeled "No" to reflect no preprocessing was used); we compare to this baseline to see what is the net benefit w.r.t. discrimination-reduction of our proposed methods and how much accuracy we have to trade in for that reduction.
   (b) We **remove the sensitive attribute and its most correlated attributes** before learning ("No_SA" for No Sex Attribute). In this way we get many baseline classifiers, depending on how many of the correlated attributes we remove.

We analyze our proposed algorithms in two scenarios:

1. S is part of the training set, but cannot be used during prediction. In these experiments we only use the information about S for evaluating the discrimination measurement, but S is not considered for prediction. Notice that this setup respects all our assumptions.
2. S is part of the training set and can be used at prediction time. This setup actually violates our assumption (A2) that S should not be used during prediction but has been added for reference.

Datasets. In our experiments we used the **Census Income** and **Communities and Crimes** datasets, available in the UCI ML repository [1]. The **Census Income** dataset has 48,842 instances and contains demographic information of people. The associated prediction task is to determine whether a person makes over 50K per year or not; i.e., income class *High* or *Low* will be predicted. We denote income class *High* as $+$ and income class *Low* as $-$. Each data object is described by 14 attributes, of which 8 are categorical and 6 are numerical attributes. We excluded the attribute *fnlwgt* from our experiments (as suggested in the documentation of the dataset). The other attributes in the dataset include: age, type of work, education, years of education, marital status, occupation, type of relationship (husband, wife, not in family), sex, race, native country, capital gain, capital loss, and weekly working hours. We use *Sex* as discriminatory attribute. In our sample of the dataset, 16,192 citizens have $Sex = f$ and 32,650 have $Sex = m$. The discrimination is 19.45%:

$$P(X(Class) = + \mid X(Sex) = m) - P(X(Class) = + \mid X(Sex) = f) \;=\; 19.45\%$$

The *Communities and Crimes* dataset has 1994 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 100K population. In our experiments we use only predictive attributes which are numeric. We add a sensitive attribute *Black* to divide the communities by thresholding the numerical attribute *racepctblack* at 0.06. We discretized the class attribute to divide the data objects into major and minor violent communities.

We also apply our proposed techniques to the **Dutch census datasets of the year 2001** [11]. The *Dutch Census 2001* dataset has 189,725 instances representing aggregated groups of inhabitants of the Netherlands in 2001. The dataset is described by 13 attributes namely *sex, age, household position, household size, place of previous residence, citizenship, country of birth, education level, economic status (economically active or inactive), current economic activity, marital status, weight, and occupation*. We removed the records of underage people, some middle level professions and people with unknown professions, leaving 60,420 instances for our experiments. We use the attribute *occupation* as a class attribute with values "high level" (prestigious) and "low level" professions. *sex* is the sensitive attribute.

Experimental setup. The goal is to learn a classifier that has minimal discrimination and maintains high accuracy. All reported accuracy numbers in the paper were obtained using **10-fold cross-validation** and reflect the true accuracy; that is, on **unaltered data (no preprocessing is applied)**. Figure 5 shows a detailed representation of our experimental setup. We can observe in Fig. 5 that we apply, in each iteration of the cross-validation, our proposed preprocessing methods only to the folds for training and not to the test fold. We use this preprocessed training set for learning a classifier and evaluate this learnt classifier over the test fold of this iteration. The predictions for the test fold are stored. We repeat this process for all folds and append all predictions on the test sets over all folds. Based on the predictions and the true class we calculate the final accuracy and discrimination scores. It is also important to notice that **no parameter tuning** was performed; our preprocessing methods are parameter-free and all base learners were ran in Weka with their default parameter settings.

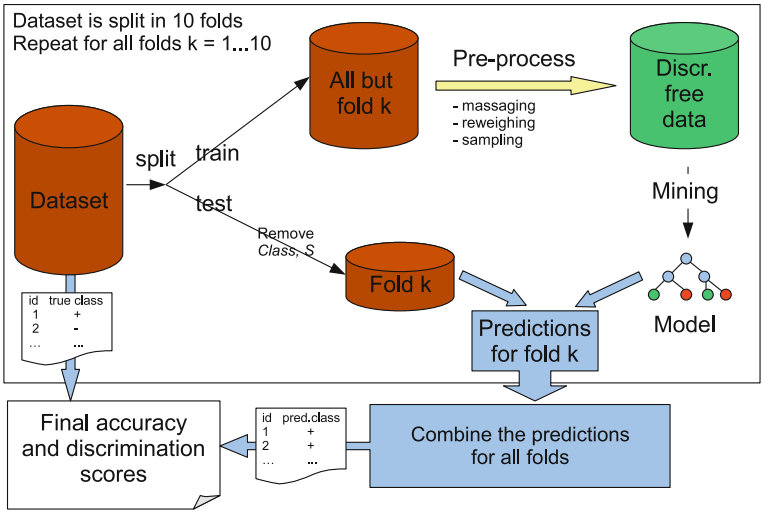


Fig. 5 10-fold cross-validation experimental setup

Table 5 Performance of classifiers trained on discriminatory data; with and without the sensitive attribute

The results clearly confirm the existence of a redlining effect

Dataset	With S (%)	Without S (%)
German credit	11.09	9.32
Census income	16.48	16.65
Communities and crimes	40.14	38.07
Dutch 2001 census	34.91	17.92

All datasets and the source code of all implementations reported upon in this section are available at https://sites.google.com/site/faisalkamiran/. Experiments over the rather small German Credit dataset available in the UCI ML repository have not been included here but can be found in [13].

6.1 Redlining

Our first experiment concerns the redlining effect, i.e., removing the attribute S from the dataset does not always result in the removal of the discrimination, because of indirect discrimination due to other attributes that correlate with S. For all datasets we show in Table 5 the discrimination of a classifier (a decision tree) learned on unaltered training data, with and without the sensitive attribute. The results clearly motivate our work: classifiers learned on biased data produce biased classifiers, even if the sensitive attribute is removed during training.

6.2 Census income dataset

In Fig. 6a,b, respectively, the discrimination and accuracy results for all algorithms under comparison are given. On the X axis are the names of the data preprocessing techniques used to make the training dataset discrimination-free. The resultant discrimination has been given on the Y axis of Fig. 6a and the accuracy on the Y axis of Fig. 6b. We observe that the classi-

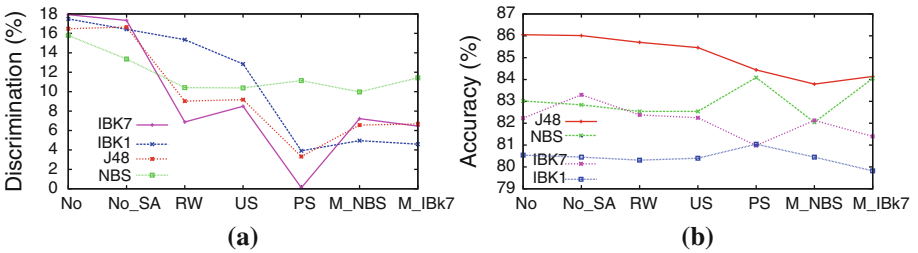


Fig. 6 The results of 10-fold CV for **Census Income dataset** when S is used in the learning phase but not for prediction. **a** Baseline discrimination = 19.45. **b** Baseline accuracy = 76.3

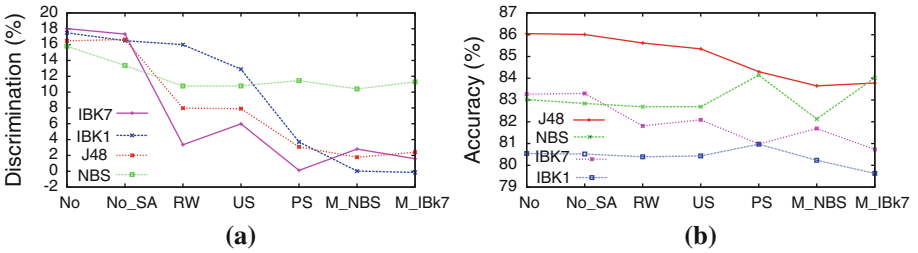


Fig. 7 The results of 10-fold CV for **Census Income dataset** when S is used for both learning and prediction. **a** Baseline discrimination = 19.45. **b** Baseline accuracy = 76.3

fiers learned on the preprocessed data produce less discriminatory results as compared to the baseline algorithms; in Fig. 6a we see that IBk7 classifies the future data objects with 17.93% discrimination which is lowered only slightly if the Sex attribute is removed. If *Preferential Sampling* is applied, however, the discrimination goes down to 0.11%. On the other hand, We observe in Fig. 6b that the loss in accuracy is modest in comparison with the large reduction in discrimination. The discrimination always goes down when we apply our classifiers with non-discrimination constraints, while accuracy remains at a high level. In these experiments, we omitted S from our test datasets.

Figure 7a,b represent the results of the same experiment, except that this time S can be used at prediction time. These two experiments produce very similar results. We observe that the combination of J48 as base learner for *Massaging* produces promising results. *PS* gives excellent results when it is used with unstable classifiers, e.g., J48. When *PS* is used with J48, the discrimination level decreases from 16.48 to 3.32% while the accuracy level decreases from 86.05 to 84.3%. Figure 7b shows the resultant accuracy for all these methods. We find that the *Reweighing* approach maintains a high accuracy level.

Figure 8a,b offer a good overview that allows us to quickly assess which of the combinations are DA-optimal (discrimination-accuracy-optimal) among the classifiers learned in our experiments. Figure 8a represents a graphical representation of the experiments when the attribute *Sex* is not used at prediction time. Figure 8b shows the results of the experiments when *Sex* is used at prediction time. Each pictogram in these figures represents a particular combination of a classification algorithm (shown by the outer symbol) and a preprocessing technique (shown by the inner symbol). For *Massaging*, the inner symbol represents the ranker that was used. On the X axis we see the discrimination and on the Y axis, the accuracy. Thus, we can see the trade-off between accuracy and discrimination for each combination. The closer we are to the top left corner the higher accuracy and the lower discrimination

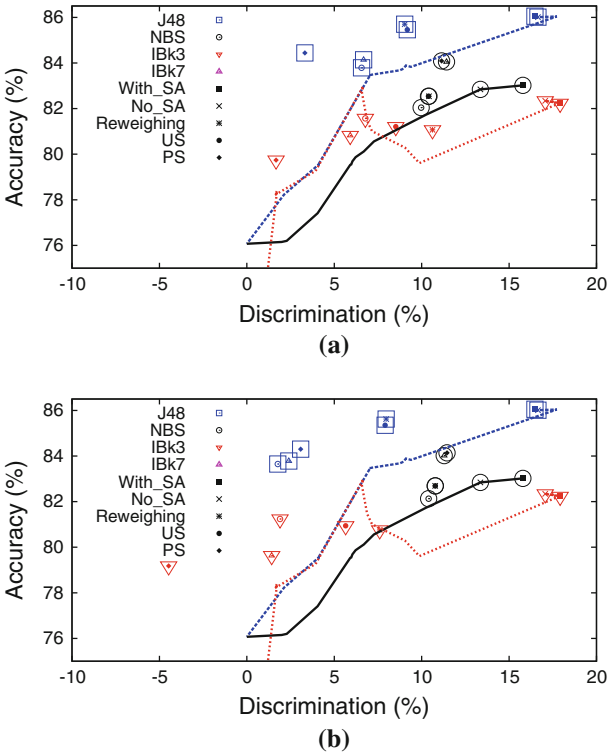


Fig. 8 Accuracy-discrimination trade-off comparison for the **Census Income dataset**. *Outer and inner symbol of each data point shows the corresponding base learner and preprocessing technique, respectively. Three lines* represent the baselines for three classifiers J48, NBS, and IBK3 (*top* to *bottom*). **a** *S* is used in the learning phase but not for prediction. **b** *S* is used for both learning and prediction

we obtain. The three lines in the figure represent the baselines: three classifiers (J48, NBS, and IBk3) learned on the original dataset (the most top-right point in each line, denoted with *With_SA* symbol), the original dataset with the *Sex* attribute removed (denoted with *No_SA* symbol), the original dataset with the *Sex* attribute and the one (two, three, and so on) most correlated attribute(s) removed (that typically correspond to the further decrease in both accuracy and discrimination). We observe that the top left area in the figure is occupied by the data points corresponding to the performance of *Massaging* and *PS* approaches. The *Reweighing* and *US* approaches fall behind *Massaging* but also show reasonable performance. From Fig. 8a,b we can see that our approaches compare favorably to the baselines in the sense that almost all combinations dominate the baseline solutions.

6.3 Other datasets

We repeated all the experiments with the other datasets as well: the Dutch 2001 Census and the Communities and Crimes datasets. The results of these experiments are shown, respectively, in Figs. 9 and 10. We observe that our proposed discrimination-aware classification methods outperform the traditional classification methods and baseline approaches w.r.t. the accuracy-discrimination trade-off; in both datasets the discrimination is considerably lowered from initially around 40%, at the cost of only very little accuracy.

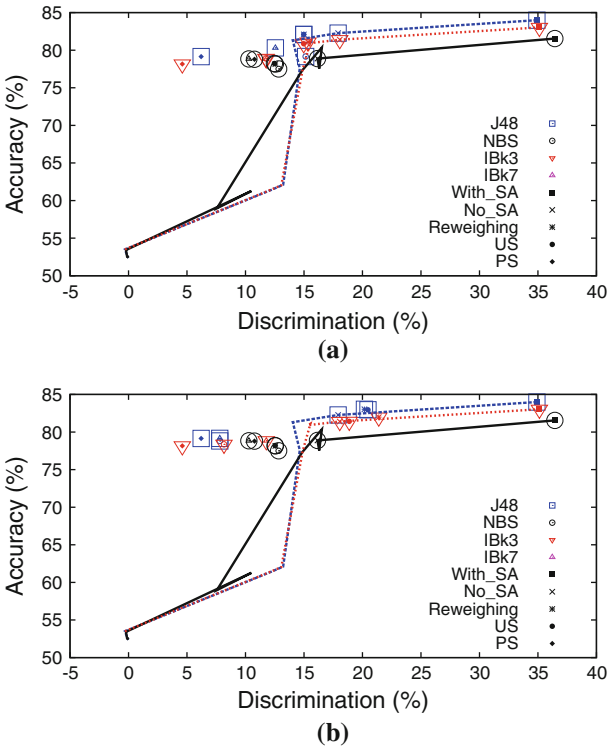
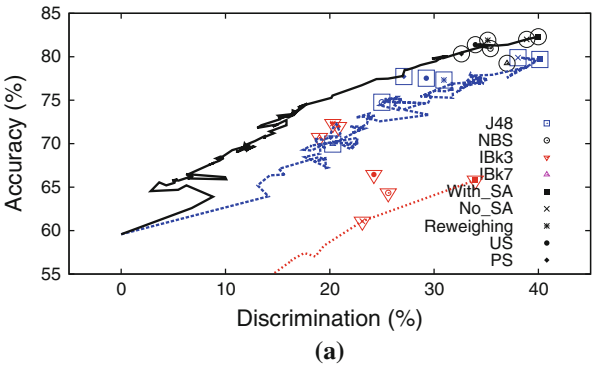


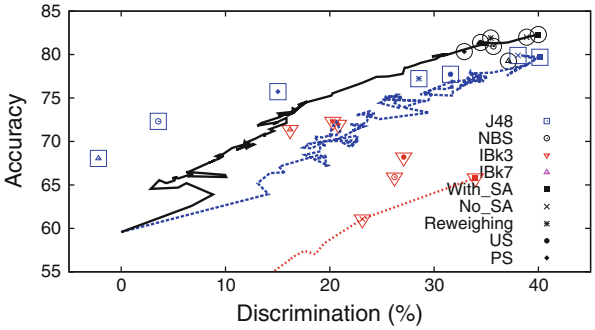
Fig. 9 Accuracy-discrimination trade-off comparison for the **Dutch 2001 Census dataset**. *Outer and inner symbol of each data point shows the corresponding base learner and preprocessing technique, respectively. Three lines* represent the baselines for three classifiers J48, IBK3, and NBS (*top* to *bottom*). **a** *S* is used in the learning phase but not for prediction. **b** *S* is used for both learning and prediction

6.4 How to choose a base classifier for massaging?

From the different experiments it is not clear which base classifier is preferable for the *Massaging* method, although it seems that the effect is better transferred to future classification in case of unstable classifiers such as, e.g., decision trees, in the sense that both the discrimination level and the accuracy go down more than for a stable (noise-resistent) classifier such as, e.g., Naíve Bayes. We conducted additional controlled experiments to further explore this issue. In our controlled experiments, we used a k-nearest neighbor classifier as a base classifier for the *Massaging* method. This classifier has the advantage that we can influence its stability with the parameter k: the higher k, the more stable it becomes. Figure 11 represent the results of the experiments with IBk as base learner and NBS as ranker for the *Massaging* approach. We changed the value of k for IBk from 1 to 19 (only odd values) to change its stability as a base classifier. We observe that the resultant discrimination and accuracy increase both with increasing k. From these controlled experiments, we make the following observation: if minimal discrimination is the first priority, an unstable classifier, i.e., a classifier more sensitive to noise as base learner, is the better option and if high accuracy is the main concern, a stable classifier might be more suitable.



(a)



(b)

Fig. 10 Accuracy-discrimination trade-off comparison over the **Communities and Crimes dataset**. *Outer and inner symbol of each data point shows the corresponding base learner and preprocessing technique, respectively. Three lines* represent the baselines for three classifiers NBS, J48, and IBK3 (*top* to *bottom*). **a** *S* is used in the learning phase but not for prediction. **b** *S* is used for both learning and prediction

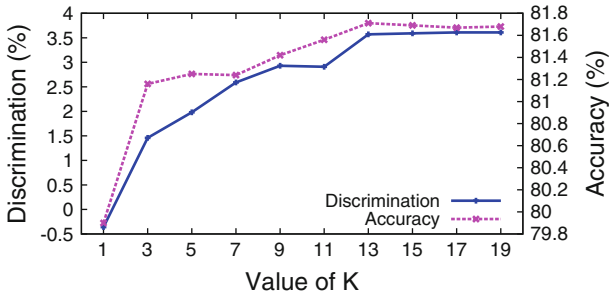


Fig. 11 Accuracy and discrimination comparison with NBS as a ranker and IBk as a base learner with different values of *k*

6.5 Comparison with other, non-preprocessing techniques

For completeness, Fig. 12 gives a comparison of our proposed pre-processing methods with the other current state-of-art methods, i.e., the methods of [15] (labeled DA-Trees in the graph) and [4] (labeled Three-NB in the graph). We only depicted the results for which the discrimination is between 0 and 4%. We can observe that our proposed methods clearly outperform the best result picked from [4] and have comparable performance as the discrimination-aware trees of [15]. Our methods, however, have a much wider applicability as they can be used with any classifiers while the methods of [15] and [4] are restricted to, respectively, decision trees and Naïve Bayes classifiers. If these classifier types are not performing

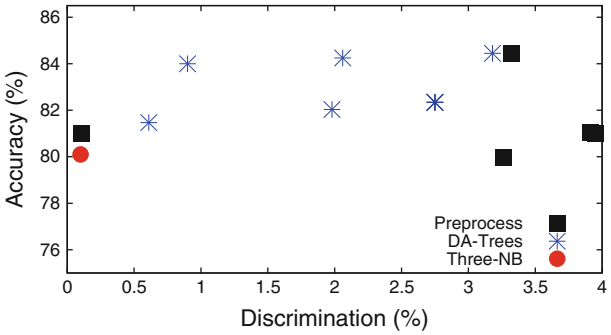


Fig. 12 Accuracy and discrimination comparison of our proposed method to the other current state-of-the-art methods. All methods are evaluated over the Census Income dataset

well on a given dataset, the discrimination-aware methods based upon them will also give poor performance.

6.6 Conclusions of the experiments

From the results of our experiments we draw the following conclusions:

1. Just removing the sensitive attribute from the dataset is not enough to ensure discrimination-aware classification due to the redlining effect.
2. Our proposed preprocessing methods consistently outperform the baseline methods w.r.t. the accuracy-discrimination trade-off.
3. The proposed methods for discrimination-aware classification can be combined with any classifier. The preprocessing is more effective when training unstable classifiers.

7 Related work

Despite the abundance of related works, none of them satisfactorily solves the classification with non-discrimination constraints problem. We consider related work in Discrimination-Aware Data Mining itself, cost-sensitive classification, constraint-based classification, and sampling techniques for unbalanced datasets.

In *Discrimination-Aware Data Mining* two main directions can be distinguished: detection of discrimination [21–24], and the direction followed in this paper, namely learning classifiers if the data are discriminatory [4,15]. A central notion in the 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 focus also on the case where the discriminatory attribute is not present in the dataset and background knowledge for the identification of discriminatory guidelines has to be used. The works on discrimination-aware classification in which our paper falls, however, assume that the discriminatory data are given but the discrimination should be avoided in future predictions. As such our work can be seen as a logical following step after the detection of discrimination. In the current paper, we concentrate on preprocessing techniques after which the normal classifiers can be trained. Another option is to learn classifiers on discriminatory data, and adapt the learning process itself of e.g., decision trees [15] or Bayesian learners [4].

In a recent paper, the authors of [19] propose a variant of k-NN classification for the discovery of discriminated objects. They 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 Massaging technique [13], 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 [19] propose to continue relabeling until all labels are consistent. From a legal point of view, the cleaned dataset obtained by [19] is probably more desirable as it contains less "illegal inconsistencies." For the task of discrimination-aware classification, however, it is unclear if 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 [32] address a similar problem and propose methods to build classifiers when data come from multiple sources (one of the reasons for discrimination, discussed in scenario 2 of Sect. 1). They mainly focus, however, to get high accuracy scores and do not take the discrimination aspect into account.

In *Constraint-Based Classification*, next to a training dataset also some constraints on the model have been given. Only those models that satisfy the constraints are considered in model selection. For example, when learning a decision tree, an upper bound on the number of nodes in the tree can be imposed. Our proposed classification problem with non-discrimination constraints clearly fits into this framework. Most existing works on constraint-based classification, however, impose purely syntactic constraints limiting, e.g., model complexity, or explicitly enforcing the predicted class for certain examples. The difference with our work is that for the syntactic constraints, the satisfaction does not depend on the data itself, but only on the model and most research concentrates on efficiently listing the subset of models that satisfy the constraints. In our case, however, satisfaction of the constraints depends on the data itself and hence requires a different approach. One noteworthy exception is *monotone classification* [10,18]. In monotone classification, next to the normal labeled training data, additionally a function is given for which the predictions should be monotone. An example of such a constraint could be that when assigning a loan based on a number of scores, the assigned label should be monotone in the scores; e.g., if one person gets assigned the loan, and another person scores higher while all other fields are equal to the first person, then the second person should receive the loan as well. Whereas the discrimination criterion is global, the monotonicity criterion is local in the sense that it can be checked by looking at pairs of tuples only. Also, in many cases, the monotonicity can and will be checked syntactically.

In *Cost-Sensitive and Utility-Based learning* [5,12,20,28,31], it is assumed that not all types of prediction errors are equal and not all examples are as important. For example, if the classification task is to predict if an e-mail is spam, the cost of a false positive, i.e., wrongly filtering out a righteous e-mail as spam, is many times higher than the cost of a false negative, i.e., letting through a spam e-mail. The type of error (false positive versus false negative) determines the cost. Sometimes costs can also depend on individual examples. In cost-sensitive learning the goal is no longer to optimize the accuracy of the prediction, but rather the total cost. Our *Reweighing* technique can be seen as an instance of cost-sensitive learning in which, e.g., an object of class $+$ with $X(S) = b$ gets a higher weight and hence an error for this object becomes more expensive. Domingos proposes a method named

MetaCost [9] 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 *Massaging* with respect to relabeling the training data, but *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. These approaches, however, do not guarantee that the desired level of discrimination is reached as again, and they are local.

In *Sampling Techniques for Unbalanced Datasets.* [7], 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. [8] also utilize a wrapper [16] 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. [17] present an innovative approach that augments the minority class by adding synthetic points in distance spaces then use Support Vector Machines for classification. These sampling methods show some similarity with our reweighing and sampling techniques; by increasing the number of samples in one group (the minority class/the deprived community members with a positive label), we try to increase the importance of this group such that the classifier learned on the re-sampled dataset is forced to spend more attention to this group. Making an error on this group will hence be reflected in more severe penalties than in the original dataset, leading to a desired bias toward more easily assigning the minority class label or the positive label to the discriminated group, respectively.

8 Discussion and conclusion

We have presented the classification with non-discrimination constraints problem. Three approaches toward the problem, based upon pre-processing the training dataset, were proposed: *Massaging*, *Reweighing* and *Sampling*. All approaches remove the discrimination from the training data and subsequently a classifier is learned on this unbiased data. Experimental evaluation shows that indeed these preprocessing approaches allow for removing discrimination from the dataset more efficiently than simple methods such as, e.g., removing the sensitive attribute from the training data. All methods have in common that to some extent accuracy must be traded-off for lowering the discrimination. This trade-off was studied and confirmed theoretically.

As future work we are interested in extending the discrimination model itself; in many cases, the non-discrimination constraints as introduced in this paper are too strong: often it is acceptable from an ethical and legal point of view to have a correlation between the gender of a person and the label given to him or her, as long as it can be explained by other attributes. Consider, e.g., the car insurance example: suppose that the number of male drivers involved in two or more accidents in the past is significantly higher than the number of female drivers with two or more accidents. In such a situation it is perfectly acceptable for a car insurance broker to base his or her decisions on the number of previous accidents, even though this will result in a higher number of men than women being denied from getting a car insurance. This discrimination is acceptable because it can be explained by the attribute "Number of car crashes in the past." Similarly, using the attribute "Years of driving experience" may result in acceptable age discrimination. Therefore, it would be interesting to refine our model to *Conditional non-discrimination Constraints*. A promising direction could be to extend the

work [19] 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, e.g., giving them a higher weight.

Furthermore, in this paper, we restricted ourselves to a binary classification problem and one binary sensitive attribute. We can extend our current settings to a multiple class problem by simply assuming one class as the desired class value and the rest of the class values as the not-desired category and vice versa. Nevertheless, often there will be a more subtle gradation in desirability between the classes that need 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; such as, e.g., *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. As a last potential future extension we mention numerical sensitive attributes; e.g., we want the outcome of a university admission procedure to be independent of the gross income of the parents.

In conclusion, this paper only touches the tip of the iceberg. Much remains to be done to extend the solutions to include more sensitive attributes, take into account explanatory attributes, deal with numerical sensitive attributes, etc. We believe discrimination-aware classification is a relevant and interesting research area with many open problems.

Acknowledgments We thank the anonymous reviewers for their insightful comments and the many suggestions that contributed substantially to the improvement of the document.

References

1. Asuncion A, Newman D (2007) UCI machine learning repository
2. Attorney-General's Department C (1984) Australian sex discrimination act 1984. via. http://www.comlaw.gov.au/Details/C2010C00056
3. Calders T, Kamiran F, Pechenizkiy M (2009) Building classifiers with independency constraints. In: IEEE ICDM workshop on domain driven data mining. IEEE press
4. Calders T, Verwer S (2010) Three naive bayes approaches for discrimination- free classification. Data Min Knowl Discov 21(2):277–292
5. 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, pp 164–168
6. Chao EL, Rones PL (2007) Women in the labor force: a databook. US Department of Labor and Bureau of Labor Statistics, Washington, DC
7. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) Smote: synthetic minority over-sampling technique. J Artif Intell Res 16:321–357
8. Chawla NV, Hall LO, Joshi A (2005) Wrapper-based computation and evaluation of sampling methods for imbalanced datasets
9. Domingos P (1999) Metacost: a general method for making classifiers cost-sensitive. In: Proceedings of ACM SIGKDD conference on knowledge discovery and data mining, pp 155–164
10. Duivesteijn W, Feelders A (2008) Nearest neighbour classification with monotonicity constraints. In: Proceedings of ECML/PKDD European conference on machine learning and principles and practice of knowledge discovery in databases. Springer, pp 301–316

11. Dutch Central Bureau for Statistics (2001) Volkstelling. http://easy.dans.knaw.nl/dms
12. Elkan C (2001) The foundations of cost-sensitive learning. In: Proceedings of IJCAI international joint conference on artificial intelligence, pp 973–978
13. Kamiran F, Calders T (2009a) Classifying without discriminating. In: Proceedings of IEEE IC4 international conference on computer, Control & Communication. IEEE press
14. Kamiran F, Calders T (2009b) Discrimination-aware classification. In: BNAIC Benelux conference on artificial intelligence
15. Kamiran F, Calders T, Pechenizkiy M (2010) Constructing decision trees under non-discriminatory constraints. In: Proceedings of IEEE ICDM international conference on data Mining. IEEE press
16. Kohavi R, John GH (1997) Wrappers for feature subset selection. Artif Intell 97(1–2):273–324
17. 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
18. Kotlowski W, Dembczynski K, Greco S, Slowinski R (2007) Statistical model for rough set approach to multicriteria classification. In: Proceedings of ECML/PKDD European conference on machine learning and principles and practice of knowledge discovery in databases. Springer
19. 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
20. Margineantu D, Dietterich T (1999) Learning decision trees for loss minimization in multi-class problems. Technical report. Department of Computer Science, Oregon State University
21. Pedreschi D, Ruggieri S, Turini F (2008) Discrimination-aware data mining. In: Proceedings of ACM SIGKDD conference on knowledge discovery and data mining
22. Pedreschi D, Ruggieri S, Turini F (2009) Measuring discrimination in socially-sensitive decision records. In: Proceedings of SIAM conference on data mining
23. Ruggieri S, Pedreschi D, Turini F (2010a) Dcube: discrimination discovery in databases. In: Proceedings of ACM SIGMOD international conference on management of data, pp 1127–1130
24. Ruggieri S, Pedreschi D, Turini F (2010b) Integrating induction and deduction for finding evidence of discrimination. Artif Intell Law, 1–43
25. The European Court of Justice E (2011) The European court of justice ruling. via. http://ec.europa.eu/ireland/press_office/news_of_the_day/ecj-ruling-sex-discrimination-in-insurance-contracts_en.htm
26. The US department of Justice U (2011) The us federal legislation. via. http://www.justice.gov/crt
27. Turner M, Skidmore F (1999) Mortgage lending discrimination: a review of existing evidence. Urban Institute Monograph Series on Race and Discrimination. Urban Institute Press
28. Turney P (2000) Cost-sensitive learning bibliography. Institute for Information Technology, National Research Council, Ottawa
29. US Department of Justice U (1974) Us equal credit opportunity act. via. http://www.fdic.gov/regulations/laws/rules/6500-1200.html
30. US Empl. Opp. Comm. E (1963) Us equal pay act. via. http://www.eeoc.gov/laws/statutes/epa.cfm
31. Wang B, Japkowicz N (2009) Boosting support vector machines for imbalanced data Sets. Knowl Inf Syst, 1–20
32. Wang H, Wang S (2010) Mining incomplete survey data through classification. Knowl Inf Syst, 1–13

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 October 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. His research interests includes constraints-based classification, privacy preserving, and graph mining.

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.