Chapter 12
Techniques for Discrimination-Free Predictive Models

Faisal Kamiran, Toon Calders, and Mykola Pechenizkiy

Abstract. In this chapter, we give an overview of the techniques developed ourselves for constructing discrimination-free classifiers. In discrimination-free classification the goal is to learn a predictive model that classifies future data objects as accurately as possible, yet the predicted labels should be uncorrelated to a given sensitive attribute. For example, the task could be to learn a gender-neutral model that predicts whether a potential client of a bank has a high income or not. The techniques we developed for discrimination-aware classification can be divided into three categories: (1) removing the discrimination directly from the historical dataset before an off-the-shelf classification technique is applied; (2) changing the learning procedures themselves by restricting the search space to non-discriminatory models; and (3) adjusting the discriminatory models, learnt by off-the-shelf classifiers on discriminatory historical data, in a post-processing phase. Experiments show that even with such a strong constraint as discrimination-freeness, still very accurate models can be learnt. In particular, we study a case of income prediction, where the available historical data exhibits a wage gap between the genders. Due to legal restrictions, however, our predictions should be gender-neutral. The discrimination-aware techniques succeed in significantly reducing gender discrimination without impairing too much the accuracy.

Faisal Kamiran
Lahore Leads University, Pakistan
e-mail: faisal.kamiran@gmail.com

Toon Calders · Mykola Pechenizkiy
Eindhoven University of Technology, The Netherlands
e-mail: {t.calders,m.pechenizkiy}@tue.nl

B. Custers et al. (Eds.): Discrimination & Privacy in the Information Society, SAPERE 3, pp. 223–239.
springerlink.com

12.1 Introduction

Classifier construction is one of the most popular data mining and machine learning techniques (see also Chapter 2 of this book). We assume that a training set in which labels are assigned to the instances is given. The labels indicate the *class* the training examples belong to, and will hence often be called the *class labels*. The training examples are represented by tuples over a set of attributes; that is, every example will be described by values for the same set of attributes. The attribute containing the label will be called the *class attribute*. The label of an example is hence its value for the class attribute. In Table 12.1 an example training set is given. Every example corresponds to a person and is described by the attributes *gender*, *ethnicity*, *highest degree*, *job type*, and the *class* attribute determining whether or not this person belongs to the class of people with a high income (label '+'), or a low income (label '–'). A classifier construction algorithm learns a predictive model for labeling new, unlabeled data. For the given example, a classifier construction algorithm would learn a model for predicting if a person has a high income or not, based upon this person's gender, ethnicity, degree, and job type. Many algorithms for learning various classes of classification models have been proposed during the last decades. The quality of a classifier is measured by its predictive accuracy when classifying previously unseen examples. To assess the accuracy of a classifier, usually a labeled test-set is used; test samples from which the label is removed are classified by the model and the predicted label is compared to the true label.

For the vast majority of these classification techniques maximizing *accuracy* is the only objective; i.e, when the classifier is applied on new data, the percentage of correctly labeled instances should be as high as possible. As explained in detail in Chapter 3 of this book, however, blindly optimizing for high accuracy may lead to undesirable side-effects such as discriminatory classifiers. In this chapter we study the following fictitious case: a bank wants to attract new, preferably rich customers.For this purpose, the dataset of Table 12.1 of its current clients is gathered and labeled according to their income. On the basis of this dataset, a classifier is learnt and applied on the profiles of some prospective clients. If the classifier predicts that the candidate has a high income, a special promotion will be offered to him or her. Such promotional schemes targeting particularly profitable groups are not uncommon in commercial settings. In the dataset of Table 12.1, however, we can clearly observe that the positive label is strongly correlated to males and to the native people. As a result, the promotional scheme will mainly benefit the group of native males, potentially leading to ethical and legal issues. We will use this scenario as a running example.

In this chapter, we concentrate on the very specific case in which the input data for training a classifier can be discriminatory; for instance due to historical discrimination in decision making. And, it is either forbidden by law, or ethically unacceptable, that a classifier learns and applies this discrimination on new instances. We assume that the class label that needs to be predicted can take two values: $+$ and $-$. Furthermore, there is only one sensitive attribute S that can take two values; one for the deprived community (f for "female"), and one for the favored community (m for

"male"). This setting represents the simplest possible of all situations and marks the starting point of the recent discrimination-aware research. For a discussion on more elaborated settings which builds upon this base case, but involves a more complex ecology of attributes, see Chapter 8 of this book.

First we motivate the problem of discrimination-free classification by relating it to existing anti-discrimination laws that prohibit discrimination in housing, employment, financing, insurance, and wages on the basis of race, color, national origin, religion, sex, familial status, and disability (Section 12.2.1). For a more in-depth discussion on anti-discrimination and privacy legislation, we refer the interested reader to Chapter 4 of this book. we give a measure for discrimination on which the problem of *classification without discrimination* will be based (Section 12.2.2). Then, we show how to learn accurate classifiers on discriminatory training data that do not discriminate in their future predictions (Section 12.3). Particularly, we discuss three types of techniques that lead to discrimination-free classifiers. The three classes of techniques and where in the classifier learning process they take place is illustrated in Figure 12.1.

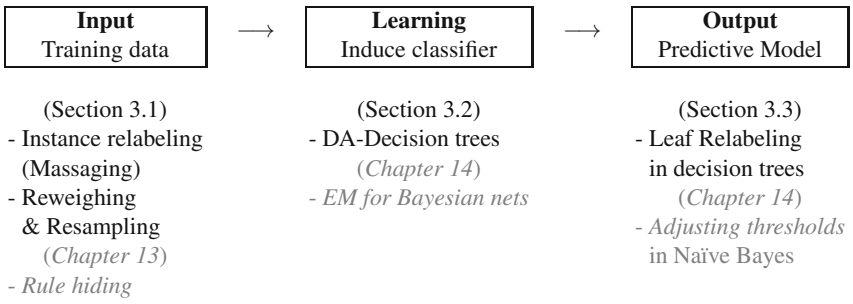

Fig. 12.1 Graphical illustration of the three classes of discrimination-free techniques for classification

The first class of techniques removes the discrimination from the input data, either by selectively relabeling some of the instances (we call this *massaging*); for instance, in the example above, some of the unsuccessful females could be labeled as successful and some of the successful males as unsuccessful, or by resampling the input data; that is, some of the successful males are removed from the input data, and some of the successful females' records get duplicated, or by reweighing, that is assigning higher weights for unsuccessful females and lower weight for successful males(Calders, Kamiran, & Pechenizkiy,2009; Kamiran & Calders, 2009a). Another approach that belongs to this class is described in Chapter 13 of this book; based on a collection of discriminative rules detected by discrimination discovery techniques as described in Chapter 5 of this book, rule hiding techniques from privacy preserving data mining (Chapter 11 of this book) are used to suppress the discriminative rules in the input data.

Table 12.1 Sample relation for the income class example

Sex	Ethnicity	Highest Degree	Job Type	Class
m	native	university	board	+
m	native	high school	board	+
m	native	university	education	+
m	non-native	university	healthcare	+
m	non-native	none	healthcare	-
f	non-native	high school	board	-
f	native	university	education	-
f	native	none	healthcare	+
f	non-native	high school	education	-
f	native	university	board	+

The second class of techniques is based upon the modification of the classifier learning procedure itself. We show how a decision tree learning algorithm can be adapted for inducing discrimination-free predictive models. Technical details of this approach can be found in (Kamiran et al., 2010a). Another approach that belongs to this class, a non-discriminating Bayesian classifier, can be found in Chapter 14 of this book.

The third class of techniques is based upon the post-processing of the learnt models. We explain one decision tree leaves relabeling approach that allows to make an already induced decision tree, with an off-the-shelf approach like C4.5 on biased historical data, discrimination-free (Kamiran et al., 2010b). Another technique in this class, but for Bayesian models is presented in Chapter 14 of this book.

We illustrate the behavior of these different types of techniques in Section 12.4 using the well-known *Adult* dataset (Frank & Asuncion, 2010). The goal associated with this dataset is to predict, for promotional purposes, whether a person falls into the high or the low income class. The dataset, however, exhibits a significant gender-gap with respect to income; there are substantially less females with a high income than males. Nevertheless, as sketched in the example above, we want to learn a classifier which is gender-neutral. The sensitive attribute is thus gender, and the deprived community are the females, the favored community – the males. For the discussed techniques, we show that they clearly outperform the traditional classification approaches for this task; without trading in too much accuracy, the discrimination in the learnt classifier's predictions is reduced to an acceptable level.

12.2 Problem Statement: Discrimination-Aware Classification

The input to our problem consists of a dataset in tabular format, such as the one in Table 12.1. Every row in the table represents one instance, and there is a special column *Class*, indicating the class label that we need to learn to predict for new instances. Based upon the dataset it is expected that a model is learnt that can predict the class based upon the other attributes of a previously unseen instance. Further-

more, in the discrimination-aware paradigm, we assume that a *sensitive attribute*, here "sex" and a *sensitive attribute value*, in this case "female" are set to indicate a subset of the instances which should not be discriminated against. The goal is now to learn a predictive model that will classify future instances as accurately as possible into the high or low income class, under the constraint that the predictions should not be discriminative with respect to the sensitive attribute sex.

In the example dataset of Table 12.1, we can see that 4 out of 5 males have the positive class label, whereas for the females, only 2 out of 5 have the positive class label. Nevertheless, our classifier should learn a predictive model which will, overall, assign to the same proportion of males and females the positive class. Notice that in the problem statement we do not consider the potential existence of other attributes that can explain (part of) the discrimination. For a discussion on explanatory attributes and how they influence the problem we refer to Chapter 8 of this book. In this chapter we concentrate only on the case in which none of the other attributes can be used to justify the discrimination.

Before the formal definition of the discrimination-free classification we give a discussion of anti-discrimination legislation followed by an explanation of how the discrimination should be measured.

12.2.1 Motivation: Links to 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 etc. For instance, the Australian Sex Discrimination Act 1984 (Australian Law, 1984) prohibits discrimination in work, education, services, accommodation, land, clubs on the grounds of marital status, pregnancy or potential pregnancy, and family responsibilities. The US Equal Credit Opportunity Act 1974 (US Legislation, 1968) 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. Similarly there are many other laws which prohibit discriminatory practices. Our discrimination-aware classification paradigm clearly applies to these situations. If we are interested to apply classification techniques and our available *historical* data contains discrimination, it will be illegal to use traditional classifiers without taking the discrimination aspect into account.

The problem of classification with non-discrimination constraints is not a trivial one. The straightforward solution of removing the sensitive attribute from the training-set does in most cases not solve this problem at all. Consider, for example, the German Credit Dataset available in the UCI ML-repository (Frank & Asuncion, 2010). This dataset contains demographic information of people applying for loans and the outcome of the scoring procedure. The rating in this dataset correlates with the age of the applicant. Removing the *age* attribute from the data, however, does

not remove the age-discrimination, as many other attributes such as, *own_house*, indicating if the applicant is a home-owner, turn out to be good predictors for *age*. A parallel can be drawn with the practice of *redlining*: denying inhabitants of particular racially determined areas from services such as loans. It describes the now abolished practice of marking a red line on a map to delineate the area where banks would not invest; later the term was used for indirect discrimination against a particular group of people (usually by race or sex) no matter the geography[1].

12.2.2 Measuring Discrimination

There are many different ways in which discrimination could be quantified, and each of them has its own advantages and disadvantages. Here, in this chapter, and in our earlier works (Calders et al., 2009; Kamiran & Calders, 2010; Kamiran et al., 2010b; Kamiran & Calders, 2009a,b; Kamiran et al., 2010a; Calders & Verwer, 2010), we define the level of discrimination in a dataset as the difference between the probability that someone from the favored group gets a positive class and the probability that someone from the deprived community gets a positive class. For alternative measures of discrimination, see Chapters 5 and 6 of this book.

For the running example of Table 12.1, the discrimination with respect to the deprived community *Sex=female* is 4/5 - 2/5 = 40%. Formally, for a *sensitive attribute S*, *deprived community (sensitive attribute value) f*, *favored community m*, the discrimination in D with respect to the group $S = f$, denoted $disc_{S=f}(D)$, is defined as:

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

When measuring the discrimination of a classifier, we want to assess how the classifier will act on new, previously unseen examples. We assume a setting in which one example comes at a time, and the classifier needs to assign a label to them immediately. In order to assess the level of discrimination of the classifier when it would be applied to unseen examples, we use a test-set; that is, following standard machine learning practice, before learning a classifier, we split the dataset in two parts; one for learning the classifier, and one for measuring its quality. The examples of the test-set (with their labels removed) are passed one by one to the classifier and its decisions are recorded. After that, the discrimination of the classifier can be assessed as follows. The discrimination of the classifier C with respect to the group $S = f$ on a test dataset D_{test}, denoted $disc_{S=f}(C, D_{test})$, is defined as:

[1] Source: `http://en.wikipedia.org/wiki/redlining`, November 17th, 2011.

$$disc_{S=f}(C, D_{test}) := \frac{|\{X \in D_{test} \mid X(S) = m, C(X) = +\}|}{|\{X \in D_{test} \mid X(S) = m\}|} - \frac{|\{X \in D_{test} \mid X(S) = f, C(X) = +\}|}{|\{X \in D_{test} \mid X(S) = f\}|} .$$

12.3 Techniques for Discrimination-Free Classification

In this section we discuss different techniques for discrimination-aware classification. First, we discuss data pre-processing techniques to make the training data unbiased before learning a classifier. Second, we discuss the adaptation of a classifier learning procedure itself to make it discrimination-free. Third, we discuss the modification of the post-processing phase of a learnt classifier to make it unbiased.

12.3.1 Pre-processing Techniques

The first kind of solutions are based on removing the discrimination from the training dataset. If we can remove discrimination directly from the source data, a classifier can be learnt on a cleaned, discrimination-free 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, as the classifier will no longer generalize the discrimination. The first approach we discuss here is called *massaging* the data (Kamiran & Calders, 2009a). It is based on changing the class labels in order to remove the discrimination from the training data. The second approach is less intrusive as it does not change the class labels in the training data. Instead, weights are assigned to the data objects to make the dataset discrimination-free. This approach is called *reweighing* (Calders et al., 2009). Since reweighing requires the learner to be able to work with weighted tuples, we propose another variant, in which we re-sample the dataset in such a way that the discrimination is removed. We refer to this approach as *Sampling* (Kamiran & Calders, 2010).

12.3.1.1 Massaging

In *massaging* we change the class labels in the training set; some objects of the deprived community change from class $-$ to $+$, and the same number of objects of the favored community change from $+$ to $-$. In this way the discrimination decreases, yet the overall class distribution is maintained; the same number of people has the positive class as before. 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. Notice that we do not randomly pick the objects to relabel. Instead, first we learn a regular, possibly discriminative (i.e. not discrimination-free) classifier. This classifier, although not acceptable as a final result, still provides useful information. Based on this classifier we can see, for the deprived and favored communities separately, which instances are closest to the *decision boundary*. Many classifiers assign a probability of being in the positive class to the instances, and if

Table 12.2 Sample relation for the income class example with positive class probability

Sex	Ethnicity	Highest Degree	Job Type	Class	Prob
m	native	university	board	+	.99
m	native	high school	board	+	.90
m	native	university	education	+	.92
m	**non-native**	**university**	**healthcare**	**+**	**.76**
m	non-native	none	healthcare	-	.44
f	non-native	high school	board	-	.09
f	**native**	**university**	**education**	**-**	**.66**
f	native	none	healthcare	+	.66
f	non-native	high school	education	-	.02
f	native	university	board	+	.92

this probability exceeds 0.5, the object is assigned to the positive class. The objects close to the decision boundary are those with a probability close to 0.5. We select these objects first to relabel.

Example 1. *Consider again the dataset D given in Table 12.1. We want to learn a classifier to predict the class of objects for which the predictions are non-discriminatory towards Sex* $= f$. *In this example we rank the objects by their positive class probability given by a* Naive Bayes *classification model. In Table 12.2 the positive class probabilities as given by this ranker are added to the table for reference (calculated using the "NBS" classifier of Weka (Hall et al., 2009)).*

In the second step, we arrange the data separately for female *applicants with class* $-$ *in descending order and for* male *applicants with class* $+$ *in ascending order with respect to their positive class probability. Relabeling one promotion candidate and one demotion candidate makes the data discrimination-free. Hence, we relabel the top promotion candidate; that is, the highest scoring female with a negative class label, and the top demotion candidate; that is, the lowest scoring male with a positive class label (the bold examples in Table 12.2). After the labels for these instances are changed, the discrimination decreases from* 40% *to* 0%. *The resulting dataset is used as a training set for classifier induction.*

12.3.1.2 Reweighing and Resampling

The *massaging* approach is rather intrusive as it changes the class labels of the objects. Our second approach does not have this disadvantage. Instead of relabeling the objects, different weights are attached to them. For example, the deprived community objects with $X(Class) = +$ get higher weights than the deprived community objects with $X(Class) = -$ and the favored community objects with $X(Class) = +$ get lower weights than the favored community objects with $X(Class) = -$. We refer to this method as *massaging*. 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 would have been unbiased; that is, S and *Class* were statistically independent, the expected probability of being non-native and having the positive class $P_{exp}(f \wedge +)$ would be:

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

For instance in the example dataset of Table 12.1, 50% of people are female, and 60% of people have a positive class. Therefore, if the dataset was non-discriminatory, one would expect also 60% of females to have the positive class, which gives in total $50\% \times 60\% = 30\%$ of people being female and having the positive class. In reality, however, the observed probability in D,

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

might be different. If the expected probability is higher than the observed probability value, it shows the bias towards class '$-$' for those objects X with $X(S) = f$. Continuing the example, in the dataset of Table 12.1, we observe that only 2 people in the dataset are female and have a positive class label, so the observed probability of female and positive is 20%, which is considerably lower than the expected 30%, thus indicating discrimination.

To compensate for the bias, we assign weights to objects. If a particular group is under-represented, we give members of this group a higher weight, making them more important in the classifier training process. The weight we assign to an object is exactly the expected probability divided by the observed probability. In the example this would mean that we assign a weight of 30% divided by 20% = 1.5 to females with a positive class label. In this way we assign a weight to every object according to its S- and *Class*-values. We call the dataset D with the added weights, D_W. It can be proven that the resulting dataset D_W is unbiased; that is, if we multiply the frequency of every object by its weight, the discrimination is 0. On this balanced dataset the discrimination-free classifier is learnt.

Since not every classification algorithm can directly work with weights, we may also use the weights when resampling the dataset; that is, we randomly select objects from our training set to form a new dataset. When forming the new dataset, some objects may be omitted and some may be duplicated. In the sampling procedure, the weight of an object represents its relative chance of being chosen from the dataset; that is, an object with a weight of 2.4 in every selection step has a 4 times higher probability of being chosen than an object with a weight of 0.6. This variant is called *resampling*.

Example 2. *Consider again the dataset in Table 12.1. The weight for each data object is computed according to its S- and Class-value, e.g. for instances with values* $X(Sex) = f$ *and* $X(Class) = +$ *:*

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

Similarly the weights of all other combinations is as follows:

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

12.3.1.3 Related Approaches

The authors of (Luong et al., 2011) propose a variant of k-NN classification for the discovery of discriminated objects. They consider a data object as discriminated if there exists a significant difference of treatment among its neighbors belonging to the deprived community and its neighbors not belonging to it (that is, 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 (Kamiran & Calders, 2009a), 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 (Luong et al., 2011) propose to continue relabeling until all labels are consistent. From a legal point of view, the cleaned dataset obtained by (Luong et al., 2011) 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 authors of (Hajian, Domingo-Ferrer, & Martinez-Balleste, 2011; Hajian, Domingo-Ferrer, & Martínez-Ballesté, 2011) also propose methods similar to massaging to preprocess the training data in such a way that only potentially non-discriminatory rules can be extracted. For this purpose they modify all the items in a given dataset that lead to the discriminatory classification rules by applying rule hiding techniques on either given, or discovered discriminative rules. For an extensive description of this technique, see Chapter 13 of this book.

12.3.2 Changing the Learning Algorithms

In this section, we discuss the discrimination-aware techniques in which we modify the classification model learning process itself to produce discrimination-free classifiers. For this purpose, we discuss the discrimination-aware decision trees construction in which we modify the decision tree construction procedure to make them discrimination-free.

12.3.2.1 Discrimination-Aware Decision Tree Induction

Traditionally, when constructing a decision tree (Quinlan, 1993), we iteratively refine a tree by splitting its leaves until a desired objective is achieved. Consider the dataset given in Table 12.1. Suppose we want to learn a tree over this dataset in

Table 12.3 Gini Index for different possible splits of the data from Table 12.2

Condition	left branch		right branch		Gini Index
	# pos	# neg	# pos	# neg	
sex=m	4	1	3	2	0.4
ethnicity=native	5	1	1	3	0.32
diploma=none	1	1	5	3	0.48
…	…	…	…	…	…

order to predict the *Class*. Initially, we start with a tree consisting of only one node, predicting the majority class '+'. Then, iteratively, we refine the tree by considering all possible splitting criteria, and evaluating which split is the best. Selecting the best split is done by observing how the split condition separates the positive class from the negative class. A split that is better at separating the classes will score higher on the quality measure. For the dataset of Table 12.1, the different splits are as follows: The split $sex = m$ would divide the dataset into those instances that satisfy the condition (the left branch), including 4 positive and 1 negative instance, and those instances that do not satisfy the condition (the right branch), having 3 negative and 2 positive examples. Based on these figures, a degree of impurity can be computed, in this case, based upon the Gini index (Lerman & Yitzhaki, 1984): to compute the Gini-index of a split, we first separate the dataset according to the split criterion. For each partition, the relative frequencies of the positive and negative class, f_+ and f_- respectively, are counted. The Gini-index is then the weighted average of the Gini-score $1-(f_+^2+f_-^2)$. If a partition is pure, this implies that either $f_+ = 1$ and $f_-^2 = 0$, or $f_+ = 0$ and $f_-^2 = 1$. In both cases, the partition contributes $1-(f_+^2+f_-^2) = 0$ to the gini-score of the split. The contribution of a partition is the highest if it is maximally impure; i.e., $f_+ = f_-^2 = 0.5$. For the example split $sex = m$, the partition containing the males contributes $1-((1/5)^2+(4/5)^2) = 8/25$, while the partition with the females contributes $1-((2/5)^2+(3/5)^2) = 12/25$. The Gini-index for the split is now the weighted average over the two partitions, being: $0.5(8/25)+0.5(12/25) = 10/25 = 0.4$.

The better the split separates positive from negative, the lower the impurity. From all splits the one with the lowest impurity is selected. The dataset is split in two parts, according to the splitting criterion and the procedure continues on both parts until a stopping condition is met. In (Kamiran et al., 2010b, 2010a) we show how the splitting criterion can be changed in such a way that not only the impurity with respect to the class label can be incorporated, but also the level of discrimination introduced by the split. In particular, we do not only compute how good the split predicts the class label, but also how good it predicts the sensitive attribute, using the same gini-index, but now with the relative frequencies of the deprived and favored communities in the partitions of the split. The good split will then be the one that achieves a high purity with respect to the class label, but a low purity with respect to the sensitive attribute. In the running example this means that we want splits that are good for distinguishing high income from low income people, without separating

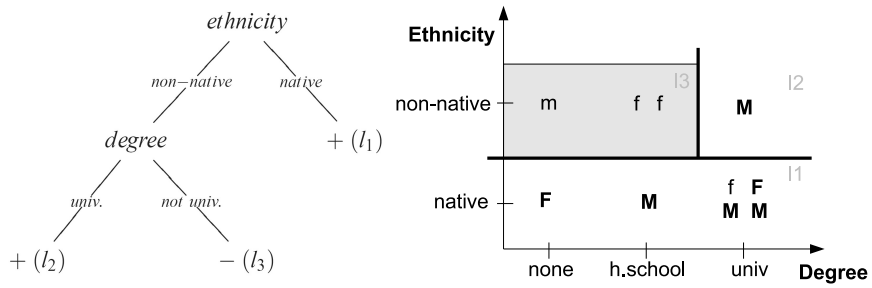

Fig. 12.2 Decision tree with the partitioning induced by it. The bold capital letters in the partitioning denote the positive examples, the lowercase letters the negative examples. m/M denotes a male, f/F denotes a female. The grey background denotes regions where the majority class is −. The discrimination of the tree is 20%.

too much the males from the females. In that way we can guide the iterative tree refinement procedure, disallowing steps that would increase discrimination in the predictions or explicitly adding a penalty term for increasing discrimination into the quality scores of the splits.

12.3.2.2 Related Approaches

Also for other learning algorithms a similar approach could be applied by embedding the anti-discrimination constraints deeply into the learning algorithm. Another example of such an approach is described in Chapter 14 of this book, where a Naïve Bayes model is learnt which explicitly models the effect of the discrimination. By learning the most probable model that leads to the observed data, under the assumption that discrimination took place, one can reverse-engineer the effect of the discrimination and hence filter it out when making predictions.

12.3.3 Post-Processing the Induced Models

Our third and last type of discrimination-aware techniques is based upon the modification of the post-processing phase of the learnt model. We discuss the decision tree leaf relabeling approach of (Kamiran et al., 2010b) where we assume that a tree is already given and the goal is to reduce the discrimination of the tree by changing the class labels of some of the leaves.

12.3.3.1 Decision Tree Leaf Relabeling

The rationale behind this approach is as follows. A decision tree partitions the space of instances into non-overlapping regions. See, for example, Figure 12.2. In this figure (left) a fictitious decision tree with 3 leaves is given, labeled l_1 to l_3. The right

part of the figure shows the partitioning induced by the decision tree. For example, the third leaf in the tree corresponds to all non-native people without a university diploma. The leaves can hence be seen as non-overlapping "profiles" dividing up the space of all instances. Every example fits exactly one profile, and with every profile exactly one class is associated. When a new example needs to be classified by a decision tree, it is given the majority class label of the region/profile it falls into. If some of the profiles are very homogeneous with respect to the sensitive attribute; for instance, containing only members of the deprived community, then this may lead to discriminative predictions. In l_3, for instance, two thirds of the instances are from the deprived community. The relabeling technique now consists of changing the labels for those regions where this results in the highest reduction in discrimination while trading in as little accuracy as possible. Conceptually this method corresponds to merging neighboring regions to form larger, less discriminative profiles. The process of relabeling continues until the discrimination is removed.

Example 3. *Consider the example decision tree given in Figure 12.2. The discrimination of the decision tree is* 20%. *Suppose we want to reduce the discrimination to* 5%. *For each of the leaves it is given how much the discrimination changes (*Δ*disc) when relabeling the node, and how much the accuracy decreases (*Δ*acc). The node for which the tradeoff between discrimination reduction versus lowered accuracy is most beneficial, is selected first for relabeling.*

Node	Δacc	$\Delta disc$	$\frac{\Delta disc}{\Delta acc}$
l_1	-40%	0%	0
l_2	-10%	10%	1
l_3	-30%	10%	$1/3$

In this particular case, the reduction algorithm hence pick l_2 *to relabel; that is, the split on degree is removed and leaves* l_2 *and* l_3 *are merged.*

12.3.3.2 Related Approaches

The idea of model correction has been explored in different settings, particularly in cost-sensitive learning, learning from imbalanced data, and context sensitive or context-aware learning. Concrete examples of model correction include Naive Bayes prior correction (also in Chapter 14 of this book) and posterior probabilities correction based on a confusion matrix (Morris & Misra, 2002); nearest neighbor based classification or identification correction based on current context, e.g. in driver-route identification (Mazhelis, Zliobaite, & Pechenizkiy, 2011) or in context-sensitive correction of phone recognition output (Levit, Alshawi, Gorin, & Nöth, 2003). The tree node relabeling ideas have been used in recognizing textual entailments (Heilman & Smith, 2010) and probabilistic context-free grammar parsing (Johnson, 1998). But these are not related to the idea of decision tree learning. However, we are not aware of other approaches directly related to the discussed idea of leaf relabeling in decision trees applicable to our settings.

12.4 Experiments

The different techniques discussed in this chapter have been experimented with extensively. We refer the interested reader for the detailed discussion of the experimental studies and results to (Kamiran et al., 2010b,a; Kamiran & Calders, 2012; Kamiran, 2011). In this section we give an overview of the most important empirical results for the *Adult* dataset. This 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; that is, income class *High* or *Low* has to be predicted. 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 consider *Sex* as sensitive attribute. In our sample of the dataset, 16 192 citizens have $Sex = f$ and 32 650 have $Sex = m$. The discrimination with respect to $Sex = m$ in the historical data is 19.45%: $P(X(Class) = + \mid X(Sex) = m) - P(X(Class) = + \mid X(Sex) = f) = 19.45\%$. The goal is to learn a classifier that has minimal discrimination and maintains high accuracy.

Figure 12.3 shows the result of experiments when we learn decision trees after applying our proposed discrimination-aware preprocessing techniques on the training data (label 'Preprocessing'), with discrimination-aware splitting criteria (label 'Learner-adaptation'), with leaf relabeling (label 'Postprocessing'), a Naïve Bayes model of Chapter 14 of this book (label '3-NaiveBayes') and learnt without any discrimination-aware technique (label 'Zero-treatment'). We observe in Figure 12.3 that the discrimination-aware techniques discussed in this chapter reduce the discrimination significantly while maintaining a high accuracy as compared to the ordinary methods. For instance, a traditional decision tree without using any discrimination removal method classifies the future data objects with 16.65%

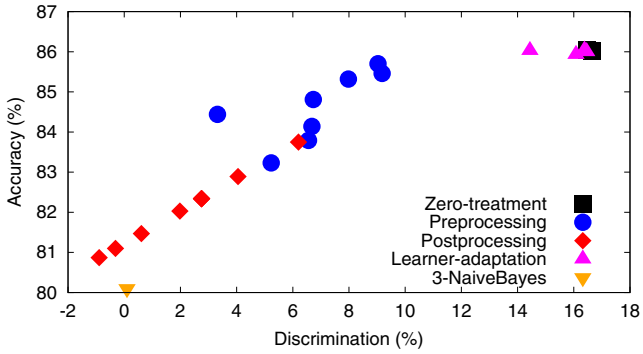

Fig. 12.3 Comparison of techniques discussed in Section 12.3.1 (label Preprocessing), Section 12.3.2 (label Learner-adaptation), Section 12.3.3 (label Postprocessing), Naïve Bayes model of Chapter 14 (label 3-NaiveBayes), and ordinary methods (label Zero-treatment) over the Adult dataset.

discrimination and 86.01% accuracy even though the sensitive attribute was not used at the prediction time. We observe in our experiments that learning a decision tree with modified splitting criterion, that is, using the second type of discrimination-aware classification alone does not significantly reduce the discrimination. However, when the decision trees are learnt on cleaner data obtained with discrimination-aware pre-processing techniques, the discrimination is reduced to 3.32% while keeping the accuracy at 84.44%. The decision trees with leaf relabeling were able in our experiment to reduce the discrimination to 0% while keeping a reasonably high accuracy. Figure 12.3 also shows that our proposed methods outperform the discrimination-aware Naïve Bayes model of Chapter 14 of this book with respect to the accuracy-discrimination trade-off.

12.5 Discussion and Conclusion

In this chapter we discussed the idea of discrimination-aware classification and introduced a procedural way to calculate the discrimination in a given dataset and in the predictions of a classifier. We also discussed three types of techniques to learn the discrimination-free classifiers which include data preprocessing techniques, an adapted classifier learning procedure and an approach for postprocessing of a learnt decision tree by changing the labels of some of its leaves to make the final predictive model discrimination-free. Finally, we presented empirical validation results showing that the discrimination-aware classification methods predict labels for the previously unseen data objects with no or significantly lower discrimination and with the minimal loss of accuracy.

Depending on the situation one of the proposed techniques may be better than another. First of all, if none of the other attributes is correlated to the sensitive attribute, clearly it suffices to just remove this attribute. Unfortunately this is seldomly the case, and even if it is the case, no guarantees can be given that no such correlations exist. The presented preprocessing techniques have the advantage that they make input data discrimination-free which can then be used by any classification algorithm, yet have the disadvantage of giving no guarantee about the degree of discrimination in the final classifier. The model post-processing techniques do not have this disadvantage; in principle the postprocessing is continued until a discrimination-free classifier (on a validation set) is obtained. The model post-processing techniques as well as the learner adaptation techniques on their turn, however, have the disadvantage of being model and even algorithm specific; for every classifier new algorithms will have to be invented. In the experiments it was further shown that the learner adaptation approach did not work as expected, unless it was combined with the post-processing techniques. This surprising failure calls for more research to better understand the reasons for it.

Despite of showing some promising results on discrimination-free classifier construction, our study is far from complete. For instance, often there is a much more complex ecology of attributes than what is assumed in the chapter. In the chapter

we assume there is just one sensitive attribute, dividing the objects into one disadvantaged and one advantaged group. Often, however, there may be more than two groups, each of which are advantaged/disadvantaged to a different level. Consider, e.g., different ethnic minorities being treated in different ways. Furthermore, there may be multiple of such sensitive attributes; e.g., gender, age, and ethnicity. Removing gender-discrimination by the preprocessing techniques may introduce an age-discrimination. Furthermore, it could be the case that even if discrimination does not manifest itself at the general level, in some specialized niches or contexts, there might be discrimination present. Chapter 5 of this book deals with the detection of such subtle contexts for discrimination. Also, as discussed in Chapter 8 of this book, not all difference in acceptance rates between an advantaged and a disadvantaged group is due to discrimination. If people in the disadvantaged group are more likely to be lowly educated, as a result their salaries will be lower on average, without this difference necessarily indicating a discrimination. As a conclusion, the area of discrimination-aware classification remains a rich source of inspiration and application area for novel techniques in the data mining area, and we hope to see significant contributions in future to this ethically and societally important research area, leading towards providing companies and practitioners with the necessary toolkit for data-driven discrimination-free decision making.

References

Australian Law. Australian Sex Discimination Act 1984. Australian sex discimination act 1984. via: (1984) http://www.comlaw.gov.au/Details/C2010C00056

Calders, T., Kamiran, F., Pechenizkiy, M.: Building Classifiers with Independency ConstraintsBuilding classifiers with independency constraints. In: Saygin, Y., et al. (eds.) ICDM Workshops 2009, IEEE International Conference on Data Mining Workshops, Miami, Florida, USA, December 6, pp. 13–18 (2009)

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

Frank, A., Asuncion, A.: UCI Machine Learning Repository. UCI machine learning repository (2010), http://archive.ics.uci.edu/ml

Hajian, S., Domingo-Ferrer, J., Martinez-Balleste, A.: Discrimination prevention in data mining for intrusion and crime detectionDiscrimination prevention in data mining for intrusion and crime detection. In: IEEE Symposium on Computational Intelligence in Cyber Security (CICS)IEEE Symposium on Computational Intelligence in Cyber Security (CICS), pp. 47–54 (2011)

Hajian, S., Domingo-Ferrer, J., Martínez-Ballesté, A.: Rule protection for indirect discrimination prevention in data miningRule protection for indirect discrimination prevention in data mining. In: Modeling Decision for Artificial Intelligence, pp. 211–222 (2011)

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.: The WEKA data mining software: An updateThe WEKA data mining software: An update. ACM SIGKDD Explorations Newsletter 11, 110–118 (2009)

Heilman, M., Smith, N.A.: Tree edit models for recognizing textual entailments, paraphrases, and answers to questionsTree edit models for recognizing textual entailments, paraphrases, and answers to questions. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 1011–1019. USAAssociation for Computational Linguistics, Stroudsburg (2010)

Johnson, M.: PCFG models of linguistic tree representationsPCFG models of linguistic tree representations. Comput. Linguist. 24, 613–632 (1998)

Kamiran, F.: Discrimination-aware ClassificationDiscrimination-aware classification. Doctoral dissertation, Eindhoven University of Technology, The Netherlands (2011)

Kamiran, F., Calders, T.: Classifying without discriminatingClassifying without discriminating. In: 2nd IEEE International Conference on Computer, Control and Communication (IC4), pp. 1–6 (2009a)

Kamiran, F., Calders, T.: Discrimination-Aware ClassificationDiscrimination-aware classification. In: 21st Benelux Conference on Artificial Intelligence (BNAIC), pp. 333–334 (2009b)

Kamiran, F., Calders, T.: Classification with No Discrimination by Preferential Sampling-Classification with no discrimination by preferential sampling. In: Proceedings Machine Learning Conference of Belgium and The Netherlands, BENELEARN (2010)

Kamiran, F., Calders, T.: Data Preprocessing Techniques for Classification without DiscriminationData preprocessing techniques for classification without discrimination. Knowledge and Information Systems (2012) (to Appear)

Kamiran, F., Calders, T., Pechenizkiy, M.: Discrimination Aware Decision Tree LearningDiscrimination aware decision tree learning Tech. Rep. No. CS 10-13. Eindhoven University of Technolgy. 16 Faisal Kamiran, Toon Calders, and Mykola Pechenizkiy (2010a)

Kamiran, F., Calders, T., Pechenizkiy, M.: Discrimination Aware Decision Tree LearningDiscrimination aware decision tree learning. In: IEEE International Conference on Data Mining IEEE International Conference on Data Mining, pp. 869–874 (2010b)

Lerman, R., Yitzhaki, S.: A Note on the Calculation and Interpretation of the Gini IndexA note on the calculation and interpretation of the gini index. Economics Letters 15(3-4), 363–368 (1984)

Levit, M., Alshawi, H., Gorin, A.L., Nöth, E.: Context-sensitive evaluation and correction of phone recognition outputContext-sensitive evaluation and correction of phone recognition output. In: Proc. of the 8th European Conference on Speech Communication and Technology, EUROSPEECH 2003, ISCA (2003)

Luong, B., Ruggieri, S., Turini, F.: k-NN as an Implementation of Situation Testing for Discrimination Discovery and Preventionk-NN as an Implementation of Situation Testing for Discrimination Discovery and Prevention. In: Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 502–510 (2011)

Mazhelis, O., Žliobaitė, I.e., Pechenizkiy, M.: Context-Aware Personal Route Recognition. In: Elomaa, T., Hollmén, J., Mannila, H. (eds.) DS 2011. LNCS, vol. 6926, pp. 221–235. Springer, Heidelberg (2011)

Morris, A., Misra, H.: Confusion matrix based posterior probabilities correctionConfusion matrix based posterior probabilities correction Idiap-RR No. Idiap-RR-53-2002. IDIAP (2002)

Quinlan, J.: C4. 5: programs for machine learning. Morgan Kaufmann (1993)

US Law. The US Equal Credit Opportunity Act.The US equal credit opportunity act. via: (1968), http://www.fdic.gov/regulations/laws/rules/6500-1200.html