Discrimination aware decision tree learning

Citation for published version (APA):
Kamiran, F., Calders, T. G. K., & Pechenizkiy, M. (2010). *Discrimination aware decision tree learning*. (Computer science reports; Vol. 1013). Technische Universiteit Eindhoven.

Document status and date:
Published: 01/01/2010

Document Version:
Publisher's PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:

• A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:

www.tue.nl/taverne

Take down policy
If you believe that this document breaches copyright please contact us at:

openaccess@tue.nl

providing details and we will investigate your claim.

Download date: 21. Sep. 2023

Discrimination Aware Decision Tree Learning*

Faisal Kamiran, Toon Calders and Mykola Pechenizkiy
Email: {t.calders,f.kamiran,m.pechenizkiy}@tue.nl

Eindhoven University of Technology,
The Netherlands

Abstract. Recently, the following problem of discrimination aware classification was introduced: given a labeled dataset and an attribute B, find a classifier with high predictive accuracy that at the same time does not discriminate on the basis of the given attribute B. This problem is motivated by the fact that often available historic data is biased due to discrimination, e.g., when B denotes ethnicity. Using the standard learners on this data may lead to wrongfully biased classifiers, even if the attribute B is removed from training data. Existing solutions for this problem consist of "cleaning away" the discrimination from the dataset before a classifier is learned. In this paper we study an alternative approach in which the non-discriminatory constraint is pushed deeply into a decision tree learner by changing its splitting criterion and pruning strategy by using a novel leaf relabeling approach. Experimental evaluation shows that the proposed approach advances the state-of-the-art in the sense that the learned decision trees have a lower discrimination than models provided by previous methods with only little loss in accuracy.

1 Introduction

In this paper we consider the case where we plan to use data mining for decision making, but we suspect that our available historical data contains discrimination. Applying the traditional classification techniques on this data will produce biased models. Due to anti-discriminatory laws or simply due to ethical concerns the straightforward use of classification techniques is not acceptable. The solution is to develop new techniques which we call *discrimination aware* – we want to learn a classification model from the potentially biased historical data such that it generates accurate predictions for future decision making, yet does not discriminate with respect to a given discriminatory attribute.

The concept of discrimination aware classification can be illustrated with the following example [3]:

A recruitment agency (employment bureau) has been keeping the track of various parameters of job candidates and advertised job positions. Based on this data, the company wants to learn a model for partially automating the match-making between a job

* Supporting material for the paper: F. Kamiran and T. Calders and M. Pechenizkiy. Discrimination Aware Decision Tree Learning. In *IEEE International Conference on Data Mining*. IEEE press, 2010.

and a job candidate. A match is labeled as successful if the company invited the applicant for an interview. It turns out, however, that the historical data is biased; for higher board functions, male candidates have been favored systematically. A model learned directly on this data will pick up this discriminatory behavior and apply it for future predictions.

From an ethical and legal point of view it is of course unacceptable that a model discriminating in this way is deployed; instead, it is preferable that the class assigned by the model is independent of this discriminatory attribute. It is desirable to have a mean to "tell" the algorithm that it should not discriminate the job applicants in future recruitment on the basis of, in this case, the content of the *sex* attribute.

As was already shown in previous works, the straightforward solution of simply removing the attribute B from the training data does not work, as other attributes may be correlated with the suppressed sensitive attribute. It was observed that classifiers tend to pick up these relations and discriminate indirectly [9, 3].

It can be argued that in many real-life cases discrimination can be explained; e.g., it may very well be that females in an employment dataset overall have less years of working experience, justifying a correlation between the gender and the class label. Nevertheless, in this paper we assume this not to be the case. We assume that the data is already divided up into strata based on acceptable explanatory attributes. Within a stratum, gender discrimination can no longer be justified.

Problem statement. In the paper we assume the following setting (cfr. [3]): a labeled dataset D is given, and *one Boolean* discriminatory attribute B (e.g., gender) is specified. The task is to learn a classifier that accurately predicts the class label and whose predictions do not discriminate w.r.t. the sensitive attribute B. We measure discrimination as the probability of getting a positive label for the instances with $B = 0$ minus the probability of getting a positive label for the instances with $B = 1$. For the above recruitment example, the discrimination is hence the ratio of males that are predicted to be invited for the job interview, minus the ratio of females predicted to be invited.

We consider discrimination aware classification as 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 to reduce discrimination [3]. Discrimination and accuracy have to be measured on an unaltered test set that was not used during the training of the classifier.

Proposed solution. We propose the following two techniques for incorporating discrimination awareness into the decision tree construction process:

- **Dependency-Aware Tree Construction.** When evaluating the splitting criterion for a tree node, not only its contribution to the accuracy, but also the level of dependency caused by this split is evaluated.
- **Leaf Relabeling.** Normally, in a decision tree, the label of a leaf is determined by the majority class of the tuples that belong to this node in the training set. In leaf relabeling we change the label of selected leaves in such a way that dependency is lowered with a minimal loss in accuracy. We show a relation between finding the optimal leaf relabeling and the combinatorial optimization problem KNAPSACK [2]. Based on this relation an algorithm is proposed.

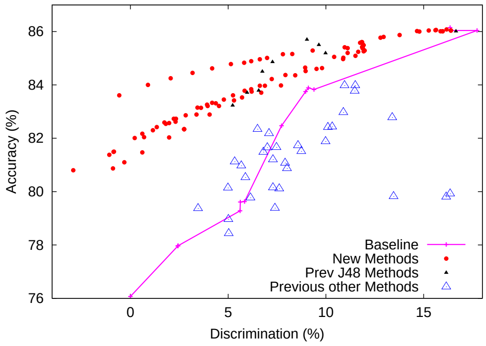
The choice of the decision trees as the type of classifier is arbitrary and simply reflects their popularity. We believe that the proposed techniques can be generalized to other popular classification approaches that construct a decision boundary by partitioning of the instance space, yet this direction is left for further research.

Experiments. We have performed an extensive experimental study, the results of which show what generalization performance we can achieve while trying to have as little discrimination as possible. The results also show that the introduced discrimination aware classification approach for decision tree learning improves upon previous methods that are based on dataset cleaning (or so-called Massaging) [3]. we have also studied the performance of different combinations of the existing data cleaning methods and the new decision tree construction methods.

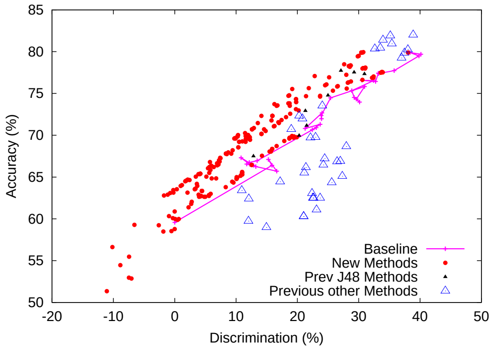
List of Contributions:

1. The theoretical study of discrimination-accuracy trade-off for the discrimination aware classification.
2. *Algorithms.* The development of two new techniques for constructing discrimination aware decision trees: changing the splitting criterion and leaf relabeling. For leaf relabeling a link with KNAPSACK is proven and exploited.
3. *Experiments.* In the experimental section we show the superiority of our new discrimination aware decision tree construction algorithm w.r.t. existing solutions that are based upon "cleaning away" discrimination from the training data before model induction [3] or modifying a naive Bayes classifier [4]. Also, where applicable, combinations of the newly proposed methods with the existing methods have been tested.
4. *Sanity check.* Of particular interest is an experiment in which a classifier is trained on census data from the Netherlands in the 70s and tested on census data of 2001. In these 30 years, gender discrimination w.r.t. unemployment decreased considerably, creating a unique opportunity for assessing the quality of a classifier learned on biased data on (nearly) discrimination-free data. The experiments show that the discrimination-aware decision trees do not only outperform the classical trees w.r.t. discrimination, but also w.r.t. predictive accuracy.
5. An experimental validation of the new techniques showing that their performance w.r.t. the trade-off between accuracy and discrimination outperforms the current state-of-the-art techniques for discrimination aware classification. Figure 1 is given as an evidence which shows the results of experiments conducted over three datasets to compare our newly proposed solutions to the current stat-of the-art techniques. Figure 1 demonstrates clearly that our new techniques give low discrimination scores by maintaining the high accuracy.

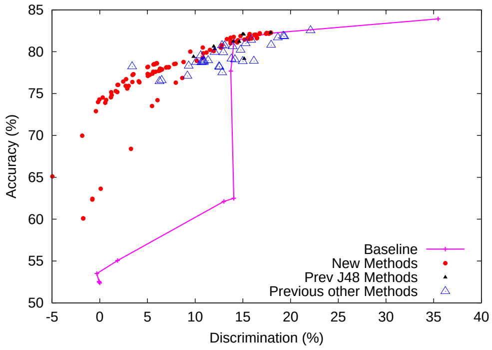
Outline. The rest of the paper is organized as follows. The motivation for the discrimination problem is given in Section 2. Related work is given in Section 3. In Section 4 we formally define the problem statement. In Section 5, the two different approaches towards the problem are discussed. These solutions are empirically evaluated in Section 6. Section 7 concludes the paper and gives directions for further work.



(a) Census Income Dataset



(b) Communities Dataset



(c) Dutch 2001 Census Dataset

Fig. 1. Comparison of our proposed methods with current stat-of-the-art methods

2 Motivation

Discrimination is a sociological term that 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.

Many anti-discrimination laws, e.g., the Australian Sex Discrimination Act 1984, the US Equal Pay Act of 1963 and the US Equal Credit opportunity act have been enacted to eradicate the discrimination and prejudices. It is quite intuitive that if some discriminatory practice is banned by law, nobody would like to practise it anymore due to heavy penalties. However, if we plan to use data mining for decision making, particularly a trained classifier, and our available *historical* data contains discrimination, then the traditional classification techniques will produce biased models. Due to the above mentioned laws or simply due to ethical concerns such use of existing classification techniques is unacceptable. The solution is to develop new techniques which we call *discrimination aware* – we want to learn a classification model from the potentially biased historical data such that generates accurate predictions for future decision making yet does not discriminate with respect to a given sensitive attribute.

We further explore the concept of discrimination with some real world examples [1]: the United Nations had concluded that women often experience a ”glass ceiling” and that there are no societies in which women enjoy the same opportunities as men. The term ”glass ceiling” is used to describe a perceived barrier to advancement in employment based on discrimination, especially sex discrimination.

The China’s leading headhunter, Chinahr.com, reported in 2007 that the average salary for white-collar men was 44,000 yuan ($6,441), compared with 28,700 yuan ($4,201) for women. Even some women who have done well in business complain that a glass ceiling limits their chances of promotion. A recent Grant Thornton survey found that only 30 percent of senior managers in China’s private enterprises are female. In United States, in 2004 the median income of full-time, year-round (FTYR) male workers was $40,798, compared to $31,223 for FTYR female workers, i.e., women’s wages were 76.5% of men’s wages. The US Equal Pay Act of 1963 aimed at abolishing wage disparity based on sex. Due to enactment of this anti-discriminatory law, women’s pay relative to men’s rose rapidly from 1980 to 1990 (from 60.2% to 71.6%), and less rapidly from 1990 to 2004 (from 71.6% to 76.5%).

As illustrated by the next example, the problem of discrimination aware classification can be further generalized:

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.

[1] Source: `http://en.wikipedia.org/wiki/Discrimination`, May 17th, 2010

For example, if many surveys from hospitals in a particular area A are supplied by an enquirer who more quickly than the others diagnoses anxiety symptoms, faulty conclusions such as "Patients in area A suffer from anxiety symptoms more often than other patients" may emerge.

In such cases it is highly likely that the input data will contain discrimination due to high degree of dependency between the data and class attributes on the one hand, and on the data source on the other. Simply removing the information about the source may not resolve the problem, as the data source may be tightly connected to other attributes in the data. For example, a survey about the food quality in restaurants may have been distributed geographically over enquirers. When learning from this data which characteristics of a restaurant are good indicators for the food quality, one may overestimate the impact of region if not all enquirers were equally strict in their assessments. The main claim on which discrimination aware classification is based is therefore that explicitly taking into account non discriminatory constraints in the learning process avoids the classifiers to overfit to such artifacts in the data.

Unfortunately, the straightforward solution of simply removing the attribute B from the training data does not work, as other attributes may be correlated with the suppressed sensitive attribute [9, 3]. For example, ethnicity may be strongly linked to address. Consider, for example, the German Dataset available in the UCI ML-repository [1]. 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, e.g., *own_house*, indicating if the applicant is a home-owner, turn out to be good predictors for *age*. Similarly removing the *sex* and *ethnicity* for the job-matching example or *enquirer* for the survey example from the training data often does not solve this, as other attributes may be correlated with the suppressed attributes. For example, area can be highly correlated with enquirer. Blindly applying an out-of-the-box classifier on the medical-survey data without the enquirer attribute may still lead to a model that discriminates indirectly based on the locality of the hospital. In the context of racial discrimination, this effect of indirect relationships and its exploitation are often referred to as redlining. In the literature [9], this problem was confirmed on the German Dataset available in the UCI ML-repository [1].

3 Related Work

In a series of recent papers [14, 15, 9, 3, 4, 10], the topic of discrimination in data mining received quite some attention. The authors of [14, 15] concentrate mainly on identifying the discriminatory rules that are present in a dataset, and the specific subset of the data where they hold, rather than on learning a discrimination aware classifier for future predictions. Discrimination-aware classification and its extension to independence constraints, were first introduced in [9, 3] where the problem of discrimination is handled by "cleaning away" the discrimination from the dataset before applying the traditional classification algorithms. They propose two approaches *Massaging* and *Reweighing* to clean away the data. Massaging changes the class labels of selected objects in the training data in order to obtain a discrimination free dataset while the Reweighing method selects a biased sample to neutralize the impact of discrimination. Authors of [4] propose three approaches for making the naive Bayes classifiers discrimination-free: these three approaches include modifying the probability of the decision being positive, training one model for every sensitive attribute value and balancing them, and adding a latent variable in the Bayesian model that represents the unbiased label and optimizing the model parameters for likelihood using expectation maximization. In the current paper we propose not to change the dataset, but the algorithms instead, in this case a decision tree learner.

There are many relations with the traditional classification techniques but due to space restrictions, we only discuss the most relevant links. Despite the abundance of related works, none of them satisfactory solves the discrimination aware classification problem. 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 [13]. Our proposed discrimination aware classification problem 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. One noteworthy exception is monotonic classification [11, 5], where the aim is to find a classification that is monotone in a given attribute. Of all existing techniques in classification, monotone classification is probably the closest to our proposal. In Cost-Sensitive and Utility-Based learning [8, 12], it is assumed that not all types of prediction errors are equal and not all examples are as important. 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. Nevertheless it is unclear how these techniques can be generalized to non-discriminatory constraints. For example, satisfaction of monotonic constraints does not depend on the data distribution, whereas for the non-discriminatory constraints it clearly does, and the independency between two attributes cannot easily be reduced to a cost on individual objects.

4 Problem Statement

We formally introduce the problem of discrimination aware classification and we explore the trade-off between accuracy and discrimination.

4.1 Non-discriminatory Constraints

We assume a set of attributes $\{A_1, \ldots, A_n\}$ and their respective domains $dom(A_i)$, $i = 1 \ldots n$ have been given. A *tuple* over the schema $S = (A_1, \ldots, A_n)$ is an element of $dom(A_1) \times \ldots \times dom(A_n)$. We denote the component that corresponds to an attribute A of a tuple x by $x.A$. A dataset over the schema $S = (A_1, \ldots, A_n)$ is a finite set of tuples over S and a labeled dataset is a finite set of tuples over the schema $(A_1, \ldots, A_n, \mathit{Class})$. $dom(\mathit{Class}) = \{+, -\}$.

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	+

As usual, a classifier C is a function from $\prod_{i=1}^{n} dom(A_i)$ to $\{+, -\}$. Let B be a binary attribute with domain $dom(B) = \{0, 1\}$. The *discrimination of* C *w.r.t.* B *in dataset* D, denoted $disc_B(C, D)$ is defined as :

$$disc_B(C,D) := \frac{|\{x \in D \mid x.B = 0, C(x) = +\}|}{|\{x \in D \mid x.B = 0\}|} - \frac{|\{x \in D \mid x.B = 1, C(x) = +\}|}{|\{x \in D \mid x.B = 1\}|} .$$

(When clear from the context we will omit B and D from the notation.) A positive discrimination means that tuples with $B = 1$ are less likely to be classified as positive by the classifier C than others.

The discrimination of D w.r.t. B, denoted $disc_B(D)$, is defined as:

$$disc_B(D) := \frac{|\{x \in D \mid x.B = 0, x.Class = +\}|}{|\{x \in D \mid x.B = 0\}|} - \frac{|\{x \in D \mid x.B = 1, x.Class = +\}|}{|\{x \in D \mid x.B = 1\}|} .$$

For $\epsilon \in [0, 1]$, the formula $disc_B(C, D) \leq \epsilon$ is called a *non-discriminatory constraint*.

Example 1 *In Table 1, an example dataset is given. This dataset contains the Sex, Ethnicity, Highest Degree of* 10 *job applicants, the Job Type they applied for and the Class defining the outcome of the selection procedure. In this dataset, the discrimination ratio between* Sex *and Class will be* $disc_{Sex=f}(D) := \frac{4}{5} - \frac{2}{5} = 40\%$ *. It means that the data object with* $Sex = f$ *will have* 40% *less chance of getting a job than the one with* $Sex = m$*.*

4.2 Discrimination Aware Classification

The problem of discrimination aware classification can now be stated as follows.

Problem 1 (Discrimination aware classification). Let a labeled dataset D and a sensitive attribute B be given. The discrimination aware classification problem is to learn a classifier such that (a) The accuracy of C is high, and (b) the discrimination of C w.r.t. B is low. (Both accuracy and discrimination are to be computed with respect to an unaltered test set).

Notice that a more natural formulation may have been to require high accuracy of the learned classifier C on the "ground truth"; i.e., the correct data without the discrimination. Neither this ground truth, however, nor the exact process that lead to the discrimination are available in general. In this context it is natural to ask for a classifier without discrimination, for which the predictions, nevertheless, stay close to the labels given in the training set. In this way, our criteria reflect that we assume that it is more likely that only few of the labels changed during the introduction of discrimination than that many labels changed.

The formulation of the problem statement is rather informally requiring "high" accuracy and "low" discrimination. This ambiguity in not arbitrary, but due to the trade-off which exists between the accuracy and the resulting discrimination of a classifier. In general, lowering the discrimination will result in lowering the accuracy as well and vice versa. In the next subsection we go deeper into this issue with a discussion on *Discrimination-Accuracy optimal classifiers*.

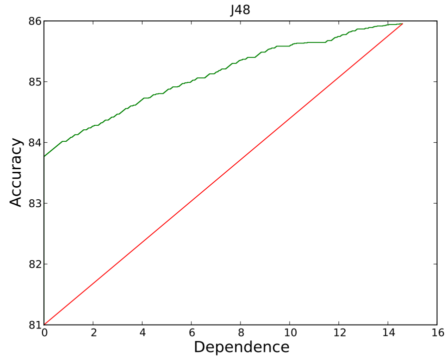
In the remainder of the paper we make the following three assumptions:

(A) There is only one non-discriminatory constraint. The sensitive attribute is B and $dom(B) = \{0, 1\}$.
(B) The prime intention is learning the most accurate decision tree for which the discrimination is close to 0. Essentially we envision a scenario in which a maximally allowable discrimination ϵ is specified.
(C) As it is assumed that the discrimination on B is an artifact, the learned classifier should not use the attribute B at prediction time. Only at learning time we can use the attribute B.

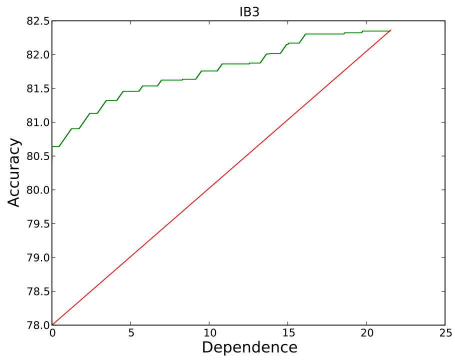
4.3 Accuracy - Discrimination Trade-Off

Before going into the proposed solutions, we first theoretically study the trade-off between discrimination and accuracy in a general setting. Let $\mathcal{C}$ be a set of classifiers. We will call a classifier C *optimal w.r.t. discrimination and accuracy (DA-optimal) in* $\mathcal{C}$ if for every other classifier C' in $\mathcal{C}$ either $disc(C) < disc(C')$, or $acc(C') < acc(C)$.

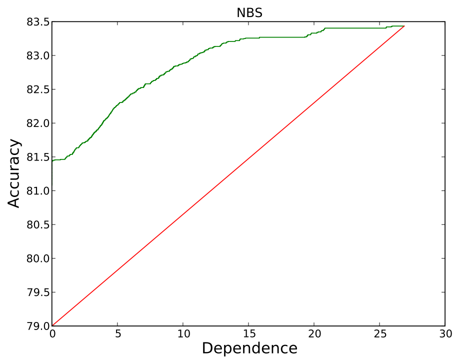
In this section we will theoretically study the following scenario: suppose a discriminatory classifier C (or ranker R) is learned using a traditional learning method. What are the DA-optimal classifiers we can construct by changing the labels that C assigns in a post-processing phase? In this study we will assume that we can use attribute B in the post-processing phase. It is easy to see that otherwise without the attribute B, the only thing we can do that is guaranteed to change the discrimination of a classifier, is to always make the same prediction. The importance of this theoretical study is that it establishes the following important fact: *no matter what we do, if we only use a given classifier without any additional information on the data distribution, the best*



(a) Decision tree AUC=87



(b) 3-NN AUC = 76



(c) Naive Bayes AUC = 89

Fig. 2. Trade-off between accuracy and dependence (discrimination) for the DA-optimal classifiers in $\mathcal{C}_R$ (Curved green line) and $\mathcal{C}_C$ (straight red line)

trade-off between accuracy and discrimination we can hope for is linear. If we use extra information; e.g., we have a ranking function, we can do better than a linear trade-off.

Post-processing the classifier output. Suppose we have a classifier C for which we want to change the predictions in order to reduce the discrimination on attribute B. We use the following probabilities for keeping the label assigned by C:

	$C = -$	$C = +$
$B = 0$	p_{0-}	p_{0+}
$B = 1$	p_{1-}	p_{1+}

We denote the resulting classifier by $C[p_{0-}, p_{0+}, p_{1-}, p_{1+}]$. If we have to label a new example x with $B = b$, it will be assigned to class $c = C(x)$ by $C[p_{0-}, p_{0+}, p_{1-}, p_{1+}]$ with probability p_{bc}, and to the other class with probability $1 - p_{bc}$. For example, $C[1, 1, 1, 1]$ denotes the classifier C itself, and $C[0, 0, 0, 0]$ the classifier that never assigns the same class as C. We denote the set of all classifiers we can construct from C in this way, by $\mathcal{C}_C$:

$$\mathcal{C}_C \ := \ \{C[p_{0-}, p_{0+}, p_{1-}, p_{1+}] \mid 0 \leq p_{0-}, p_{0+}, p_{1-}, p_{1+} \leq 1\} \ .$$

Example 1. For example, suppose we keep $p_{0-} = p_{1-} = p_{1+} = 1$, and change $p_{0+} = 0.8$. Since $C[1, 0.8, 1, 1]$ does not change the class assignments for tuples with $B = 1$, the positive class probability for these tuples remains the same. On the other hand, some of the positively labeled examples with $B = 0$ are changed, while the negative ones remain. As a result, the positive class probability for the tuples with $B = 0$ decreases, and the total effect is that discrimination decreases at the cost of some accuracy.

The next theorem generalizes the example and shows what is the optimal we can obtain in this way. In the analysis we will implicitly assume that the probability of an example being relabeled does not depend on its true class, given its B-value and the class assigned to it by C. This assumption holds in the limit as the true label C is unknown at the moment C' is applied. With these definitions we get the following theorem:

Theorem 1 *Let C be a classifier with $disc(C) > 0$. A classifier C' with $disc(C') \geq 0$ is DA-optimal in $\mathcal{C}_C$ iff*

$$acc(C) - acc(C') \ = \ \alpha(disc(C) - disc(C'))$$

with

$$\alpha \ := \ \min\left(P[B = 0]\frac{tp_0 - fp_0}{tp_0 + fp_0}, P[B = 1]\frac{tn_1 - fn_1}{tn_1 + fn_1}\right)$$

tp_b (tn_b,fp_b,fn_b), $B = 0, 1$ denotes the true positive (true negative, false positive, false negative) rate for the tuples with $B = b$; e.g., tp_1 is the probability that C assigns the correct label to a positive example with $B = 1$.

Proof. For the classifier $C' = C[p_{0-}, p_{0+}, p_{1-}, p_{1+}]$, the true positive rate for $B = 0$, tp'_0, will be:

$$tp'_0 \ = \ p_{0+} tp_0 + (1 - p_{0-}) fn_0 \ ,$$

as there are two types of true positive predictions: on the one hand true positive predictions of C that were not changed in C' (probability p_{0+}) and on the other hand false negative predictions of C that were changed in C' (probability $1 - p_{0-}$). For the other quantities similar identities exist. Based on these equations we can write accuracy and discrimination of C' in function of tp_b, tn_b, fp_b, fn_b for $B = 0, 1$ (p_b denotes $P[B = b]$):

$$
\begin{aligned}
acc(C') &= p_0(tp'_0 + tn'_0) + d_1(tp'_1 + tn'_1) \\
&= p_0(p_{0+}tp_0 + (1 - p_{0-})fn_0 + p_{0-}tn_0 + (1 - p_{0+})fp_0) \\
&\quad + p_1(p_{1+}tp_1 + (1 - p_{1-})fn_1 + p_{1-}tn_1 + (1 - p_{1+})fp_1) \\
disc(C') &= (tp'_1 + fp'_1) - (tp'_0 + fp'_0) \\
&= (p_{1+}tp_1 + (1 - p_{1-})fn_1 + p_{1+}fp_1 + (1 - p_{1-})tn_1) \\
&\quad - (p_{0+}tp_0 + (1 - p_{0-})fn_0 + p_{0+}fp_0 + (1 - p_{0-})tn_0)
\end{aligned}
$$

The formulas can be simplified by the observation that the DA-optimal classifiers will have $p_{0+} = p_{1-} = 1$; i.e., we never change a positive prediction for a tuple having $B = 0$ to a negative one or a negative prediction for a tuple having $B = 1$ into a positive one. The theorem now follows from analyzing when the accuracy is maximal for fixed discrimination. □

We see a linear trade-off between discrimination and accuracy in the theorem. This linear trade-off could be interpreted as a negative result: if we rely only on the learned classifier and try to undo the discrimination in a post-processing phase, the best we can do is trading in accuracy linearly proportional to the decrease in discrimination we want to achieve. The more balanced the classes are, the higher the price we need to pay per unit of discrimination reduction.

Classifiers based on rankers. On the bright side, however, most classification models actually provide a score or probability $R(x)$ for each tuple x of being in the positive class, instead of only a class label. For example, a Naive Bayes classifier computes a score for every example and for decision tree we often also have access to (an approximation of) the class distribution in every leaf. Such a score allows us for a more careful choice about which tuples to change the predicted label for: instead of using a uniform weight for all tuples with the same predicted class and B-value, the score can be used as follows: We dynamically set different cut-off c_0 and c_1 for respectively tuples with $B = 0$ and $B = 1$; for a ranker R, the classifier $R(c_0, c_1)$ will predict $+$ for a tuple x if $x.B = 0$ and $R(x) \geq c_0$ and if $x.B = 1$ and $R(x) \geq c_1$. In all other cases, $-$ is predicted. The class of all classifiers $R(c_0, c_1)$ will be denoted $\mathcal{C}_R$. 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 actually sometimes a change will 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, however, is far beyond the scope of this paper. Instead we tested this trade-off empirically. The results of this study are shown in Figure 2. In this figure the DA-optimal classifiers in the classes

Algorithm 1: Decision Tree Induction

Parameters: Split evaluator $gain$, purity condition $pure$
Input Dataset D over $\{A_1, \ldots, A_n, Class\}$, att_list
Output Decision Tree $DT(D, att\_list)$
Create a node N
if $pure(D)$ or att_list is empty **then**
    Return N as a leaf labeled with majority class of D
end if
Select $test\_att$ from att_list and $test$ s.t. $gain(test\_att, test)$ is maximized
Label node N with $test\_att$
for Each outcome S of $test$ **do**
    Grow a branch from node N for the condition $test(test\_att) = S$
    Let D_S be the set of examples x in D for which $test(x.test\_att) = S$ holds
    if D_S is empty **then**
        Attach a leaf labeled with the majority class of D
    else
        Attach the node returned by $Decision\_Tree(D_S, att\_list - \{test\_att\})$
    end if
end for

$\mathcal{C}_R$ (curves) and $\mathcal{C}$ (straight line) are shown for the Census-Income dataset [1]. The three classifiers are a Decision Tree (J48), a 3-Nearest Neighbor model (3NN), and a Naive 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.

Conclusion. In this section the accuracy-discrimination trade-off is clearly illustrated. It is theoretically shown that if we rely on only post-processing the output of the classifiers, the best we can hope for is a linear trade-off between accuracy and discrimination. Notice also that the classifiers proposed in this section violate our assumption C; the classifiers $C[p_{0-}, p_{0+}, p_{1-}, p_{1+}]$ use the attribute B at prediction time.

5 Solutions

In this section we propose two solutions to construct decision trees without discrimination. The first solution is based on the adaptation of splitting criterion for tree construction to build a discrimination-aware decision tree. The second approach is post-processing of decision tree with discrimination-aware pruning and relabeling of tree leaves.

5.1 Discrimination-Aware Tree Construction

Traditionally, when constructing a decision tree, we iteratively refine a tree by iteratively splitting its leaves until a desired objective is achieved, as shown in Algorithm 1. The optimization criteria used are usually trying to optimize the overall accuracy of the tree, e.g., based on the so-called *information gain*. Suppose that a certain split divides the data D into $D_1, \ldots, D_k$. Then, the information gain is defined as:

$$IGC := H_{Class}(D) - \sum_{i=1}^{k} \frac{|D_i|}{|D|} H_{Class}(D_i) \ ,$$

where H_{Class} denotes the entropy w.r.t. the class label. In this paper, however, we are not only concerned with accuracy, but also with discrimination. Therefore, we will change the iterative refinement process by also taking into account the influence of newly introduced split on the discrimination of the resulting tree. Our first solution is changing the attribute selection criterion as in step 5 of Algorithm 1. To measure the influence of the split on the discrimination, we will use the same information gain, but now w.r.t. the sensitive attribute B instead of the class $Class$. This gain in sensitivity to B will be denoted IGS. The IGS is defined as:

$$IGS := H_B(D) - \sum_{i=1}^{k} \frac{|D_i|}{|D|} H_B(D_i) \ ,$$

here H_B describes the entropy w.r.t. sensitive attribute. Based on these two measures IGC and IGS, we introduce three alternative criteria for determining the best split:

IGC-IGS: We only allow for a split if it is non-discriminatory, i.e., we select an attribute which is homogeneous w.r.t. class attribute but heterogenous w.r.t. sensitive attribute. We subtract the gain in discrimination from the gain in accuracy to make the tree homogeneous w.r.t. class attribute and heterogenous w.r.t. sensitive attribute.

IGC/IGS: We make a trade-off between accuracy and discrimination by dividing the gain in accuracy by gain in discrimination.

IGC+IGS: We add up the accuracy gain and the discrimination gain. It means, we want to construct a homogeneous tree w.r.t. both accuracy and the sensitive attribute. IGC+IGS will lead to good results in combination with the relabeling technique we show next.

5.2 Relabeling

In this section 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. Let T be a decision tree with n leaves. Such a decision tree partitions the example space into n non-overlapping regions. See Figure 3 for an example; in this figure (left) a decision tree with 6 leaves is given, labeled l_1 to l_6. The lower part of the figure shows the partitioning induced by the decision tree. When a new example needs to be classified by a decision tree, it is given the majority class label of the region it falls into; i.e., the leaves are labeled with the majority class of their corresponding region. The *relabeling* technique, however, will now change this strategy of assigning the label of the majority class. Instead, we try to relabel the leaves of the decision tree in such a way that the discrimination decreases while trading in as little accuracy as possible. We can compute the influence of relabeling a leaf on the accuracy and discrimination of the tree on

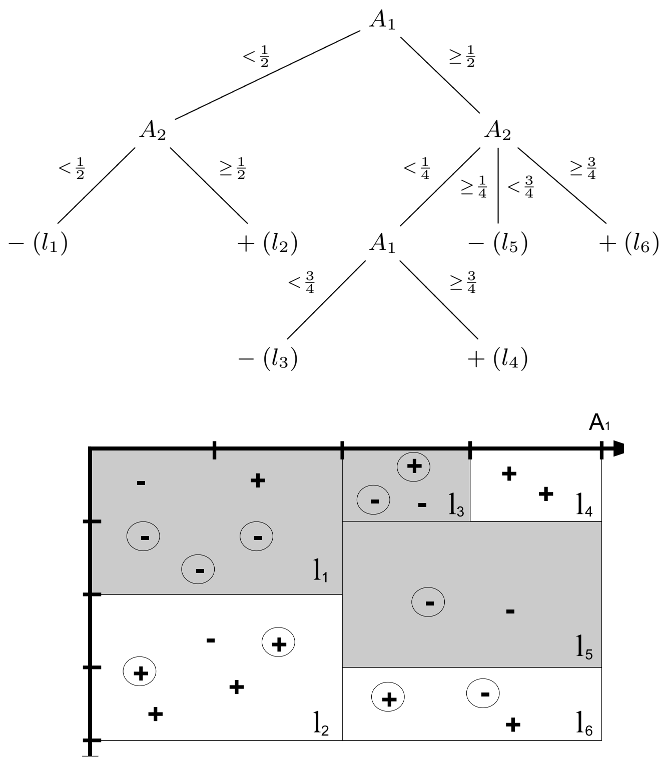


Fig. 3. Decision tree with the partitioning induced by it. The $+$ and $-$ symbols in the partitioning denote the examples that were used to learn the tree. Encircled examples have $B = 1$. The grey background denotes regions where the majority class is $-$

a dataset D as follows. Let the joint distributions of the class attribute C and the sensitive attribute B for respectively the whole dataset and for the region corresponding to the leaf be given by the following contingency table (For the dataset additionally the frequencies have been split up according to the predicted labels by the tree):

Dataset

Class →	$-$	$+$	
Pred. →	$-/+$	$-/+$	
$B=1$	U_1/U_2	V_1/V_2	b
$B=0$	W_1/W_2	X_1/X_2	$\overline{b}$
	N_1/N_2	P_1/P_2	1

Leaf 1

	$-$	$+$	
$B=1$	u	v	b
$B=0$	w	x	$\overline{b}$
	n	p	a

Hence, e.g., a fraction a of the examples end up in the leaf we are considering for change, of which n are in the negative class and p in the positive. Notice that for the leaf we do not need to split up u, v, w, and x since all examples in a leaf are assigned to the same class by the tree.

With these tables it is now easy to get the following formulas for the accuracy and discrimination of the decision tree *before* the label of the leaf l is changed:

$$acc_T = N_1 + P_2$$
$$disc_T = \frac{W_2 + X_2}{\bar{b}} - \frac{U_2 + V_2}{b}$$

The effect of relabeling the leaf now depends on the majority class of the leaf; on the one hand, if $p > n$, the label of the leaf changes from $+$ to $-$ and the effect on accuracy and discrimination is expressed by:

$$\Delta acc_l = n - p$$
$$\Delta disc_l = \frac{u + v}{b} - \frac{w + x}{\bar{b}}$$

on the other hand, if $p < n$, the label of the leaf changes from $-$ to $+$ and the effect on accuracy and discrimination is expressed by:

$$\Delta acc_l = p - n$$
$$\Delta disc_l = -\frac{u + v}{b} + \frac{w + x}{\bar{b}}$$

Notice that relabeling leaf l does not influence the effect of the other leaves and that Δacc_l is always negative.

Example 2 *Consider the dataset and tree given in Figure 3. The contingency tables for the dataset and leaf l_3 are as follows:*

Dataset

Class $\rightarrow$	$-$	$+$	
Pred. $\rightarrow$	$-/+$	$-/+$	
$B = 1$	$\frac{5}{20}/\frac{1}{20}$	$\frac{1}{20}/\frac{3}{20}$	$\frac{1}{2}$
$B = 0$	$\frac{3}{20}/\frac{1}{20}$	$\frac{1}{20}/\frac{5}{20}$	$\frac{1}{2}$
	$\frac{8}{20}/\frac{2}{20}$	$\frac{2}{20}/\frac{8}{20}$	1

Leaf l_3

	$-$	$+$	
$B = 1$	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{2}{20}$
$B = 0$	$\frac{1}{20}$	0	$\frac{1}{20}$
	$\frac{2}{20}$	$\frac{1}{20}$	$\frac{3}{20}$

The effect of changing the label of node l_3 from $-$ to $+$ hence is: $\Delta acc_l = -\frac{1}{20}$ and $\Delta disc_l = -\frac{1}{10}$.

The central problem now is to select exactly this set of leaves that is optimal w.r.t. reducing the discrimination with minimal loss in accuracy, as expressed in the following *Optimal relabeling problem* (RELAB):

Problem 1 (RELAB) *Given a decision tree T, a bound $\epsilon \in [0, 1]$, and for every leaf l of T, Δacc_l and $\Delta disc_l$, find a subset L of the set of all leaves $\mathcal{L}$ satisfying*

$$rem\_disc(L) \ := \ disc_T + \sum_{l \in L} \Delta disc_l \leq \epsilon$$

that minimizes

$$lost\_acc(L) \ := \ -\sum_{l \in L} \Delta acc_l \ .$$

We will now show that the RELAB problem is actually equivalent to the following well-known combinatorial optimization problem:

Problem 2 (KNAPSACK). Let a set of items $\mathcal{I}$, a weight $w(i)$ and a profit $p(i)$, both positive integers, for every item $i \in \mathcal{I}$, and an integer bound K be given. Find a subset $I \subseteq \mathcal{I}$ subject to $\sum_{i \in I} w(i) \leq K$ that maximizes $\sum_{i \in I} p(i)$.

The following theorem makes the connection between the two problems explicit.

Theorem 2 *Let T be a decision tree, and $\epsilon \in [0, 1]$ and for every leaf l of T, Δacc_l and $\Delta disc_l$ have been given.*

The RELAB problem with this input is equivalent to the KNAPSACK problem with the following inputs:

- $\mathcal{I} = \{\, l \in \mathcal{L} \mid \Delta disc_l < 0 \,\}$
- $w(l) = -\alpha \Delta disc_l$ *for all* $l \in \mathcal{I}$
- $p(l) = -\alpha \Delta acc_l$ *for all* $l \in \mathcal{I}$
- $K = \alpha \left(\sum_{l \in \mathcal{I}} disc_l - disc_T + \epsilon \right)$

Where α is the smallest number such that all $w(l), p(l)$, and K are integers.

Any optimal solution L to the RELAB problem corresponds to a solution $I = \mathcal{I} \setminus L$ for the KNAPSACK problem and vice versa.

Proof. Let L be an optimal solution to the RELAB problem. Suppose $l \in L$ has $\Delta disc_l \geq 0$. Then, $rem\_disc(L \setminus \{l\}) \leq rem\_disc(L) \leq \epsilon$, and, since Δacc_l is always negative, $lost\_acc(L \setminus \{l\}) \leq lost\_acc(L)$. Hence, there will always be an optimal solution for RELAB with $L \subseteq \mathcal{I}$. The equivalence of the problems follows easily from multiplying the expressions for rem_disc and $lost\_acc$ with α and rewriting them, using $\sum_{l \in \mathcal{I}} w(l) = \sum_{l \in L} w(l) + \sum_{l \in I} w(l)$ for $I = \mathcal{I} \setminus L$. $\square$

Algorithm 2: *Relabel*

Input Tree T with leaves $\mathcal{L}$, $\Delta acc(l), \Delta disc(l)$ for every $l \in \mathcal{L}$, $\epsilon \in [0, 1]$
Output Set of leaves L to relabel

```
I := { l ∈ L | Δdisc_l < 0 }
L := {}
while rem_disc(L) > ε do
    best_l := arg max_{l ∈ I \ L} (disc_l / acc_l)
    L := L ∪ {l}
end while
return L
```

From this equivalence we can now derive many properties regarding the intractability of the problem, approximations, and guarantees on the approximation. Based on the connection with the KNAPSACK problem, the greedy Algorithm 2 is proposed for approximating the most optimal relabeling. The following corollary gives some computational properties of the RELAB problem and a guarantee for the greedy algorithm.

Corollary 1

1. *RELAB is **NP**-complete.*
2. *RELAB allows for a* fully polynomial approximation scheme (FPTAS) *[2].*
3. *An optimal solution to RELAB can be found with a dynamic programming approach in time* $\mathcal{O}(|D|^3|\mathcal{I}|)$ [2]
4. *The difference in accuracy of the optimal solution and the accuracy of the tree given by Algorithm 2 is at most* $\frac{rem\_disc(L)-\epsilon}{\Delta disc_l}\Delta acc_l$ *where* l *is the last leaf that was added to* L *by Algorithm 2.*

Proof. Membership in **NP** follows from the reduction of RELAB to KNAPSACK. Completeness, on the other hand follows from a reduction from PARTITION to RELAB. Given a multiset $\{i_1, \ldots, i_n\}$ of positive integers, the PARTITION problem is to divide this set into two subsets that sum up to the same number. Let $N = i_1 + \ldots + i_n$. Consider a database D with $3N$ tuples and a decision tree T with the following leafs: T has 2 big leafs with N tuples with $B = 0$ and $Class = 0$, and n leafs with respectively $i_1, \ldots, i_n$ tuples, all with $B = 1$ and $Class = 1$. The accuracy of the tree is 100%. It is easy to create such an example. The discrimination of the tree T equals $100\% - 50\% = 50\%$. Changing one of the big leafs will lead to a drop in accuracy of $1/3$ and a drop in discrimination of 50%, to 0%. Changing the jth positive leaf will lead to a drop in accuracy of $i_j/3N$ and a drop in discrimination of i_j/N. The partition problem has a solution if and only if the optimal solution to the RELAB problem for the tree T with $\epsilon = 0$ has $lost\_acc = 1/6$.

Point 2 follows directly from the reduction of RELAB to KNAPSACK. 3 follows from the fact that α is at most $|D|(|D|B)(|D|\overline{B}) \leq |D|^3$ and the well known dynamic programming solution for KNAPSACK in time $\mathcal{O}(K|\mathcal{I}|)$. 4 follows from the relation between KNAPSACK and the so-called *fractional KNAPSACK-problem* [2]. The difference between the optimal solution and the greedy solution of Algorithm 2 is bounded above by the accuracy loss contributed by the part of l that overshoots the bound ϵ. This "overshoot" is $\frac{\epsilon-rem\_disc(L)}{\Delta disc_l}$. The accuracy loss contributed by this overshoot is then obtained by multiplying this fraction with $-\Delta acc_l$. □

The most important result in this corollary is with no doubt that the greedy Algorithm 2 approximates the optimal solution to the RELAB problem very well. In this algorithm, in every step we select the leaf that has the least loss in accuracy per unit of discrimination that is removed. This procedure is continued until the bound ϵ has been reached. The difference with the optimal solution is proportional to the accuracy loss that corresponds to the fraction of discrimination that is removed too much.

Example 3 *Consider again the example decision tree and data distribution given in Figure 3. The discrimination of the decision tree is* 20%. *Suppose we want to reduce the*

[2] Notice that this bound in 3. is not inconsistent with the **NP**-completeness of 1., as RELAB does not take the dataset D as input, but only the Δ's.

discrimination to 5%*. The* Δ*'s and their ratio are as follows:*

Node	Δacc	$\Delta disc$	$\frac{\Delta disc}{\Delta acc}$
l_1	-15%	-10%	$2/3$
l_2	-15%	-10%	$2/3$
l_3	-5%	-10%	2
l_4	-10%	-20%	2
l_5	-10%	0%	0
l_6	-5%	10%	-2

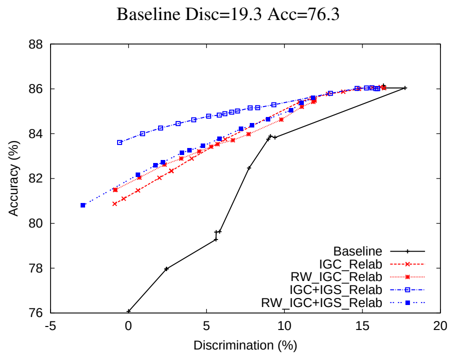
The reduction algorithm will hence first pick l_3 *or* l_4*, then* l_1 *or* l_2*, but never* l_5 *or* l_6*.*

6 Experiments

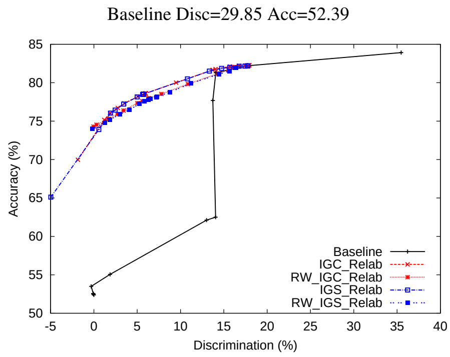
All datasets and the source code of all implementations reported upon in this section are available at `http://www.win.tue.nl/~fkamiran/code`.

In this section we show the results of experiments with the new discrimination-aware splitting criteria and the leaf relabeling for decision trees. As we observe that the discrimination-aware splitting criteria by themselves do not lead to significant improvements w.r.t. lowering discrimination, we have omitted them from the experimental validation. However, the new splitting criteria IGC+IGS is an exception: sometimes, when used in combination with leaf relabeling, it outperforms the leaf relabeling with original decision tree split criterion IGC. IGC+IGS in combination with relabeling outperforms other splitting criteria because this criterion tries to make tree leaves homogeneous w.r.t. both class attribute and sensitive attribute. The more homogeneous w.r.t. the sensitive attribute the leaves are, the less number of leaves we will have to relabel to remove the discrimination from the decision tree. So the use of this criterion with leaf relabeling reduces the discrimination by making the minimal possible changes in our decision tree. For the relabeling approach, however, the results are very encouraging, even when the relabeling is applied with normal splitting criterion IGC. We compare the following techniques (between brackets their short name):

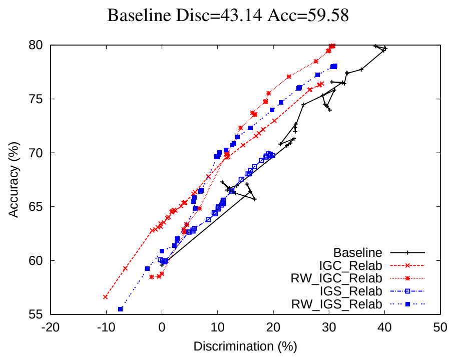
1. The baseline solutions (Baseline) that consist of removing B and its k most correlated attributes from the training dataset before learning a decision tree, for $k = 0, 1, \ldots, n$. In the graphs this baseline will be represented by a black continuous line connecting the performance figures for increasing k.
2. We also present a comparison to the previous state-of-the-art techniques, shown in Table 4, which includes discrimination aware *naive Bayesian* approaches [4], and the pre-processing methods *Massaging* and *Reweighing* [9, 3] that are based on cleaning away the discrimination from the input data before a traditional learner is applied.
3. From the proposed methodes we show the relabeling approach in combination with normal decision tree splitting criteria (IGC_Relab) and with new splitting criteria IGC+IGS (IGC+IGS_Relab).
4. Finally we also show some hybrid combinations of the old and new methods; we present the results of experiments where we first applied the *Reweighing* technique



(a) Census Income Data



(b) Dutch Census 2001 Data



(c) Communities Data

Fig. 4. Accuracy-discrimination trade-off for different values of epsilon $\epsilon \in [0, 1]$ is plotted. We change the value of epsilon from the baseline discrimination in the dataset (top right points of lines) to the zero level (bottom left points of these lines).

of [3] on the training data to learn a tree with low discrimination (either with the normal or the new splitting criterion). On this tree we then apply relabeling to remove the last bit of discrimination from it (RW_IGC_Relab and RW_IGC+IGS_Relab). The other combinations led to similar results and are omitted from the comparison.

Experimental Setup. We apply our proposed solutions on the Census Income dataset [1], the Communities dataset [1], and two Dutch census datasets of 1971 and 2001 [6, 7]. 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*. All the attributes are categorical except *weight* (representing the size of the aggregated group) which we exclude from our experiments. We use the attribute *occupation* as a class attribute where the task is to classify the instances into "high level" (prestigious) and "low level" professions. We remove the records of underage people, some middle level professions and people with unknown professions, leaving 60 420 instances for our experiments. The Dutch 1971 Census dataset consists of 159 203 instances and has the same features except the missing attribute *place of previous residence* and the extra attribute *religious denominations*. After removing the records of people under the age of 19 years and records with missing values, we use 99 772 instances in our experiments. We use the attribute *sex* as sensitive attribute.

The Communities dataset has 1 994 instances which give information about different communities and crimes within the United States. Each instance is described by 122 predictive attributes which are used to predict the total number of violent crimes per 100K population while 5 non predictive attributes are also given which can be used only for extra information. In our experiments we use only predictive attributes which are numeric. We add a sensitive attribute *black* to divide the communities according to race and discretize the class attribute to divide the data objects into major and minor violent communities.

The Census Income dataset has 48 842 instances. This dataset contains demographic information about people and 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. Each data object is described by 14 attributes, including: *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 $Sex = f$ as sensitive attribute. In this dataset, 16 192 citizens have $Sex = f$ and 32 650 have $Sex = m$. The discrimination of the class w.r.t. $Sex = f$ is $disc_{Sex=f}(D) := \frac{9918}{32650} - \frac{1769}{16192} = 19.45\%$.

Testing the Proposed Solutions. The reported figures are the averages of a 10-fold cross-validation experiment. Every point represents the performance of one learned decision tree on original test data excluding the sensitive attribute from it. Every point in the graphs corresponds to the discrimination (horizontal axis) and the accuracy (vertical axis) of a classifier produced by one particular combination of techniques. Ideally, points should be close to the top-left corner. The comparisons show clearly that relabeling succeeds in lowering the discrimination much further than the baseline approach.

Table 2. The detail of working and not working males and females in the Dutch 1971 Census dataset.

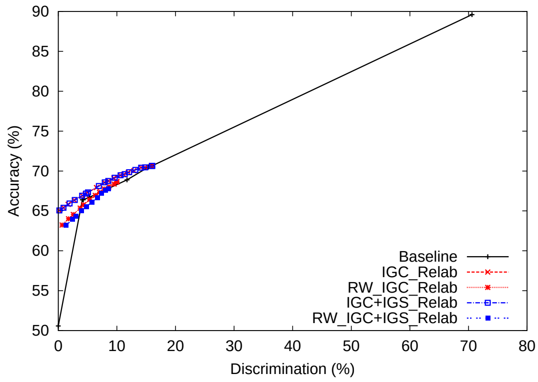
	Job=Yes (+)	Job=No (-)	
Male	38387 (79.78%)	9727 (20.22%)	48114
Female	10912 (21.12%)	40746 (78.88%)	51658
Disc = 79.78 - 21.12 = 58.66%			

Table 3. The detail of working and not working males and females in the Dutch 2001 Census dataset.

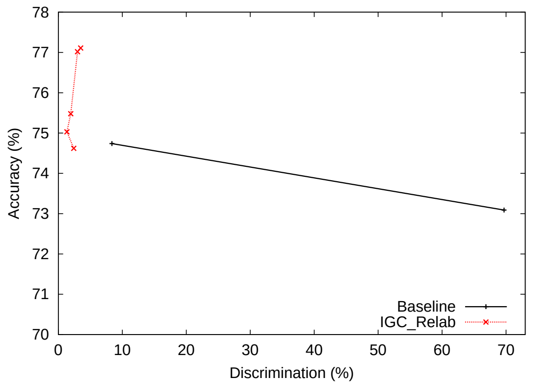
	Job=Yes (+)	Job=No (-)	
Male	52885 (75.57%)	17097 (24.43%)	69982
Female	37893 (51.24%)	36063 (48.768%)	73956
Disc = 75.57 - 51.24 = 24.23%			

Figure 4 shows a comparison of our discrimination aware techniques with the baseline approach over three different datasets. We observe that the discrimination goes down by removing the sensitive attribute and its correlated attribute but its impact over the accuracy is very severe. On the other hand the discrimination aware methods classify the unseen data objects with minimum discrimination and high accuracy for all values of ϵ. We also run our proposed methods with both *Massaging* and *Reweighing* but we only present the results with *Reweighing* because both show similar behavior in our experiments.

It is very important to notice here that we measure the accuracy score here over the discriminatory data but ideally we expect a non-discriminatory test data. If our test set is non discriminatory, we expect our discrimination aware methods to outperform the traditional method w.r.t. both accuracy and discrimination. In our experiments, we mimic this scenario by using the Dutch 1971 Census data as a training set and the Dutch 2001 Census dataset as a test set. We use the attribute *economic status* as class attribute because this attribute uses similar codes for both 1971 and 2001 dataset. The use of *occupation* as class attribute was not possible in these experiments because its coding is different in both datasets. This attribute *economic status* determines whether a person has some job or not, i.e., is economically active or not. We remove some attributes like *current economic activity* and *occupation* from these experiments to make both datasets consistant w.r.t. codings. Tables 2 and 3 show that in Dutch 1971 Census data, there is more discrimination toward female and their percentage of unemployment is higher than in the Dutch 2001 Census data. Now if we learn a traditional classifier over 1971 data and test it over the same dataset using 10-fold cross validation method, it will give excellent performance as shown in Figure 5 (a). When we apply this classifier to 2001 data without taking the discrimination aspect into account, it performs very poorly and accuracy level goes down from 89.6% (when tested on 71 data; Figure 5 (a)) to 73.09% (when tested on 2001 data; Figure 5 (b)). Figure 5 makes it very obvious that our discrimination aware technique not only classify the future data without discrimination but they also work more accurately than the traditional classification methods when tested over non-discriminatory data. In Figure 5 (b), we only show the results of IGC_Relab



(a) Dutch 1971 Census data is as test set.



(b) Dutch 2001 Census data is used as test set.

Fig. 5. The results of experiments when Dutch 1971 Census dataset is used as train set while the test set is different for both plots.

because other proposed methods also give similar results. Figure 5 (b) shows that if we change the value of ϵ from 0 to 0.04 the accuracy level increases significantly from 74.62% to 77.11%. We get the maximum accuracy at $\epsilon = 0.04$ because the Dutch 2001 Census data is not completely discrimination free.

In order to assess the statistical relevance of the results, in Table 4 the exact accuracy and discrimination figures together with their standard deviations have been given. As can be seen, the deviations are in general much smaller than the differences between the points. Table 4 also gives a comparison of our proposed methods with the other state-of-the-art methods on the Census Income dataset. We select the best results of the competitive methods to compare with. We observe that out proposed method outperform the others approaches w.r.t. accuracy-discrimination trade off.

From the results of our experiments we draw the following conclusions: (1) Our proposed methods give high accuracy and low discrimination scores when applied to non-discriminatory test data. In this scenario, our methods are the best choice, even if we are

Table 4. The results of experiments over the Census Income dataset with their standard deviations. ($\epsilon = 0.01$)

Method	Disc (%)	Acc (%)
IGC_Relab	0.31 ± 1.10	81.10 ± 0.47
IGC+IGS_Relab	0.90 ± 1.50	84.00 ± 0.46
RW_IGC_Relab	0.59 ± 1.17	81.66 ± 0.60
RW_IGC+IGS_Relab	0.63 ± 1.29	82.27 ± 0.67
Massaging	6.59 ± 0.78	83.82 ± 0.22
Reweighing	7.04 ± 0.74	84.84 ± 0.38
Naive Bayesian Approach	0.10	80.10

only concerned with accuracy. (2) The improvement in discrimination reduction with the relabeling method is very satisfying. The relabeling reduces discrimination to almost 0 in almost all cases if we decrease the value of ϵ to 0. (3) The relabeling methods outperform the baseline in almost all cases. As such it is fair to say that the straightforward solution is not satisfactory and the use of dedicated discrimination-aware techniques is justified. (4) Our methods significantly improve the current stat-of-the-art techniques w.r.t. accuracy-discrimination trade off.

7 Conclusions

In this paper we presented the construction of a decision tree classifier without discrimination. This is a different approach of addressing the discrimination-aware classification problem. Most of the previously introduced approaches were focused on "removing" undesired dependencies from the training data and thus can be considered as "preprocessors". In this paper on the contrary, we propose the construction of decision trees with with non-discriminatory constraints. Especially relabeling, for which an algorithm based on the KNAPSACK problem was proposed, showed promising results in an experimental evaluation. It was shown to outperform the other discrimination aware techniques by giving much lower discrimination scores and maintaining the accuracy high. Moreover, it is shown that if we are only concerned with accuracy, our method is the best choice when training set is discriminatory and test set is non-discriminatory. 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, non-discriminatory constraints as introduced in this paper are too strong: consider for example that often it is acceptable from an ethical and legal point of view to have a correlation between the sex of a person and the label given to him or her, as long as it can be explained by other attributes. Consider, e.g., a 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 insured. 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.

8 Acknowledgments

This work was supported by the Netherlands Organisation for Scientific Research (NWO) grant *Data mining without Discrimination* (KMVI-08-29) and by Higher Education Commission (HEC) of Pakistan.

References

1. A. Asuncion and D. Newman. UCI machine learning repository. 2007.
2. G. Ausiello, P. Crescenzi, G. Gambosi, V. Kann, A. Marchetti-Spaccamela, and M. Prosati. *Complexity and Approximation. Combinatorial Optimization Problems and Their Approximability Properties*. Springer, 2003.
3. T. Calders, F. Kamiran, and M. Pechenizkiy. Building classifiers with independency constraints. In *IEEE ICDM Workshop on Domain Driven Data Mining*. IEEE press., 2009.
4. T. Calders and S. Verwer. Three naive bayes approaches for discrimination-free classification (accepted for publication). In *Proc. ECML/PKDD*, 2010.
5. W. Duivesteijn and A. Feelders. Nearest neighbour classification with monotonicity constraints. In *Proc. ECML/PKDD'08*, pages 301–316. Springer, 2008.
6. Dutch Central Bureau for Statistics. Volkstelling, 1971.
7. Dutch Central Bureau for Statistics. Volkstelling, 2001.
8. C. Elkan. The foundations of cost-sensitive learning. In *Proc. IJCAI'01*, pages 973–978, 2001.
9. F. Kamiran and T. Calders. Classifying without discriminating. In *Proc. IC409*. IEEE press.
10. F. Kamiran and T. Calders. Classification with no discrimination by preferential sampling. In *Proc. BENELEARN*, 2010.
11. W. Kotlowski, K. Dembczynski, S. Greco, and R. Slowinski. Statistical model for rough set approach to multicriteria classification. In *Proc. ECML/PKDD'07*. Springer, 2007.
12. D. Margineantu and T. Dietterich. Learning decision trees for loss minimization in multi-class problems. Technical report, Dept. Comp. Science, Oregon State University, 1999.
13. S. Nijssen and E. Fromont. Mining optimal decision trees from itemset lattices. In *Proc. ACM SIGKDD*, 2007.
14. D. Pedreschi, S. Ruggieri, and F. Turini. Discrimination-aware data mining. In *Proc. ACM SIGKDD*, 2008.
15. D. Pedreschi, S. Ruggieri, and F. Turini. Measuring discrimination in socially-sensitive decision records. In *Proc. SIAM DM*, 2009.