Discrimination-aware classification

Discrimination-aware Classification

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de
Technische Universiteit Eindhoven, op gezag van de
rector magnificus, prof.dr.ir. C.J. van Duijn, voor een
commissie aangewezen door het College voor
Promoties in het openbaar te verdedigen
op dinsdag 11 oktober 2011 om 16.00 uur

door

Faisal Kamiran

geboren te Burewala, Pakistan

Dit proefschrift is goedgekeurd door de promotor:

prof.dr. P.M.E. De Bra

Copromotor:
dr. T.G.K. Calders

A catalogue record is available from the Eindhoven University of Technology Library
ISBN: 978-90-386-2789-2
Kamiran, Faisal
Discrimination-aware Classification. -
Eindhoven : Technische Universiteit Eindhoven, 2011.
NUR 984
Subject headings: data mining; databases; artificial intelligence
CR Subject Classification: H.2.8, I.5.2, I.2.6

Eerste promotor: prof.dr. P.M.E De Bra (Technische Universiteit Eindhoven)

Copromotor: dr. T.G.K (Toon) Calders (Technische Universiteit Eindhoven)

Kerncommissie:
prof.dr.ir. W.P.M. van der Aalst (Technische Universiteit Eindhoven)
prof.dr. B. Goethals (Universiteit Antwerpen)
prof.dr. D. Pedreschi (Universiteit Pisa)

The research reported in this thesis is supported by Higher Education Commission of Pakistan and has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.
SIKS Dissertation Series No. 2011-29

Printing: Ijaz Riaz Printers
Cover design: Farukh Manzoor

Acknowledgements

In February 2008, I was about to leave for the Netherlands for my doctoral studies in the Databases and Hyper Media group. It was a difficult decision to take; on the one hand the challenge of joining a top research institution, and on the other hand to leave my family back home in Pakistan. I opted for my doctoral studies and left for Netherland with great optimism. However, soon my optimism evaporated and I discovered that it was not easy at all to live away from my family. Moreover, it was very demanding to meet the international research standards and I did have some gaps in my research background. I had no other option except to bury myself in studies and that's what I did. I started to work really hard on my research to divert my attention from home sickness and to meet the international standards. My research became my refuge abroad.

Firstly, I would like to thank my supervisors dr. Toon Calders and prof. dr. Paul De Bra for their continuous guidance during my PhD studies and research. Toon, I always say you are a great advisor and a refined researcher as well. I learned a great deal from you; how to think, write and deliver scientifically. During my PhD research I also found you a scientist of a high caliber. One thing I would like to mention in this regard: the selection of my research topic "Discrimination-Aware Data Mining"; almost fifteen research papers have already been published at top data mining venues on this topic. I am also very impressed with your teaching methodology, you made this complex work easy and relaxed for me. Publishing research papers at top venues and doing high quality research was a dream for me but your motivation, encouragement and intelligent supervision made it very interesting and manageable. Paul, it was an uphill task for me to complete my PhD, but you made it a smooth sailing for me and to tell you the truth, it would have been impossible without your help and support. Your lenient policy gave me a lot of independence. You always made things easier which gave me the opportunity to fully concentrate on my research. Your support and encouragement to visit the top data mining events and other research groups was a main source of motivation for me.

I also need to pay my gratitude to my core committee members: prof. Wil van der

Aalst, prof. Dino Pedreschi and prof. Bart Goethals, who agreed to be a part of my core committee, reading and approving the thesis. I would like to thank prof. van der Aalst in particular for his valuable comments, which really helped me to improve the quality of my dissertation. I am also highly grateful to the Higher Education Commission of Pakistan (HEC) for funding my doctoral research in the Netherlands. I would like to thank Tauheed Ahmed for his kind support in the finalization of my agreement with HEC.

I also want to convey my special thanks to the other Information Systems group members for their friendly and supportive attitude during last four years. I would like to thank Indrė Žliobaitė; Indrė, I really enjoyed your collaboration, the discussions we had and your help and sincerity in my job hunt. I would like to thank Mykola Pechenizkiy for our research collaborations and support in getting myself registered for conferences. I would extend my thanks to Asim Karim, Dino Pedreschi, Sicco Verwer, Hoang Thanh Lam for their collaborations during my research work. I also thank to my office-mate J.C. Prabhakara for his continuous support and valuable suggestions. I want to pay my regards and thanks to other colleagues Evgeny Knutov, Jorn Bakker, George Fletcher, and Natalia Stash for joining the data mining meetings, lunches and soccer games. I would like to thank Riet van Buul and Ine van der Ligt for their help during last four years.

I think it is justified to devote a small paragraph to express my gratitude to the people really helped me when I was away from my family. In this regard, I am thankful to Toon Calders and Paul De Bra for allowing me to visit my family so frequently which made my life a lot easier. I would like to thank Mykola Pechenizkiy and Katja Vasilyeva again for their love and support in this context. Special thanks to Riet van Buul, for making the presence of my wife and daughter possible at my defense ceremony. Riet, it's only you, who made it possible, I am indebted to you for this effort, it has joined us in a lifelong bond.

I would also like to thank my Pakistani friends, who helped me to settle down in the Netherlands. Akhter Hussain, I really enjoyed your cooking expertise and discussion on different matters. M. Atif, you were a blessing as a neighbor, thank you for your help and sincere advice on many occasions. Yahya, thanks for entertaining me during the time of utter despair and your help during my poor cooking sessions. I also thank Naveed Ahmed Gill, Saeed Ahmed, Dil Ahmed, Asif Mahmood, Adnan Haider, Abu Zar Sibtain Shah and other colleagues, who supported me to take a decision for doctoral studies in the Netherlands. The list is very long and it is not possible to mention all of them. I thank you all, who supported me during last four years.

My deepest gratitude is to my family who has enabled me to be what I am today. My mother, Saleem Akhtar is a strong individual, who always shielded us from the extreme situations my family went through after the death of my father. It was

her courage and love which made us strong to fight the hazardous passages. My success is the result of her prayers and invaluable affection. My father, Maqbool Ahmed, always wanted me to be a well-read person, he could not see what he always dreamt, but I believe he is seeing it in heavens. My elder brother M. Amir and his wife Saima Amir who supported my wife and kids during my stay in Netherland, I can't give words to their love and affection for me and my kids. I am really indebted to you. I am also thankful to my sisters for their care and unconditional love.

Last but not the least, my love and thanks to my wife, Rashida Faisal and kids; Affan Faisal, Navian Ahmed, Rania Faisal. They have shown a great courage and determination during the last four years. Their strength and sincerity of my wife was my driving force during my stay in the Netherland. My wife went through a great personal loss, as she lost her mother during this period. I would like to pay regard to my late mother-in-law, who was a source of inspiration in my family. I am also thankful to my in-laws for their moral support and encouragement to my wife in my absence. If my wife and kids would have behaved otherwise, my doctoral studies would never have completed. It was their love and strength which made me go through this period and I was able to complete my studies and research quite well in time.

Thank you all!

Contents

Chapter 1

Introduction to the Discriminaton-aware Classification Problem

Due to the advancement of technology for data generation and data collection terabytes of data are being generated daily in many organizations. With the rapid increase in the volumes of data, it is important to have data mining techniques to discover the hidden useful patterns from the large volumes of data.

Data mining refers to the extraction of knowledge from large amounts of data. It is a process that finds important pieces of knowledge from huge amounts of raw data. More precisely we define data mining as *the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data sets [41].*

Many people refer to data mining with slightly different terms such as knowledge mining from databases, pattern analysis, data archeology, knowledge extraction and knowledge discovery in databases (KDD). Alternatively, others view data mining as a step in the process of knowledge discovery in databases. Data mining is an interdisciplinary field which is closely related to database systems, statistics, machine learning, visualization, and information science.

Data mining is a relatively new field that has received a lot attention in the last few years from the research community. Many data mining techniques have been developed till now. These methods can be broadly classified into clustering (dividing a given dataset into logical homogeneous groups), pattern mining (the discovery of trends, or patterns in a given dataset) and classification (learning to predict class of data objects based on already labeled examples). Clustering and pattern mining are often referred to as unsupervised methods as they only require data, whereas supervised methods, such as classification, require labeled data. Clustering is often used in situations in which we have no idea about categories of the data objects and we try to automatically assign the objects to groups on the basis of the similarity of their characteristics. The second type of data mining technique, pattern mining, is used for detecting patterns and associations in the given dataset. This technique is again unsupervised in the sense that no target label is given. Instead, it aims at the detection of unusual relations between attribute- values. In this thesis, we focus on classification techniques only [40, 35, 84].

1.1 Classification

Databases are often rich of hidden information that can be used for intelligent decision making. Classification is an important form of data analysis that can be used to build models describing important data classes. Categorization of data into different classes by classification helps us with a better understanding of the

data. The goal of classification is to accurately predict the target class for each data object with unknown class label. In the classification model learning process, we start with a given data set with already known class labels. For example, a classification model to classify loan applicants as low or high credit risk could be developed based on observed data for many loan applicants over a period of time. In addition to the credit rating history, the data may have some other useful information about the loan applicants such as employment history, postal code, age, income, occupation, and weekly working hours. A classification model will select the most important attributes to infer classification rules for future decision making. In this loan application example, credit rating would be the *target* or *class*, the other attributes would be the *predictors* or *features*, and the data for each loan applicant would be considered as one data object. For instance in Figure 1.1, we show a simple decision tree learnt from labeled historical data to classify future loan applicants into high and low risk classes.

In the model building process, a classification algorithm finds relationships between the values of the predictors and the values of the target. For example, the decision tree given in Figure 1.1 determines the credit risk category on the basis of age, income and employment of loan applicants. Different classification algorithms use different techniques for finding these relationships. The relationships are summarized in a model, which can then be applied to label objects of which the class assignments is unknown.

Typically the historical (given) data for learning a classification model is divided into two data sets: one for building the model which is referred to as *training set* and an other for testing the model which is referred to as *test set*. The performance of a classifier is judged by its accuracy scores over the test set. *Accuracy* refers to the percentage of correct predictions made by the model when compared with the actual class labels in the test data.

Classification has many applications in business modeling, marketing, credit analysis, and biomedical and drug response modeling. The desired accuracy scores vary from one application to the other. For instance, 90% accuracy may be considered very high in a credit rating application but may be very low in designing a model to predict if one is suffering from cancer or not.

Many classification methods have been proposed by researchers in data mining, machine learning, pattern recognition, and statistics [40]. We will focus on the following data classification techniques: decision tree classifiers [68], bayesian classifiers [29], and k-nearest-neighbor [29].

In this thesis we only focus on binary classification; the target attribute has only two possible values; for example, high credit rating or low credit rating.

Postal-code	Age	Employment	Income	Credit-history	Credit-risk
A	30	No	15K	Good	High
B	35	Yes	50K	Excellent	Low
A	23	No	—	—	High
C	40	Yes	25K	Fair	Low
—	—	—	—	—	—
—	—	—	—	—	—

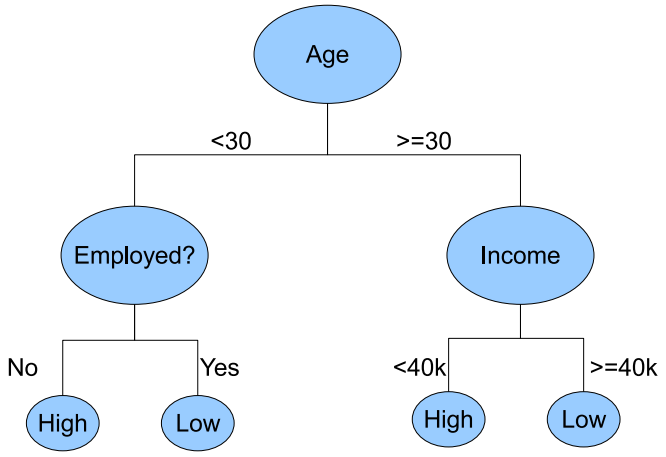


Figure 1.1 A simple decision tree leant from the data set given in the above table.

1.2 Discrimination-aware Classification

The word discrimination originates from the Latin word *discriminare*, which means to *distinguish between*. Discrimination is usually studied in social sciences [42] where it refers to the unfair treatment of individuals of a certain group based solely on the basis of their affiliation with 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.

Classifier construction is one of the most researched topics within the data mining and machine learning communities. Literally thousands of algorithms have been proposed. The quality of the learned models, however, depends critically on the quality of the training data. No matter which classifier inducer is applied, if the training data is incorrect, poor models will result. In this work, we use discrimination in its social sense; we do not want our learnt models to make socially discriminating future decisions. We study cases in which the input data is discriminatory and we want to learn a discrimination-free classifier for future classification. Now we discuss different scenarios where the discrimination-aware classification paradigm is applicable:

Scenario 1: historical discrimination. Such cases occur naturally when, e.g., the decision process leading to the labels was biased due to discrimination as illustrated by the next example [19]: *Throughout the years, an employment bureau recorded various parameters of job candidates. Based on these parameters, the company wants to learn a model for partially automating the match-making between a job and a job candidate. A match is labeled as successful if the company hires the applicant. It turns out, however, that the historical data is biased; for higher board functions, Caucasian males are systematically being favored.* A model learned directly on this data will learn this discriminatory behavior and apply it over future predictions. From an ethical and legal point of view it is of course unacceptable that a model discriminating in this way is deployed.

Scenario 2: multiple data sources. Next to data generated by a deliberately biased process, discrimination in training data also appears naturally when data is collected from different sources; e.g., surveys with subjective questions taken by different enquirers (leading to an indirect discrimination based on the geographical area covered by enquirers). We illustrate this kind of discrimination by this example: *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 record answers to these questions in different ways. Generalizing directly from the training set consisting of all surveys without taking into account these differences among the enquirers may easily result in misleading findings. For example, if many surveys from hospitals*

in 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 this example the non-discrimination constraints are a useful tool to avoid over-fitting the classifier to artifacts by requiring that the learned classifier does not correlate with the enquirer. Other similar cases could be: *different scores given by different reviewers, movie ratings of different persons, student grades given by different examiners etc.*

Scenario 3: sensitive attribute as a proxy. In some cases the discrimination in the input data appears when the sensitive attribute serves as a proxy of features that are not present in the dataset. With respect to this last case, we quote [82]: *"If lenders think that race is a reliable proxy for factors they cannot easily observe that affect credit risk, they may have an economic incentive to discriminate against minorities. Thus, denying mortgage credit to a minority applicant on the basis of minorities on average-but not for the individual in question-may be economically rational. But it is still discrimination, and it is illegal."*

In all these cases it is desirable to have a means to "tell" the algorithm that it should not discriminate on the basis of the sensitive attributes, e.g., sex, ethnicity. Such attributes upon which we do not want the classifier to base its predictions, we call *sensitive attributes*. So in *Discrimination-aware Classification*, we want to learn non-discriminatory classification models from potentially biased historical data such that they generate accurate predictions for future decision making, yet do not discriminate with respect to a given sensitive attribute.

1.2.1 Research Question

Our research question may be stated as: "Is it possible to learn accurate classifiers based upon discriminatory training data that do no discriminate in their predictions?" It raises many sub-questions:

- How can we measure discrimination? (Sections 2.2 and 5.2)
- What is relationship between accuracy and discrimination? (Section 2.4)
- Can we solve the problem by just removing the sensitive attribute from the training data? (Section 1.2.3)
- Can we learn discrimination-free classifiers by removing the discrimination from the training data and then learning classifiers over it? (Chapter 3)
- Can we directly learn discrimination-free models from biased data? (Chapter 4)

1.2.2 Motivation and Anti-discrimination Laws

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

The Australian Sex Discrimination Act 1984 [1]: This act prohibits discrimination in work, education, services, accommodation, land, clubs on the grounds of marital status, pregnancy or potential pregnancy, and family responsibilities. This act defines sexual harassment and other discriminatory practices on different grounds and declares them unlawful. The main objectives of this act are as follows:

(a) *to give effect to certain provisions of the Convention on the Elimination of All Forms of Discrimination Against Women; and*

(b) *to eliminate, so far as possible, discrimination against persons on the ground of sex, marital status, pregnancy or potential pregnancy in the areas of work, accommodation, education, the provision of goods, facilities and services, the disposal of land, the activities of clubs and the administration of Commonwealth laws and programs; and*

(ba) *to eliminate, so far as possible, discrimination involving dismissal of employees on the ground of family responsibilities; and*

(c) *to eliminate, so far as possible, discrimination involving sexual harassment in the workplace, in educational institutions and in other areas of public activity; and*

(d) *to promote recognition and acceptance within the community of the principle of the equality of men and women.*

Moreover, this law prohibits indirect and unintentional discrimination. In such cases, it is the responsibility of the accused party to prove that his/her intention was not to discriminate the aggrieved party. We further discuss such kind of discrimination in Chapter 5 and refer it to as *the conditional discrimination*. The importance to avoid from the indirect and unintentional discrimination is very well illustrated from this part of the act: *in a proceeding under this Act, the burden of proving that an act does not constitute discrimination because of section 7B lies on the person who did the act.* Section 7B of this act describes indirect discrimination: *a person does not discriminate against another person by imposing, or proposing to impose, a condition, requirement or practice that has, or is likely to have, the disadvantaging effect mentioned in subsection 5(2), 6(2) or 7(2) if the condition, requirement or practice is reasonable in the circumstances.*

The US Equal Pay Act 1963 [9]: This act requires that men and women in the

same workplace be given equal pay for equal work. The jobs need not to be identical, but they must be substantially equal. This law covers all forms of pay including salary, overtime pay, bonuses, stock options, profit sharing and bonus plans, life insurance, vacation and holiday pay, cleaning or gasoline allowances, hotel accommodations, reimbursement for travel expenses, and benefits. If there is an inequality in wages between men and women, employers may not reduce the wages of either sex to equalize their pay. The act describes it as follows: *No employer having employees subject to any provisions of this section shall discriminate, within any establishment in which such employees are employed, between employees on the basis of sex by paying wages to employees in such establishment at a rate less than the rate at which he pays wages to employees of the opposite sex in such establishment for equal work on jobs the performance of which requires equal skill, effort, and responsibility, and which are performed under similar working conditions, except where such payment is made pursuant to (i) a seniority system; (ii) a merit system; (iii) a system which measures earnings by quantity or quality of production; or (iv) a differential based on any other factor other than sex: Provided, that an employer who is paying a wage rate differential in violation of this subsection shall not, in order to comply with the provisions of this subsection, reduce the wage rate of any employee.*

This act aimed at abolishing wage disparity based on sex. According to the US Bureau of Labor Statistics, women's salaries vis-à-vis men's have risen dramatically since the enactment of this equal pay act, from 62% of men's earnings in 1970 to 80% in 2004 [22]. This real world case illustrates our Scenario 1 (Section 1.2 where our historical data is discriminatory due to a biased data generation process and we are supposed to build discrimination-free classifiers from it.

The US Equal Credit Opportunity Act 1974 [8]: This act 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 [11].

European Council Directive 2004: Even though there is clear historical evidence showing higher accident rates for male drivers, insurance companies are not allowed to discriminate based on gender in many countries. We can illustrates this prohibition by the following ruling of European Court of Justice [2]: *The European Court of Justice decided on March 1, 2011 that, from 21 December 2012, it will no longer be legal under EU law to charge women less for insurance than men. The verdict means that different priced premiums for men and women drivers will now be considered to be in breach of the EU's anti-discrimination rules.* This ruling is the implementation of European Council Directive 2004/113/EC of 13 December

2004 requiring the principle of equal treatment between men and women in the access to and supply of goods and services (adopted unanimously by the EU Council of Ministers). It prohibits direct and indirect sex discrimination outside of the labor market.

All of the anti-discrimination laws prohibit discriminatory practices in future. It means that our discrimination-aware classification paradigm clearly applies to these situations. If we are interested to apply classification techniques, and our available *historical* data contains discrimination, it will be illegal to use traditional classifiers without taking the discrimination aspect into account due to these anti-discrimination laws.

1.2.3 Redlining

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 [14]. 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 (Section 1.2 scenario 1) or *enquirer* for the survey example (Section 1.2 scenario 2) from the training data often does not solve the discrimination problem, 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.

A parallel can be drawn with the practice of *redlining*: denying inhabitants of certain racially determined areas from services such as loans. It describes the practice of marking a red line on a map to delineate the area where banks would not invest; later the term was applied to discrimination against a particular group of people (usually by race or sex) no matter the geography. During the heyday of redlining, the areas most frequently discriminated against were black inner city neighborhoods. Through at least the 1990s this practice meant that banks would often lend to lower income whites but not to middle or upper income blacks[1], i.e., the deci-

[1] Source: `http://en.wikipedia.org/wiki/Redlining`, March 7th, 2011

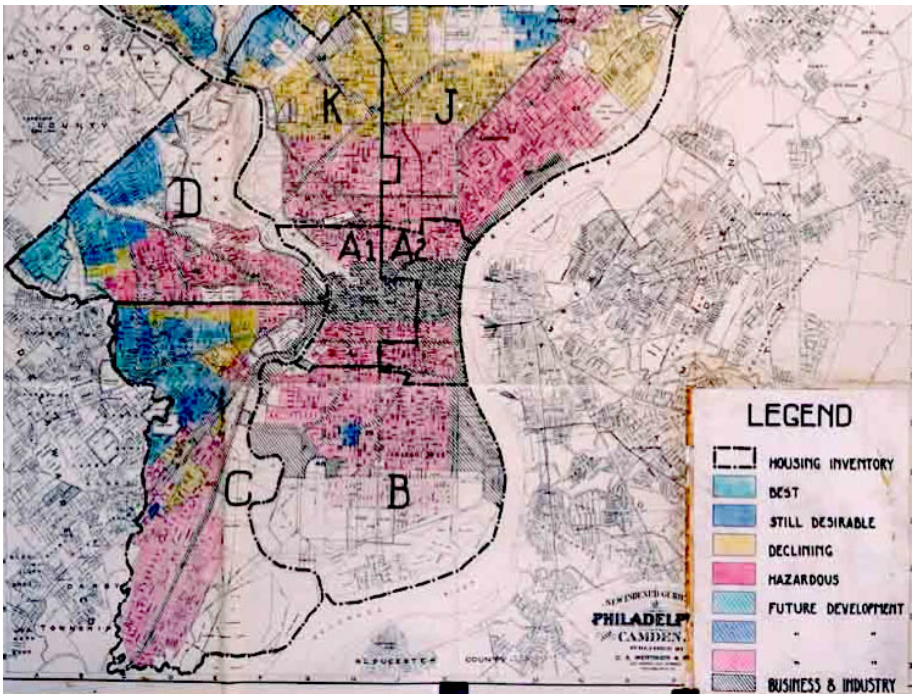


Figure 1.2 A house owners' loan corporation 1936 security map of Philadelphia showing redlining of lower income neighborhoods. Households and businesses in the red zones could not get mortgages or business loans.

sions of banks were discriminatory towards black loan applicants. Figure 1.2 shows a house owner loan corporation (HOLC) 1936 map[2] which illustrates that instead of directly using the ethnicity for loan decision making, different areas were used for decision making. Certain areas which were mostly inhabited by low income blacks and other such ethnic groups were marked in red over the map. Table 1.1 shows the description of different areas in the section J of the map. It shows that even the house value of areas, with negro majority, was reasonably high but their residential area was marked hazardous (red) on the map for loans. At the same time, the neighborhoods with native whites were considered desirable or highly desirable for the loans, even though the house values were not that high in their areas. So it shows indirect discrimination towards colored people by using their residential areas.

[2]Source: `http://cml.upenn.edu/redlining/HOLC_1936.html`, March 7th, 2011

Table 1.1 Description of different areas shown in section J of the map of Figure 1.2 and their impact over loan applications.

Area	House value ($)	Inhabitants	Loan Category
Red	2000-10000	Negro (predominating)	Hazardous
Yellow	2000-6000	Laborers and workers	Def_declining
Blue	3500-7000	White collar native whites	Still desirable
Green	5000-8000	Good class native whites	Best

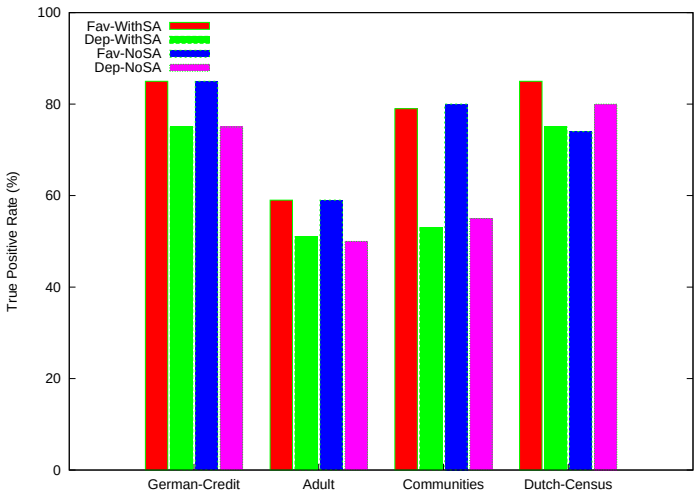
Redlining and Real World Datasets: We further explore this impact of redlining over some dataset which we use in our experiments. We observe that the removal of the discriminatory attribute does not solve the problem of discrimination because the learned model still discriminates due to the redlining effect. The discrimination goes down only in those datasets where the sensitive attribute is weakly correlated to other attributes in the data. We will discuss this effect in more detail later.

Figure 1.3 (a) gives the True Positive (TP) rate and Figure 1.3 (b) gives the True Negative (TN) rate for both favored, e.g., male and deprived, e.g., female communities. Furthermore, in both figures we give the results of experiments when we learn decision tree learners over the Adult Dataset [14] with and without using the sensitive attribute.We calculate the TP rate for the favored community by

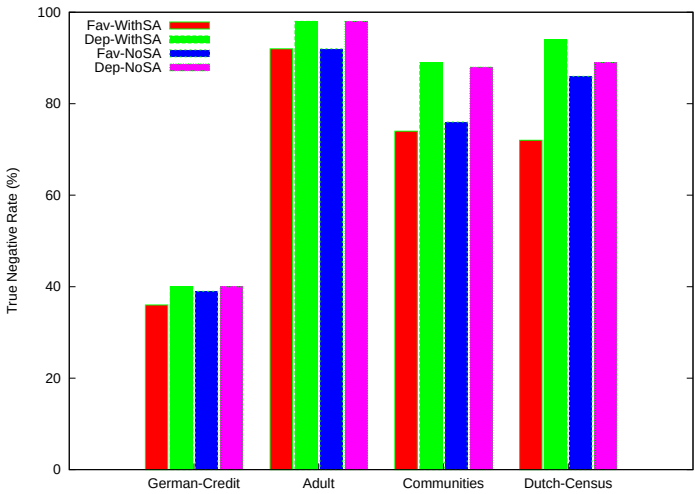
$$P(a\ classifier\ assigns\ positive\ label|positive, favored\ community)\ .$$

Similarly we calculate the TP rate for the deprived community and the TN rates for the favored and deprived communities by replacing the positive class with negative class. We make following important observation from Figure 1.3:

- We observe that the true positive rate for the favored community is higher than that of the deprived community while the true negative rate for the favored community is lower than that of the deprived community. This difference is due to the effect of discrimination, the classifier learnt over discriminatory data shows a biased attitude towards the deprived community and tends to assign more negative class labels to them.
- We can observe from the results of these experiments that the deprived community gets more disadvantage than its actual share.
- Just the removal of sensitive attribute does not solve this problem and we will have to use some sophisticated techniques to neutralize this discriminatory effect.



(a) True positive (TP) rate comparison



(b) True Negative (TN) rate comparison

Figure 1.3 TP rate for the favored community and TN rate for the deprived community are both higher; the removal of sensitive attribute has a very little effect due to redlining effect.

1.3 Solutions

Our proposed solutions to the discrimination problem fall into two broad categories. First, we propose pre-processing methods to remove the discrimination

from the training dataset. On this cleaned dataset then a classifier can be learned. Our rationale for this approach is that, since the classifier is trained on discrimination-free data, it is likely that its predictions will be (more) discrimination-free as well. The empirical evaluation confirms this statement. In these preprocessing methods, our first approach, called *Massaging the data*, is based on changing the class labels in order to remove the discrimination from the training data. The second approach, called *Reweighing*, is less intrusive as it does not change the class labels. Instead, weights are assigned to the data objects to make the dataset discrimination-free. Since reweighing requires the learner to be able to work with weighted tuples, we also propose a third pre-processing method in which we re-sample the dataset in such a way that the discrimination is removed. We refer to this approach as *Sampling*.

Second, we propose solutions to the discrimination problem by directly pushing the non-discrimination constraints into classification models and by post-processing learned models. We propose two solutions to construct decision trees without discrimination. The first solution is based on the adaptation of the splitting criterion for tree construction to build a discrimination-aware decision tree. The second solution is post-processing of decision trees with discrimination-aware pruning and relabeling of tree leaves, for which an algorithm based upon a reduction to the KNAPSACK [60] problem is given. It is shown to outperform the other discrimination aware techniques by giving significantly lower discrimination scores and maintaining high accuracy.

We further studied the discrimination-aware classification paradigm in the presence of explanatory attributes that correlate with the sensitive attribute, e.g., decline from the job may be explained by the low education level. In such a case, as we show, not all discrimination can be considered bad, therefore we introduce a new way of measuring discrimination, by explicitly splitting it up into *explainable* and *bad* discrimination and only remove bad discrimination.

1.3.1 Validation

For the validation of our proposed method, we used the well-known data mining tool Weka [39] which is an open source software issued under the GNU General Public License [80]. Weka is a collection of machine learning algorithms for data mining tasks. Different data mining methods for data pre-processing, classification, regression, clustering, association rules, and visualization have been implemented and added to Weka. For the fair comparison of our developed method to the standard data mining techniques we have incorporated our proposed solutions for

discrimination-aware classification problem into Weka. We refer to this version of Weka with discrimination-aware classification methods as *Discrimination-aware Weka*. It does not only give us an opportunity to fairly compare our developed techniques to the standard ones but it also enables us to use our method in arbitrary combination with the standard methods. In this way we explore how our methods affect the performance of the current state-of-the-art methods when used both in combination with or in isolation of the standard data mining techniques.

Experimental Set-up:

All reported empirical results in this thesis were obtained using **10-fold cross-validation** and reflect the true accuracy; that is, on **unaltered data (no discrimination removal technique is applied)**. Figure 1.4 shows a detailed representation of our experimental setup. We can observe in Figure 1.4 that we apply, in each iteration of the cross-validation, our proposed discrimination removal methods only to the folds for training and not to the test fold. We use this preprocessed training set for learning a classifier or directly learn a non-discriminatory classification model and evaluate this learnt classifier over the test fold of this iteration. The predictions for the test-fold are stored. We repeat this process for all folds and append all predictions on the test sets over all folds. Based on the predictions and the true class we calculate the final accuracy and discrimination scores. It is also important to notice that **no parameter tuning** was performed; all experiments were done in Weka with their default parameter settings.

Datasets: In our experiments we used the **Adult** dataset and the **Communities and Crimes** dataset which are available in the UCI ML-repository [14] and two **Dutch Census datasets of 1971 and 2001** [31, 32].

Adult Dataset

The Adult dataset has 48 842 instances and contains demographic information of people. The associated prediction task is to determine whether a person makes over 50K per year or not; i.e., income class *High* or *Low* will be predicted. We denote income class *High* as a desired class and income class *Low* as not desired class. Each data object is described by 14 attributes which include 8 categorical and 6 numerical attributes. We excluded the attribute *fnlwgt* from our experiments (as suggested in the documentation of the dataset). This dataset is a collection of 51 (US) state samples and people with similar demographic characteristics get similar values for this attribute *fnlwgt* in each state. This attribute is only useful if we work with a sample from only one state because people from multiple states would have inconsistent values for this attribute. The other attributes in the dataset include:

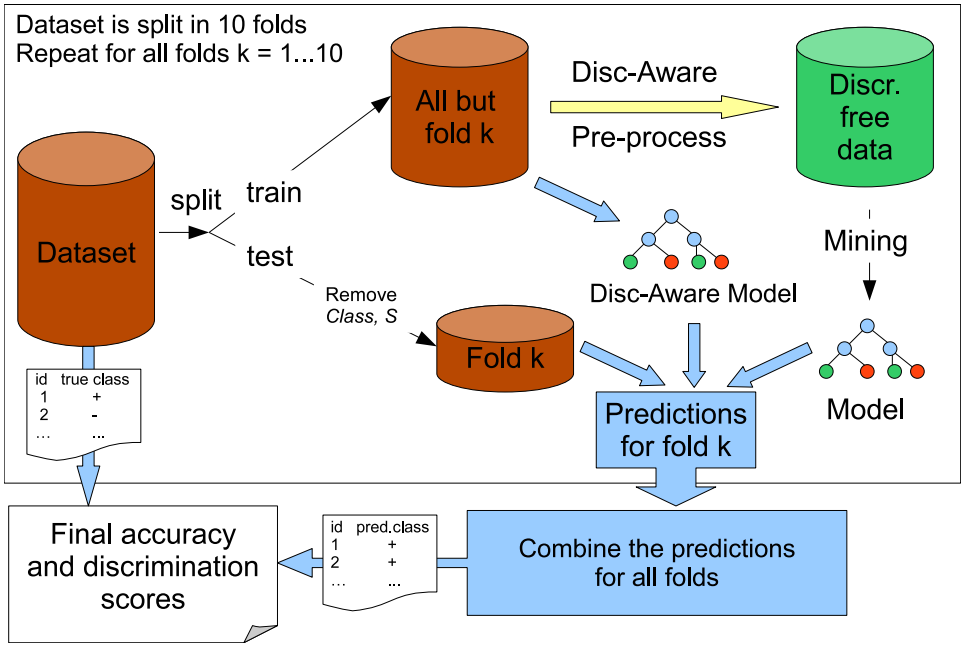


Figure 1.4 10-fold cross-validation experimental setup.

age, type of work, education, years of education, marital status, occupation, type of relationship (husband, wife, not in family), sex, race, native country, capital gain, capital loss and weekly working hours. We use Sex as discriminatory attribute. In our sample of the dataset, 16 192 citizens have $Sex = f$ and 32 650 have $Sex = m$.

Communities and Crimes Dataset

The Communities and Crimes 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.

Dutch Census Datasets

We also apply our proposed techniques to two Dutch census datasets of 1971 and 2001 [31, 32]. The *Dutch Census 2001* dataset has 189 725 instances representing aggregated groups of inhabitants of the Netherlands in 2001. The dataset is described by 13 attributes namely *sex, age, household position, household size, place of previous residence, citizenship, country of birth, education level, economic status (economically active or inactive), current economic activity, marital status, weight and occupation*. We removed the records of underage people, some middle level professions and people with unknown professions, leaving 60 420 instances for our experiments. We use the attribute *occupation* as a class attribute with values "high level" (prestigious) and "low level" professions. We use the attribute *sex* as sensitive attribute. The *Dutch 1971 Census dataset* is comparable to the Dutch 2001 census dataset and consists of 159 203 instances. It has the same features except for the attribute *place of previous residence* which is not present in the 1971 dataset, and an extra attribute *religious denominations*. After removing the records of people under the age of 19 and records with missing values, 99 772 instances remained for our experiments. All the attributes are categorical except *weight* (representing the size of the aggregated group) which we excluded from our experiments.

All datasets and the source code of all implementations reported upon in this thesis are available at `https://sites.google.com/site/faisalkamiran/`.

1.3.2 Practical Relevance

A recently started collaboration with *WODC* (the study center of the Dutch Department of Justice), and *CBS* (the Dutch Central Bureau for Statistics) is an important source of motivation to study the problem of discrimination. These agencies support policy making on the basis of demographic and crime information they have. Their interest emerges from the possibility of correlations between ethnicity and criminality that can only be partially explained by other attributes due to data incompleteness (e.g., latent factors). Learning models and classifiers directly on such data could lead to discriminatory recommendations to the decision makers. Removing the ethnicity attributes would not solve the problem due to the redlining effect, but rather aggravate it, as the discrimination still would be present, only it would be better hidden. In such situations our discrimination-aware data mining paradigm clearly applies.

1.4 A Quick Overview of the Thesis

Figure 1.5 gives a quick overview of the organization of this thesis. In Chapter 2 we formally define the problem statement and make a theoretical analysis of the trade-off between accuracy and discrimination.

In Chapter 3, we propose three data pre-processing techniques for the solution of the discrimination problem. These solutions are empirically evaluated over real world datasets. The discrimination-aware techniques discussed in this chapter are published in: IEEE conference on computer, control and communication [46]; Benelux conference on artificial intelligence [47]; the annual machine learning conference of Belgium and The Netherlands [48], and domain driven data mining workshop of IEEE international conference on data mining [19].

In Chapter 4, we advances our solution to the discrimination problem by directly incorporating the non-discrimination constraints into the classification model learning. In this chapter our solution to the problem is based on the modifying the splitting criterion of a decision tree learner. We also introduce a decision tree leaf relabeling approach to make an already built decision tree discrimination-free. We draw a parallel between our leaf relabeling approach and the well-know combinatorial problem Knapsack. These methods are published in IEEE international conference on data mining [49]. Later a detailed version is published as a technical report at Eindhoven university of technology [50]

In Chapter 5 we extend our problem to the conditional non-discrimination problem.

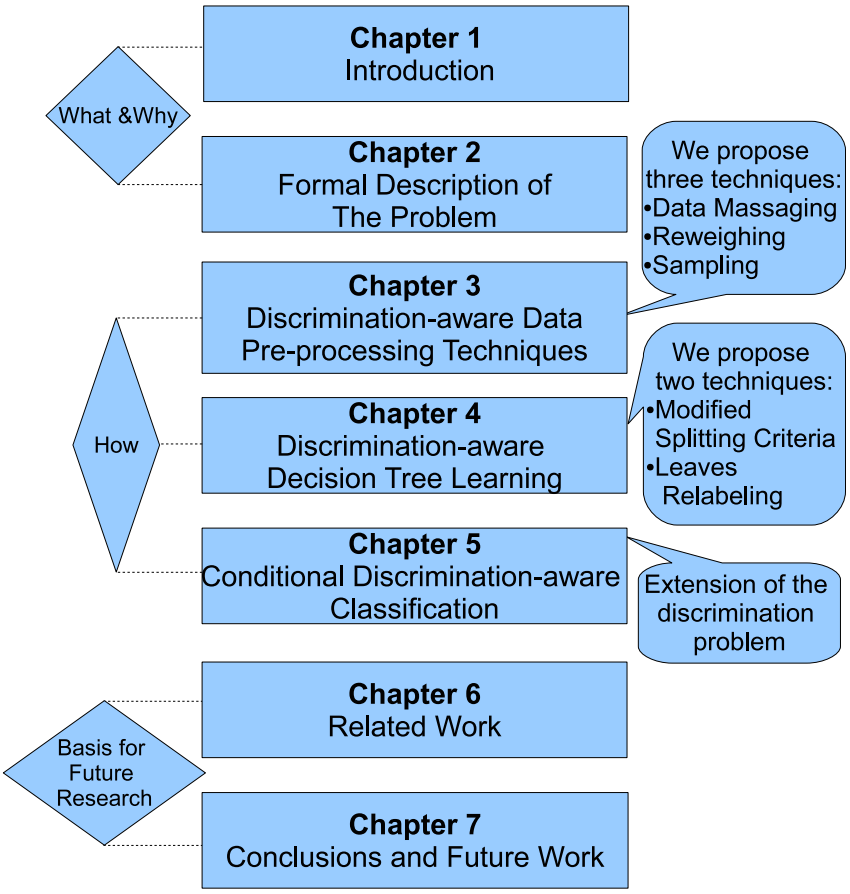


Figure 1.5 Thesis overview.

We discuss the discrimination problem from a different perspective. We introduce that not all the discrimination is always bad. A part of the discrimination may be acceptable in some situations. We refer to this acceptable discrimination as explainable discrimination. We develop local variants of the global massaging and

sampling methods to solve the conditional non-discrimination problem. This work is accepted for publication in IEEE international conference on data mining [86].

In Chapter 6 we give a comprehensive overview of the related work of the discrimination problem and Chapter 7 concludes the work and gives directions for further research.

Chapter 2

Formal Description of the Discrimination-aware Classification Problem

In this chapter we give a formal description to the discrimination-aware classification problem, introduce the important notations that we use through out this thesis, and introduce methods to quantify the discrimination in a given dataset or in the predictions of a classification model. We also give a discrimination model to unveil the regions with high discrimination level and use this model to support the rationale of our proposed methods in the next chapters. Finally, we analytically study the relationship of discrimination and accuracy.

2.1 Preliminaries

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

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

2.2 Discrimination Measurement

We define the discrimination in the following way:

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

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

That is, the difference of the probability of being in the positive class between the tuples having $X(S) = w$ in D and those having $X(S) = b$ in D.

Table 2.1 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	+

(When clear from the context we will omit $S = b$ from the subscript.)

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

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

where $C(X)$ denotes the prediction of the classifier C for a data object X. The discrimination in classifiers's predictions is the difference of the probability of being assigned the positive class by the classifier between the tuples having $X(S) = w$ in D and those having $X(S) = b$ in D. (When clear from the context we will omit $S = b$ from the subscript.)

Example 1 *In Table 2.1, an example dataset is given. This dataset contains the Sex, Ethnicity, and 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 w.r.t. the attribute Sex and Class is: $disc_{Sex=f}(D) := \frac{4}{5} - \frac{2}{5} =$*

40% *. It means that in the dataset, a female is, in absolute numbers,* 40% *less likely to have a job than a male.*

Example 2 *Now we use our discrimination measure to calculate the discrimination in the Adult dataset discussed in Section 1.3.1 of Chapter 1. In the Adult dataset the associated prediction task is to determine whether a person makes over 50K per year or not; i.e., income class* High *or* Low *will be predicted. We denote income class* High *as* $+$ *and income class* Low *as* $-$*. We use the attribute* Sex *as sensitive attribute. If we apply discrimination to calculate the bias toward females for* $+$ *class, we the discrimination is as high as 19.45%:*

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

2.2.1 Motivation for the Discrimination Measure

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

t-test (assuming unknown and potentially different variances) is:

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

where s_1 and s_2 denote the empirical standard deviations for the acceptance of the two groups and n_1 and n_2 denote the respective size of the groups. The statistical test, however, only tells us if there is discrimination, but does not indicate the severity of discrimination. For instance, if we calculate information gain between sex and job decisions. It will just tell us whether the decision making is dependent over the sex of the applicants or not. It will not quantify that how much dependency of decision making over sex is due to discrimination. In this respect notice that the test statistic for the hypothesis $h_0 : \ m_1 - m_2 = d_0$ is:

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

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

2.3 Discrimination-aware Classification

The problem we study in the thesis is now as follows:

Problem 1 ***Classifier with non-discrimination constraint:*** *Given a labeled dataset D, an attribute S, and a value $b \in dom(S)$, learn a classifier C such that:*

(a) the accuracy of C for future predictions is high; and

(b) the discrimination of new examples classified by C is low.

Clearly there will be a trade-off between the accuracy and the discrimination of the classifier. In general, lowering the discrimination will result in lowering the accuracy and vice versa. This trade-off is further elaborated upon in the Section 2.4.

2.3.1 Discrimination Model

In this section we discuss how the discrimination affects the decision making and which regions or objects are the most vulnerable from the discriminatory effect. For this purpose, we analyse the discrimination problem in relation to experimental findings in social sciences reported in [42] we assume that discrimination happens in the following way. The historical data originates from human decision making, which can be considered as a classifier C. That classifier consists of three main parts:

1. a function from attributes to a score $r = f(X')$, where $X' = X \setminus \{S\}$, i.e., X' does not include the sensitive attribute;

2. a discrimination bias function $B(S) = \begin{cases} d & \text{if } S = w \\ -d & \text{if } S = b \end{cases}$;

3. the final decision making function $y = \mathcal{C}\left(f(X') + B(S)\right)$.

According to this model a decision is made in the following way. First the qualifications of a candidate are evaluated based on attributes in X' and a preliminary score is obtained $r = f(X')$. The qualifications are evaluated objectively. Then the discrimination bias is introduced by looking at the gender of a candidate and either adding or subtracting a fixed bias from the qualification score, to obtain $r^* = f(X') + B(s) = f(X') \pm d$. The final decision is made by $\mathcal{C}(r^*)$. Decision making can have two major forms: online and offline. With the offline decision the candidates are ranked based on their scores r^*, and n candidates that have the highest scores are accepted. With the online decision an acceptance threshold t is set, the incoming candidates that have the score $r^* > t$ are accepted.

This discrimination model has two important implications. First, the decision bias is more likely to influence the individuals that are close to the decision boundary according to their score r. If an individual is far from the decision boundary, then adding or subtracting the discriminatory bias d does not influence the final decision. This observation is consistent with experimental findings how discrimination happens in practice [42].

Second, there might be attributes within X that are correlated with the sensitive attribute S. These attributes will affect the initial score r. When observing the decisions it would seem due to correlation that the decision is using S, i.e., $B(S)$ will already be present in the initial score r.

2.3.2 Assumptions

In this thesis we are making two strong assumptions:

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

A2 Ideally the learned classifier should not use the attribute S to make its predictions. However, we show in our experiments that our proposed discrimination-aware methods give promising results with and without using the sensitive attribute.

2.4 Theoretical Analysis of the Accuracy - Discrimination Trade-Off

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

Definition 3 *Let C and C' be two classifiers. We say that C* dominates *C' if the accuracy of C is larger than or equal to the accuracy of C', and the discrimination of C is at most as high as the discrimination of C'. C* strictly dominates *C' if one of these inequalities is strict.*

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

For reasons of simplicity, in our theoretical exposition we assume that a dataset D is given against which discrimination and accuracy of all classifiers is measured. This assumption is not limiting our theoretical results since all our results still obtain when the cardinality of D is infinite; i.e., we can think of D as a perfect

description of the true underlying probability distribution. We will use $\mathcal{C}_{all}$ to denote the set of all classifiers and $\mathcal{C}^*_{all}$ to denote the set of all classifiers C such that $P(X(Class) = +|X \in D) = P(C(X) = +|X \in D)$; i.e., all classifiers that have the same overall probability of assigning the positive label as observed in D.

2.4.1 Perfect Classifiers

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

Let D_b and D_w be defined as follows:

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

and let d_b and d_w be respectively $|D_b|$ and $|D_w|$. d denotes $|D|$. The following theorem gives us some insight in the trade-off between accuracy and discrimination in perfect classifiers, namely those that are DA-optimal in the set of *all* classifiers, and those that are DA-optimal in the set of all classifiers that does not change the class distribution:

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

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

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

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

Let C be a DA-optimal classifier. We denote the number of true negatives, true positives, false positives and false negatives of C by respectively tn, tp, fp, and fn; e.g., $tp = |\{X \in D \mid X(Class) = C(X) = +\}|$. tp_b denotes the number of

true positives that have $S = b$. tp_b, fp_b, ..., and fn_w are defined similarly. With these conventions, we can express the accuracy and discrimination of C as follows:

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

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

Minimize

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

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

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

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

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

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

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

- **Case 1:** $[p_b - tp_b > fp_w]$

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

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

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

- **Case 2:** $[p_b - tp_b \leq fp_w]$

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

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

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

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

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

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

$$fp_b = fn_w \ .$$

From these equalities the theorem now easily follows. □

As was claimed before, there is a trade-off between the accuracy of the DA-optimal classifiers and their discrimination. This trade-off is linear; lowering the discrimination level by 1% results in an accuracy decrease of $\min(d_b, d_w)\%$ and an accuracy decrease of $2d_b d_w\%$ if the class distribution needs to be maintained. These DA-optimal classifiers can be constructed from the perfect classifier.

2.4.2 Imperfect Classifiers

In the last theorem we assumed a perfect classifier. In most cases, however, we will only have an imperfect classifier at our disposal. We will now assume that we have such an imperfect classifier C of which we want to reduce its discrimination by randomly changing some of its predictions. The probability with which we will change a prediction of an instance X, will depend on $X(S)$ and $X(Class)$ only. We will denote these four probabilities by p_{b+}, p_{b-}, p_{w+}, and p_{w-}. The resulting classifier is denoted $C[p_{b+}, p_{b-}, p_{w+}, p_{w-}]$; i.e., $C[p_{b+}, p_{b-}, p_{w+}, p_{w-}](X)$ equals $C(X)$ with probability $p_{X(S)C(X)}$. Notice that the accuracy and discrimination of

this random classifier in fact represents the expected accuracy and discrimination of all deterministic classifiers with $p_{b+}, p_{b-}, p_{w+}, p_{w-}$ correspondence with C. We will denote the class of all classifiers that can be derived from C in this way by $\mathcal{C}_C$. $\mathcal{C}_C^*$ will denote all classifiers C' in $\mathcal{C}_C$ for which it holds that $P(C'(X) = +) = P(C(X) = +)$ The following theorem characterizes the DA-optimal classifiers of $\mathcal{C}_C$ and of $\mathcal{C}_C^*$.

Theorem 2 *The classifier C' is DA-optimal in $\mathcal{C}_C$ iff*

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

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

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

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

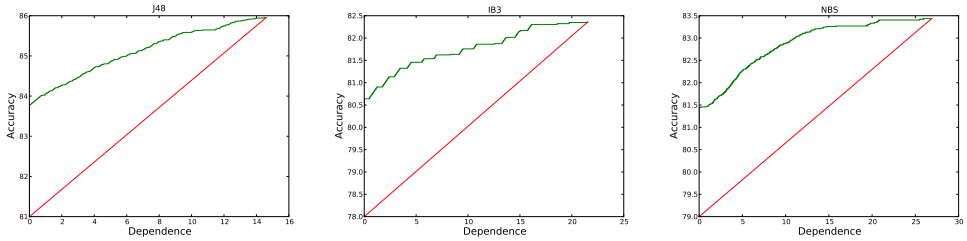
We assume, without loss of generality, that $acc(C) \geq 0.5$; if this is not the case, we switch all predictions of C to obtain a new classifier with an accuracy of $1 - acc(C)$. Let now C' be any classifier with $corr(C, C') = \gamma$; i.e., C and C' agree (correspond) on a fraction γ of the dataset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

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

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

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

On the other hand, we can use Theorem 1 to find the relation between the maximal correspondence with C and the discrimination of the classifier C'; the maximal



(a) Decision tree (b) IBk (c) Naive Bayes

Figure 2.1 Trade-off between accuracy and discrimination (dependence) for the DA-optimal classifiers in $\mathcal{C}_R$ and $\mathcal{C}_C$.

reduction in discrimination linked to the minimal reduction in correspondence is as follows:

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

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

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

Combining these two facts, leads directly to the theorem. □

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

Classifiers based on rankers. On the bright side, however, most classification models actually provide a score or probability for each tuple for being in the positive class instead of only giving the class label. Such a scoring classifier, called a ranker, actually ranks the objects according to its assessment of the probability that the object is in the positive class. The score allows us for a more careful choice of objects of which to change the prediction: instead of using a uniform chance for all tuples with the same predicted class and S-value, the score can be used as follows. Suppose we have such a scoring classifier R that assigns to all objects a score. We can dynamically set different cut-off c_b and c_w for respectively tuples with $S = b$ and $S = w$ to obtain the classifier $R(c_b, c_w)$ that will predict $+$ for a tuple X if $X(S) = b$ and $R(X) \geq c_b$ and if $X(S) = w$ and $R(X) \geq c_w$. Otherwise $-$ is predicted. We denote the class of all classifiers $R(c_b, c_w)$ by $\mathcal{C}_R$. Pseudocode

for $\mathcal{C}_R$ is given in Algorithm 1. Intuitively one expects that slight changes to the discrimination will only incur minimal changes to the accuracy, as the tuples that are being changed are the least certain ones and hence sometimes a change will actually result in a better accuracy. The decrease in accuracy will thus no longer be linear in the change in discrimination, but its rate will increase as the change in discrimination increases, until in the end it becomes linear again, because the tuples we change will become increasingly more certain leading to a case similar to that of the perfect classifier. A full analytical exposition of this case, however, is far beyond the scope of this thesis. Instead we tested this trade-off empirically. The results of this study are shown in Figure 2.1. In this figure the DA-optimal classifiers in the classes $\mathcal{C}_R$ (curves) and $\mathcal{C}$ (straight line) are shown for the Adult dataset, given in Section 1.3.1 of Chapter 1. The three classifiers are a Decision Tree (J48), an instance based classification model with three neighbors (IBk), 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.

Algorithm 1: $\mathcal{C}_R$

Input: (X, c_b, c_w, R)
Output: $X(Class)$

```
if X(S) = b and R(X) ≥ c_b then
    return +
else
    return −
end if
if X(S) = w and R(X) ≥ c_w then
    return +
else
    return −
end if
```

Conclusion. In this section the accuracy-discrimination trade-off is clearly illustrated. It is theoretically shown that if we rely on classifiers, and not on rankers, the best we can hope for is a linear trade-off between accuracy and discrimination. For important classes of classifiers the DA-optimal classifiers were explicitly constructed. Notice, however, that the theoretical solutions proposed in this section violate our assumption A2; the classifiers $C[p_{b+}, p_{b-}, p_{w+}, p_{w-}]$ and $R(c_b, c_w)$

heavily use the attribute S to make their predictions. Therefore these optimal solutions are not suitable for our purposes.

Chapter 3

Data Pre-processing Techniques for Classification without Discrimination

In the previous chapter, we have formally described that the *Discrimination-Aware Classification Problem* is applied to a given situation in which our training data contains discrimination (e.g., gender or racial) and we are supposed to learn a classifier that optimizes accuracy, but does not discriminate in its predictions on the test data. Such situations occur naturally as artifacts of the data collection process when the training data is collected from different sources with different labeling criteria, when the data is generated by a biased decision process, or when the *sensitive attribute* serves as a proxy for unobserved features. In many situations, a classifier that detects and uses the racial or gender discrimination is undesirable for legal reasons.

3.1 Discrimination-aware Techniques

In this section we propose three solutions to learn a non-discriminating classifier that uses the attribute S only during learning not at prediction time. All solutions are based on removing the discrimination from the training dataset. Then a classifier can be learned on this cleaned dataset. Our rationale for this approach is that, since the classifier is trained on discrimination-free data, it is likely that its predictions will be (more) discrimination-free as well. The empirical evaluation in Section 3.2 will confirm this statement. The first approach we present, called *Massaging the data*, is based on changing the class labels in order to remove the discrimination from the training data. A preliminary version of this approach was presented in [46]. The second approach is less intrusive as it does not change the class labels. Instead, weights are assigned to the data objects to make the dataset discrimination-free. This approach will be called *Reweighing*. Since reweighing requires the learner to be able to work with weighted tuples, we propose another solution without this requirement, in which we re-sample the dataset in such a way that the discrimination is removed. We will refer to this approach as *Sampling*. Two ways of sampling will be presented and tested.

3.1.1 Massaging

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

$X(Class) = -$ will be called the *promotion candidates* and the set dem of objects X with $X(S) = w$ and $X(Class) = +$ will be called the *demotion candidates*.

We will not randomly pick promotion and demotion candidates to relabel. Instead we use the rationale of our discrimination model given in Section 2.3.1 of Chapter 2 that the data objects close to the decision boundary are more vulnerable to the effect of discrimination and will select these objects to relabel. For this purpose we learn a ranker R on the training data for ranking the object according to their positive class probability. We assume that higher scores indicate a higher chance to be in the positive class. With this ranker, the promotion candidates are sorted according to descending score by R and the demotion candidates according to ascending score. When selecting promotion and demotion candidates, first the top elements will be chosen. In this way, the objects closest to the decision border are selected first to be relabeled, leading to a minimal effect on the accuracy. This modification of the training data is continued until the discrimination becomes zero. The number M of pairs needed to be modified to make a dataset D discrimination-free can be calculated as follows. If we modify M pairs, the resulting discrimination will be:

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

To reach zero discrimination, we hence have to make:

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

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

Example 3 *Consider again the dataset D given in Table 2.1. We want to learn a classifier to predict the class labels of data 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 3.1 the positive class probabilities as given by this ranker are added to the table for reference (calculated using the "NBS" classifier of Weka). In the second step, we arrange the data separately for* female *applicant with class $-$ in descending order and for* male *applicants with class $+$ in ascending order with respect to their positive class probability. The ordered promotion and demotion candidates are given in Table 3.2.*

Table 3.1 Sample job-application relation with positive class probability.

Sex	Ethnicity	Highest Degree	Job Type	Cl.	Prob
m	native	h. school	board	+	98%
m	native	univ.	board	+	89%
m	native	h. school	board	+	98%
m	**non-nat.**	**h. school**	**healthcare**	**+**	69%
m	non-nat.	univ.	healthcare	-	30%
f	non-nat.	univ.	education	-	2%
f	**native**	**h. school**	**education**	**-**	40%
f	native	none	healthcare	+	76%
f	non-nat.	univ.	education	-	2%
f	native	h. school	board	+	93%

Table 3.2 Promotion candidates (negative objects with $Sex = f$ in descending order) and demotion candidates (positive objects with $Sex = m$ in ascending order)

Sex	Ethnicity	Highest Degree	Job Type	Cl.	Prob
f	**native**	**h. school**	**education**	**-**	**40%**
f	non-nat.	univ.	education	-	2%
f	non-nat.	univ.	education	-	2%
Sex	Ethnicity	Highest Degree	Job Type	Cl.	Prob
m	**non-nat.**	**h. school**	**healthcare**	**+**	**69%**
m	native	univ.	board	+	89%
m	native	h. school	board	+	98%
m	native	h. school	board	+	98%

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

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

So, relabeling one promotion candidates and one demotion candidates list makes the data discrimination-free. Hence, we will relabel the top promotion candidate; i.e., the highest scoring female with a negative label, and the top demotion candidate; i.e., the lowest scoring male with a positive label. After the labels for these instances are changed, the discrimination will decrease from 40% *to* 0%*. the resulting dataset will be used as a training set for classifier induction.* □

Algorithm. The pseudo-code of our algorithm is given in Algorithm 2 and Algorithm 3. Algorithm 2 describes changing the class labels and classifier learning, and Algorithm 3 the sorting of the promotion and demotion lists.

Algorithm 2: *Learn Classifier on Massaged Data*

Input: Labeled dataset D, sensitive attribute S and value b, desired class $+$
Output: Classifier C, learned on D

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

Algorithm 3: *Rank*

Input: Labeled dataset D, Sensitive attribute S and value b, desired class $+$
Output: Ordered promotion list pr and demotion list dem

1: Learn a ranker R for prediction $+$ using D as training data
2: $pr := \{X \in D \mid X(S) = b, X(Class) \neq +\}$
3: $dem := \{X \in D \mid X(S) = w, X(Class) = +\}$
4: Order pr descending w.r.t. the scores by R
5: Order dem ascending w.r.t. the scores by R
6: **return** (pr, dem)

3.1.2 Reweighing

The *Massaging* approach is rather intrusive as it changes the labels of the objects. Our second approach does not have this disadvantage. Instead of relabeling the

objects, different weights will be attached to them. For example, objects with $X(S) = b$ and $X(Class) = +$ will get higher weights than objects with $X(S) = b$ and $X(Class) = -$ and objects with $X(S) = w$ and $X(Class) = +$ will get lower weights than objects with $X(S) = w$ and $X(Class) = -$. We will refer to this method as *Reweighing*. Again we assume that we want to reduce the discrimination to 0% while maintaining the overall positive class probability. We now discuss the idea behind the weight calculation.

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

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

In reality, however, the observed probability in D,

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

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

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

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

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

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

Example 4 *Consider again the dataset in Table 2.1. The weight for each data object is computed according to its S- and $Class$-value. We calculate the weight of a data object with $X(S) = f$ and $X(Class) +$ as follows. We know that 50% objects have $X(S) = f$ and 60% objects have $Class$-value $+$, so the expected probability of the object should be:*

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

Table 3.3 Sample job-application relation with weights.

Sex	Ethnicity	Highest Degree	Job Type	Cl.	Weight
m	native	h. school	board	+	0.75
m	native	univ.	board	+	0.75
m	native	h. school	board	+	0.75
m	non-nat.	h. school	healthcare	+	0.75
m	non-nat.	univ.	healthcare	-	2
f	non-nat.	univ.	education	-	0.67
f	native	h. school	education	-	0.67
f	native	none	healthcare	+	1.5
f	non-nat.	univ.	education	-	0.67
f	native	h. school	board	+	1.5

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

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

Similarly the weights of all other combinations is as follows:

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

The weight of each data object of the Table 2.1 is given in Table 3.3.

Algorithm.

The pseudocode of the algorithm describing our *Reweighing* approach is given in Algorithm 4.

3.1.3 Sampling

Since not all classifier learners can directly incorporate weights in their learning process, we also propose a *Sampling* approach. The dataset with weights is transformed by sampling into a normal dataset which can be used by all algorithms.

Algorithm 4: *Reweighing*

Input: $(D, S, Class)$
Output: Classifier learned on reweighted D

for $s \in \{b, w\}$ **do**
  for $c \in \{-, +\}$ **do**
    Let $W(s,c) := \dfrac{|\{X \in D \mid X(S) = s\}| \times |\{X \in D \mid X(Class) = c\}|}{|D| \times |\{X \in D \mid X(Class) = c \text{ and } X(S) = s\}|}$
  end for
end for
$D_W := \{\}$
for X in D **do**
  Add $(X, W(X(S), X(Class)))$ to D_W
end for
Train a classifier C on training set D_W, taking onto account the weights
return Classifier C

By sampling the objects with replacement according to the weights, we make the given dataset discrimination-free.

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

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

Consider Figure 3.1, representing a dataset with 40 data points. The data points in the positive class are represented by $+$, the data points of the negative class by $-$. The projection on the horizontal axis represents the probability of each data object to be in the positive class: the more to the right is the point, the higher its positive class probability. This probability comes, e.g., from a ranker we learned on the training data. This probability will only be of interest for our second sampling method, the preferential sampling, and can for the moment being be ignored. The data points plotted in the upper half of the graph, respectively the lower half, represent the deprived, respectively the favored community. In the case of discrim-

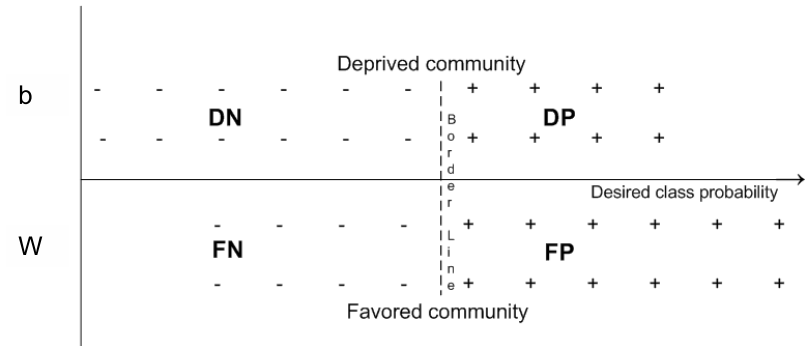


Figure 3.1 A figure with 40 data points to show the sampling method.

ination, the relative size of DN versus DP will be larger than the relative size of FN versus FP.

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

Sample Size	**DP**	**DN**	**FP**	**FN**
Actual	8	12	12	8
Expected	10	10	10	10

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

Uniform Sampling

As the name already suggests, in *US* all the data objects of the same group have the same chance of being duplicated or skipped; if we need to sample n objects from a group P, *US* will apply uniform sampling with replacement. In Figure 3.2 a possible re-sampling of the dataset is given; the bold elements are duplicated while the encircled objects are removed. Algorithm 5 gives a formal description of the *US* method.

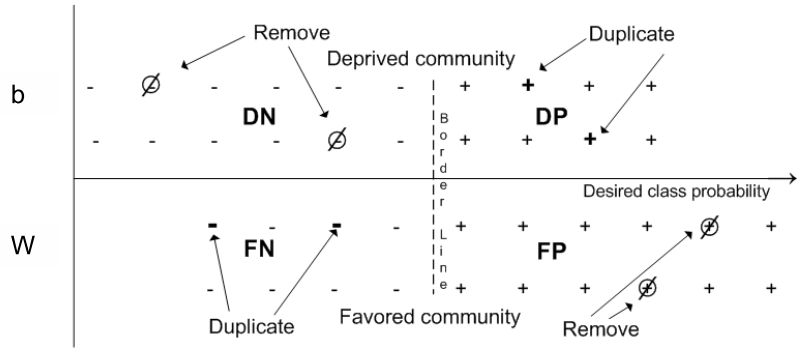


Figure 3.2 Pictorial representation of the *Uniform Sampling* scheme. The re-substituted data points are in bold while the encircled ones are skipped.

Algorithm 5: *Uniform Sampling*

Input: $(D, S, Class)$
Output: Classifier C learned on resampled D

for $s \in \{b, w\}$ **do**
  for $c \in \{-, +\}$ **do**
    Let $W(s,c) := \dfrac{|\{X \in D \mid X(S) = s\}| \times |\{X \in D \mid X(Class) = c\}|}{|D| \times |\{X \in D \mid X(Class) = c \text{ and } X(S) = s\}|}$
  end for
end for
Sample uniformly $\lfloor W(b,+) \times |DP| \rfloor$ objects from DP;
Sample uniformly $\lfloor W(w,+) \times |FP| \rfloor$ objects from FP;
Sample uniformly $\lfloor W(b,-) \times |DN| \rfloor$ objects from DN;
Sample uniformly $\lfloor W(w,-) \times |FN| \rfloor$ objects from FN;
Let D_{US} be the set of all samples generated in steps 6 to 9
return Classifier C learned on D_{US}

Preferential Sampling

In *Preferential Sampling (PS)* we again use the intuition of our discrimination model given in Section 2.3.1 of Chapter 2 that data objects close to the decision boundary are more prone to have been discriminated or favored due to discrimination in the dataset and give preference to them for sampling. To identify the borderline objects, *PS* starts by learning a ranker on the training data. *PS* uses this

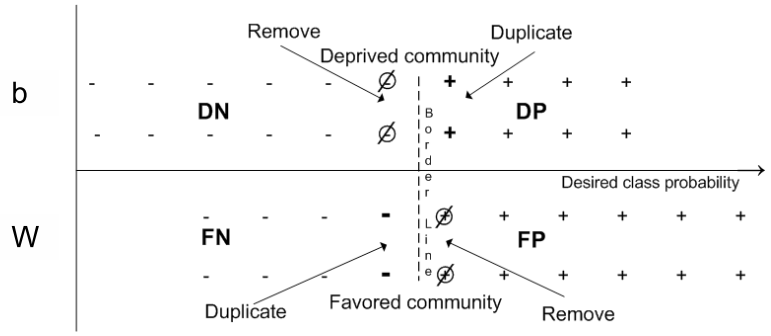


Figure 3.3 Pictorial representation of *Preferential Sampling* scheme. The re-substituted data points are in bold while the encircled ones are skipped.

ranker to sort the data objects of DP and FP in ascending order, and the objects of DN and FN in descending order w.r.t. the positive class probability. Such arrangement of data objects makes sure that the higher up in the ranking an element occurs, the closer it is to the boundary.

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

- Decreasing the size of a group is always done by removing the data objects closest to the boundary; i.e., the top elements in the ranked list.

- Increasing the sample size is done by duplication of the data object closest to the boundary. When an object has been duplicated, it is moved, together with its duplicate, to the bottom of the ranking. We repeat this procedure until the desired number of objects is obtained.

In most cases, only a few data objects have to be duplicated or removed. The exact algorithm is given in Algorithm 6.

Algorithm 6: *Preferential Sampling*

Input: $(D, S, Class)$
Output: Classifier C learned on resampled D

for $s \in \{b, w\}$ **do**
  for $c \in \{-, +\}$ **do**
    Let $W(s,c) := \frac{|\{X \in D \mid X(S) = s\}| \times |\{X \in D \mid X(Class) = c\}|}{|D| \times |\{X \in D \mid X(Class) = c \text{ and } X(S) = s\}|}$
  end for
end for
Learn a ranker R for predicting $+$ using D as training set
$D_{PS} := \{\}$
Add $\lfloor W(b,+) \rfloor$ copies of DP to D_{PS}
Add $\lfloor W(b,+) - \lfloor W(b,+) \rfloor \times |DP| \rfloor$ lowest ranked elements of DP to D_{PS}
Add $\lfloor W(b,-) \times |DN| \rfloor$ lowest ranked elements of DN to D_{PS}
Add $\lfloor W(w,+) \times |FP| \rfloor$ highest ranked elements of FP to D_{PS}
Add $\lfloor W(w,-) \rfloor$ copies of FN to D_{PS}
Add $\lfloor W(w,-) - \lfloor W(b,-) \rfloor \times |FN| \rfloor$ highest ranked elements of FN to D_{PS}
return Classifier C learned on D_{PS}

3.2 Experiments

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

1. The preprocessing techniques introduced in the chapter:
   (a) The **Massaging** approach with different rankers. We consider five different rankers: one based on a Naïve Bayes classifier (M_NBS), one based on decision tree learner (M_J48) and three based on a nearest neighbor classifier for respectively 1, 3 and 7 neighbors (M_IBk1, M_IBk3, and M_IBk7). These rankers are used to relabel the dataset to make it discrimination-free.
   (b) **Reweighing** (RW) and **Uniform Sampling** (US); these methods are parameter-free as they do not rely on a ranker.
   (c) **Preferential Sampling** (PS) with a rankers based on a Naïve Bayes classifier.

   This gives a total of 8 preprocessing methods to clean away the discrimina-

tion of the input data. On the cleaned data, different base classifiers were trained: a Naïve Bayes Classifier (NBS), two nearest neighbor classifiers with respectively 1 and 7 neighbors (IBk1 and IBk7), and a decision tree learner: the Weka implementation of the C4.5 classifier (J48). This gives a total of $8 \times 4 = 32$ combinations. Many more combinations have been tested (including, e.g., Adaboost) but we restricted ourselves to the choices above as they present a good summary of the obtained results; for the other classifiers similar results were obtained.

2. Two **baseline** approaches:

   (a) An **out-of-the-box classifier** not taking any anti-discrimination measures into account in any way (labeled "No" to reflect no preprocessing was used); we compare to this baseline to see what is the net benefit w.r.t. discrimination-reduction of our proposed methods and how much accuracy we have to trade in for that reduction.

   (b) We **remove the sensitive attribute and its most correlated attributes** before learning ("No_SA" for No Sex Attribute). In this way we get as many baseline classifiers, depending on how many of the correlated attributes we remove. The continuous lines in the figures show these baseline results.

We analyze our proposed algorithms in two scenarios:

- S is part of the training set, but cannot be used during prediction. In these experiments we only use the information about S for evaluating the discrimination measurement, but S is not considered for prediction. Notice that this set-up respects all our assumptions.

- S is part of the training set and can be used at prediction time. This set-up actually violates our assumption (A2) that S should not be used during prediction but has been added for reference.

Experimental set-up. In our experiments we use apply our proposed methods on the **Adult dataset**, on the **Communities and Crimes dataset**, on the **two Dutch census datasets of 1971 and 2001** [31, 32] and use the same experimental set-up as discussed in Section 1.3.1 of Chapter 1.

3.2.1 Redlining

Our first experiment concerns the redlining effect. For all datasets we show in Table 3.4 the discrimination of a classifier learned on unaltered training data, with and without the sensitive attribute. The results clearly motivate our work: classifiers learned on biased data produce biased classifiers, even if the sensitive attribute is removed during training.

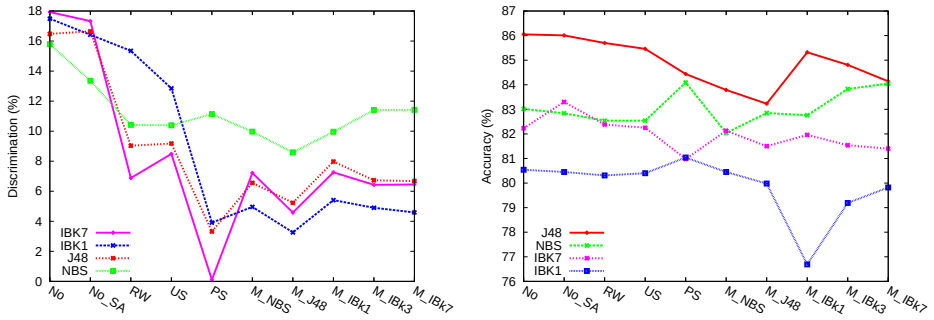
Table 3.4 Discrimination scores of classifiers trained on discriminatory data; with and without the sensitive attribute. The results clearly confirm the existence of a redlining effect.

Dataset	**With** S	**Without** S
German Credit	11.09%	9.32%
Adult	16.48%	16.65%
Communities and Crimes	40.14%	38.07%
Dutch 2001 Census	34.91%	17.92%

3.2.2 Adult Dataset

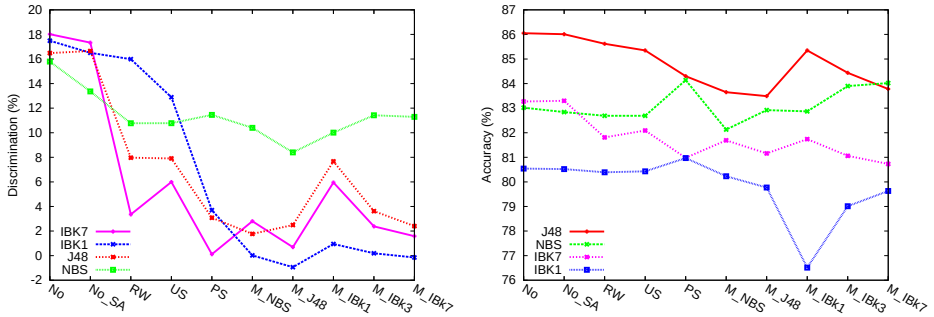
In Figures 3.4(a) and 3.4(b), respectively the discrimination and accuracy results for all algorithms under comparison are given. On the X-axis are the names of the data preprocessing techniques used to make the training dataset discrimination-free. The resultant discrimination has been given on the Y-axis of Figure 3.4(a) and the accuracy on the Y-axis of Figure 3.4(b). We observe that the classifiers learned on the preprocessed data produce less discriminatory results as compared to the baseline algorithms; in Figure 3.4(a) we see that IBk7 classifies the future data objects with 17.93% discrimination which is lowered only slightly if the Sex attribute is removed. If *Preferential Sampling* is applied, however, the discrimination goes down to 0.11%. On the other hand, We observe in Figure 3.4(b) that the loss in accuracy is modest in comparison with the large reduction in discrimination. The discrimination always goes down when we apply our classifiers with non-discrimination constraints, while accuracy remains at a high level. In these experiments, we omitted S from our training and test datasets.

Figures 3.5(a) and 3.5(b) represent the results of the same experiment, except that this time S can be used at prediction time. These two experiments produce very similar results. We observe that the combination of J48 as base learner and Naive Bayes as a ranker for *Massaging* produces promising results. IBk as a ranker for



(a) Baseline discrimination=19.45% (b) Baseline accuracy=76.3%

Figure 3.4 The results of 10-fold CV for the **Adult dataset** when S is used in the learning phase but not for prediction.



(a) Baseline discrimination=19.45% (b) Baseline accuracy=76.3%

Figure 3.5 The results of 10-fold CV for the **Adult dataset** when S is used for both learning and prediction.

the *Massaging* filter is also one of the best choices. *PS* gives excellent results when it is used with unstable classifiers, e.g., J48. When *PS* is used with J48, the discrimination level decreases from 16.48% to 3.32% while the accuracy level decreases from 86.05% to 84.3%. Figure 3.5(b) shows the resultant accuracy for all these method. We find that the *Reweighing* approach and some combinations of the *Massaging* approach maintain a high accuracy level while the accuracy drops to some extent with other combinations of *Massaging*. Clearly, the choice of base learner and ranker (for *Massaging*) plays a very important role in discrimination-free classification.

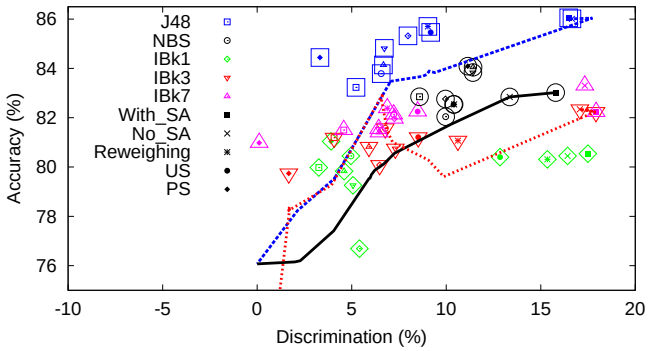
Figure 3.6(a) represents a graphical representation of the experiments when the attribute *Sex* is not used at prediction time. Figure 3.6(b) shows the results of

the experiments when *Sex* is used at prediction time. Each pictogram in these figures represents a particular combination of a classification algorithm (shown by the outer symbol) and a preprocessing technique (shown by the inner symbol). For *Massaging*, the inner symbol represents the ranker that was used. On the X-axis we see the discrimination and on the Y-axis, the accuracy. Thus, we can see the trade-off between accuracy and discrimination for each combination. The closer we are to the top left corner the higher accuracy and the lower discrimination we obtain. The three lines in the figure represent three classifiers (J48, NBS and IBk3 from the top to bottom) learned on the original dataset (the most top-right point in each line, denoted with $With\_SA$ symbol), the original dataset with the *Sex* attribute removed (denoted with No_SA symbol), the original dataset with the *Sex* attribute and the one (two, three, and so on) most correlated attribute(s) removed (that typically correspond to the further decrease in both accuracy and discrimination).

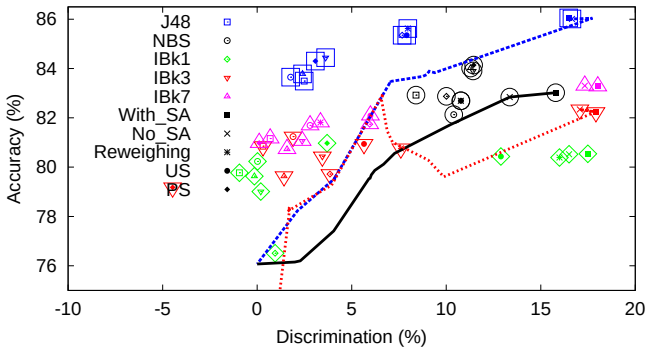
Figures 3.6(a) and 3.6(b) offer a good overview that allows us to quickly assess which of the combinations are DA-optimal (discrimination-accuracy-optimal) among the classifiers learned in our experiments. We observe that the top left area in the figure is occupied by the data points corresponding to the performance of *Massaging* and *PS* approaches. The *Reweighing* and *US* approaches fall behind *Massaging* but also show reasonable performance. From Figures 3.6(a) and 3.6(b) we can see that our approaches compare favorably to the baselines in the sense that almost all combinations dominate the baseline solutions.

3.2.3 Dutch Census Datasets

We repeated all the experiments over the Dutch 2001 Census dataset. The results of these experiments are shown in the Figure 3.7. We observe that our proposed discrimination-aware classification methods outperform the traditional classification method w.r.t. accuracy discrimination trade-off. Figure 3.7 shows that our proposed methods classify the unseen data objects with low discrimination and high accuracy. The discrimination is lowered from 38% to almost 0% at the cost of a very little accuracy. All the methods we tried in our experiments give excellent results w.r.t. accuracy-discrimination trade-off on this dataset when applied in combination with discrimination-aware techniques and clearly outperform the baseline approaches.



(a) S is used in the learning phase but not for prediction.

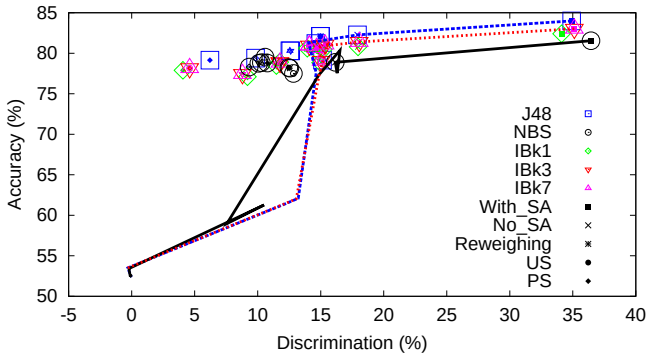


(b) S is used for both learning and prediction.

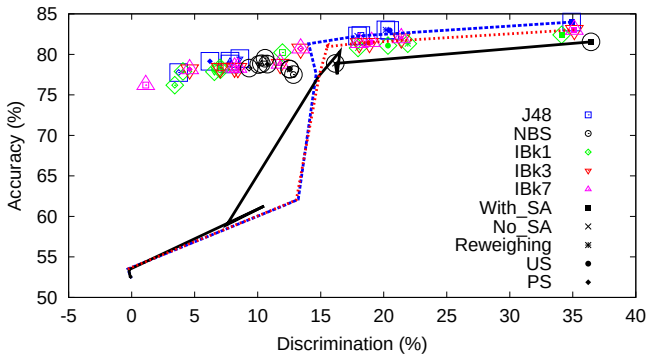
Figure 3.6 Accuracy-discrimination trade-off comparison for the **Adult dataset**. *Outer and inner symbol of each data point shows the corresponding base learner and preprocessing technique respectively*. Three lines represent the baselines for three classifiers J48, NBS, IBK3 (top to bottom).

3.2.4 Communities and Crimes Dataset

We repeated the same experiment over the Communities and Crimes dataset and found similar results. Figure 3.8 gives an overview of the results. We observe that



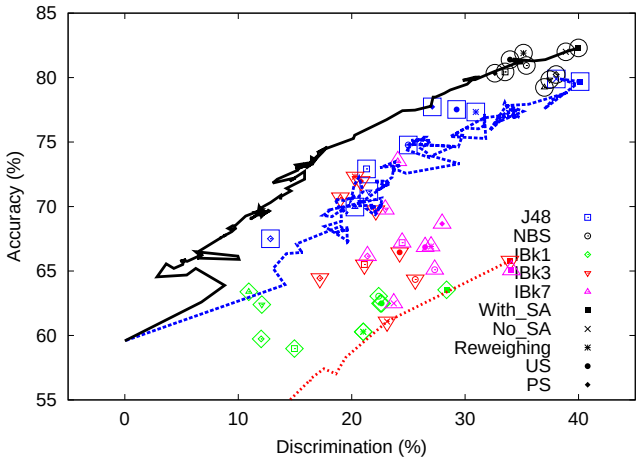
(a) S is used in the learning phase but not for prediction.



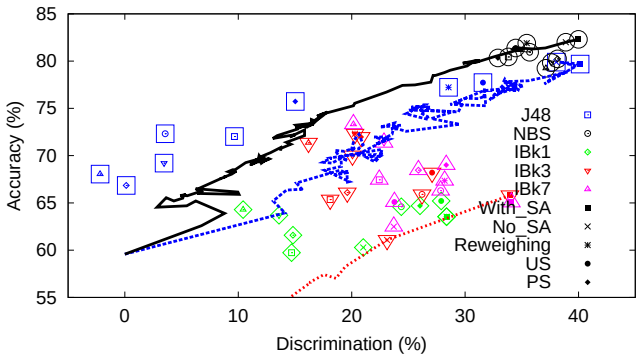
(b) S is used for both learning and prediction.

Figure 3.7 Accuracy-discrimination trade-off comparison for the **Dutch 2001 Census dataset**. *Outer and inner symbol of each data point shows the corresponding base learner and preprocessing technique respectively*. Three lines represent the baselines for three classifiers J48, IBK3, NBS (top to bottom).

our proposed solutions outperform the baseline approaches. Naive Bayes works extremely well on this dataset. When we remove discrimination from the training data, the effect is transferred to future classification in case of unstable classifiers



(a) S is used in the learning phase but not for prediction.



(b) S is used for both learning and prediction.

Figure 3.8 Accuracy-discrimination trade-off comparison over the **Communities and Crimes dataset**. (*Outer and inner symbol of each data point shows the corresponding base learner and preprocessing technique respectively*. Three lines represent the baselines for three classifiers NBS, J48, IBK3 (top to bottom).

and both the discrimination level and the accuracy goes down more than for a stable (noise-resistent) classifier.

3.2.5 How to Choose Ranker and Classifier for Massaging

From the different experiments, we make the following observation: if minimal discrimination is the first priority, an unstable classifier; i.e., a classifier more sensitive to noise, as base learner is the better option and if the high accuracy is the main concern, a stable classifier might be more suitable. To substantiate this hypothesis further, we conducted additional experiments where we used a k-nearest neighbor classifier. This classifier has the advantage that we can influence its stability with the parameter k: the higher k, the more stable it becomes. Figure 3.9 represent the results of the experiments with IBk as base learner and NBS as ranker for the *Massaging* approach. We changed the value of k for IBk from 1 to 19 (only odd values) to change its stability as a base classifier. We observe that the resultant discrimination and accuracy increase both with increasing k, which supports our claim.

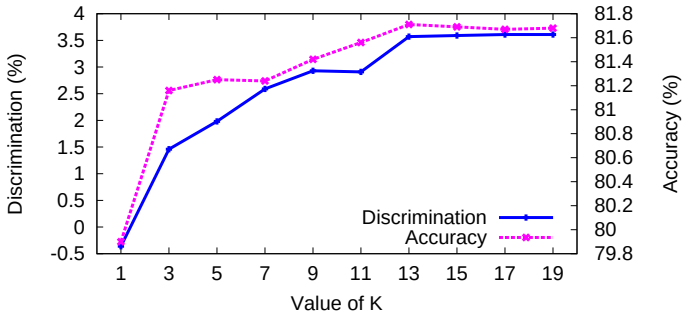


Figure 3.9 Accuracy and discrimination comparison with NBS as a ranker and IBk as a base learner with different values of k.

3.2.6 Sanity Check

In our current setting of the discrimination problem, we assume that our training set is discriminatory while our future test set is expected to be non-discriminatory. Unfortunately, this ideal scenario is not readily available for experiments but in this chapter we try to mimic this scenario by using the Dutch 1971 census data as

Table 3.5 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 3.6 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%			

a training set and the Dutch 2001 census data as a test set. In our experiments, we use the attribute *economic status* as class attribute because this attribute uses similar codes for both the 1971 and the 2001 dataset. This attribute determines whether a person has a job or not; i.e., is economically active or not. We remove some attributes like *current economic activity* and *occupation* from the experiments to make both datasets consistent w.r.t. codings. Tables 3.5 and 3.6 show that in Dutch 1971 Census data, the percentage of unemployment among females is much higher than in the Dutch 2001 Census data. This difference shows that the gender-related inequality in access to the job market reduced from the 70s to 2001.

If we now learn a traditional classifiers over the 1971 data and test it over the 2001 data without taking the discrimination aspect into account, it will classify the future data with low accuracy and high discrimination. In contrast, our discrimination-aware classification methods classify the future data objects with low discrimination and maintain a significantly high level of accuracy. However, we also observe that the *Massaging* method with some rankers overshoots the discrimination and results in low accuracy scores. One reason for this low accuracy scores is that our test set is not completely discrimination-free. In future we plan to repeat these experiments over the unbiased test set for further exploration. We are also interested to propose discrimination-aware methods that reduce the discrimination level to a desired level only, not more than that. It is important to notice in these experiments that when the test set is discriminatory, our prosed methods always lose accuracy but in this case, when the test set is relatively less discriminatory, it is not always the case and many times our methods affect the accuracy positively.

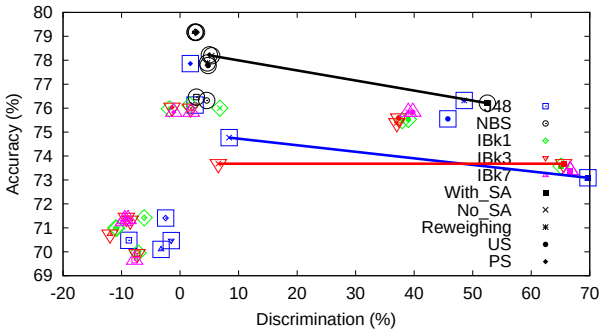


Figure 3.10 Accuracy and discrimination comparison when we use discriminatory training set (the Dutch 1971 census dataset) and non-discriminatory test set (the Dutch 2001 Census dataset). Three lines represent the baselines for three classifiers NBS, J48, IBK3 (top to bottom).

3.2.7 Conclusions of the Experiments

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

1. Our proposed methods give comparable accuracy and low discrimination scores on average when applied to non-discriminatory test data.
2. Just removing the sensitive attribute from the dataset is not enough to ensure discrimination-aware classification due to redlining effect.
3. All the proposed methods consistently outperform the baseline methods w.r.t. the accuracy-discrimination trade-off.
4. Our proposed preprocessing methods for discrimination-aware classification can be combined with any classifier. The effect of this preprocessing is better captured when training unstable classifiers.

3.3 Conclusion

In this chapter we presented the classification with non-discriminatory constraints problem. Three approaches towards the problem were proposed: *Massaging*, *Reweigh-*

ing and *Sampling* the dataset. All approaches remove the discrimination from the training data and subsequently a classifier is learned on this unbiased data. Experimental evaluation shows that indeed this approach allows for removing discrimination from the dataset more efficiently than simple methods such as, e.g., removing the sensitive attribute from the training data. We also empirically show that when the test set is non- (or less) discriminatory, our proposed methods do not always influence the accuracy negatively.

Chapter 4

Discrimination-aware Decision Tree Learning

In the previous chapter we proposed three solutions to the discrimination-aware classification problem based on modifying the input data. We propose two solutions to construct decision trees without discrimination. The first solution is based upon the adaptation of the splitting criterion in the decision tree learner and the second approach is based upon the post-processing of decision tree with discrimination-aware pruning and relabeling of tree leaves. Before going into the details of our proposed solutions to the discrimination problem we revise decision tree learning.

4.1 Decision Tree

A decision tree is a flow-chart like structure which is fairly easy to interpret and allows for easy identification of significant variables. Each internal node (non-leaf node) in a decision tree denotes a test on an attribute. We refer to this attribute as *test attribute* and is denoted by $test\_att$. Each branch represents an outcome of the test and corresponds to a value of the test attribute. In case of numeric attributes, we split the whole range of values at certain point and each branch corresponds to the one range of values. Each leaf node of the tree is labeled with a certain class label. The topmost node in a tree is the root node. Learned trees can also be represented as sets of if-then-else rules to improve human readability.

Decision trees classify future data objects by sorting them down the tree from the root to some leaf node. An instance is classified by starting at the root node of the tree, testing the attribute specified by this node, then moving down the tree branch corresponding to the value of the attribute in the given data object. This process is then repeated for the subtree rooted at the new node and continues up to the leaves. Each leaf is labeled with the majority class of its data objects and any future instance ending on this leaf will get the class label of this leaf.

Figure 4.1 gives a simple decision tree built over the dataset given in the same figure. This tree assigns accept (+) or reject (-) class to job applicants based upon their features. If we want to use this decision tree to classify a new job applicant, it will first check the job type of the application. If the job type is board, the applicant will be assigned the positive class and if the job type is education the applicant will be assigned the negative class label. The applicants, who have applied for health care jobs, need further evaluation on the basis of their highest degree attribute to get a class label.

The core decision tree algorithms ID3 (Iterative Dichotomiser) [68] and C4.5 (a successor of ID3) [69] use a top-down recursive divide-and-conquer approach through

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	+

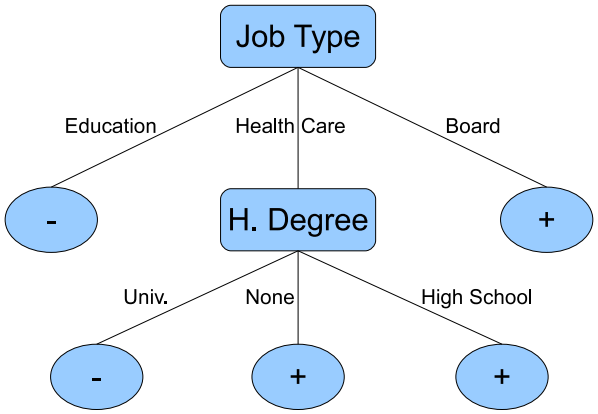


Figure 4.1 A simple decision tree leant from the data set given in the above table.

the space of all sub-trees. Later on many extensions to these basic algorithms have been proposed. In this chapter, however, we work with C4.5 and do experiments with its Weka implementation (known as J48). Decision tree inducers start with a training set of tuples and their associated class labels. The training set is recursively partitioned into smaller subsets as the tree is being built. The pseudocode of a basic decision tree algorithm is given in Algorithm 7.

Decision trees are very popular classification models in data mining and machine learning because they require only modest resources to learn and fast in perfor-

Algorithm 7: Decision Tree Induction

Parameters: Split evaluator $gain$
Input: Dataset D over $\{A_1, \ldots, A_n, Class\}$.
Output: Decision Tree DT

Create a node N
if $pure(D)$ is empty **then**
  return N as a leaf labeled with majority class of D
else
  Select $test\_att$ s.t. $gain(test\_att, test)$ is maximized
  Label node N with $test\_att$
  for Each outcome i of $test(test\_att)$ **do**
    Grow a branch from node N for the condition $test(test\_att) = i$
    Let D_i be the set of examples X in D for which $test(X.test\_att) = i$ holds
    Attach the node returned by *Decision_Tree*(D_i) at the end of the branch
  end for
end if

mance. Decision trees are also robust to noisy data, errors in the attribute values, and errors in class labels of historical data objects. Classification methods have been successfully applied to a broad range of tasks, e.g., to diagnose medical cases, credit risk analysis, astronomy [62, 40].

4.1.1 Split Criteria

In decision trees construction, the most important question is which attribute should be selected as a test attribute. The splitting criterion is a heuristic for selecting the test attribute that "best" separates a given data partition, D, of labeled data objects into different classes. Our main intention is to divide the dataset, D, into smaller *pure partitions*, i.e., most of the data object in each data partition should belong to the same class. For this purpose there are many splitting criteria to select the best attribute to split the decision trees into subtrees or leaves, e.g., information gain, gain ratio, gini index.

We only discuss information gain in detail, for other splitting criteria the reader is referred to [40]. We revise some of the formal notation given in Chapter 2. We refer the dataset, D, as a data partition. Suppose the class attribute $Class$ has k distinct values defining k distinct classes, $c_1, c_2, \ldots, c_k$. Let D_{ci} be the set of tuples of

class $Class_i$ in D. Let $|D|$ and $|D_{ci}|$ denote the number of tuples in D and D_{ci}, respectively.

Information Gain

In this section we discuss information gain as the test attribute selection measure. We select an attribute with the highest information gain as the test attribute. We start building a decision tree from the root node. At the root node our data partition equal to the whole training data and our attribute list consists all the attributes in D. In every iteration of the algorithm, we want to select the test attribute for node N that minimizes the information needed to classify the data objects in the resulting data partitions (line 5 of Algorithm 7). The information required to classify a data object in D is:

$$Info^{Class}(D) := \sum_{i=1}^{k} -P_i \log_2 P_i \ , \tag{4.1}$$

where P_i is the probability that a given data object in D belongs to class $Class_i$ and is estimated by $\frac{|D_{ci}|}{|D|}$. Equation 4.1 gives the average amount of information required to classify a data object in data partition D, also known as the *entropy* of D.

Now, we calculate the amount of information needed to classify a data object if we split our data D into smaller partitions according to the values of an attribute F. Let the attribute F has v distinct values, $f_1, f_2, \ldots, f_v$. We make a separate partition of D for each value of F, i.e., D is partitioned into v subsets, $D_1, D_2, \ldots, D_v$, where D_j contains those tuples in D that have outcome f_j of F. Ideally each smaller partition should be pure, i.e., all data objects in this partition belong to one class. If these smaller partitions are still not pure, it mean that we are still in need of more information to arrive at an exact classification, i.e. to make every partition pure. We calculate this information in the following way:

$$Info_F^{Class}(D) := \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info^{Class}(D_j) \ . \tag{4.2}$$

The term $\frac{|D_j|}{|D|}$ acts as the weight of the j^{th} partition. $Info_F^{Class}(D)$ is the expected information required to classify a tuple from D based on the partitioning by F. The smaller the expected information (still) required, the greater the purity of the partitions. Information gain is defined as the difference between the original information requirement (i.e., based on just the proportion of classes) and the new

requirement (i.e., obtained after partitioning by F). That is,

$$IGC(F) = Info^{Class}(D) - Info_F^{Class}(D) \tag{4.3}$$

$IGC(F)$ gives us the expected reduction in the information requirement by partitioning the data w.r.t. the values of F. It is important to notice that we denote this gain by IGC (information gain for class attribute) because in the next sections we also use IGS where we want to make our data partitions pure w.r.t. the sensitive attributes. We calculate IGS for an attribute F in the following way:

$$IGS(F) = Info^{S}(D) - Info_F^{S}(D) \ . \tag{4.4}$$

We select the attribute with the highest information gain as the test attribute at node N [40].

Example 5 *Let us select the best attribute from the data set given in Figure 4.1. We first use Equation (4.1) to compute the expected information needed to classify a tuple in D:*

$$Info^{Class}(D) := -\frac{5}{10}\log_2\frac{5}{10} - \frac{5}{10}\log_2\frac{5}{10} = 1 \ .$$

Next, we need to compute the expected information requirement for each attribute. Lets start with the attribute Job Type. The attribute Job Type has three values; board, health care, and education. If we partition our data w.r.t. to the values of the Job Type attribute, the board partition consists of 3 + class objects (it is pure), the education partition has 3 − class data objects (it is also pure), and the health care partition has 2 + class objects and one − class objects (not fully pure). We use the formula given in Equation 4.2 to calculate how much more information will be required to achieve an exact classification if we partition our data over the attribute Job Type. The average amount of information will be:

$$\begin{aligned} Info_{Job\ Type}^{Class}(D) &= \frac{4}{10} \times (-\frac{4}{4}\log_2\frac{4}{4}) \\ &\quad + \frac{3}{10} \times (-\frac{3}{3}\log_2\frac{3}{3}) \\ &\quad + \frac{3}{10} \times (-\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}) \\ &= 0 + 0 + 0.305 = 0.305 \end{aligned}$$

So, the gain in information from this partitioning would be

$$\begin{aligned} IGC(Job\ Type) &= Info^{Class}(D) - Info_{Job\ Type}^{Class}(D) \\ &= 1 - 0.305 = 0.695 \end{aligned}$$

Similarly, we can compute $IGC(H.Degree) = 0.285$, $IGC(Ethnicity) = 0.256$, *and* $IGC(Sex) = 0.125$. *As the attribute Job Type has the highest information gain, it is selected as the test attribute. The root node is labeled with this attribute and three branches w.r.t. the values of this attribute are grown. As the board partition and the education partition are pure, we assign class labels to these nodes and convert these nodes to leaves of the trees with certain class labels. The same procedure is repeated at the resultant node for the value health care. For this partition, the attribute Highest Degree has the highest information gain and is selected as the test attribute. In this way, we get a final tree as shown in Figure 4.1 [62, 40].*

4.1.2 Pruning

Most of the decision tree learning algorithms continue to grow branches of the tree to make every leaf as pure as possible. This greedy approach leads to difficulties when the training data is noisy or too small to produce a representative sample of the true target class. In either of these cases, we end up with the trees that overfit the training examples. Tree pruning methods address this problem of overfitting the data. Pruned trees are smaller, less complex and, thus, easier to comprehend. They are usually faster and better at correctly classifying the test data than unpruned trees.

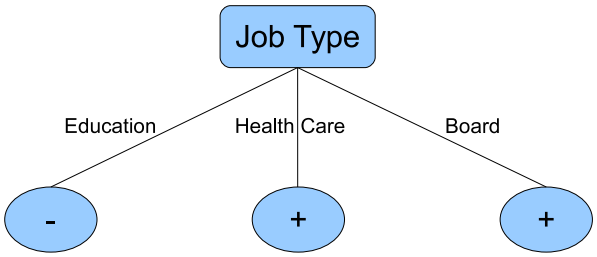


Figure 4.2 A pruned decision tree

Tree pruning has two types; prepruning and postpruning. In the prepruning approach, we prune a tree by stoping its construction early. We do not grow the branches of the tree further from a node, if the further partitioning falls below a certain threshold. Different measures, e.g., information gain are used to assess this threshold which is often hard to choose. In this case the node becomes a leaf and labeled with the majority class of its instances. In postpruning, we first grow a

full tree and then convert some of its branches into leaves. In practice this pruning method is more successful and more frequently used. For both pruning methods it is very important to decide the most optimal size of the tree. There are different techniques to determine the optimal size of the tree, e.g.; use a separate set of examples, called validation set, to evaluate the utility of post-pruning nodes from the tree. Figure 4.2 shows a pruned tree when apply pruning to the decision tree given in Figure 4.1 [62, 40].

4.2 Discrimination-Aware Tree Construction

In discrimination-aware classification 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.

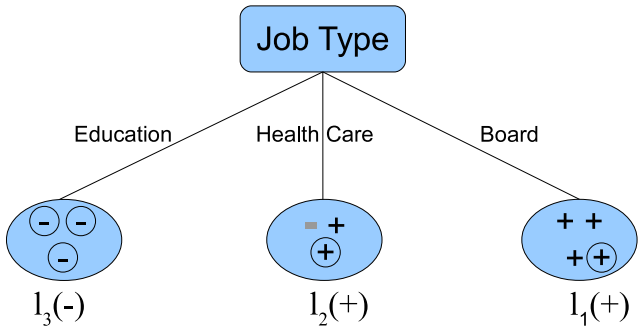


Figure 4.3 A pruned decision tree with the details of $+$ and $-$ objects in each leaf. The data objects with $Sex = f$ are encircled. The instance shown in gray color represent the misclassified instances. Acc=90% Disc=60%

Our first solution is changing the attribute selection criterion as in step 5 of Algorithm 7. To measure the influence of the split on the discrimination, we use the same information gain, but now w.r.t. the sensitive attribute S instead of the class $Class$ and denote it by IGS. IGS selects the attribute as the test attribute with the highest information gain (w.r.t. the sensitive attribute) to partition the data into smaller pure (w.r.t. the sensitive attribute) data partitions. We define the IGS in

Equation 4.4 in the following way:

$$IGS(F) = Info^S(D) - Info^S_F(D) \ .$$

$Info^S(D)$ and $Info^S_F(D)$ are calculated in a similar way as given in Equation 4.1 and 4.2 except now our objective is to make the smaller partitions of D pure w.r.t. the sensitive attribute. $IGS(F)$ gives us the expected reduction in the information requirement by partitioning the data w.r.t. the value of F.

Example 6 *Before going into the details of our proposed discrimination-aware criteria, we calculate the resultant accuracy and discrimination of the decision tree, constructed in our running example, using the traditional splitting criterion (IGC). Figure 4.3 shows a pruned tree constructed over the dataset given in Figure 4.1 using the splitting criterion. This tree has three leaves l_1, l_2, l_3 where leaves l_1 and l_2 are labeled with $+$ class and leave l_3 is labeled with $-$ class. When we test this leaned tree over the dataset given in Figure 4.1, it makes only one error in leaf l_2 and classifies a male applicant with $-$ class to $+$ class. Its accuracy is 90%. Now we calculate the discrimination in the prediction of this tree. It classifies all the male applicants into the positive class while it assigns the positive class to only two female applicants. It means that the discrimination is 60%. Even though this tree has a high accuracy but it is not a desirable classification model due to its high discrimination. It is important to mention here that in this example we use the same dataset for training and testing the decision tree just for the purpose of easy explanation. In our experiments, however, we train our decision tree on the training data and test it on unseen test data. All the reported discrimination and accuracy scores are over unseen test data.*

Based on these two measures IGC and IGS, we introduce two alternative criteria for determining the best split:

IGC-IGS: We only allow for a split if causes to reduce the discrimination, i.e., we select an attribute which is homogeneous w.r.t. class attribute but heterogenous w.r.t. the sensitive attribute. For this purpose, we subtract the gain in discrimination from the gain in accuracy. We further explain the impact of this discrimination-aware splitting criterion by using our running example of job application data.

Example 7 *Table 4.1 gives the information gain values for attributes. Figure 4.4 shows a decision tree built using IGC-IGS splitting criterion. When we use IGC-IGS as splitting criterion, the attribute Ethnicity is selected as the test attribute because it has the highest value for IGC-IGS criterion, as given in Table 4.1. This*

Table 4.1 Information gain values by using different split criteria. The value in bold shows the corresponding attribute as the test attribute for the corresponding split criterion.

Attribute	IGC	IGS	IGC-IGS	IGC/IGS
Job Type	**0.695**	**0.5**	0.195	1.39
H. Degree	0.285	0.115	0.17	2.47826087
Ethnicity	0.256	0.0001	**0.2559**	**2560**

tree has 80% accuracy and 0% discrimination. It means the with a loss of a little accuracy, our decision tree becomes discrimination-free. The decision tree given in Figure 4.4 supports our intention to use IGC-IGS as a splitting criterion that we want to make the leaves of the learnt tree pure w.r.t. the class attribute and impure w.r.t. the sensitive attribute. Both leaves of the tree have the same number of male and female applicants.

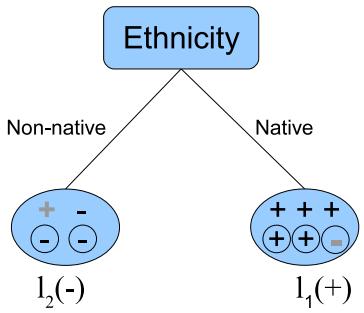


Figure 4.4 A pruned decision tree with the details of $+$ and $-$ objects in each leaf. The data objects with $Sex = f$ are encircled. The instances shown in gray color represent the misclassified instances. Acc=80% Disc=0%

IGC/IGS: We make a trade-off between accuracy and discrimination by dividing the gain in accuracy by the gain in discrimination. This splitting criterion also aims to make the decision trees pure w.r.t. to class attribute and impute w.r.t. the sensitive attribute.

When we use this splitting criterion to learn a decision tree in our running job

applicants example, we get the same tree as shown in Figure 4.4. However, the decision tree learnt by IGC/IGS is not always similar to one that uses IGC-IGS as splitting criterion.

As we show in our experiments that these discrimination-aware splitting criteria reduce the discrimination from the learnt classifiers to some extent but do not make the classifiers entirely discrimination-free. For this purpose we introduce a decision tree leaf relabeling approach in the next section.

4.3 Relabeling

In this section we discuss our second solution to the discrimination problem which is based the relabeling of some leaves of the learned tree. In *relabeling* we assume that a tree is 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 4.5 for an example; in this figure (top) 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 the 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. For example, in the tree we can compute the influence of relabeling a leaf on the accuracy and discrimination of the tree on a dataset D as follows. Let the joint distributions of the class attribute $Class$ and the sensitive attribute S 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. →	$-/+$	$-/+$	
$S = b$	U_1/U_2	V_1/V_2	F
$S = w$	W_1/W_2	X_1/X_2	M
	N_1/N_2	P_1/P_2	1

$Leaf\ l_1$

	$-$	$+$	
$S = b$	u	v	f
$S = w$	w	x	m
	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

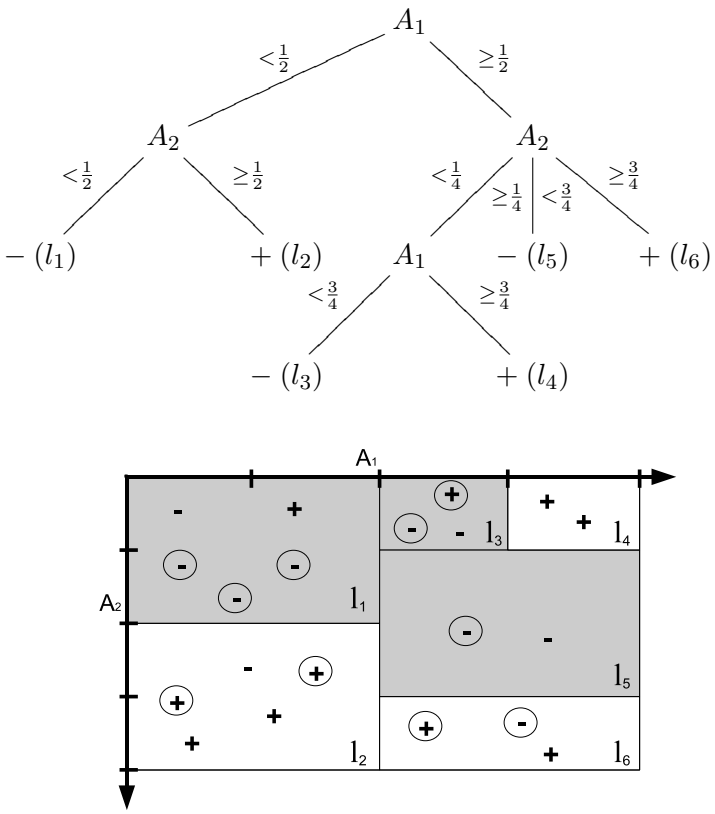


Figure 4.5 Decision tree with the partitioning induced by it. The $+$ and $-$ symbols in the partitioning denote the examples that were used to learn the tree. Encircled examples have $S = b$. The grey background denotes regions where the majority class is $-$.

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, denoted by acc_T and $disc_T$ respectively, *before* the label of the leaf l is changed:

$$\begin{aligned} acc_T &= N_1 + P_2 \\ disc_T &= \frac{W_2 + X_2}{M} - \frac{U_2 + V_2}{F} \end{aligned}$$

We will now study what will be the influence of relabeling one of the leaves. 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:

$$\begin{aligned} \Delta acc_l &= n - p \\ \Delta disc_l &= \frac{u+v}{f} - \frac{w+x}{m} \end{aligned}$$

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:

$$\begin{aligned} \Delta acc_l &= p - n \\ \Delta disc_l &= -\frac{u+v}{f} + \frac{w+x}{m} \end{aligned}$$

Notice that relabeling leaf l does not influence the effect of the other leaves because every leaf in a learned decision tree is independent of other leaves in the tree and that Δacc_l is always negative or zero.

Example 8 *Consider the dataset and tree given in Figure 4.5. The contingency tables for the dataset and leaf l_3 are as follows:*

$Dataset$

Class $\rightarrow$	$-$	$+$	
Pred. $\rightarrow$	$-/+$	$-/+$	
$S = b$	$\frac{5}{20}/\frac{1}{20}$	$\frac{1}{20}/\frac{3}{20}$	$\frac{1}{2}$
$S = w$	$\frac{3}{20}/\frac{1}{20}$	$\frac{1}{20}/\frac{5}{20}$	$\frac{1}{2}$
	$\frac{8}{20}/\frac{2}{20}$	$\frac{2}{20}/\frac{8}{20}$	1

$Leaf\ l_3$

	$-$	$+$	
$S = b$	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{2}{20}$
$S = w$	$\frac{1}{20}$	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 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 2 (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 3 (KNAPSACK) *Let a set of items $\mathcal{I}$, an integer bound K, and for every item $i \in \mathcal{I}$, a weight $w(i)$ and a profit $p(i) > 0$ 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 3 *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:

- $\mathcal{I} = \{\, l \in \mathcal{L} \mid \Delta disc_l < 0 \,\}$
- $w(l) = -\alpha \Delta disc_l \qquad$ *for all* $l \in \mathcal{I}$
- $p(l) = -\alpha \Delta acc_l \qquad$ *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 $I = \mathcal{I} \setminus L$. □

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 8 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.

Algorithm 8: *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

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

Corollary 1

1. *RELAB is **NP**-complete.*
2. *RELAB allows for a* fully polynomial approximation scheme (FPTAS) *[15].*
3. *An optimal solution to RELAB can be found with a dynamic programming approach in time $\mathcal{O}(|D|^3|\mathcal{I}|)$* [1]
4. *The difference in accuracy of the optimal solution and the accuracy of the tree given by Algorithm 8 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 8.*

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 $S = b$ and $Class = -$, and n leafs with respectively $i_1, \ldots, i_n$ tuples, all with $S = w$ and $Class = +$. 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 $lost\_acc = 1/6$.

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

Point 2 follows directly from the reduction of RELAB to KNAPSACK. 3 follows from the fact that α is at most $|D|(|D|S)(|D|\overline{S}) \leq |D|^3$ 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* [15]. The difference between the optimal solution and the greedy solution of Algorithm 8 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 8 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 9 *Consider again the example decision tree and data distribution given in Figure 4.5. The discrimination of the decision tree is* 20%*. Suppose we want to reduce the 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*. In this example we relabel leaf* l_3 *to make the decision tree discrimination-free, as shown in Figure 4.6. After relabeling of leaf* l_3*, as the leaf* l_3 *and* l_4 *have same labels, they are merged to make a single leaf.*

Optimal Relabeling With IGC+IGS Split: We introduce a new splitting criterion in which we add up the accuracy gain and the discrimination gain. It means that we want to construct a homogeneous tree w.r.t. both accuracy and the sensitive attribute. Our rationale behind this splitting criterion is that it leads to pure leaves w.r.t. the class attribute to achieve high accuracy and pure leaves w.r.t. the sensitive attribute which enable us make minimal changes when we relabel a the decision

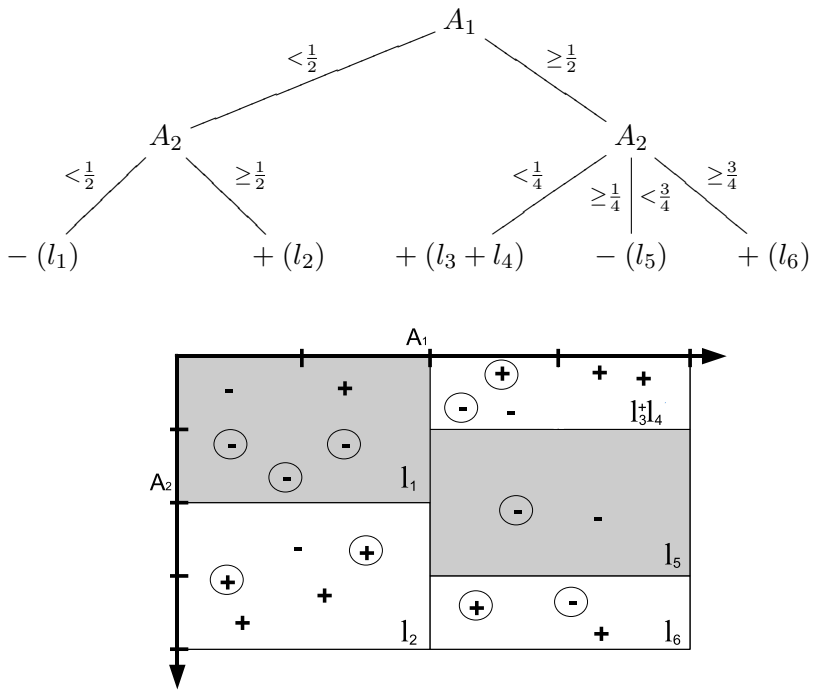


Figure 4.6 A decision tree after Relab. The leaf l_3 is relabeled and leaves l_3 and l_4 are merged.

tree leaves to make the decision tree discrimination-free. So IGC+IGS will lead to good results in combination with the relabeling technique as we show in our experiments.

4.4 Experiments

Datasets: We apply our proposed solutions on the **Adult** dataset, the **Communities** dataset, and two **Dutch census** datasets of 1971 and 2001 given in Section 1.3.1 of Chapter 1.

In this section we show the results of experiments with the new discrimination-aware splitting criteria and the leaf relabeling for decision trees. We observe that the discrimination-aware splitting criteria by themselves do not lead to significant improvements w.r.t. lowering discrimination, as shown in Figure 4.7. Figure 4.7

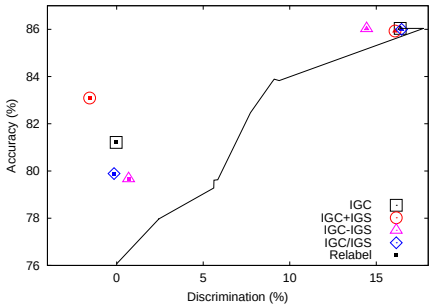


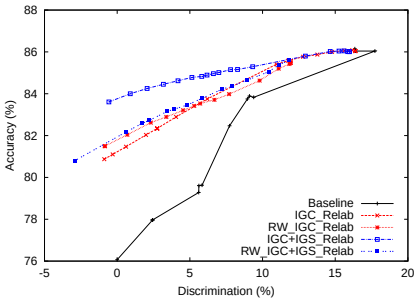
Figure 4.7 Results of the experiments when decision trees are learnt by using different splitting criteria in-combination with and without leaf relabeling. The continuous line shows baseline.

shows the results when we learn a decision tree with modified split criteria, the discrimination level does not go down that much. For split criterion IGC-IGS (label IGC-IGS), the discrimination level goes down but not up to the desired level (0%), in combination with relabeling, however, both IGC-IGS and IGC/IGS (label IGC/IGS) reduce the discrimination to 0% but at the cost of a lot of accuracy. 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 the 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. In the rest of the experiments, we only discuss the results of relabeling approach with IGC and IGC+IGS splitting criteria in more detail.

We compare the following techniques (between brackets their short name):

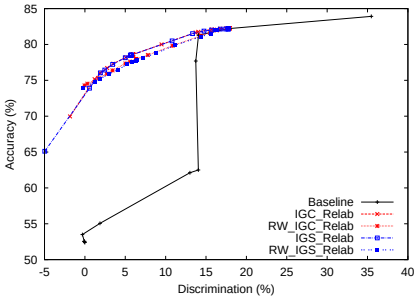
1. The baseline solutions (Baseline) that consist of removing S and its k most correlated attributes from the training dataset before learning a decision tree, for $k = 0, 1, \ldots, n$. In the graphs this baseline will be represented by a black continuous line connecting the performance figures for increasing k.

2. We also present a comparison to the previous state-of-the-art techniques,



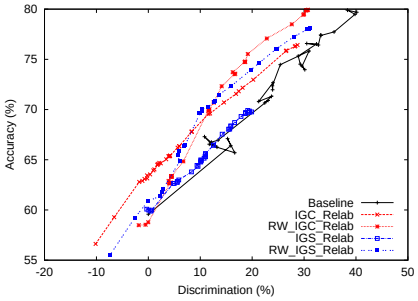
Baseline Disc=19.3 Acc=76.3

(a) Adult Data



Baseline Disc=29.85 Acc=52.39

(b) Dutch Census 2001 Data



Baseline Disc=43.14 Acc=59.58

(c) Communities Data

Figure 4.8 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).

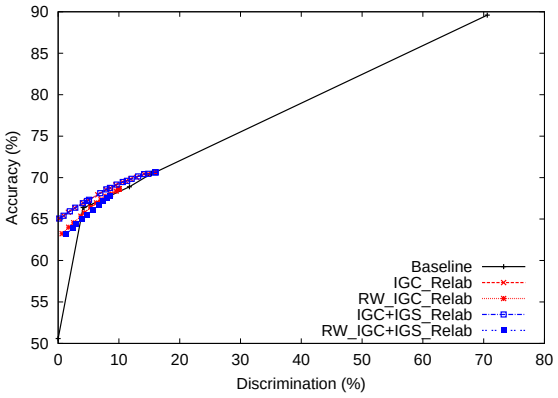
shown in Table 4.2, which includes discrimination-aware *Naive Bayesian* approaches [20], and the pre-processing methods *Massaging* and *Reweighing*, given in Chapter 3, that are based on cleaning away the discrimination from the input data before a traditional learner is applied.

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).

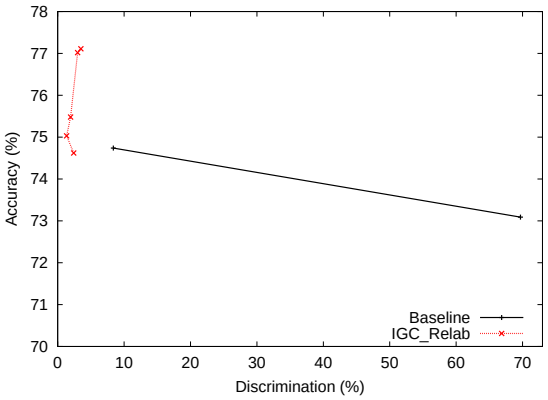
4. Finally we also show some hybrid combinations of the old and new methods; we present the results of experiments where we first applied the *Reweighing* technique on the training data to learn a tree with low discrimination (either with the normal or the new splitting criterion). On this tree we then 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.

4.4.1 Testing the Proposed Solutions

The reported figures are the averages of a 10-fold cross-validation experiment. In each experiment, we use the training data to learn and relabel our decision tree, and test it over unseen test data. 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. Figure 4.8 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 ran our proposed methods with both *Massaging* and *Reweighing* but we only present the results with *Reweighing* because both show similar behavior in our experiments.



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



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

Figure 4.9 The results of experiments when Dutch 1971 Census dataset is used as train set while the test set is different for both plots.

4.4.2 Sanity Check

It is very important to notice here that we measure the accuracy score here over the discriminatory data but ideally we expect 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. To validate this claim, we use the same experimental setup in our experiments as discussed in Section 3.2.6 of Chapter 3. 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 4.9 (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 4.9 (a)) to 73.09% (when tested on 2001 data; Figure 4.9 (b)). Figure 4.9 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 4.9 (b), we only show the results of IGC_Relab because other proposed methods also give similar results. Figure 4.9 (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 4.2 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 4.2 also gives a comparison of our proposed methods with the other state-of-the-art methods on the Adult dataset. We select the best results of the competitive methods to compare with. We observe that our proposed method outperform the others approaches w.r.t. the 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 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. the accuracy-discrimination trade off.

Table 4.2 The results of experiments over the Adult 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

4.5 Conclusion

In this chapter we have presented the construction of a decision tree classifier without discrimination. This is a different approach of addressing the discrimination-aware classification problem. The discrimination-aware techniques introduced in Chapter 3 are focused on "removing" discrimination from the training data and thus can be considered as "preprocessors". In this chapter on the contrary, we propose the construction of decision trees with non-discrimination constraints. Especially relabeling, for which an algorithm based on the KNAPSACK problem is proposed, showed promising results in an experimental evaluation. It is shown that the discrimination-aware decision trees 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 the training set is discriminatory and the 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 in Chapter 2.

Chapter 5

Conditional Discrimination-aware Classification

The discrimination-aware classification techniques discussed in Chapters 3 and 4 aim at removing all discrimination and do not take into account the fact that a part of the discrimination may be explainable by other attributes. For instance in the Adult dataset [14] one can observe that females have a lower annual income than males on average. However, one can also observe that females work fewer hours per week on average; see Table 5.1.

Table 5.1 Summary statistics of the Adult dataset

	hours per week	annual income (K$)
female	36.4	10.9
male	42.4	30.4
all data	40.4	23.9

Assume the task is to build a classifier to determine a salary, given an individual. The previous works would correct the decision making in such a way that males and females would get on average the same income, say 20 K$, leading to a reverse discrimination as it would result in male employees being assigned a lower salary than female employees for the same amount of working hours. Making the probabilities of acceptance equal for both would lead to favoring the group which is being deprived. In many real world cases, if the difference in the decision can be justified, it is not considered as bad discrimination.

In this chapter, we show that some of the differences in decisions across the sensitive groups can be explainable and hence tolerable. We take a step forward in designing discrimination-free classifiers and extend our discrimination problem setting. We argue that only the part of the discrimination which is not explainable by other characteristics should be removed. We observe that in such cases, the previously discussed discrimination-aware classification techniques tend to remove all the discrimination by ignoring the explainable part of the discrimination.

Therefore, in this chapter we analytically quantify how much of the difference in the decision making across the sensitive groups is objectively explainable, and how much is not. We aim at removing discrimination, but only if it is not explained by other attributes. We refer to the discrimination-aware classification under this condition as *conditional discrimination-aware classification*. With our analytical results we develop two new techniques for handling the unexplainable discrimination when one of the attributes is considered to be an explanatory attribute for the discrimination. Our proposed techniques are based on pre-processing the data before training a classifier so that only the discrimination that is not explainable is

removed. These two techniques are called *local massaging* and *local preferential sampling*. Finally we give an experimental evaluation which demonstrates that the new techniques remove exactly the bad discrimination, allowing the differences in decisions to be present as long as they are explainable.

5.1 Formal Setting

In general there is no objective truth which attribute is more reasonable to use as the explanation for the discrimination. Some attributes, such as relationships ('wife' or 'husband') are not a good explanation for low income if we want to remove the gender-discrimination, but for others it is situation-dependent. For instance, in many cases, high or low education is an appropriate reason to have different acceptance rates between ethnic groups. We assume that the convention of which attributes can be used as an explanation is given externally by law or by domain experts. We refer to such attributes as *the explanatory attributes*.

In this chapter we assume there is only one explanatory attribute $E \in A$ that is correlated with the sensitive attribute S, and at the same time gives objective information about the $Class$. Both relations can be measured in the data, for instance, as the information gain about S given E, and about $Class$ given E.

In the discrimination-aware classification techniques discussed in the previous chapters, the discrimination was considered to be present if the probabilities of acceptance for the favored community w and the deprived community b were not equal, i.e., $P(X(Class) = +|X(S) = w) \neq P(X(Class) = +|X(S) = b)$. Discrimination was measured as the difference between the two probabilities

$$D_{all} = P(X(Class) = +|X(S) = w) - P(X(Class) = +|X(S) = b).$$

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

$$D_{all} = D_{explainable} + D_{bad}. \tag{5.1}$$

In this study we are interested to remove and thus measure D_{bad}, which from Eq. (5.1) is

$$D_{bad} = D_{all} - D_{explainable}. \tag{5.2}$$

For that we need to find an expression for $D_{explainable}$. We will follow that same formal notations as given in Chapter 2. Additionally, in this chapter we make following assumptions.

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

5.2 Explainable and Bad Discrimination

To illustrate the difference between the explainable and bad discrimination, consider a toy example about the admission procedure of a fictitious university[1]. Gender is the sensitive attribute; male (m) and female (f) are the sensitive groups, against which discrimination may occur. There are two programs: medicine (med) and computer science (cs) with potentially different acceptance standards. Program is considered to be the explanatory attribute. In this example, we assume that the differences in acceptance statistics between male and female that can be attributed to different participation grades into the programs are acceptable. All applicants take a test for which their score is recorded (test). The acceptance (+) decision is made personally for each candidate during the final interview. Figure 5.1 shows the setting.

There are four relations between variables in this example. Relation (1) shows that the final decision whether to accept partially depends on the test score. Notice that the test scores are assumed to be independent from gender or program. Relation (2) shows that the probability of acceptance depends on the program. For example, the competition to medicine may be higher, thus less applicants are accepted in total. Relation (3) shows that the choice of program depends on gender. For instance, the larger part of the female candidates may apply to medicine, while more males apply to computer science. Relation (4) shows that acceptance also depends on gender, which is a bias in the decision making that is clearly a case of bad discrimination. Now we discuss different examples, where some of these dependencies are not present, to illustrate the effect of combinations of bad discrimination and explainable discrimination. In some cases, one kind of discrimination persists while in other both exist together.

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

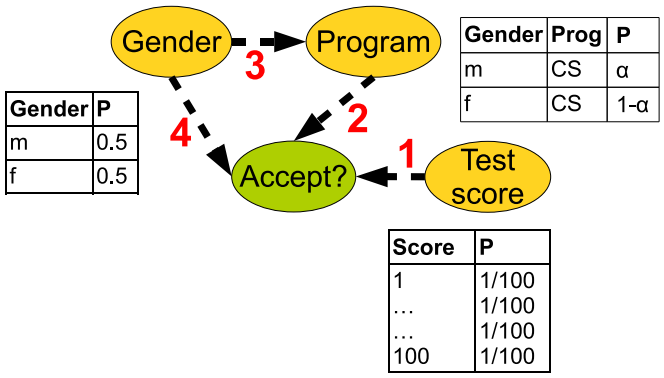


Figure 5.1 Toy example. Tables show the probability distribution of each attribute.

5.2.1 How Much Discrimination is Explainable?

We will now discuss a couple of scenarios to show the different combinations of bad and explainable discrimination. In our first scenario we give an example to show that all the discrimination is explainable while in our second scenario we give an example to show a case with both explainable and bad discrimination.

Only explainable discrimination (Example 1): Assume there are 2000 applicants, 1000 males and 1000 females. Each program receives the same number of applicants, but medicine is more popular among females. Assume further that medicine is more competitive. The situation is described in Table 5.2. Within each program males and females are treated equally. But when we count the final scores, it appears that 36% of males were accepted, but only 24% of females. The difference is explained by the fact that more females applied to the more competitive program. Thus, there is no bad discrimination. Such a case is reported in

Table 5.2 No bad discrimination.

	medicine		computer	
	female	male	female	male
number of applicants	800	200	200	800
acceptance rate	20%	20%	40%	40%
accepted (+)	160	40	80	320

the Berkely study [17]. Examination of aggregate data on graduate admissions to

the University of California, Berkeley, for fall 1973 shows a clear but misleading pattern of bias against female applicants. Apparently there is 9% discrimination (D_{All}) towards female applicants, i.e., overall 44% of males and 35% of female applicants are admitted. However, the examination of pooled data w.r.t. different departments, shows that there is a small but statistically significant bias in favor of females. It means that the overall low admission rate for females is explainable by their tendency to apply to graduate departments that were more difficult for applicants of either sex to enter. The case concluded that there was no discrimination.

Both explainable and bad discrimination (Example 2): Assume a similar situation, but now there is a bias in the decision making favoring males, as shown in Table 5.3. The programs obviously have different aggregated acceptance rates, medicine 17% and computer science 43%. It appears that in total 19% of females and 41% of males are accepted. We want to know which part of this difference is explainable by program, and which part is due to bad discrimination.

Table 5.3 Bad discrimination is present.

	medicine		computer	
	female	male	female	male
number of applicants	800	200	200	800
acceptance rate	15%	25%	35%	45%
accepted (+)	120	50	70	360

First we need to settle what would have been the correct acceptance rates $P^\star(+|med)$ and $P^\star(+|cs)$ within each program e_i if both genders would have been treated equally. Then we can find which part of the difference between the genders is explainable, and treat the remaining part as bad discrimination that needs to be removed. Finding the correct acceptance rates, however, is challenging, as there is no unique way to do it. One can only assume what would have happened if there was no gender bias in acceptance decisions. Would all the acceptance rate have been as the ones for males now, or all as for females? Or an average of the two?

In this study we refer to the discrimination model given in Section 2.3.1 of Chapter 2 to find the correct acceptance rates. Under this model, it is reasonable to assume that roughly the same fraction of men would benefit from the bias (those that are at most d below the acceptance threshold), as there are females that have a disadvantage due to the bias (those that are at most d above the threshold), because within the programs, males and females are assumed to be equally capable. Under this assumption we should take the average of the acceptance probability of males and females, resulting in 20% for medicine and 40% for CS. In contrast, if we would fix the number of positive labels in the groups to the number observed in the discrim-

inatory data, we would get $170/1000 = 17\%$ for medicine and $440/1000 = 44\%$ for computer science. Following the rationale of the discrimination model of Section 2.3.1 of Chapter 2, however, these numbers are skewed and would result in programs mainly populated by females to be perceived as being more selective, leading to redlining. It means that it would transfer the discrimination from gender to program; program with lots of females would receive an overall lower acceptance rate.

Thus we assume that the acceptance thresholds would have been fixed as the average of the historical acceptance thresholds for males and females. This choice is motivated by the case that the candidates come one by one, and that any of the candidates that is sufficiently qualified would get a position, or salary level, or loan. Hence, there is no resource constraint and we can assume that the number of positive outputs only depends on which instances qualify.

An alternative scenario would be to assume that all the candidates come in batch within a deadline. Then the candidates are ranked and a fixed number of the best candidates are offered a position. Whether to keep the number of accepted persons fixed or to keep the acceptance threshold fixed depends on the application domain. For instance, in case of scholarships, job application, university acceptance fixing the number of persons may be more reasonable, since the applicants come in batch at the deadline. In case of deciding to grant a credit or what salary level to apply, fixing the threshold makes more sense (accept all individuals that pass qualification requirements), since the individuals come one by one. We argue that the choice of acceptance scenario is situation dependent and hence not part of the non-discrimination techniques' design.

Table 5.4 illustrates the calculation of the explainable part for the discrimination toward females as given in Table 5.3. We find the correct acceptance rate within each program as the average of male and female acceptance rates. To find the correct acceptance rate we first need to find the acceptance thresholds, i.e., the rates at which males and females are accepted if they apply for the programs at the same rates. Then we take an average of the two. We illustrate the calculation in Table 5.4. With the new rates we get the same situation as in the Example 1, where 36% of the male and 24% of the female were accepted. Thus, $D_{explainable} = 36\% - 24\% = 12\%$. Therefore we get $D_{all} = 41\% - 19\% = 22\%$. From Eq.(5.2) we get
$D_{bad} = D_{all} - D_{explainable} = 22\% - 12\% = 10\%$. Thus, there is 10% of bad discrimination in the data.

The explainable discrimination is the difference between acceptance of males and females, if every individual with value e_i in the explanatory attribute would have

Table 5.4 Calculating the explainable difference

	medicine		computer	
	female	male	female	male
number of applicants	800	200	200	800
acceptance rate (Example 2)	15%	25%	35%	45%
corrected acceptance rate	20%		40%	
accepted explainable	160	40	80	320

the same chance

$$P^\star(+|e_i) := \frac{P(+|e_i, m) + P(+|e_i, f)}{2} \tag{5.3}$$

to be accepted, independent of the gender:

$$\begin{aligned} D_{explainable} &= \sum_{i=1}^{k} P(e_i|m)P^\star(+|e_i) - \sum_{i=1}^{k} P(e_i|f)P^\star(+|e_i) \\ &= \sum_{i=1}^{k} \left(P(e_i|m) - P(e_i|f)\right) P^\star(+|e_i) \end{aligned} \tag{5.4}$$

where we assume $dom(E) = e_1, \ldots, e_k$, $P(e_i|m)$ and $P(e_i|f)$ are observed from data, and $P^\star_c(+|e_i)$ is calculated as in Eq.(5.3).

The bad discrimination can thus be computed as the difference between $D_{all} = P(+|m) - P(+|f)$ and $D_{explainable}$:

$$D_{bad} = \left(P(+|m) - P(+|f)\right) - \sum_{i=1}^{k} \left(P(e_i|m) - P(e_i|f)\right) P^\star(+|e_i) \tag{5.5}$$

5.2.2 Illustration of the Redlining Effect

Now that we formalized what is bad and explainable discrimination, our next step is to analyze under what circumstances a trained classifier risks to capture bad discrimination. The effect of redlining over discrimination-aware classification is already discussed in Chapters 3 and 4 in quite detail. In this section however, we further explore it with reference to conditional discrimination-aware classification. We continue to assume that there is one explanatory attribute, which is correlated with the sensitive attribute. We measure D_{bad} as defined in Eq.(5.2).

For our analysis we use synthetic data based on our toy example introduced in Figure 5.1. We generate 10 000 male and 10 000 female instances. The (integer) test

scores $T \in [1, 100]$ are assigned uniformly for any individual. In all experiments all probabilities in the Belief network (given in Figure 5.1) are fixed, except for the probabilities $P(e_i|S)$: for $\alpha \in [0, 1]$, we will generate data with: $P(med|f) = \alpha$, $P(cs|f) = 1 - \alpha$, $P(med|m) = 1 - \alpha$, and $P(cs|m) = \alpha$. In this way we can study the influence of the strength of the relationship between gender and program on the discrimination, while the total number of people applying for med (resp. cs) remains the same. The farther away α is from 0.5, the stronger the dependency between the explainable and sensitive attribute becomes. Hence, the closer $P(med|f)$ will be to 0.5, the less explainable discrimination there will be. The probability of acceptance is varied in the three experiments; in the first one it depends on program only (all discrimination is explainable), in the second experiment it depends on gender only (all discrimination is bad), while in the third experiment it depends on both (discrimination is partly explainable and partially bad). For all experiments, we plot the all-discrimination and the bad discrimination in function of $\alpha = p(med|f)$, as well as for a classifier learned on this data after gender was removed. These experiments show that our discrimination measure is in-line with the intuition, and that removing gender from the training data does not necessarily remove the bad discrimination, because of the *redlining effect*.

Case I: no bad discrimination, everything is explainable. Several observations

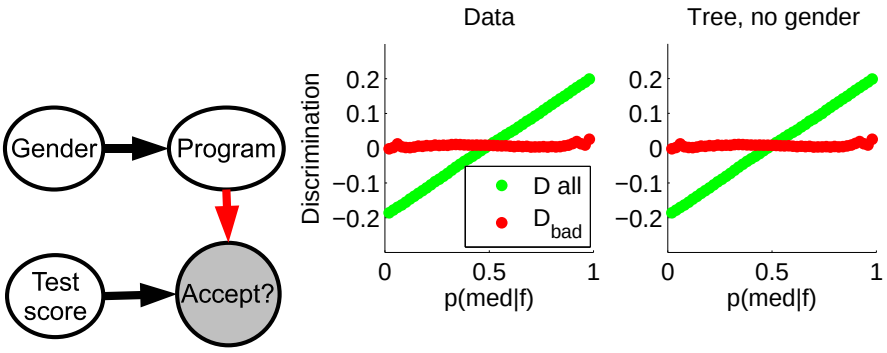


Figure 5.2 Case I: no bad discrimination. All the discrimination is explainable by the different choice of program by males and females.

can be made from the plots. Since the label in the historical data does not depend on the gender, there is no discrimination problem in the learned classifier. The

probabilities of accepting a male and a female are different if gender and program is strongly correlated, but all the difference is explainable. The probabilities of accepting a male and a female are equal if the choices of programs are the same for both genders, i.e., $P(med|f) = P(med|m) = P(med) = 0.5$.

Figure 5.2 illustrates the situation when all difference in acceptance is explainable by program. It corresponds to Example 1 in Table 5.2. We plot discrimination as a function of $P(med|f)$, which determines how strongly the sensitive attribute (gender) is related with the explanatory attribute (program). We fix $P(med|m) = 1 - P(med|f)$ to keep the same number of applicants.

First we plot all discrimination D_{all} and bad discrimination D_{bad} in the testing data with the original labels. In addition, we plot the resulting discriminations with the predicted labels by a decision tree. A decision tree is learned on the training data from which gender has been removed. Thus, the training data includes only the program and the test score.

Case II: only bad discrimination, program does not explain the label. Figure 5.3 illustrates an opposite case of the first one. Here all discrimination is bad, therefore in the plots D_{all} and D_{bad} are the same. In this case there is no direct relation between program and label. However, when the gender attribute is removed, the learned decision tree captures the discriminatory decisions indirectly through program; i.e., the *redlining effect* appears. The effect is strong when gender and program are strongly correlated. There is no redlining ($D_{bad} = 0$) if program and gender are independent, i.e., $P(med|f) = P(med|m) = P(med) = 0.5$.

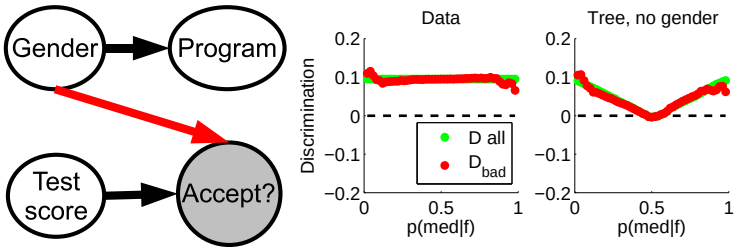


Figure 5.3 Case II: only bad discrimination. The acceptance depends on the gender of candidates.

Notice that in this extreme case the classifier can be easily made discrimination-free by removing both gender and program from the input space, without losing

any useful information.

Case III: explanatory and bad discrimination. Figure 5.4 illustrates the situation when explanatory discrimination and bad discrimination are together. It corresponds to Example 2 in Table 5.3. The data contains 10% discrimination in the decision (red line), while the probability of accepting a male or a female vary depending on the application ratios by each gender.

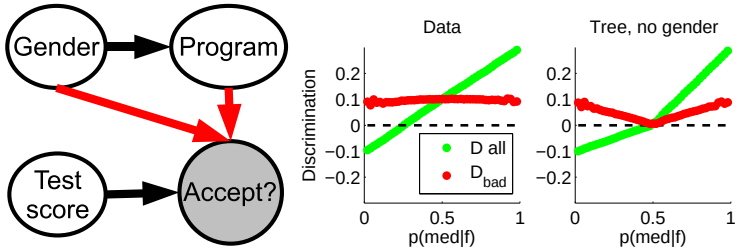


Figure 5.4 Case III: explanatory and bad discrimination together. The acceptance partially depends on the gender of candidates.

The learned decision tree shows interesting results. The actual bad discrimination captured by the tree is the same as in Case II (red line). The overall probabilities of acceptance differ (green line) due to the direct relation between program and label. The figure with the decision tree results actually illustrates the Simpson's paradox [79], in which a relation present in different groups is reversed when the groups are combined. We can see that if very few females apply to medicine ($p(med|f)$ is close to zero), which is more competitive program, then D_{all} (green line) indicates that females are favored, while in fact they are deprived, as 10% of bad discrimination is present (red line).

To sum up, the experiments with simulated data demonstrate the following effects:

- removing the sensitive attribute does not remove discrimination if the sensitive attribute is correlated with other attributes (Cases II and III);
- if the input attribute is at the same time correlated with the sensitive attribute and the class label and is nominated as explainable, not all difference in probabilities of acceptance is bad, some difference is explainable; removing all the difference in such case would result in reverse discrimination;
- Case III shows that there is a need for advanced training strategies to remove discrimination, and at the same time to leave the objective information,

which could be captured by one and the same variable.

5.3 How to Remove Bad Discrimination When Training a Classifier?

Removing the sensitive attribute only works if no other attribute is correlated with the sensitive attribute. In real life scenarios, however, this condition does not hold and more involved strategies to remove discrimination are required.

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

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

2. $P_c(+|e_i) = P^\star(+|e_i)$, where $P^\star(+|e_i)$ is defined in Eq. 5.3. This means that the prediction is consistent with the original distribution of the data.

Recall that, as we discussed before, the first condition in isolation is insufficient due to the *redlining effect*; a classifier that only takes this condition into account would under-estimate the positive class probability of a group in which females are over-represented.

We distinguish two main strategies that could make classifiers free from bad discrimination. The first strategy is to remove the relation between the sensitive attribute and the class label from the training data, which is the source of the bad discrimination (relation (1) in Figure 5.5). Note that removing the relation is not the same as removing the sensitive attribute itself, it means making $P(+|med, f) = P(+|med, m) = P^\star(+|med)$. We can do that, for instance, by modifying the original labels of the training data.

There is an alternative. The data can be split into smaller groups based on the explanatory attribute. Then individual classifiers can be trained for each of the group. It would remove the relation between the sensitive and the explanatory attributes. It would also require to correct the training labels in each groups, otherwise the *redlining effect* will manifest. In addition, it would significantly reduce the data available for training a classifier, which is undesirable.

In this section we concentrate on the first type of strategy, which is simpler than the second type and do not have the drawbacks of the third type.

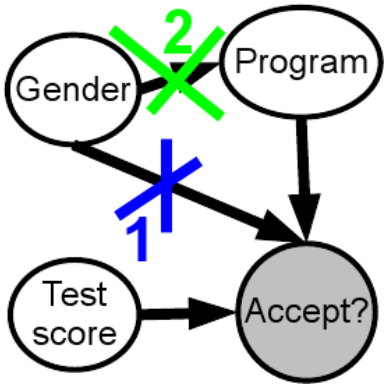


Figure 5.5 To remove bad discrimination (due to gender) remove the relation 1 or remove the relation 2.

Hence, we aim to modify the labels of the training data so that the outputs of the trained classifiers would satisfy $P_c(+|e_i, f) = P_c(+|e_i, m) = P^\star(+|e_i)$. The first design choice is to fix the desired probabilities of acceptance $P_c(+|e_i)$, which would have been correct. In this study we choose $P_c^\star(+|e_i)$ to be the average of male and female acceptance rates, Eq. (5.3). After finding $P^\star(+|e_i)$ for all $e_i \in E$, the remaining part is to modify the training data so that $P'(+|e_i, f) = P'(+|e_i, m) = P^\star(+|e_i)$. ($P'$ denotes the probability in the modified data). In this work we propose two techniques: local massaging and local preferential sampling. 'Local' emphasizes the contrast with previous works given in Chapter 3 that used global modification and thus risked overshooting.

Local Massaging

In local massaging, for every partition in the training data induced by the explanatory attribute, we will change labels until $P'(+|m, e_i)$ and $P'(+|f, e_i)$ both become equal to $P^\star(+|e_i)$. The massaging technique changes the values of labels of the selected instances. To this end, massaging selects the instances that are close to the decision boundary and switches their labels. For the selection of instances close to boundaries, we learn a ranker on each partition to order the instances according to their probability of acceptance. This choice is motivated by the discrimination model, presented in Section 2.3.1 of Chapter 2, which implies that discrimination is worse for objects closer to the decision boundary. In the admission example this implies that in both the medicine and the cs group, we change the labels of negatively classified females and positively classified males that are the closest to the

boundary until P' becomes as desired. This technique is the local variant of the massaging technique proposed in Chapter 3.

Suppose females have been discriminated and this discrimination is reflected in the historical data. The massaging technique will select a number of females that were almost accepted, and switch their labels to positive. It would also select a number of males that were most likely to be rejected, but have not been rejected, and switch their label to negative.

Our new local massaging uses the same principles as the massaging technique of Chapter 3. However, local massaging works on the partitioned data, within each program separately. In addition, it also modifies and controls the number of accepted males and females according to Eq. (5.3), to ensure no redlining. A procedure for local massaging is illustrated in Figure 5.6 and the pseudo-code is given in Algorithm 9.

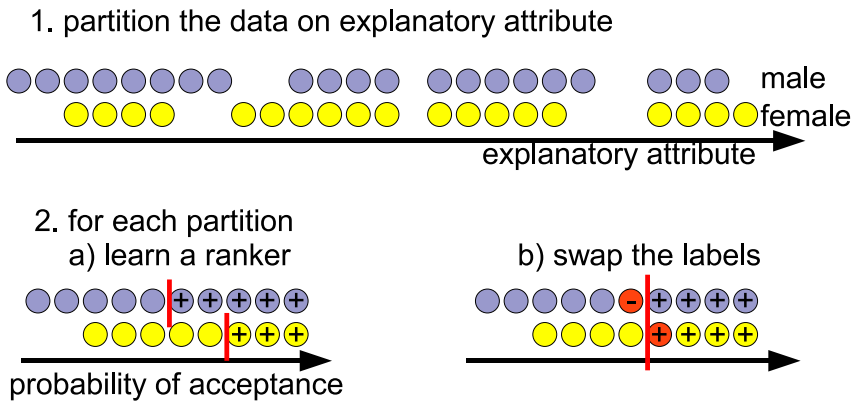


Figure 5.6 Local massaging.

Local Preferential Sampling

The preferential sampling technique, discussed in Section 3.1.3 of Chapter 3, does not modify the training instances or labels. Instead it modifies the composition of the training set. It deletes and duplicates training instances so that the labels of new training set contain no discrimination and satisfy the criteria in Eq. (5.3).

Our new local preferential sampling applies the same principles of preferential sampling but now locally to partitions of the data. It modifies and controls the number of accepted male and female according to Eq. (5.3), to ensure no redlin-

Algorithm 9: Local massaging

Input : dataset $(\mathbf{X}, S, E, Class)$
Output: modified labels $\hat{Class}$

```
1 PARTITION (X, E) (Algorithm 10);
2 for each partition X^(i) do
3 |   learn a ranker H_i : X^(i) → Class^(i);
4 |   rank males using H_i;
5 |   relabel DELTA (male) males that are the closest to the decision
  |   boundary from + to − (Algorithm 11);
6 |   rank females using H_i;
7 |   relabel DELTA (female) females that are the closest to the decision
  |   boundary from − to +
8 end
```

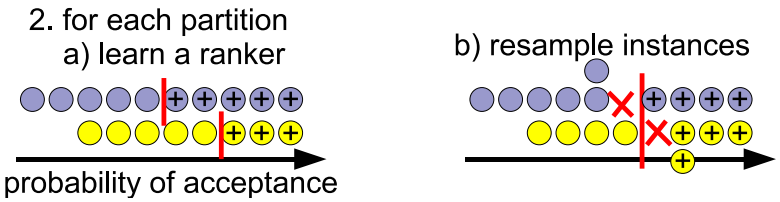


Figure 5.7 Local preferential sampling.

ing. The procedure for local preferential sampling is presented in Figure 5.7 and summarized in Algorithm 12.

Algorithm 10: subroutine PARTITION($\mathbf{X}$, E)

```
1 find all unique values of E: {E_1, E_2, ..., E_k};
2 for i = 1 to k do
3 |   make a group X^(i) = {X : E = e_i};
4 end
```

5.4 Experiments

In this section we compare the performance of the proposed local discrimination handling techniques with their global counterparts. The objective is to minimize

Algorithm 11: subroutine DELTA(gender)

return $G_i|P(+|e_i, \mathsf{gender}) - P^\star(+|e_i)|$,
where $P^\star(+|e_i)$ comes from (Eq. (5.3)),
G_i is the number of gender people in $X^{(i)}$;

Algorithm 12: Local preferential sampling

Input : dataset $(\mathbf{X}, S, E, Class)$
output: resampled dataset (a list of instances)

PARTITION ($\mathbf{X}$, E);
for *each partition* $X^{(i)}$ **do**
    learn a ranker $\mathcal{H}_i : X^{(i)} \rightarrow Class^{(i)}$;
    rank males using $\mathcal{H}_i$;
    delete $\frac{1}{2}$DELTA (male) males $+$ that are the closest to the decision boundary;
    duplicate $\frac{1}{2}$DELTA (male) males $-$ that are the closest to the decision boundary;
    rank females using $\mathcal{H}_i$;
    delete $\frac{1}{2}$DELTA (female) females $-$ that are the closest to the decision boundary;
    duplicate $\frac{1}{2}$DELTA (female) females $+$ that are the closest to the decision boundary;
end

(the absolute value of) *bad* discrimination while keeping accuracy as high as possible. It is important not to overshoot and end up with reverse discrimination. The experiments have the following goals:

1. present a motivation for conditional discrimination-aware classification research,
2. explore how well the proposed techniques remove bad discrimination as compared to the existing techniques for global non-discrimination, and
3. analyze the effects of removing discrimination on the final classification accuracy.

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

5.4.1 Data

In our experiments we apply our proposed methods on the **Adult** dataset discussed in Section 1.3.1 of Chapter 1. Some instances from the Adult dataset with missing values of explanatory attributes are removed and the Adult dataset we use in our experiments consists of $47,696$ instances, which are described by 13 attributes and a class label. Originally 6 of the 13 attributes were numeric attributes, which we discretized. Gender is the sensitive attribute, income is the label. We repeat the experiment several times, where any of the other attributes in turn is selected as explanatory. Figure 5.8 (top) shows the discrimination in the dataset. The blue line shows D_{all}, while the red line shows D_{bad}. The horizontal axis denotes the index of the explanatory attribute.

We observe the following from the plots. First, there are several attributes that are not strongly correlated with gender, such as workclass, education, occupation, race, capital loss, native country. This implies that picking any of those attributes as explanatory will not change the situation w.r.t. discrimination much. For instance, we know from biology that race and gender are independent. Thus, race cannot explain gender discrimination; the discrimination is either bad or it is due to some other attributes. Indeed, we see that all discrimination with race (attribute $\#7$) as explanatory attribute is bad.

On the other hand, it can be seen that the relationship attribute explains a lot of D_{all}. Whether relationship is an acceptable argument to justify differences in income is

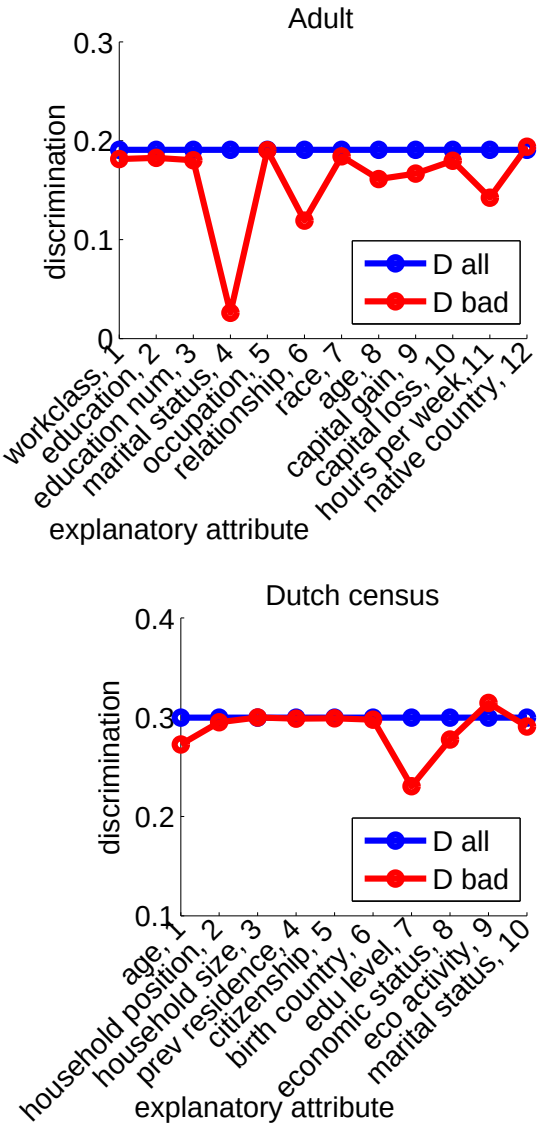


Figure 5.8 Discrimination in the datasets.

a question for lawyers to answer. In this case it is unlikely since this attribute contains values 'wife', 'husband' which clearly capture gender information. From a data mining perspective, if we treat it as acceptable, a lot of the difference is explained by this attribute.

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

The **Dutch Census dataset of 2001**, given in 1.3.1 of Chapter 1, represents aggregated groups of inhabitants of the Netherlands in 2001. We formulated a binary classification task to classify the individuals into 'high income' (prestigious) and 'low income' professions, using occupation as the class label. Individuals are described by 11 categorical attributes (including gender). We remove the records of under-aged people, several professions in the middle level and people with unknown professions, leaving 60 420 instances. Gender is the sensitive attribute.

Figure 5.8 (bottom) presents the discrimination contained in the data. The difference is much less apparent than in the Adult data. This means that in many cases the other attributes are not that strongly correlated with the sensitive attribute. Just removing the sensitive attribute should therefore perform reasonably well. Nevertheless, education level, age and economic activity (i.e., economic status) present cases for conditional non-discrimination, thus we explore this dataset in our experiments.

5.4.2 Motivation for Experiments

To give a motivation for our approaches we illustrate that the discrimination-aware techniques discussed in chapters 3 and 4 do not solve the conditional non discrimination problem.

Removing the Sensitive Attribute

First we test the following baseline approach. We explore what happens if the sensitive attribute is removed from the training data. We learn a decision tree with the J48 classifier (Weka implementation) on all the data except the gender attribute. Figure 5.9 shows the resulting discriminations, when the tree is tested using 10-fold cross validation.

The *redlining effect* is clearly present, especially in the Adult data. Even though the sensitive attribute is removed, the bad discrimination is still present.

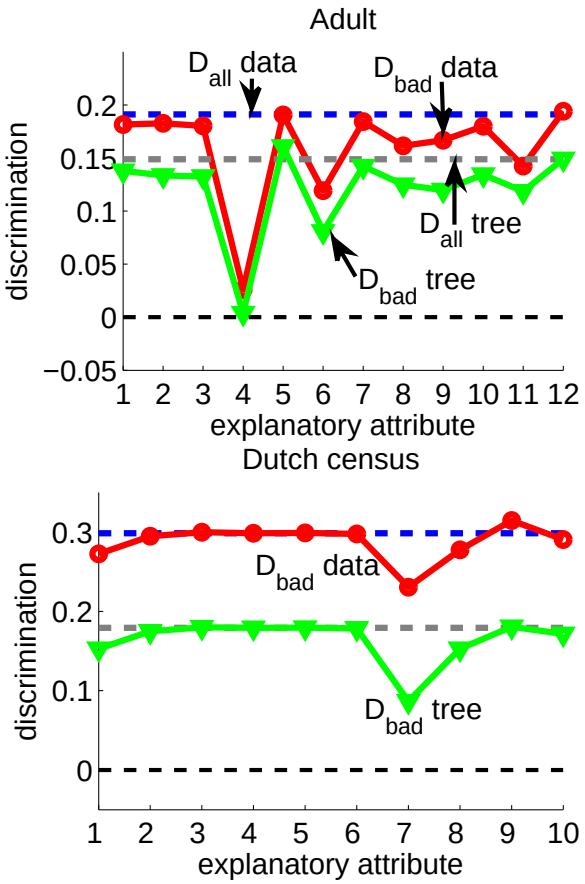


Figure 5.9 The discrimination w.r.t. different explanatory attribute when we learn a decision after removing the sensitive attribute from the dataset.

Global Techniques

Next we show to what extent two global techniques, discussed in Chapter 3, remove bad discrimination. Global massaging modifies the labels of the training data to make the probabilities of acceptance equal for the two sensitive groups. Global preferential sampling, samples the training data so that non-discrimination constraints for the label distribution are satisfied. Both methods aim at making D_{all} equal to 0, which is not the same as removing D_{bad} and will actually reverse the discrimination, as is illustrated in Figure 5.10. The global techniques do not take into account, that the distributions of the sensitive groups may differ and thus some

of the differences in probabilities are explainable.

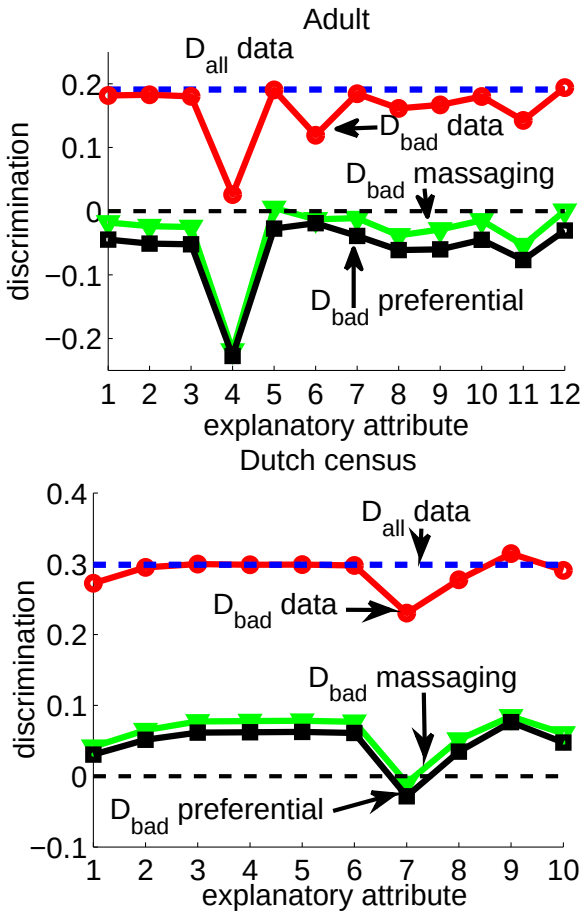


Figure 5.10 Discrimination when we make our training set discrimination-free with the *global* discrimination-aware techniques.

As expected, the massaging and preferential sampling techniques work well for removing all discrimination. For the Adult data, $D_{all} = 0$ after massaging. But, if we treat 'marital status' as the explanatory attribute, the result is a reverse bad discrimination. The same, but on a smaller scale, holds for several other explanatory attributes; especially for 'hours per week' and 'age'. On the Dutch Census data, both techniques overshoot if conditioned on 'education level'.

The results show that a reverse bad discrimination is introduced when global discrimination handling techniques are applied, illustrating the need for local meth-

ods.

When are the Local Techniques Essential?

Existing techniques fail mostly when the difference between D_{all} and D_{bad} in the data is large. For instance, the Adult data in Figure 5.8 shows sharp negative peaks when 'marital status' or 'relationship' are the explanatory attributes. In these cases, specific techniques for handling conditional discrimination are essential.

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

$$IGC(E) = Info^{Class}(D) - Info_E^{Class}(D), \text{ and}$$
$$IGS(E) = Info_S(D) - Info_E^S(D).$$

$IGC(E)$ and $IGS(E)$ represent the information gain, given in Section 4.1.1 of Chapter 4, for the class attribute and the sensitive attribute w.r.t. the values of E respectively. The information gains for the Adult and the Dutch census datasets are plotted in Figure 5.11. The figure confirms the intuition that the stronger the relation with the explanatory attribute (higher information gain) the larger the share of the total discrimination that is explainable. See Figure 5.8 for the discriminations.

5.4.3 Non-discrimination Using Local Techniques

We analyze how the proposed local strategies handle discrimination. We expect them to remove exactly the bad discrimination and nothing more. We test the performance with decision trees (J48) via 10-fold cross validation.

Figure 5.12 shows the resulting discrimination after applying local massaging and local preferential sampling. Both local techniques perform well on the Adult data. Bad discrimination is reduced to nearly zero, except for relationship as explanatory attribute when massaging is applied to the Adult dataset. Our techniques do not produce reverse discrimination as in the case of global massaging.

In the case of the Dutch census dataset, the proposed solutions do not perform that well, as the sensitive attribute is not very strongly correlated with any other attribute in the dataset, because local techniques are primarily designed to handle high correlations with the sensitive attribute that make redlining possible.

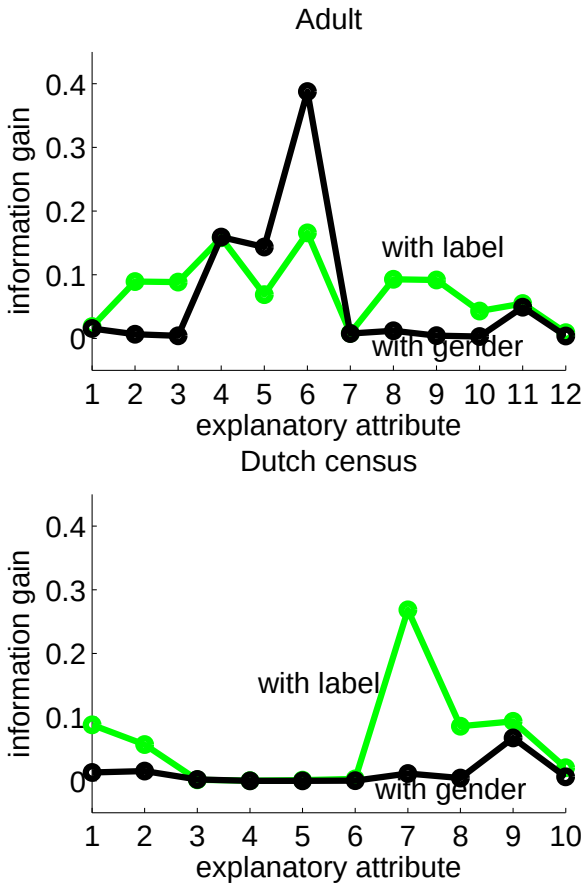


Figure 5.11 Relations between sensitive, explanatory attributes and labels.

We emphasize that in a case where the base classifier can also serve as an accurate ranker, there is a simpler local approach to use our discrimination measure. Different rankers can be learned for males and females and used directly for classification simply setting the thresholds to $p^\star(+|e_i)$ as in Eq. (5.3).

5.4.4 Accuracy with the Local Techniques

When classifiers become discrimination-free, they may lose some accuracy, as measured on the historical data. Let us look at the resulting accuracies, when local massaging and local preferential sampling techniques are applied. Figure 5.13 presents the testing accuracy of a decision tree (J48) when all the attributes are

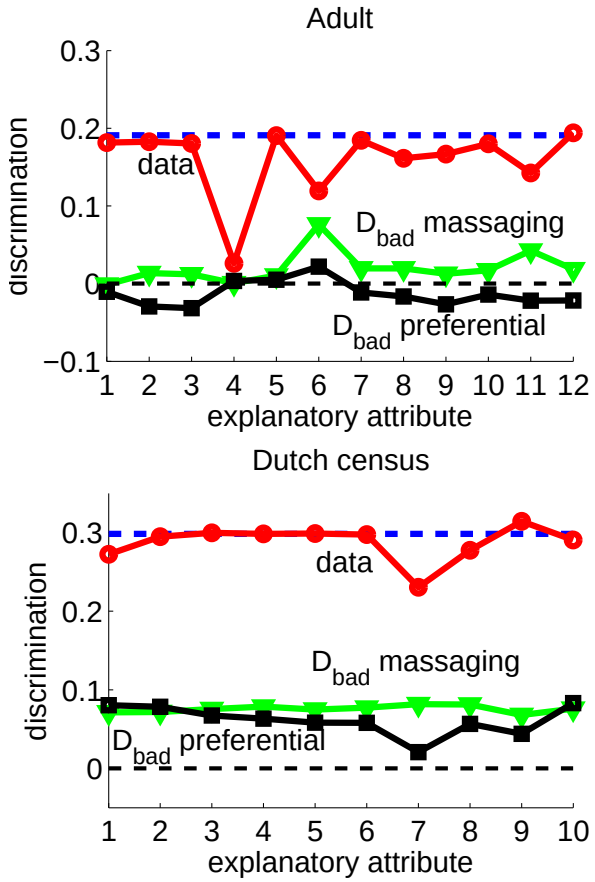


Figure 5.12 Discrimination when we make our training set discrimination-free with *the local* discrimination-aware techniques and then learn a decision tree over it.

used, and the results of our local techniques. The accuracy of the local techniques decreases as the testing is done on the data that contains discrimination. Nevertheless, the absolute accuracy remains high; it drops only by 5% at most. Our experiments demonstrate that the local massaging and the local preferential sampling classify future data with reasonable accuracy and maintain low discrimination.

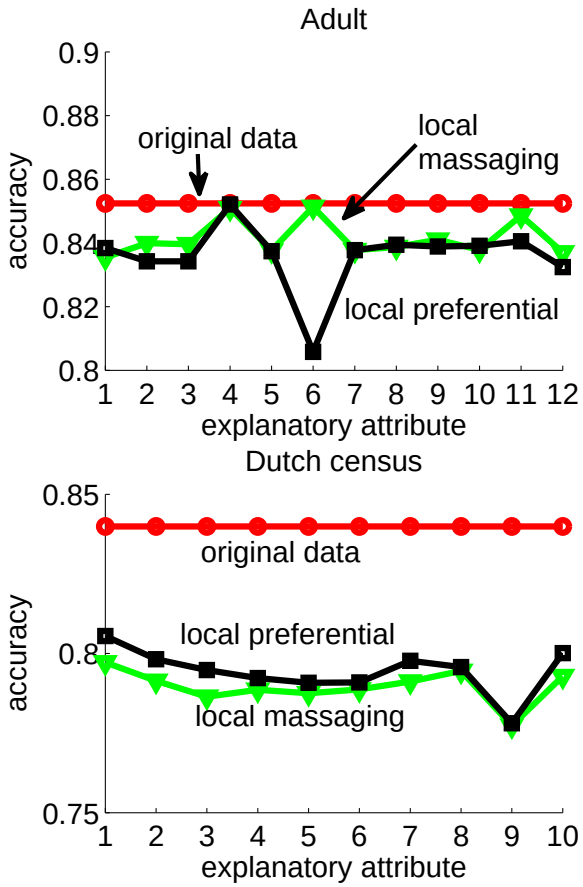


Figure 5.13 Accuracy when we make our training set discrimination-free with *the local* discrimination-aware techniques and then learn a decision tree over it.

5.5 Conclusion

In this chapter we considered the discrimination-aware classification paradigm in the presence of an explanatory attribute that is correlated with the sensitive attribute. In such a case, as we showed, not all discrimination can be considered bad and the previous techniques tend to overshoot and start reverse discrimination. Therefore, we introduced a new way of measuring discrimination, by explicitly splitting it up into explainable and bad discrimination. Local variants of the relabeling and preferential sampling were introduced and experimentally evaluated. The experiments demonstrated the usefulness of the new local techniques, espe-

cially in cases when the sensitive attribute is highly correlated with the explanatory attribute.

Chapter 6

Related Work

Since the American Civil War the term *discrimination* evolved in American English usage as an understanding of prejudicial treatment of an individual based solely on their race, later generalized as membership in a certain socially undesirable group or social category [37]. The concept of discrimination is relatively new in data mining but it has been studied in social sciences for a long time. We broadly categorize the related work of the discrimination-aware classification problem into the related works in social sciences and the related works in data mining.

6.1 Social Sciences

The term *social sciences* is commonly used to refer the disciplines that study different aspects of society. It includes criminology, economics, law, sociology, education, geography, archaeology, business administration, linguistics, political science, communication, history, and psychology. The problem of discrimination is pervasive in almost every society for a long time. The general forms of discrimination: racism, sexism, ableism, ageism, casteism, classism, colorism, linguicism, and rankism are referred to the social discrimination on the basis of race, gender, disabilities, age, caste, social class, skin or eye color, language, and rank of a person respectively.

Here we discuss the most important types of social discrimination. The term *racism* [81] often refers to the belief that one group of people with a particular biological make up or race is superior to other groups with different biological make ups. This racial discrimination has existed throughout the human history. It has continued to persist both in westen and eastern societies. Joe R. Feagin, a U.S. sociologist and social theorist, argues that the United States can be characterized as a "total racist society" [36]. He gives a comprehensive overview of the racial discriminatory practices in the United Stated. He argues that every major sector of the US society, e.g., economy, politics, law, and education, is still under the control of white elites who generally make the most important decisions. Discriminatory practices often lead to inter-groups conflicts and anti-discrimination movements. Joe R. Feagin further discusses that the civil rights movements of 1950's and 1960's forced the elite class whites to make modest changes in the racist system and to introduce anti-discrimination laws. However, he argues that in spite of these anti-discriminatory regulation, white males are still mostly holding top positions. It is not only the US where the racial discrimination used to persist; racial discrimination has existed in almost every other society as well. As Joe R. Feagin claims that the discriminatory practices are being diminished with passage of time to some extent due to anti-discrimination movements. This observation is inline with our research problem

where we want to integrate anti-discrimination constraints into the technological solution to make future decision making discrimination-free to match the future trends in society.

The term *sexism* [55] refers to the practice of social discrimination on the basis of sex or gender. Sexism stems from a concept that men and women are being identified with particular occupations. Women are often restricted to certain profession. Women are often victims of rape, domestic violence, prostitution, and income disparity. Moreover, in some cultures females are brought up in such a way that they accept the superiority of males. For example, according to different UNICEF surveys, the percentage of women aged between 15 and 49 who thought that a husband is justified in hitting or beating his wife under certain circumstances, was 90% in Jordan, 85.6% in Guinea, 85.4% in Zambia, 85% in Sierra Leone, 81.2% in Laos, and 81% in Ethiopia [5]. These facts show that gender discrimination is happening in many cultures and that with the advancement of technology it is important to explore it in the modern technological fields to eradicate the discriminatory practices from society. A lot of work has already been done to overcome this unfair treatment towards women but there is still a lot to do.

6.1.1 Definition of Discrimination in the Legal Domain

In this section we discuss the most important related works of the discrimination problem from legal domain. There are many civil right laws to prohibit the practice of discrimination. The United Nations use the definition of racial discrimination laid out in the *International Convention on the Elimination of All Forms of Racial Discrimination* [78], adopted in 1966: *any distinction, exclusion, restriction or preference based on race, color, descent, or national or ethnic origin which has the purpose or effect of nullifying or impairing the recognition, enjoyment or exercise, on an equal footing, of human rights and fundamental freedoms in the political, economic, social, cultural or any other field of public life.(Part 1 of Article 1).*

In the United States there are many anti-discrimination laws [11] to prevent the discriminatory practices from the society, e.g., the equal credit opportunity act [8], equal pay act [9], the civil rights act [7], and the fair housing act [10].

The US equal pay act aims at the eradication of social discrimination in employment. These parts of the law emphasise the equality of employees : *No employer having employees subject to any provisions of this section [section 206 of title 29 of the United States Code] shall discriminate, within any establishment in which such employees are employed, between employees on the basis of sex by paying wages to employees in such establishment at a rate less than the rate at which he*

pays wages to employees of the opposite sex in such establishment for equal work on jobs[...] the performance of which requires equal skill, effort, and responsibility, and which are performed under similar working conditions, except where such payment is made pursuant to (i) a seniority system; (ii) a merit system; (iii) a system which measures earnings by quantity or quality of production; or (iv) a differential based on any other factor other than sex [9].

Similarly in the European Union [4, 34] and the UK [6] there are many laws which prohibit discrimination and ensure the equal treatment to the people. In addition to the anti-discrimination laws, there are many organization which are working to protect the civil rights of citizens. For instance, the European Network Against Racism (ENAR) [3] is a network of European NGOs working to combat racism in all EU member states and represents more than 700 NGOs throughout the European Union. ENAR is fighting against racism, racial discrimination, xenophobia and related intolerance, to promote equality of treatment between European Union citizens and third country nationals, and to link local/regional/national initiatives with European Union initiatives.

On the research side, the author of [26] is a legal expert and argues that characteristics of group profiles may strongly influence the achievements of individuals when their abilities are judged by the group characteristics. He proposes to develop new ethical, legal, and technological standards that adequately recognize the possible harmful consequences of group profiles on individual achievements. The authors of [42, 45, 57] argue that though there are many anti-discrimination laws, the discriminatory practices are still pervasive in the form of stereotyping and unconscious biases on the basis of sensitive attributes, such as race and gender.

6.1.2 Economic Discrimination

In this section we present an overview of the different works in the economic domain to explore and discourage the discriminatory practices in society which will emphasize the need to study the discrimination problem more elaborately.

Discrimination in economics includes the biased treatment of individuals of certain groups in employment, wages, access to market, and access to services and goods. The term *economic discrimination* was first used in the British Railway Clauses Consolidation Act of 1845 to prohibit a common carrier from charging one person more for carrying freight than was charged to another customer for the same service. In 19th century English and American common law, discrimination was meant to indicate improper distinctions in economic transactions. For example, discrimination occurred if a hotelier refused to give rooms to a patron [12].

The authors of [63, 56] give an excellent review and existing evidence of the discrimination in mortgage lending and claim that minorities are more than twice as likely to be denied a mortgage as whites. The authors of [82] conclude in their study that minority home buyers in the United States are discriminated by mortgage lending institutions. They argue that discrimination is not only present at the time of approval or rejection of loan but also focused on the advertising, outreach, the referral stage and on the loan administration stage. They prove their claim of unequal treatment of minorities and whites by statistical analysis of data assembled by the Federal Reserve Bank of Boston. The authors of [71] analyze loan-approval and loan-performance data and devise tests for detecting discrimination in the contemporary mortgage market. They provide an in-depth review of the 1996 Boston Fed Study and its critics, along with new evidence that the minority-white loan-approval disparities in the Boston data represent discrimination. Their analysis also reveals several major weaknesses in the current fair-lending enforcement system, e.g., insulation of some discriminating lenders from investigation. They devise new procedures to overcome these weaknesses and show how the procedures can also be applied to discrimination in loan-pricing and credit-scoring.

Gary S. Becker [16] gives a detailed overview of the forces that determine discrimination in the market place, employer and employee discrimination, consumer discrimination, and changes in the discrimination over time in his book *The Economics of Discrimination* [16]. He develops a useful model for analyzing the economic effects of discrimination. He treats negro and white sectors of the United States as if they were separate countries in an international trade model and he assumes that the white sector owns a higher ratio of capital to labor than the negro sector does. The discrimination affects the dealings of negroes and whites in a similar way as tariff barriers impede trade between two countries. Gary S. Becher's work got a lot of attention from the research community and resulted in many critical reviews [27, 70, 54, 25, 76] on this book which proposed new directions to study the economic discrimination.

The economic discrimination has many negative impacts over the development of society. It often leads to conflicts among different social groups, poverty, and unemployment. Harry Gilman discusses the effects of economic discrimination on unemployment [38]. He examines the difference in the levels of unemployment rates between white and nonwhite experienced male workers in the United States. He presents unemployment figures to support his claims. Messner [61] presents the effects of economic discrimination against social groups on national rates of homicide. He proposes that countries with high discrimination will exhibit comparatively high levels of homicide, and that the effects of discrimination will exceed those of income inequality. The author uses INTERPOL and the World Health

Organization homicide data to support his claim.

The economic condition is very important for every individual. If the someone is discriminated in economically, it affects the overall growth of that individual or the whole community very badly. For instance, if people from a certain ethnic group are restricted to low profile jobs, it will affect the health and education of that community and their offspring. The community will be deprived of the facilities which are accessible to other citizens. Such kind of economic disparity among different groups will lead to social problems, such as unemployment, crimes, and anti-discrimination movements.

6.2 Data Mining

In this section we discuss the related work in discrimination-aware data mining itself, cost-sensitive classification, constraint-based classification, and sampling techniques for unbalanced datasets.

6.2.1 Discrimination-aware Data Mining

In *Discrimination-Aware Data Mining*, two main research directions can be distinguished: the detection of discrimination [64, 66, 73, 65, 74, 58, 72] and the direction followed in this thesis, namely learning discrimination-free classifiers if the data is discriminatory.

The authors of [64, 72, 66] introduce similar concepts of discrimination and undesirable dependencies between the class and some sensitive attributes in data mining as we do. These works, however, concentrate on identifying the discriminatory rules that are present in a dataset rather than on learning a classifier that avoids this discrimination in future predictions. A central notion in their work on identifying discriminatory rules is that of the *context* of the discrimination. That is, specific regions in the data are identified in which the discrimination is particularly high. We further explain the concept of context of discrimination with an example. For instance the probability that a person living in New York or with black ethnicity would get a loan is reasonably high, however, when a black applicant from a certain neighborhood in New York applies for the loan, there is a little chace for the acceptance. It means that the bias towards to the black people for loan denial is considerably high in this context, i.e., when they come from this certain neighborhood. This work also focuses on the case where the discriminatory attribute is not present in the dataset and background knowledge for the identification of discrim-

inatory guidelines has to be used. The authors of [73] give an implementation of the techniques of [64, 72] to provide guidance in the legal issues about discrimination hidden in data, and through several legally-grounded analyses to unveil discriminatory situations. The works on discrimination-aware classification, however, assumes that the discriminatory data is given and the task is to construct a discrimination-free classifier. As such our work can be seen as an orthogonal to the detection of discrimination work.

The authors of [65, 74] present a reference model for finding evidence of discrimination in datasets of historical decision records in socially sensitive tasks, e.g., access to credit, mortgage, insurance, labor market and other benefits. In their reference model, they assume the decision support system as a black box which takes a case consisting of attribute-value pairs, e.g., application data as input and provides its decision label. After extracting the frequent classification rules from the historical decisions made by the automated decision support system and then they use key legal concepts to unveil the discriminatory classification patterns. They formally describe the process of direct and indirect discrimination discovery in a rule-based framework, by modeling protected-by-law groups, such as minorities or disadvantaged segments, and contexts where discrimination occurs. They establish an excellent connection of the discrimination problem with the legal domain and propose a detailed range of discrimination measures to encompass different notions of discrimination from the judicial literature. Their work mainly aims to support the anti-discrimination analysts to discover direct and indirect discriminatory patterns.

The authors of [58] propose a variant of k-NN classification for the discovery of discriminated objects. They consider a data object as discriminated if there exist a significant difference of treatment among its neighbors belonging to a protected-by-law group (deprived) and its neighbors not belonging to it (favored). 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 (given in Section 3.1.1 of Chapter 3) but also differs at some key points. For example, the authors of [58] change the labels of all data object which are suspected to be discriminated while our *Massaging* approach modify the class labels of only selected data objects (the ones close to the decision boundary) to make the training set impartial.

The work of [77] also aims at finding interesting subsets of a classified example set that deviates from the overall distribution. Furthermore, similar in nature to our proposal is the work on anonymity [75]. Although the goal there is different, namely removing data that allows for the identification of individuals, the mech-

anism is the same: before the data is released for mining, it is sanitized and the altered dataset is released.

6.2.2 Constraint Based Classification

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. Similarly in discrimination-aware classification we insert a non-discrimination constraint in the classifier construction phase and want our learnt model to satisfy this non-discrimination constraint.

For example, when learning a decision tree, an upper bound on the number of nodes in the tree can be imposed. Our proposed classification problem with non-discrimination constraints clearly fits into this framework. Most existing work on constraint based classification, however, imposes purely syntactic constraints limiting, e.g., model complexity [51]. The difference with our work is that for the syntactic constraints, the satisfaction does not depend on the data itself, but only on the model and most research concentrates on efficiently listing the subset of models that satisfy the constraints. In our case, however, satisfaction of the constraints depends on the data itself and hence requires a different approach.

One noteworthy exception is *monotone classification* [30, 53]. In monotone classification, next to the normal labeled training data, additionally a function is given for which the predictions should be monotone. An example of such a constraint could be that when assigning a loan based on a number of scores, the assigned label should be monotone in the scores; e.g., if one person gets assigned the loan, and another person scores higher while all other fields are equal to the first person, then the second person should receive the loan as well. Whereas the discrimination criterium is global, the monotonicity criterium is local in the sense that it can be checked by looking at pairs of tuples only. Also, in many cases, the monotonicity can and will be checked syntactically. The authors of [67] survey the methods that have been so far proposed for generating decision trees that satisfy monotonicity constraints. They made a distinction between methods that work only for monotone datasets and methods that work for monotone and non-monotone datasets alike.

6.2.3 Cost-sensitive Learning

Most of the classification algorithms assume that all errors have the same cost. In most of data mining applications, however, the reality is different. In Cost-Sensitive and Utility-Bases learning [21, 33, 59], it is assumed that not all types of

prediction errors are equal and not all examples are as important. For example, if the classification task is to predict if an e-mail is spam, the cost of a false positive; i.e., wrongly filtering out a righteous e-mail as spam, is many times higher than the cost of a false negative; i.e., letting through a spam e-mail. The type of error (false positive versus false negative) determines the cost. Sometimes costs can also depend on individual examples. In cost-sensitive learning the goal is no longer to optimize the accuracy of the prediction, but rather the total cost.

The relation with discrimination-aware classification is best illustrated with an example: consider again the case of the females being discriminated when applying for a job. One approach to better balance the predictions of the classifier is to artificially assign a higher cost to miss-classifying a successful female, than to miss-classifying a successful male. In this way, the learning process will be biased towards classifiers that are "optimistic"towards female applicants and "pessimistic" towards male applicants.

The realization that cost-sensitive learning techniques are required in the real-world KDD applications led to substantial research work. Turney [83] provides an online bibliography on cost-sensitive learning. Domingos [28] proposes a method named MetaCost for making classifiers cost-sensitive by wrapping a cost minimizing procedure around them. MetaCost assumes that costs of misclassifying the examples are known in advance and are the same for all the examples. It is based on relabeling the training examples with their estimated minimal-cost classes, and applying the error-based learner to the new training set. It estimates probabilities with bagging and uses a variant of Breimans [18] bagging as the ensemble method. This variant differs from bagging in that the number of examples in each sample may be smaller than the training set size. MetaCost uses the estimated probabilities and known costs of misclassifying the training examples to re-label the training data. This re-labeled data is used to yield a classifier for future cost-sensitive classification.

MetaCost has some similarity with our *Massaging* technique, given in Chapter 3, with respect to re-labeling the training data but MetaCost aims at the identification of the rarest class examples only and does not address the discrimination problem. MetaCost gives the idea of uniform misclassification cost for each example however *Massaging* introduces the idea that misclassifying cost can be uniform for a certain subgroup, e.g., discriminated community, but can not be uniform for all the training examples. Zadrozny et al. [85] also ratifies the idea of non-uniform misclassifying for each example.

We observe that the above mentioned cost-sensitive procedures emphasize the identification of the minority (usually the more interesting) class and do not ad-

dress the problem of discrimination between class and some other attribute. In the discrimination problem, the deprived community may not be the minority but still may be discriminated. For example, in the German Credit dataset the foreign workers are a deprived community but their percentage is more than 90% of the whole dataset. The discrimination is always w.r.t. a certain class, e.g., loan approval for a particular community, e.g., an ethnic minority. It does not only focus on correctly identifying the instances of a certain class but also tries to identity the instances of the desired class after removing the discrimination to make the future decisions impartial.

6.2.4 Sampling

Sampling is the process of selecting units or subsets from a population so that by analyzing the sample we may produce some knowledge about the whole population. Methods to deal with the class imbalance problem usually analyze samples instead of the whole population to make the task of analysis easier and more rewarding. There are many types of sampling but we use uniform random sampling with replacement and introduce a novel sampling approach namely *Preferential Sampling*, given in Chapter 3. In uniform random sampling, each object in the population has same probability to be selected in the sample. In *Preferential Sampling*, however, some objects have higher chance of being selected. There are many research works on sampling but none of them address the discrimination problem, however, there are some works which have some connection to our *Reweighing* and *Sampling* methods.

In data mining, differences in prior class probabilities or class imbalances have been reported to hinder the performance of learned classifiers, e.g., decision trees [44, 43]. If the given dataset is imbalanced, we will have to either over-sample the minority class instances or under-sample the majority class instances. Both over-sampling and under-sampling have some disadvantages. [28] describes some limitation of these sampling methods and claims: that sampling methods may distort the distribution of training examples which may have a bad impact on the performance of some algorithms; they may reduce the amount of data available for learning, if stratification is applied by under-sampling; and they may increase the learning time, if stratification is done by over-sampling. The *Reweighing* scheme is a novel approach to over-sample the discriminated community and under-represent the favored community. The advantage of under-sampling by *Reweighing* is that it does not lose any piece of information. Rather, it assigns a weight to each training example of favored community to reduce its representation to a required level. Similarly it increased the representation of discriminated community by systemat-

ically calculating the weights of its training examples to over-sample it.

In [23], a synthetic minority over-sampling technique (SMOTE) for two-class problems that over-sampled the minority class by creating synthetic examples rather than replicating examples is proposed. Chawla et al. [24] also utilize a wrapper approach to determine the percentage of minority class examples to be added to the training set and the percentage to under-sample the majority class examples [52]. These sampling methods show some similarity with our *Reweighing* and *Sampling* techniques; by increasing the number of samples in one group (the minority class/the deprived community members with a positive label), we try to increase the importance of this group such that the classifier learned on the resampled dataset is forced to spend more attention to this group. Making an error on this group will hence be reflected in more severe penalties than in the original dataset, leading to a desired bias towards more easily assigning the minority class label or the positive label to the discriminated group, respectively.

6.3 Conclusion

The anti-discrimination works in social sciences, discussed in Section 6.1, provide a solid basis for discrimination-aware classification. All above discussed anti-discrimination laws require that there should not be discriminatory practices on the basis of sensitive characteristics of people. With the rapid technological advancement, it is imperative to push these non-discrimination constraints within the automated models. These discrimination-aware automated procedures will help the practitioners to prove the discriminatory practices in the court of law while at the same time , help companies to stay away from the discrimination accusations. Our discrimination-aware classification can be considered as a counter part of these anti-discrimination works in social sciences.

Almost every financial institution uses automated procedures to select the best customers. In our discrimination-aware classification framework, we argue that the automated procedures should make the best choices without violating anti-discrimination laws. Such violation may impose heavy fines on the financial institution by some court of law. For instance, recently L'Orėal, the French cosmetics giant, has been found guilty of racial discrimination by barring black, arabian and asian women from selling its shampoo. *France's highest court ruled that the group had broken the law by seeking an exclusively white sales team to promote Fructis Style, a hair product made by Garnier, L'Orėal's beauty division. The case hinged on a fax stipulating that Garnier's hostesses should be BBR, which stands for "bleu, blanc, rouge" the colors of the French flag. Sent by Districom, a divi-*

sion of Adecco, the temporary recruitment agency, the fax also said that Garnier's hostesses should be aged 18 to 22 and wear size 38 to 42 clothes (British sizes 8 to 12). [..] The court upheld fines to L'Orėal and Adecco of 30,000 euros already handed down by the Paris Appeal Court [13].

We can conclude that the problem of discrimination is studied well in social sciences but little research work has been done to incorporate this important social problem in technological solutions.

Chapter 7

Conclusions and Future Work

7.1 Conclusion

To conclude, we believe that discrimination-aware classification is a new and exciting area of research addressing a societally relevant problem. It already received a lot of attention in social sciences and in the legal domain [8, 9]. In many countries, several types of discrimination, such as those based on gender, race, sexual preference, and religion are forbidden by law. When humans make subjective decisions, inevitably individual discrimination cases may occur. Such cases can be brought to court for in depth analysis of circumstances. But not only humans can discriminate. Nowadays more and more decisions in lending, recruitment, grant or study applications are partially being automated based on models fitted on historical data. These discriminatory classification models are worse than the individual cases of discrimination because they become structural and systematic, backed up by misleading statistics in the case of redlining. Regulatory authorities and researchers put a lot of effort to monitor, analyze and ensure non discriminatory decision making in mortgage lending [63,70], recruitment, wages. It is also very important to take the discrimination perspective into account while automating the daily life procedures otherwise it could lead to serious consequences, e.g., heavy fines, legal penalties etc. It is in the best interest of the decision makers (e.g. banks, consultancies, universities) to be able to build discrimination-free classifiers even if the historical data is discriminatory.

In this thesis, we have studied the discrimination problem with respect to the data mining perspective. This work can be seen as a logical following step of the work of Dino et al. [64, 66, 73, 65, 74, 58, 72], as shown in Figure 7.1, who concentrated on the detection of discrimination from a given dataset. We have introduced methods to quantify the discrimination in a given dataset or in prediction of a classification model. We theoretically studied the effect of non-discrimination over accuracy of classifiers. We have proposed discrimination-aware classification methods to make future automated decision making as discrimination-free as possible. Our proposed solutions to the discrimination problem fall into three broad categories. First, we propose pre-processing methods (Chapter 3) to remove the discrimination from the training dataset. Second, we propose solutions to the discrimination problem by directly pushing the non-discrimination constraints into classification models and post-processing of built models (Chapter 4). Figure 7.2 gives us a comparison of the results of decision trees learnt after applying our proposed discrimination-aware preprocessing techniques on the training data (label Preprocess_methods), learnt with discrimination-aware splitting criteria (label SplitCrit_DT), learnt with leaf relabeling approach (label Relab_DT) and learnt without any discrimination-aware technique (label Ordinary_DT). We can conclude

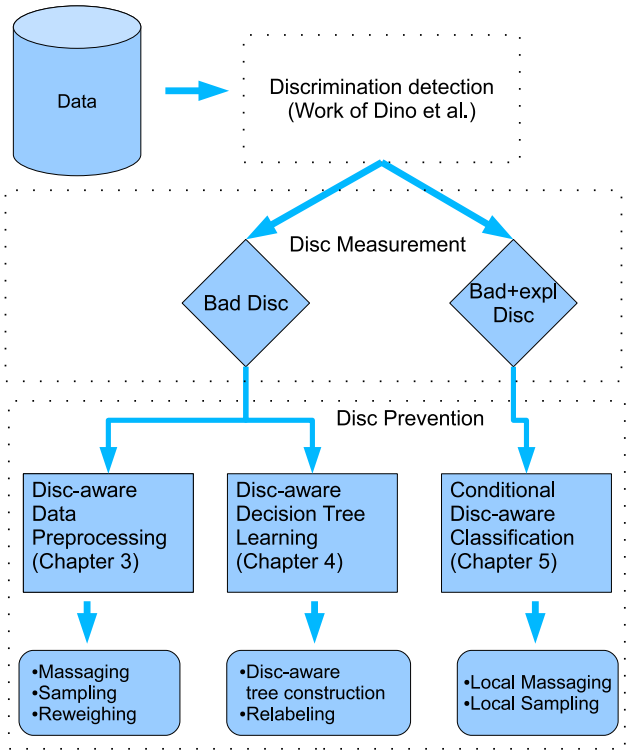


Figure 7.1 Conclusions

from the results shown in Figure 7.2 that our proposed discrimination-aware techniques reduce discrimination significantly by maintaining a high accuracy as compared to the ordinary methods. Moreover, the decision trees with leaf relabeling approach have an advantage over other methods that it reduces the discrimination to 0%.

Third, we further study the discrimination-aware classification paradigm in the presence of explanatory attributes that are correlated with the sensitive attribute, e.g., low income may be explained by the low education level. In such a case, as we show, not all discrimination can be considered bad, therefore we explicitly split the discrimination into explainable and bad discrimination. Our proposed methods in this category only remove the bad discrimination. The experimental evaluation over real world datasets, as shown in Figure 7.3, shows that our pro-

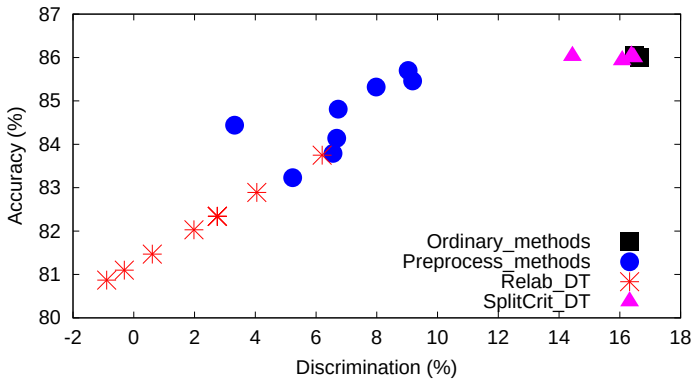


Figure 7.2 Comparison of techniques proposed in Chapter 3 (label Preprocess_methods) and Chapter 4 (label SplitCrit_DT and Relab_DT) over the Adult dataset [14].

posed discrimination-aware classification methods classify the future data objects with significantly low discrimination and high accuracy.

7.2 Future Work

According to the best of our knowledge, this thesis is the first step towards discrimination-free classifiers construction. In this thesis, we have restricted our work to one binary sensitive attribute and a binary class attribute. In future, we plan to work with arbitrary type of sensitive attributes, i.e., numerical, nominal. We also plan to use multiple attributes as sensitive attribute, e.g., simultaneous use of gender, religion and race as sensitive attributes. We plan to extend binary discrimination-aware classification problem to multi-class discrimination-aware classification problem or prediction problem.

In this conditional discrimination-aware study we considered that only one attribute at a time can be explanatory, which is a simplified scenario. In reality there may be more than one explanatory attribute. A direct extension of our work is instead of treating one attribute as explanatory to treat clusters of instances as explanatory. Our new approach of quantifying discrimination is directly applicable to such a case. The main challenge is how to logically partition the data in an unsupervised

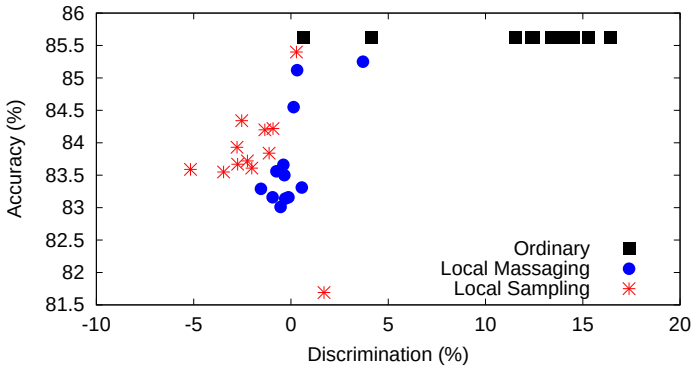


Figure 7.3 Comparison of techniques proposed in Chapter 5 (label Local Massaging and Local Sampling) with ordinary methods (label Ordinary) over the Adult dataset [14].

way.

A promising direction could be to extend the work [58] where discriminated instances are identified by finding discrepancies in labeling with its k nearest neighbors in the other community. For the definition of the distance function we could incorporate the neutrality of certain attributes such as "Number of car crashes in the past" by, e.g., giving them a higher weight. We would like to develop methods with full control over discrimination to guarantee the discrimination-free classifiers which can directly be used in legal domain and economic domain.

We would like to develop a plug-in for discrimination-aware classification methods into other data mining tools like KNIME.

In future, we plan to validate our methods over more real world scenarios. In this regards we plan to continue our collaboration with the Dutch Central Bureau of statistics.

References

[1] Australian sex discimination act 1984. via: http://www.comlaw.gov.au/Details/C2010C00056.

[2] The european court of justice ruling. via: http://ec.europa.eu/ireland/press_office/news_of_the_day/ecj-ruling-sex-discrimination-in-insurance-contracts_en.htm.

[3] European network against racism. via: http://www.enar-eu.org/.

[4] European union legislation. via: http://europa.eu/legislation_summaries/index_en.htm.

[5] Percentage of women aged 15 to 49 who think that a husband/partner is justified in hitting or beating his wife/partner under certain circumstances. http://www.childinfo.org/attitudes_data.php.

[6] United kingdom legislation. via: http://www.legislation.gov.uk/.

[7] The us civil rights act. via: http://finduslaw.com/.

[8] The us equal credit opportunity act. via: http://www.fdic.gov/regulations/laws/rules/6500-1200.html.

[9] The us equal pay act 1963. via: http://www.eeoc.gov/laws/statutes/epa.cfm.

[10] Us fair housing act. via: http://www.justice.gov/crt/about/hce/.

[11] Us federal legislation. via: http://www.justice.gov/crt.

[12] The history of economic discrimination, May 2 2011. http://en.wikipedia.org/wiki/Economic_discrimination.

[13] L'orėal fined for barring black and asian women from adverts :the telegraph report, May 2 2011. http://www.telegraph.co.uk/news/worldnews/europe/france/5635825/LOreal-fined-for-race-discrimination.html.

[14] A. Asuncion and D.J. Newman. UCI machine learning repository. 2007.

[15] 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.

[16] G.S. Becker. *The economics of discrimination*. University of Chicago Press, 1971.

[17] P.J. Bickel, E.A. Hammel, and J.W. O'connell. Sex bias in graduate admissions: Data from Berkeley. *Science (New York, N.Y.)*, 187 (4175), 1975.

[18] L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

[19] T. Calders, F. Kamiran, and M. Pechenizkiy. Building classifiers with independency constraints. In *IEEE International Conference on Data Mining Workshops*, pages 13–18. IEEE, 2009.

[20] T. Calders and S. Verwer. Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21(2):277–292, 2010.

[21] P.K. Chan and S.J. Stolfo. Toward scalable learning with non-uniform class and cost distributions: A case study in credit card fraud detection. In *Proceedings of ACM SIGKDD*, pages 164–168, 1998.

[22] E.L. Chao and K.P. Utgoff. Women in the labor force: A databook. *US Department of Labor and Bureau of Labor Statistics*, 2004.

[23] N.V. Chawla, K.W. Bowyer, L.O. Hall, and W.P. Kegelmeyer. Smote: Synthetic minority over-sampling technique. *J. Artif. Intell. Res. (JAIR)*, 16:321–357, 2002.

[24] N.V. Chawla, L.O. Hall, and A. Joshi. Wrapper-based computation and evaluation of sampling methods for imbalanced datasets. In *Proceedings of the 1st international workshop on Utility-based data mining*, pages 24–33. ACM, 2005.

[25] D. Collard. The Economics of Discrimination. *The Economic Journal*, 82(326):788–790, 1972.

[26] B. Custers. Effects of unreliable group profiling by means of data mining. In *Discovery Science*, pages 291–296. Springer, 2003.

[27] D. Dewey. The Economics of Discrimination. *Southern Economic Journal*, 24(4):494–496, 1958.

[28] P. Domingos. Metacost: A general method for making classifiers cost-sensitive. In *Proceedings of KDD*, pages 155–164, 1999.

[29] R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern classification*, volume 2. Citeseer, 2001.

[30] W. Duivesteijn and A.J. Feelders. Nearest neighbour classification with monotonicity constraints. In *Proceedings of ECML/PKDD*, pages 301–316. Springer, 2008.

[31] Dutch Central Bureau for Statistics. Volkstelling, 1971. `http://easy.dans.knaw.nl/dms`.

[32] Dutch Central Bureau for Statistics. Volkstelling, 2001. `http://easy.dans.knaw.nl/dms`.

[33] C. Elkan. The foundations of cost-sensitive learning. In *Proc. IJCAI'01*, pages 973–978, 2001.

[34] E. Ellis. *EU anti-discrimination law*. Oxford University Press, 2005.

[35] U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. *Advances in knowledge discovery and data mining*, volume 15. MIT press Cambridge, MA, 1996.

[36] J.R. Feagin. *Racist America: Roots, current realities, and future reparations*. Taylor & Francis, 2009.

[37] A. Giddens, M. Duneier, and R.P. Appelbaum. *Introduction to sociology*. WW Norton, 2000.

[38] H.J. Gilman. Economic discrimination and unemployment. *The American Economic Review*, 55(5):1077–1096, 1965.

[39] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I.H. Witten. The WEKA data mining software: An update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18, 2009.

[40] J. Han and M. Kamber. *Data mining: concepts and techniques*. Morgan Kaufmann, 2006.

[41] D.J. Hand, H. Mannila, and P. Smyth. *Principles of data mining*. The MIT press, 2001.

[42] Melissa Hart. Subjective decisionmaking and unconscious discrimination. *Alabama Law Review*, 56:741, 2005. University of Colorado, Law Legal Studies Research Paper 06-26.

[43] N. Japkowicz. The class imbalance problem: Significance and strategies. In *Proceedings of the 2000 International Conference on Artificial Intelligence (ICAI2000)*, volume 1, pages 111–117. Citeseer, 2000.

[44] N. Japkowicz and S. Stephen. The class imbalance problem: A systematic study. *Intelligent Data Analysis*, 6(5):429–449, 2002.

[45] S.L. Johnson. Unconscious racism and the criminal law. *Cornell L. Rev.*, 73:1016, 1987.

[46] F. Kamiran and T. Calders. Classifying without discriminating. In *2nd International Conference on Computer, Control and Communication (IC4)*, pages 1–6. IEEE, 2009.

[47] F. Kamiran and T. Calders. Discrimination-aware classification. In *BNAIC*, 2009.

[48] F. Kamiran and T. Calders. Classification with no discrimination by preferential sampling. In *Proc. BENELEARN*, 2010.

[49] F. Kamiran, T. Calders, and M. Pechenizkiy. Discrimination aware decision tree learning. In *IEEE International Conference on Data Mining*, pages 869–874. IEEE, 2010.

[50] F. Kamiran, T. Calders, and M. Pechenizkiy. Discrimination aware decision tree learning. Technical Report CS 10-13, Eindhoven University of Technolgy, 2010.

[51] S.S. Keerthi, O. Chapelle, and D. DeCoste. Building support vector machines with reduced classifier complexity. *The Journal of Machine Learning Research*, 7:1493–1515, 2006.

[52] R Kohavi and G.H. John. Wrappers for feature subset selection. *Artif. Intell.*, 97(1-2):273–324, 1997.

[53] W. Kotlowski, K. Dembczynski, S. Greco, and R. Slowinski. Statistical model for rough set approach to multicriteria classification. In *Proceedings of ECML/PKDD*. Springer, 2007.

[54] A.O. Krueger. The economics of discrimination. *The Journal of Political Economy*, 71(5):481–486, 1963.

[55] P. Kuhn. Sex discrimination in labor markets: the role of statistical evidence. *The American Economic Review*, 77(4):567–583, 1987.

[56] M. LaCour-Little. Discrimination in mortgage lending: A critical review of the literature. *Journal of Real Estate Literature*, 7(1):15–49, 1999.

[57] A.J. Lee. Unconscious Bias Theory in Employment Discrimination Litigation. *Harv. CR-CLL Rev.*, 40:481, 2005.

[58] B.T. Luong, S. Ruggieri, and F. Turini. k-NN as an Implementation of Situation Testing for Discrimination Discovery and Prevention. 2011.

[59] D.D. Margineantu and T.G. Dietterich. Learning decision trees for loss minimization in multi-class problems. Technical report, Dept. Comp. Science, Oregon State University, 1999.

[60] S. Martello and P. Toth. *Knapsack problems: algorithms and computer implementations*. John Wiley & Sons, Inc., 1990.

[61] S.F. Messner. Economic discrimination and societal homicide rates: Further evidence on the cost of inequality. *American Sociological Review*, 54(4):597–611, 1989.

[62] T.M. Mitchell. Machine learning. wcb. *Mac Graw Hill*, page 368, 1997.

[63] A.H. Munnell, G.M.B. Tootell, L.E. Browne, and J. McEneaney. Mortgage lending in boston: Interpreting hmda data. Working paper 92-7, Federal Reserve Bank of Boston, 1992.

[64] D. Pedreschi, S. Ruggieri, and F. Turini. Discrimination-aware data mining. In *Proceedings of ACM SIGKDD*, 2008.

[65] D. Pedreschi, S. Ruggieri, and F. Turini. Integrating induction and deduction for finding evidence of discrimination. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, pages 157–166. ACM, 2009.

[66] D. Pedreschi, S. Ruggieri, and F. Turini. Measuring discrimination in socially-sensitive decision records. In *Proceedings of SIAM Data Mining Conference*, 2009.

[67] R. Potharst and A.J. Feelders. Classification trees for problems with monotonicity constraints. *ACM SIGKDD Explorations Newsletter*, 4(1):1–10, 2002.

[68] J.R. Quinlan. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986.

[69] J.R. Quinlan. *C4. 5: programs for machine learning*. Morgan Kaufmann, 1993.

[70] M.W. Reder. The Economics of Discrimination. *The American Economic Review*, 48(3):495–500, 1958.

[71] S. Ross and J. Yinger. *The Color of Credit: Mortgage Discrimination, Research Methodology, and Fair-Lending Enforcement*. The MIT Press, 2002.

[72] S. Ruggieri, D. Pedreschi, and F. Turini. Data mining for discrimination discovery. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 4(2):1–40, 2010.

[73] S. Ruggieri, D. Pedreschi, and F. Turini. Dcube: discrimination discovery in databases. In *SIGMOD Conference*, pages 1127–1130, 2010.

[74] S. Ruggieri, D. Pedreschi, and F. Turini. Integrating induction and deduction for finding evidence of discrimination. *Artificial Intelligence and Law*, pages 1–43, 2010.

[75] P. Samarati and L. Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. In *Proceedings of the IEEE Symposium on Research in Security and Privacy*, pages 1–19, 1998.

[76] I.V. Sawhill. The economics of discrimination against women: Some new findings. *The Journal of Human Resources*, 8(3):383–396, 1973.

[77] M. Scholz. Knowledge-based sampling for subgroup discovery. In *Local Pattern Detection. Volume 3539 of Lecture Notes in Computer Science*, pages 171–189. Springer, 2005.

[78] E. Schwelb. The International Convention on the Elimination of All Forms of Racial Discrimination. *International & Comparative Law Quarterly*, 15(04):996–1068, 1966.

[79] E. H. Simpson. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society (Series B)*, 13:238–241, 1951.

[80] R. Stallman et al. GNU General Public License. *Free Software Foundation, Inc., Tech. Rep*, 1991.

[81] M.A. Stoll, S. Raphael, and H.J. Holzer. Black job applicants and the hiring officer's race. *Industrial and Labor Relations Review*, 57(2):267–287, 2004.

[82] M.A. Turner and F. Skidmore. *Mortgage lending discrimination: A review of existing evidence*. Urban Institute Monograph Series on Race and Discrimination. Urban Institute Press, 1999.

[83] P. Turney. Cost-sensitive learning bibliography. In *Institute for Information Technology, National Research Council, Ottawa, Canada,*, 2000.

[84] I.H. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann Pub, 2005.

[85] B. Zadrozny and C. Elkan. Learning and making decisions when costs and probabilities are both unknown. In *Proceedings of KDD*, pages 204–213, 2001.

[86] I. Zliobaite, F. Kamiran, and T. Calders. Handling conditional discrimination. In *IEEE International Conference on Data Mining (accepted for publication)*. IEEE, 2011.

Summary

Classifier construction is one of the most researched topics within the data mining and machine learning communities. Literally thousands of algorithms have been proposed. The quality of the learned models, however, depends critically on the quality of the training data. No matter which classifier inducer is applied, if the training data is incorrect, poor models will result. In this thesis, we study cases in which the input data is discriminatory and we are supposed to learn a classifier that optimizes accuracy, but does not discriminate in its predictions. Such situations occur naturally as artifacts of the data collection process when the training data is collected from different sources with different labeling criteria, when the data is generated by a biased decision process, or when the sensitive attribute, e.g., gender serves as a proxy for unobserved features. In many situations, a classifier that detects and uses the racial or gender discrimination is undesirable for legal reasons. The concept of discrimination is illustrated by the next example: *Throughout the years, an employment bureau recorded various parameters of job candidates. Based on these parameters, the company wants to learn a model for partially automating the matchmaking between a job and a job candidate. A match is labeled as successful if the company hires the applicant. It turns out, however, that the historical data is biased; for higher board functions, Caucasian males are systematically being favored.* A model learned directly on this data will learn this discriminatory behavior and apply it over future predictions. From an ethical and legal point of view it is of course unacceptable that a model discriminating in this way is deployed.

Our proposed solutions to the discrimination problem fall into two broad categories. First, we propose pre-processing methods to remove the discrimination from the training dataset. Second, we propose solutions to the discrimination problem by directly pushing the non-discrimination constraints into classification models and post-processing of built models.

We further studied the discrimination-aware classification paradigm in the presence of explanatory attributes that were correlated with the sensitive attribute, e.g., low income may be explained by the low education level. In such a case, as we show,

not all discrimination can be considered bad. Therefore, we introduce a new way of measuring discrimination, by explicitly splitting it up into explainable and bad discrimination and propose methods to remove the bad discrimination only.
We tried our discrimination-aware methods over real world data sets. We observed in our experiments that our methods show promising results and clearly outperform the traditional classification model w.r.t. accuracy discrimination trade-off. To conclude, we believe that discrimination-aware classification is a new and exciting area of research addressing a societally relevant problem.

Samenvatting

Het automatisch ontwerpen van beslissingsmodellen is een van de meest onderzochte onderwerpen in de *data mining* en *machine learning* onderzoeksgebieden. Letterlijk duizenden algoritmes werden reeds voorgesteld voor dit belangrijke probleem. De kwaliteit van deze modellen hangt kritisch af van de kwaliteit van de invoerdata. Indien de invoerdata niet correct of onvolledig zijn zal een zwak model geleerd worden, ongeacht welk algoritme gebruikt wordt. In dit proefschrift beschrijven we onderzoek waarbij we veronderstellen dat de invoerdata *discriminatie* bevatten, maar waarbij desondanks het uiteindelijke doel een neutraal beslissingsmodel dat niet discrimineert in haar voorspellingen is. Situaties waar dit aan de orde is, komen vaak voor de realiteit; bijvoorbeeld wanneer de data uit verschillende bronnen komen waarbij verschillende criteria gehanteerd werden, wanneer de data gegenereerd werden in een discriminerende omgeving, of wanneer er een *sensitief attribuut* is, zoals, bijvoorbeeld, geslacht of etniciteit, dat een zogenaamde *proxy* is voor andere, niet-geobserveerde attributen. Een leeralgoritme zal onvermijdelijk deze discriminerende wetmatigheden oppikken en rechtstreeks of onrechtstreeks gebruiken bij zijn voorspellingen. Dit is uiteraard vaak onaanvaardbaar vanuit legaal en ethisch perspectief. Beschouw bijvoorbeeld volgende situatieschets: doorheen de jaren heeft een rekruteringsbedrijf van al haar sollicitanten allerlei gegevens zoals woonplaats, huidige functie, geslacht, etniciteit, opleiding en leeftijd geregistreerd. Ook van de verschillende vacatures werden de gegevens bewaard, zoals het bedrijf, de sector, de beroepscategorie, het vereiste diploma en de werklocatie. Bovendien werden succesvolle koppelingen geregistreerd; dit is, welke sollicitanten voor welke vacatures werden uitgenodigd voor een sollicitatiegesprek. Op basis van deze gegevens wil het bedrijf nu deze koppeling van geschikte kandidaten aan vacatures gedeeltelijk automatiseren. Al snel, echter, blijkt dat de historische data die het rekruteringsbedrijf verzamelde een belangrijke bias bevat: voor hogere kaderfuncties werden blanke mannen systematisch bevoordeeld. Een beslissingsmodel geleerd op deze data zal dit verband oppikken en actief exploiteren in haar toekomstige voorspellingen, en aldus vrouwen en gekleurde sollicitanten discrimineren. Uiteraard is het gebruik van zulk een voor-

spellingsmodel illegaal en ethisch onaanvaardbaar.
Dit proefschrift beschrijft daarom de ontwikkeling van computationele methodes om in zulke situaties toch accurate en discriminatievrije beslissingsmodellen te leren. Onze methodes vallen uiteen in twee grote categorieën. Ten eerste beschouwen we preprocessing methodes om de discriminatie rechtstreeks uit de invoerdata te verwijderen alvorens het leeralgoritme uit te voeren. Daarnaast stellen we ook oplossingen voor die het discriminatieprobleem aanpakken door antidiscriminatie beperkingen diep in de algoritmes in te bouwen en de geleerde modellen nadien verder aan te passen in een postprocessing fase. Verder werd het paradigma van discriminatievrij leren van beslissingsmodellen uitgebreid naar situaties waarin de discriminatie gedeeltelijk verklaard kan worden door zogenaamde *verklarende attributen* die gecorreleerd zijn met de sensitieve attributen; bijvoorbeeld, een laag inkomen kan verklaard worden door een laag opleidingsniveau. In zulk een situatie kunnen we niet alle discriminatie als slecht oormerken. Daarom introduceren we nieuwe manieren om discriminatie te meten waarbij we die expliciet opsplitsen in *slechte* en *verklaarbare* discriminatie en stellen we verschillende methodes voor om enkel de slechte discriminatie te verwijderen. Al onze methodes werden toegepast op echte databestanden. Onze experimenten tonen veelbelovende resultaten en presteren duidelijk beter dan de traditionele leermethodes met betrekking tot discriminatie en accuratesse.
Wij zijn er van overtuigd dat het werk gepresenteerd in dit proefschrift de start kan zijn van een nieuw en opwindend onderzoeksgebied op een sociaal en ethisch bijzonder relevant thema.

Curriculum Vitae

Faisal Kamiran was born on January 25th, 1981 in Burewala (Punjab), Pakistan. He passed his matriculation from M.C. Model High School, Burewala in 1995. He got his bachelor degree from Bahauddin Zakria University Multan in 2000. He got his MSCS (Master in Science and Computer Science) from University of the Central Punjab (UCP), Lahore in 2005. After his MSCS, he worked in a government organization for three years.

He got a scholarship from the Higher Education Commission (HEC), Pakistan for overseas Ph.D. studies and joined the Databases and Hypermedia (DH) group as a Ph.D. student on February 22nd, 2008. Soon after joining the DH group, Faisal started his research in a relatively new area of research: "Discrimination-Aware Classification" under the supervision of dr. Toon Calders and Prof. Paul De Bra. With his hard work and excellent advising strategy of Toon Calders, this new area of research got a lot of attention from the data mining and machine learning community. They were able to publish sufficient amount of papers at high quality venues within a very short period of time. During his Ph.D., he maintained good contacts with WODC (Dutch Justice Department) and CBS (Dutch Central Bureau of Statistics) to relate and apply his developed discrimination-aware techniques to real world scenarios. Both of these institutions are interested to use his discrimination-aware methods for unbiased decision making in the Netherlands. He did extensive programming to plug-in discrimination-aware methods into well-known data mining tool Weka for the easy use of discrimination-aware methods.

SIKS Dissertatiereeks

====
1998
====

1998-1 Johan van den Akker (CWI)
DEGAS - An Active, Temporal Database of Autonomous Objects

1998-2 Floris Wiesman (UM)
Information Retrieval by Graphically Browsing Meta-Information

1998-3 Ans Steuten (TUD)
A Contribution to the Linguistic Analysis of Business Conversations within the Language/Action Perspective

1998-4 Dennis Breuker (UM)
Memory versus Search in Games

1998-5 E.W.Oskamp (RUL)
Computerondersteuning bij Straftoemeting

====
1999
====

1999-1 Mark Sloof (VU)
Physiology of Quality Change Modelling;
Automated modelling of Quality Change of Agricultural Products

1999-2 Rob Potharst (EUR)
Classification using decision trees and neural nets

1999-3 Don Beal (UM)
The Nature of Minimax Search

1999-4 Jacques Penders (UM)
The practical Art of Moving Physical Objects

1999-5 Aldo de Moor (KUB)
Empowering Communities: A Method for the Legitimate User-Driven Specification of Network Information Systems

1999-6 Niek J.E. Wijngaards (VU)
Re-design of compositional systems

1999-7 David Spelt (UT)
Verification support for object database design

1999-8 Jacques H.J. Lenting (UM)
Informed Gambling: Conception and Analysis of a Multi-Agent Mechanism for Discrete Reallocation.

====
2000
====

2000-1 Frank Niessink (VU)
Perspectives on Improving Software Maintenance

2000-2 Koen Holtman (TUE)
Prototyping of CMS Storage Management

2000-3 Carolien M.T. Metselaar (UVA)
Sociaal-organisatorische gevolgen van kennistechnologie;
een procesbenadering en actorperspectief.

2000-4 Geert de Haan (VU)
ETAG, A Formal Model of Competence Knowledge for User Interface Design

2000-5 Ruud van der Pol (UM)
Knowledge-based Query Formulation in Information Retrieval.

2000-6 Rogier van Eijk (UU)
Programming Languages for Agent Communication

2000-7 Niels Peek (UU)
Decision-theoretic Planning of Clinical Patient Management

2000-8 Veerle Coup (EUR)
Sensitivity Analyis of Decision-Theoretic Networks

2000-9 Florian Waas (CWI)
Principles of Probabilistic Query Optimization

2000-10 Niels Nes (CWI)
Image Database Management System Design Considerations,
Algorithms and Architecture

2000-11 Jonas Karlsson (CWI)
Scalable Distributed Data Structures for Database Management

====
2001
====

2001-1 Silja Renooij (UU)
Qualitative Approaches to Quantifying Probabilistic Networks

2001-2 Koen Hindriks (UU)
Agent Programming Languages: Programming with Mental Models

2001-3 Maarten van Someren (UvA)
Learning as problem solving

2001-4 Evgueni Smirnov (UM)
Conjunctive and Disjunctive Version Spaces with
Instance-Based Boundary Sets

2001-5 Jacco van Ossenbruggen (VU)
Processing Structured Hypermedia: A Matter of Style

2001-6 Martijn van Welie (VU)
Task-based User Interface Design

2001-7 Bastiaan Schonhage (VU)
Diva: Architectural Perspectives on Information Visualization

2001-8 Pascal van Eck (VU)
A Compositional Semantic Structure for Multi-Agent Systems Dynamics.

2001-9 Pieter Jan 't Hoen (RUL)
Towards Distributed Development of Large Object-Oriented Models,
Views of Packages as Classes

2001-10 Maarten Sierhuis (UvA)
Modeling and Simulating Work Practice
BRAHMS: a multiagent modeling and simulation language
for work practice analysis and design

2001-11 Tom M. van Engers (VUA)
Knowledge Management:
The Role of Mental Models in Business Systems Design

====
2002
====

2002-01 Nico Lassing (VU)
Architecture-Level Modifiability Analysis

2002-02 Roelof van Zwol (UT)
Modelling and searching web-based document collections

2002-03 Henk Ernst Blok (UT)
Database Optimization Aspects for Information Retrieval

2002-04 Juan Roberto Castelo Valdueza (UU)
The Discrete Acyclic Digraph Markov Model in Data Mining

2002-05 Radu Serban (VU)
The Private Cyberspace Modeling Electronic Environments
inhabited by Privacy-concerned Agents

2002-06 Laurens Mommers (UL)
Applied legal epistemology;
Building a knowledge-based ontology of the legal domain

2002-07 Peter Boncz (CWI)
Monet: A Next-Generation DBMS Kernel For Query-Intensive Applications

2002-08 Jaap Gordijn (VU)
Value Based Requirements Engineering: Exploring Innovative
E-Commerce Ideas

2002-09 Willem-Jan van den Heuvel(KUB)
Integrating Modern Business Applications with Objectified Legacy Systems

2002-10 Brian Sheppard (UM)
Towards Perfect Play of Scrabble

2002-11 Wouter C.A. Wijngaards (VU)
Agent Based Modelling of Dynamics: Biological and Organisational Applications

2002-12 Albrecht Schmidt (Uva)
Processing XML in Database Systems

2002-13 Hongjing Wu (TUE)
A Reference Architecture for Adaptive Hypermedia Applications

2002-14 Wieke de Vries (UU)
Agent Interaction: Abstract Approaches to Modelling, Programming and
Verifying Multi-Agent Systems

2002-15 Rik Eshuis (UT)
Semantics and Verification of UML Activity Diagrams for Workflow Modelling

2002-16 Pieter van Langen (VU)
The Anatomy of Design: Foundations, Models and Applications

2002-17 Stefan Manegold (UVA)
Understanding, Modeling, and Improving Main-Memory Database Performance

====
2003
====

2003-01 Heiner Stuckenschmidt (VU)
Ontology-Based Information Sharing in Weakly Structured Environments

2003-02 Jan Broersen (VU)
Modal Action Logics for Reasoning About Reactive Systems

2003-03 Martijn Schuemie (TUD)
Human-Computer Interaction and Presence in Virtual Reality Exposure Therapy

2003-04 Milan Petkovic (UT)
Content-Based Video Retrieval Supported by Database Technology

2003-05 Jos Lehmann (UVA)
Causation in Artificial Intelligence and Law - A modelling approach

2003-06 Boris van Schooten (UT)
Development and specification of virtual environments

2003-07 Machiel Jansen (UvA)
Formal Explorations of Knowledge Intensive Tasks

2003-08 Yongping Ran (UM)
Repair Based Scheduling

2003-09 Rens Kortmann (UM)
The resolution of visually guided behaviour

2003-10 Andreas Lincke (UvT)
Electronic Business Negotiation: Some experimental studies on the interaction between medium, innovation context and culture

2003-11 Simon Keizer (UT)
Reasoning under Uncertainty in Natural Language Dialogue using Bayesian Networks

2003-12 Roeland Ordelman (UT)
Dutch speech recognition in multimedia information retrieval

2003-13 Jeroen Donkers (UM)
Nosce Hostem - Searching with Opponent Models

2003-14 Stijn Hoppenbrouwers (KUN)
Freezing Language: Conceptualisation Processes across ICT-Supported Organisations

2003-15 Mathijs de Weerdt (TUD)
Plan Merging in Multi-Agent Systems

2003-16 Menzo Windhouwer (CWI)
Feature Grammar Systems - Incremental Maintenance of Indexes to Digital Media Warehouses

2003-17 David Jansen (UT)
Extensions of Statecharts with Probability, Time, and Stochastic Timing

2003-18 Levente Kocsis (UM)
Learning Search Decisions

====
2004
====

2004-01 Virginia Dignum (UU)
A Model for Organizational Interaction: Based on Agents, Founded in Logic

2004-02 Lai Xu (UvT)
Monitoring Multi-party Contracts for E-business

2004-03 Perry Groot (VU)
A Theoretical and Empirical Analysis of Approximation in Symbolic Problem Solving

2004-04 Chris van Aart (UVA)
Organizational Principles for Multi-Agent Architectures

2004-05 Viara Popova (EUR)
Knowledge discovery and monotonicity

2004-06 Bart-Jan Hommes (TUD)
The Evaluation of Business Process Modeling Techniques

2004-07 Elise Boltjes (UM)
Voorbeeldig onderwijs; voorbeeldgestuurd onderwijs, een opstap naar abstract denken, vooral voor meisjes

2004-08 Joop Verbeek(UM)
Politie en de Nieuwe Internationale Informatiemarkt, Grensregionale politiële gegevensuitwisseling en digitale expertise

2004-09 Martin Caminada (VU)
For the Sake of the Argument; explorations into argument-based reasoning

2004-10 Suzanne Kabel (UVA)
Knowledge-rich indexing of learning-objects

2004-11 Michel Klein (VU)
Change Management for Distributed Ontologies

2004-12 The Duy Bui (UT)
Creating emotions and facial expressions for embodied agents

2004-13 Wojciech Jamroga (UT)
Using Multiple Models of Reality: On Agents who Know how to Play

2004-14 Paul Harrenstein (UU)
Logic in Conflict. Logical Explorations in Strategic Equilibrium

2004-15 Arno Knobbe (UU)
Multi-Relational Data Mining

2004-16 Federico Divina (VU)
Hybrid Genetic Relational Search for Inductive Learning

2004-17 Mark Winands (UM)
Informed Search in Complex Games

2004-18 Vania Bessa Machado (UvA)
Supporting the Construction of Qualitative Knowledge Models

2004-19 Thijs Westerveld (UT)
Using generative probabilistic models for multimedia retrieval

2004-20 Madelon Evers (Nyenrode)
Learning from Design: facilitating multidisciplinary design teams

====
2005
====

2005-01 Floor Verdenius (UVA)
Methodological Aspects of Designing Induction-Based Applications

2005-02 Erik van der Werf (UM))
AI techniques for the game of Go

2005-03 Franc Grootjen (RUN)
A Pragmatic Approach to the Conceptualisation of Language

2005-04 Nirvana Meratnia (UT)
Towards Database Support for Moving Object data

2005-05 Gabriel Infante-Lopez (UVA)
Two-Level Probabilistic Grammars for Natural Language Parsing

2005-06 Pieter Spronck (UM)
Adaptive Game AI

2005-07 Flavius Frasincar (TUE)
Hypermedia Presentation Generation for Semantic Web Information Systems

2005-08 Richard Vdovjak (TUE)
A Model-driven Approach for Building Distributed Ontology-based Web Applications

2005-09 Jeen Broekstra (VU)
Storage, Querying and Inferencing for Semantic Web Languages

2005-10 Anders Bouwer (UVA)
Explaining Behaviour: Using Qualitative Simulation in Interactive Learning Environments

2005-11 Elth Ogston (VU)
Agent Based Matchmaking and Clustering - A Decentralized Approach to Search

2005-12 Csaba Boer (EUR)
Distributed Simulation in Industry

2005-13 Fred Hamburg (UL)
Een Computermodel voor het Ondersteunen van Euthanasiebeslissingen

2005-14 Borys Omelayenko (VU)
Web-Service configuration on the Semantic Web; Exploring how semantics meets pragmatics

2005-15 Tibor Bosse (VU)
Analysis of the Dynamics of Cognitive Processes

2005-16 Joris Graaumans (UU)
Usability of XML Query Languages

2005-17 Boris Shishkov (TUD)
Software Specification Based on Re-usable Business Components

2005-18 Danielle Sent (UU)
Test-selection strategies for probabilistic networks

2005-19 Michel van Dartel (UM)
Situated Representation

2005-20 Cristina Coteanu (UL)
Cyber Consumer Law, State of the Art and Perspectives

2005-21 Wijnand Derks (UT)
Improving Concurrency and Recovery in Database Systems by
Exploiting Application Semantics

====
2006
====

2006-01 Samuil Angelov (TUE)
Foundations of B2B Electronic Contracting

2006-02 Cristina Chisalita (VU)
Contextual issues in the design and use of information technology in organizations

2006-03 Noor Christoph (UVA)
The role of metacognitive skills in learning to solve problems

2006-04 Marta Sabou (VU)
Building Web Service Ontologies

2006-05 Cees Pierik (UU)
Validation Techniques for Object-Oriented Proof Outlines

2006-06 Ziv Baida (VU)
Software-aided Service Bundling - Intelligent Methods & Tools
for Graphical Service Modeling

2006-07 Marko Smiljanic (UT)
XML schema matching – balancing efficiency and effectiveness by means of clustering

2006-08 Eelco Herder (UT)
Forward, Back and Home Again - Analyzing User Behavior on the Web

2006-09 Mohamed Wahdan (UM)
Automatic Formulation of the Auditor's Opinion

2006-10 Ronny Siebes (VU)
Semantic Routing in Peer-to-Peer Systems

2006-11 Joeri van Ruth (UT)
Flattening Queries over Nested Data Types

2006-12 Bert Bongers (VU)
Interactivation - Towards an e-cology of people, our technological environment, and the arts

2006-13 Henk-Jan Lebbink (UU)
Dialogue and Decision Games for Information Exchanging Agents

2006-14 Johan Hoorn (VU)
Software Requirements: Update, Upgrade, Redesign - towards a Theory of Requirements Change

2006-15 Rainer Malik (UU)
CONAN: Text Mining in the Biomedical Domain

2006-16 Carsten Riggelsen (UU)
Approximation Methods for Efficient Learning of Bayesian Networks

2006-17 Stacey Nagata (UU)
User Assistance for Multitasking with Interruptions on a Mobile Device

2006-18 Valentin Zhizhkun (UVA)
Graph transformation for Natural Language Processing

2006-19 Birna van Riemsdijk (UU)
Cognitive Agent Programming: A Semantic Approach

2006-20 Marina Velikova (UvT)
Monotone models for prediction in data mining

2006-21 Bas van Gils (RUN)
Aptness on the Web

2006-22 Paul de Vrieze (RUN)
Fundaments of Adaptive Personalisation

2006-23 Ion Juvina (UU)
Development of Cognitive Model for Navigating on the Web

2006-24 Laura Hollink (VU)
Semantic Annotation for Retrieval of Visual Resources

2006-25 Madalina Drugan (UU)
Conditional log-likelihood MDL and Evolutionary MCMC

2006-26 Vojkan Mihajlovic (UT)
Score Region Algebra: A Flexible Framework for Structured Information Retrieval

2006-27 Stefano Bocconi (CWI)
Vox Populi: generating video documentaries from semantically annotated media repositories

2006-28 Borkur Sigurbjornsson (UVA)
Focused Information Access using XML Element Retrieval

====
2007
====

2007-01 Kees Leune (UvT)
Access Control and Service-Oriented Architectures

2007-02 Wouter Teepe (RUG)
Reconciling Information Exchange and Confidentiality: A Formal Approach

2007-03 Peter Mika (VU)
Social Networks and the Semantic Web

2007-04 Jurriaan van Diggelen (UU)
Achieving Semantic Interoperability in Multi-agent Systems: a dialogue-based approach

2007-05 Bart Schermer (UL)
Software Agents, Surveillance, and the Right to Privacy: a Legislative Framework for Agent-enabled Surveillance

2007-06 Gilad Mishne (UVA)
Applied Text Analytics for Blogs

2007-07 Natasa Jovanovic' (UT)
To Whom It May Concern - Addressee Identification in Face-to-Face Meetings

2007-08 Mark Hoogendoorn (VU)
Modeling of Change in Multi-Agent Organizations

2007-09 David Mobach (VU)
Agent-Based Mediated Service Negotiation

2007-10 Huib Aldewereld (UU)
Autonomy vs. Conformity: an Institutional Perspective on Norms and Protocols

2007-11 Natalia Stash (TUE)
Incorporating Cognitive/Learning Styles in a General-Purpose Adaptive Hypermedia System

2007-12 Marcel van Gerven (RUN)
Bayesian Networks for Clinical Decision Support: A Rational Approach to Dynamic Decision-Making under Uncertainty

2007-13 Rutger Rienks (UT)
Meetings in Smart Environments; Implications of Progressing Technology

2007-14 Niek Bergboer (UM)
Context-Based Image Analysis

2007-15 Joyca Lacroix (UM)
NIM: a Situated Computational Memory Model

2007-16 Davide Grossi (UU)
Designing Invisible Handcuffs. Formal investigations in Institutions and Organizations for Multi-agent Systems

2007-17 Theodore Charitos (UU)
Reasoning with Dynamic Networks in Practice

2007-18 Bart Orriens (UvT)
On the development an management of adaptive business collaborations

2007-19 David Levy (UM)
Intimate relationships with artificial partners

2007-20 Slinger Jansen (UU)
Customer Configuration Updating in a Software Supply Network

2007-21 Karianne Vermaas (UU)
Fast diffusion and broadening use: A research on residential adoption and usage of broadband internet in the Netherlands between 2001 and 2005

2007-22 Zlatko Zlatev (UT)
Goal-oriented design of value and process models from patterns

2007-23 Peter Barna (TUE)
Specification of Application Logic in Web Information Systems

2007-24 Georgina Ramrez Camps (CWI)
Structural Features in XML Retrieval

2007-25 Joost Schalken (VU)
Empirical Investigations in Software Process Improvement

====
2008
====

2008-01 Katalin Boer-Sorbn (EUR)
Agent-Based Simulation of Financial Markets: A modular,continuous-time approach

2008-02 Alexei Sharpanskykh (VU)
On Computer-Aided Methods for Modeling and Analysis of Organizations

2008-03 Vera Hollink (UVA)
Optimizing hierarchical menus: a usage-based approach

2008-04 Ander de Keijzer (UT)
Management of Uncertain Data - towards unattended integration

2008-05 Bela Mutschler (UT)
Modeling and simulating causal dependencies on process-aware information systems from a cost perspective

2008-06 Arjen Hommersom (RUN)
On the Application of Formal Methods to Clinical Guidelines, an Artificial Intelligence Perspective

2008-07 Peter van Rosmalen (OU)
Supporting the tutor in the design and support of adaptive e-learning

2008-08 Janneke Bolt (UU)
Bayesian Networks: Aspects of Approximate Inference

2008-09 Christof van Nimwegen (UU)
The paradox of the guided user: assistance can be counter-effective

2008-10 Wauter Bosma (UT)
Discourse oriented summarization

2008-11 Vera Kartseva (VU)
Designing Controls for Network Organizations: A Value-Based Approach

2008-12 Jozsef Farkas (RUN)
A Semiotically Oriented Cognitive Model of Knowledge Representation

2008-13 Caterina Carraciolo (UVA)
Topic Driven Access to Scientific Handbooks

2008-14 Arthur van Bunningen (UT)
Context-Aware Querying; Better Answers with Less Effort

2008-15 Martijn van Otterlo (UT)
The Logic of Adaptive Behavior: Knowledge Representation and Algorithms for the Markov Decision Process Framework in First-Order Domains.

2008-16 Henriette van Vugt (VU)
Embodied agents from a user's perspective

2008-17 Martin Op 't Land (TUD)
Applying Architecture and Ontology to the Splitting and Allying of Enterprises

2008-18 Guido de Croon (UM)
Adaptive Active Vision

2008-19 Henning Rode (UT)
From Document to Entity Retrieval: Improving Precision and Performance of Focused Text Search

2008-20 Rex Arendsen (UVA)
Geen bericht, goed bericht. Een onderzoek naar de effecten van de introductie van elektronisch berichtenverkeer met de overheid op de administratieve lasten van bedrijven

2008-21 Krisztian Balog (UVA)
People Search in the Enterprise

2008-22 Henk Koning (UU)
Communication of IT-Architecture

2008-23 Stefan Visscher (UU)
Bayesian network models for the management of ventilator-associated pneumonia

2008-24 Zharko Aleksovski (VU)
Using background knowledge in ontology matching

2008-25 Geert Jonker (UU)
Efficient and Equitable Exchange in Air Traffic Management Plan Repair using Spender-signed Currency

2008-26 Marijn Huijbregts (UT)
Segmentation, Diarization and Speech Transcription: Surprise Data Unraveled

2008-27 Hubert Vogten (OU)
Design and Implementation Strategies for IMS Learning Design

2008-28 Ildiko Flesch (RUN)
On the Use of Independence Relations in Bayesian Networks

2008-29 Dennis Reidsma (UT)
Annotations and Subjective Machines - Of Annotators, Embodied Agents, Users, and Other Humans

2008-30 Wouter van Atteveldt (VU)
Semantic Network Analysis: Techniques for Extracting, Representing and Querying Media Content

2008-31 Loes Braun (UM)
Pro-Active Medical Information Retrieval

2008-32 Trung H. Bui (UT)
Toward Affective Dialogue Management using Partially Observable Markov Decision Processes

2008-33 Frank Terpstra (UVA)
Scientific Workflow Design; theoretical and practical issues

2008-34 Jeroen de Knijf (UU)
Studies in Frequent Tree Mining

2008-35 Ben Torben Nielsen (UvT)
Dendritic morphologies: function shapes structure

====
2009
====

2009-01 Rasa Jurgelenaite (RUN)
Symmetric Causal Independence Models

2009-02 Willem Robert van Hage (VU)
Evaluating Ontology-Alignment Techniques

2009-03 Hans Stol (UvT)
A Framework for Evidence-based Policy Making Using IT

2009-04 Josephine Nabukenya (RUN)
Improving the Quality of Organisational Policy Making using Collaboration Engineering

2009-05 Sietse Overbeek (RUN)
Bridging Supply and Demand for Knowledge Intensive Tasks - Based on Knowledge, Cognition, and Quality

2009-06 Muhammad Subianto (UU)
Understanding Classification

2009-07 Ronald Poppe (UT)
Discriminative Vision-Based Recovery and Recognition of Human Motion

2009-08 Volker Nannen (VU)
Evolutionary Agent-Based Policy Analysis in Dynamic Environments

2009-09 Benjamin Kanagwa (RUN)
Design, Discovery and Construction of Service-oriented Systems

2009-10 Jan Wielemaker (UVA)
Logic programming for knowledge-intensive interactive applications

2009-11 Alexander Boer (UVA)
Legal Theory, Sources of Law & the Semantic Web

2009-12 Peter Massuthe (TUE, Humboldt-Universitaet zu Berlin)
perating Guidelines for Services

2009-13 Steven de Jong (UM)
Fairness in Multi-Agent Systems

2009-14 Maksym Korotkiy (VU)
From ontology-enabled services to service-enabled ontologies (making ontologies work in e-science with ONTO-SOA)

2009-15 Rinke Hoekstra (UVA)
Ontology Representation - Design Patterns and Ontologies that Make Sense

2009-16 Fritz Reul (UvT)
New Architectures in Computer Chess

2009-17 Laurens van der Maaten (UvT)
Feature Extraction from Visual Data

2009-18 Fabian Groffen (CWI)
Armada, An Evolving Database System

2009-19 Valentin Robu (CWI)
Modeling Preferences, Strategic Reasoning and Collaboration in Agent-Mediated Electronic Markets

2009-20 Bob van der Vecht (UU)
Adjustable Autonomy: Controling Influences on Decision Making

2009-21 Stijn Vanderlooy (UM)
Ranking and Reliable Classification

2009-22 Pavel Serdyukov (UT)
Search For Expertise: Going beyond direct evidence

2009-23 Peter Hofgesang (VU)
Modelling Web Usage in a Changing Environment

2009-24 Annerieke Heuvelink (VUA)
Cognitive Models for Training Simulations

2009-25 Alex van Ballegooij (CWI)
“RAM: Array Database Management through Relational Mapping”

2009-26 Fernando Koch (UU)
An Agent-Based Model for the Development of Intelligent Mobile Services

2009-27 Christian Glahn (OU)
Contextual Support of social Engagement and Reflection on the Web

2009-28 Sander Evers (UT)
Sensor Data Management with Probabilistic Models

2009-29 Stanislav Pokraev (UT)
Model-Driven Semantic Integration of Service-Oriented Applications

2009-30 Marcin Zukowski (CWI)
Balancing vectorized query execution with bandwidth-optimized storage

2009-31 Sofiya Katrenko (UVA)
A Closer Look at Learning Relations from Text

2009-32 Rik Farenhorst (VU) and Remco de Boer (VU)
Architectural Knowledge Management: Supporting Architects and Auditors

2009-33 Khiet Truong (UT)
How Does Real Affect Affect Affect Recognition In Speech?

2009-34 Inge van de Weerd (UU)
Advancing in Software Product Management: An Incremental Method Engineering Approach

2009-35 Wouter Koelewijn (UL)
Privacy en Politiegegevens; Over geautomatiseerde normatieve informatie-uitwisseling

2009-36 Marco Kalz (OUN)
Placement Support for Learners in Learning Networks

2009-37 Hendrik Drachsler (OUN)
Navigation Support for Learners in Informal Learning Networks

2009-38 Riina Vuorikari (OU)
Tags and self-organisation: a metadata ecology for learning resources in a multilingual context

2009-39 Christian Stahl (TUE, Humboldt-Universitaet zu Berlin)
Service Substitution – A Behavioral Approach Based on Petri Nets

2009-40 Stephan Raaijmakers (UvT)
Multinomial Language Learning: Investigations into the Geometry of Language

2009-41 Igor Berezhnyy (UvT)
Digital Analysis of Paintings

2009-42 Toine Bogers
Recommender Systems for Social Bookmarking

2009-43 Virginia Nunes Leal Franqueira (UT)
Finding Multi-step Attacks in Computer Networks using Heuristic Search and Mobile Ambients

2009-44 Roberto Santana Tapia (UT)
Assessing Business-IT Alignment in Networked Organizations

2009-45 Jilles Vreeken (UU)
Making Pattern Mining Useful

2009-46 Loredana Afanasiev (UvA)
Querying XML: Benchmarks and Recursion

====
2010
====

2010-01 Matthijs van Leeuwen (UU)
Patterns that Matter

2010-02 Ingo Wassink (UT)
Work flows in Life Science

2010-03 Joost Geurts (CWI)
A Document Engineering Model and Processing Framework for Multimedia documents

2010-04 Olga Kulyk (UT)
Do You Know What I Know? Situational Awareness of Co-located Teams in Multidisplay Environments

2010-05 Claudia Hauff (UT)
Predicting the Effectiveness of Queries and Retrieval Systems

2010-06 Sander Bakkes (UvT)
Rapid Adaptation of Video Game AI

2010-07 Wim Fikkert (UT)
Gesture interaction at a Distance

2010-08 Krzysztof Siewicz (UL)
Towards an Improved Regulatory Framework of Free Software. Protecting user freedoms in a world of software communities and eGovernments

2010-09 Hugo Kielman (UL)
A Politiele gegevensverwerking en Privacy, Naar een effectieve waarborging

2010-10 Rebecca Ong (UL)
Mobile Communication and Protection of Children

2010-11 Adriaan Ter Mors (TUD)
The world according to MARP: Multi-Agent Route Planning

2010-12 Susan van den Braak (UU)
Sensemaking software for crime analysis

2010-13 Gianluigi Folino (RUN)
High Performance Data Mining using Bio-inspired techniques

2010-14 Sander van Splunter (VU)
Automated Web Service Reconfiguration

2010-15 Lianne Bodenstaff (UT)
Managing Dependency Relations in Inter-Organizational Models

2010-16 Sicco Verwer (TUD)
Efficient Identification of Timed Automata, theory and practice

2010-17 Spyros Kotoulas (VU)
Scalable Discovery of Networked Resources: Algorithms, Infrastructure, Applications

2010-18 Charlotte Gerritsen (VU)
Caught in the Act: Investigating Crime by Agent-Based Simulation

2010-19 Henriette Cramer (UvA)
People's Responses to Autonomous and Adaptive Systems

2010-20 Ivo Swartjes (UT)
Whose Story Is It Anyway? How Improv Informs Agency and Authorship of Emergent Narrative

2010-21 Harold van Heerde (UT)
Privacy-aware data management by means of data degradation

2010-22 Michiel Hildebrand (CWI)
End-user Support for Access to Heterogeneous Linked Data

2010-23 Bas Steunebrink (UU)
The Logical Structure of Emotions

2010-24 Dmytro Tykhonov
Designing Generic and Efficient Negotiation Strategies

2010-25 Zulfiqar Ali Memon (VU)
Modelling Human-Awareness for Ambient Agents: A Human Mindreading Perspective

2010-26 Ying Zhang (CWI)
XRPC: Efficient Distributed Query Processing on Heterogeneous XQuery Engines

2010-27 Marten Voulon (UL)
Automatisch contracteren

2010-28 Arne Koopman (UU)
Characteristic Relational Patterns

2010-29 Stratos Idreos(CWI)
Database Cracking: Towards Auto-tuning Database Kernels

2010-30 Marieke van Erp (UvT)
Accessing Natural History - Discoveries in data cleaning, structuring, and retrieval

2010-31 Victor de Boer (UVA)
Ontology Enrichment from Heterogeneous Sources on the Web

2010-32 Marcel Hiel (UvT)
An Adaptive Service Oriented Architecture: Automatically solving Interoperability Problems

2010-33 Robin Aly (UT)
Modeling Representation Uncertainty in Concept-Based Multimedia Retrieval

2010-34 Teduh Dirgahayu (UT)
Interaction Design in Service Compositions

2010-35 Dolf Trieschnigg (UT)
Proof of Concept: Concept-based Biomedical Information Retrieval

2010-36 Jose Janssen (OU)
Paving the Way for Lifelong Learning; Facilitating competence development through a learning path specification

2010-37 Niels Lohmann (TUE)
Correctness of services and their composition

2010-38 Dirk Fahland (TUE)
From Scenarios to components

2010-39 Ghazanfar Farooq Siddiqui (VU)
Integrative modeling of emotions in virtual agents

2010-40 Mark van Assem (VU)
Converting and Integrating Vocabularies for the Semantic Web

2010-41 Guillaume Chaslot (UM)
Monte-Carlo Tree Search

2010-42 Sybren de Kinderen (VU)
Needs-driven service bundling in a multi-supplier setting - the computational e3-service approach

2010-43 Peter van Kranenburg (UU)
A Computational Approach to Content-Based Retrieval of Folk Song Melodies

2010-44 Pieter Bellekens (TUE)
An Approach towards Context-sensitive and User-adapted Access to Heterogeneous Data Sources, Illustrated in the Television Domain

2010-45 Vasilios Andrikopoulos (UvT)
A theory and model for the evolution of software services

2010-46 Vincent Pijpers (VU)
e3alignment: Exploring Inter-Organizational Business-ICT Alignment

2010-47 Chen Li (UT)
Mining Process Model Variants: Challenges, Techniques, Examples

2010-48 Milan Lovric (EUR)
Behavioral Finance and Agent-Based Artificial Markets

2010-49 Jahn-Takeshi Saito (UM)
Solving difficult game positions

2010-50 Bouke Huurnink (UVA)
Search in Audiovisual Broadcast Archives

2010-51 Alia Khairia Amin (CWI)
Understanding and supporting information seeking tasks in multiple sources

2010-52 Peter-Paul van Maanen (VU)
Adaptive Support for Human-Computer Teams: Exploring the Use of Cognitive Models of Trust and Attention

2010-53 Edgar Meij (UVA)
Combining Concepts and Language Models for Information Access

====
2011
====

2011-01 Botond Cseke (RUN)
Variational Algorithms for Bayesian Inference in Latent Gaussian Models

2011-02 Nick Tinnemeier(UU)
Work flows in Life Science

2011-03 Jan Martijn van der Werf (TUE)
Compositional Design and Verification of Component-Based Information Systems

2011-04 Hado van Hasselt (UU)
Insights in Reinforcement Learning; Formal analysis and empirical evaluation
of temporal-difference learning algorithms

2011-05 Base van der Raadt (VU)
Enterprise Architecture Coming of Age - Increasing the Performance of an Emerging Discipline.

2011-06 Yiwen Wang (TUE)
Semantically-Enhanced Recommendations in Cultural Heritage

2011-07 Yujia Cao (UT)
Multimodal Information Presentation for High Load Human Computer Interaction

2011-08 Nieske Vergunst (UU)
BDI-based Generation of Robust Task-Oriented Dialogues 2011-09 Tim de Jong (OU)
Contextualised Mobile Media for Learning

2011-10 Bart Bogaert (UvT)
Cloud Content Contention

2011-11 Dhaval Vyas (UT)
Designing for Awareness: An Experience-focused HCI Perspective

2011-12 Carmen Bratosin (TUE)
Grid Architecture for Distributed Process Mining

2011-13 Xiaoyu Mao (UvT)
Airport under Control. Multiagent Scheduling for Airport Ground Handling

2011-14 Milan Lovric (EUR)
Behavioral Finance and Agent-Based Artificial Markets

2011-15 Marijn Koolen (UvA)
The Meaning of Structure: the Value of Link Evidence for Information Retrieval

2011-16 Maarten Schadd (UM)
Selective Search in Games of Different Complexity

2011-17 Jiyin He (UVA)
Exploring Topic Structure: Coherence, Diversity and Relatedness

2011-18 Mark Ponsen (UM)
Strategic Decision-Making in complex games

2011-19 Ellen Rusman (OU)
The Mind ' s Eye on Personal Profiles

2011-20 Qing Gu (VU)
Guiding service-oriented software engineering - A view-based approach

2011-21 Linda Terlouw (TUD)
Modularization and Specification of Service-Oriented Systems

2011-22 Junte Zhang (UVA)
System Evaluation of Archival Description and Access

2011-23 Wouter Weerkamp (UVA)
Finding People and their Utterances in Social Media

2011-24 Herwin van Welbergen (UT)
Behavior Generation for Interpersonal Coordination with Virtual Humans On Specifying, Scheduling and Realizing Multimodal Virtual Human Behavior

2011-25 Syed Waqar ul Qounain Jaffry (VU)
Analysis and Validation of Models for Trust Dynamics

2011-26 Matthijs Aart Pontier (VU)
Virtual Agents for Human Communication - Emotion Regulation and Involvement-Distance Trade-Offs in Embodied Conversational Agents and Robots

2011-27 Aniel Bhulai (VU)
Dynamic website optimization through autonomous management of design patterns

2011-28 Rianne Kaptein(UVA)
Effective Focused Retrieval by Exploiting Query Context and Document Structure