Classifying without Discriminating

Faisal Kamiran

Faculty of Mathematics and Computer Science
Eindhoven University of Technology
The Netherlands
Email: f.kamiran@tue.nl

Toon Calders

Faculty of Mathematics and Computer Science
Eindhoven University of Technology
The Netherlands
Email: t.calders@tue.nl

Abstract—Classification models usually make predictions on the basis of training data. If the training data is biased towards certain groups or classes of objects, e.g., there is racial discrimination towards black people, the learned model will also show discriminatory behavior towards that particular community. This partial attitude of the learned model may lead to biased outcomes when labeling future unlabeled data objects. Often, however, impartial classification results are desired or even required by law for future data objects in spite of having biased training data. In this paper, we tackle this problem by introducing a new classification scheme for learning unbiased models on biased training data. Our method is based on massaging the dataset by making the least intrusive modifications which lead to an unbiased dataset. On this modified dataset we then learn a non-discriminating classifier. The proposed method has been implemented and experimental results on a credit approval dataset show promising results: in all experiments our method is able to reduce the prejudicial behavior for future classification significantly without loosing too much predictive accuracy.

I. INTRODUCTION

Classification models are trained on the historical data for the prediction of the class labels of unknown data samples. Often, however, the historical data is biased towards certain groups or classes of objects. For example, throughout the years, in a certain organization black people might systematically have been denied from jobs. As such, the historical employment information of this company concerning job applications will be biased towards giving jobs to white people while denying jobs from black people.

In order to reduce this type of racial discrimination, new laws requiring equal job opportunity have been enacted by the government. As such, the organization receives instructions in the form of, e.g., minimum quota for black employees. Suppose now that the company wants to partially automate its recruitment strategy by learning a classifier that predicts the most likely candidates for a job. As the historical recruitment data of the company is biased, the learned model may show unlawfully prejudiced behavior. This partial attitude of the learned model leads to discriminatory outcomes for future unlabeled data objects. This problem is exactly the one we handle in this paper: *how can we train an unbiased classifier when the training data is biased?*

This *classification without discrimination problem* can be observed in many real-world situations:

- Even though there is clear historical evidence showing higher accident rates for male drivers, insurance compa-

nies are not allowed to discriminate based on gender in many countries. In this case the historical data is biased towards assigning a higher risk class to male drivers.
- Often, salaries of women are lower than those of men. Nevertheless, when training a classifier in order to decide in which salary scale to employ a new-hire, it is undesirable to have this inequality in the learned model.

In above mentioned cases, the training data is biased. Classification models trained on such data will not fulfill the future requirements. Future data objects must follow a different class label distribution than that of the training data.

So, sometimes impartial classification results are required for future data objects in spite of having discriminatory training data. Most of the classification models, however, deal with all the attributes equally when classifying data objects and are oblivious towards the sensitivity of attributes. Simply removing the sensitive attributes from the training data in the learning of a classifier for the classification of future data objects, however, is not enough to solve this problem, because often other attributes will still allow for the identification of the discriminated community. For example, the ethnicity of a person might be strongly linked with the postal code of his residential area, leading to a classifier with indirect racial discriminatory behavior based on postal code. This effect and its exploitation is often referred to as *redlining*, stemming from the practice of denying or increasing services such as, e.g., mortgages or health care to residents in certain often racially determined areas. The term redlining was coined in the late 1960s by community activists in Chicago[1]. The authors of [1] also support this claim: even after removing the sensitive attribute from the dataset discrimination persists.

In this paper, we introduce a classification model which is learnt on biased training data but works impartially for future data. First, the discriminatory data is changed in a minimal way as to remove the existing discrimination. To this end we use a ranking function learned on the biased data. Then, based on the sanitized data, a non-discriminatory model can be learned. The fact that this model is learned on non-discriminatory data reduces the prejudicial behavior for future classification. We refer to this model as *Classification with No Discrimination (CND)*. Obviously, changing the training data

---

[1]Source: http://en.wikipedia.org/wiki/Redlining, September 30th, 2008

might result in lower accuracy scores. Nevertheless, as we try to keep the changes as minimal and least intrusive as possible, the trade-off between accuracy and non-discrimination will be minimal.

The CND method was implemented and tested on a credit score dataset displaying discriminatory behavior. Using our proposed CND method we were able to learn classifiers that no longer discriminate future data, without loosing too much accuracy.

In summary, the contributions of this paper are as follows:
1) a formal definition of the non-discriminatory classification problem. This definition involves a measure for assessing the discrimination in a dataset,
2) a proposed solution, CND, for this problem, and
3) a performance study on a credit score dataset showing promising results.

The paper is organized as follows: in Section II we define the problem formally and introduce the discrimination measure, which is illustrated with an example in Section III. In Section IV we propose a solution for the problem based on altering the training data and the results of different experiments are shown in Section V. Section VI discusses related work and Section VII concludes the work and gives some directions for future work.

II. PROBLEM FORMULATION

We assume a set of attributes

$$A = \{a_1, ..., a_m\} \ ,$$

and a binary set of class labels

$$C = \{c_1, c_2\} \ .$$

$dom(a_i)$ refers to the domain of the ith attribute. A *labeled dataset over A with labels from C* is defined as a finite set of tuples $(x_1, \ldots, x_n, c)$ with

$$x = (x_1, ..., x_m) \in dom(a_1) \times ... \times dom(a_m) \ ,$$

and the *class label* $c \in C$. We will often use $x.a_i$ to refer to the component x_i of x corresponding to the attribute a_i, and to its class label as $x.c$.

Let
$$D = \{(x^1, c^1), \ldots, (x^n, c^n)\}$$

be a labeled dataset where

$$(x^i, c^i) = (x_1^i, \ldots, x_m^i, c^i).$$

We assume that a special attribute $SA \in A$, called the *Sensitive Attribute*, and a special value $s \in dom(SA)$, called *Sensitive Attribute Value* have been given. The semantics of SA and s is that they define the discriminated community; e.g., $SA = Ethnicity$ and $s = Black$. For reasons of simplicity we will assume that the domain of SA is binary; i.e.,

$$dom(SA) = \{s, \overline{s}\}.$$

Obviously, we can easily transform a dataset with multiple attribute values for SA into a binary one by replacing all values

$v \in dom(SA)\backslash\{s\}$ with a new dedicated value $\overline{s}$. Furthermore, we assume that a *desired class* $+ \in C$ has been given. In the credit evaluation example, e.g., $+$ would be the *Good* credit class.

Let now

$$
\begin{aligned}
s &:= |\{x \in D \mid x.SA = s\}| \\
s \wedge + &:= |\{x \in D \mid x.SA = s \wedge x.c = +\}| \\
\overline{s} &:= |\{x \in D \mid x.SA \neq s\}| \\
\overline{s} \wedge + &:= |\{x \in D \mid x.SA \neq s \wedge x.c = +\}|
\end{aligned}
$$

The *discrimination in D of s towards $+$*, denoted $Disc(D, SA, s, +)$, is now defined as:

$$Disc(D, SA, s, +) := conf(\overline{s} \rightarrow +) - conf(s \rightarrow +) \ ,$$

where

$$
\begin{aligned}
conf(\overline{s} \rightarrow +) &:= \frac{\overline{s} \wedge +}{\overline{s}}, \text{ and} \\
conf(s \rightarrow +) &:= \frac{s \wedge +}{s}
\end{aligned}
$$

The goal of this paper is now to develop a classification model such that when it is trained on a biased dataset D, it does show impartial behavior on future data. For a given dataset D_f, $CND(D_f)$ denotes the labeled dataset resulting from applying CND on D_f. More formally, the *Classification with No Discrimination (CND)* problem is defined as follows:

Given
- a biased dataset D,
- a sensitive attribute SA and value s, and
- a desired class $+$,

the output is a classifier CND such that, even though the discrimination in the training data $Disc(D, SA, s, +)$ might be high,
- on unseen future data objects D_f, $Disc(CND(D_f), SA, s, +)$ must be low at the same time,
- the predictive accuracy of CND on this data D_f must be high.

That is, we will measure the quality of classifiers based not only on their predictive accuracy, but also on their discrimination on future data. Predictive accuracy as well as discrimination on unseen data can be estimated using the traditional framework of training and test data. From this definition it can be seen that when solving the CND problem we will have to trade-off accuracy to some extent in order to reduce the discrimination level. This aspect will be discussed in more detail in the experimental section.

III. EXAMPLE

Consider the fictive database given in Table I, storing credit scores decisions. Each data object has 4 attributes which are *Personal Status*, *Age*, *Housing*, *Credit History* and one class attribute *Credit Class* with class values *Good* and *Bad*. The domain of the attribute *Age* consists of the values *Young* and *Aged*. Using our discrimination identifier function for this

TABLE I
RUN-THROUGH EXAMPLE: TEST DATA FOR CLASSIFICATION.

Per_status	Age	House	Cred_hist	Class
Fem:dv/m/s	Young	own	Ex-cred paid	*Good*
Fem:dv/m/s	Young	rent	Ex-cred paid	*Bad*
Fem:dv/m/s	Young	own	Critical	*Bad*
Male:m/w	Young	own	Ex-cred paid	*Good*
Male:m/w	Young	rent	Ex-cred paid	*Bad*
Male:single	Aged	own	Critical	*Good*
Male:single	Aged	rent	Ex-cred paid	*Good*
Male:single	Aged	own	no credit	*Good*
Male:m/w	Aged	own	Critical	*Good*
Male:m/w	Aged	rent	no credit	*Bad*

database, with *Age* as *SA*, *Young* as s and *Good* as desired class, we get:

$$Disc(D, Age, Young, Good)$$
$$= Conf(Aged \rightarrow \textbf{\textit{Good}})$$
$$- Conf(Young \rightarrow \textbf{\textit{Good}}) = 40\%,$$

which shows that Aged people have 40% more chance to be assigned the credit class *Good* than *Young* people.

We know that the data of Table I is biased in favor of *Aged* people. Suppose now that, nevertheless, we want to build a classification model such that the accuracy of assigning an account holder with the correct credit class is high, but at the same time we want to remove the *Age*-discrimination, e.g., because we are interested in attracting young people to our bank. In this situation we want a classification model which is learnt on the data of Table I but classifies future loan applicants without discriminating on *Age*. This example is an instance of the CND problem. In the next section we will develop an algorithmic framework to tackle this problem and illustrate it with the example of this section.

IV. PROBLEM SOLUTION

CND assumes that historical data containing discrimination is available. Our approach will consist of first "massaging" the data to remove the discrimination with the least possible changes. To this end, the class labels of the most likely *victims*(discriminated ones) and *profiters* (favored ones) will be changed. We will use a ranker for the identification of these objects. The modified data is then used for learning a classifier with no discrimination for future decisions.

A. Massaging the Data

For massaging the data, *CND* learns a (biased) ranker for predicting the class attribute without taking into account the sensitive attribute. This ranker will then be used to rank the data objects according to their probability of being in the target class. Any ranking algorithm may be used, but for the experiments in this paper, we used a Naive Bayesian classifier for calculating the class probability of each data tuple.

Subsequently, we identify two groups of objects in the training data: on the one hand, the objects having $SA = s$ and $c \neq +$, and on the other hand the ones having $SA \neq s$ and $c = +$. The first group, denoted CP, are the candidates for *promotion*, and the second group, CD, the candidates for *demotion*. We can now reduce the discrimination in the dataset by either promoting objects in CP from class $-$ to $+$ or by demoting objects in CD from class $+$ to $-$ ($-$ represents the class "not $+$"). In order to maintain the balance between the two classes, *CND* will always do both a promotion and a demotion at the same time. In order to select the best candidate for promotion and demotion, *CND* will use the ranker. That is, the promotion list will be ranked in decreasing order of probability for class $+$, whereas the demotion list is ordered in increasing order of probability for class $+$. In this way, we will always choose the most likely "good" promotion candidate and "bad" demotion candidate first.

The modification of the training data is continued until the discrimination in it becomes zero. The number M of modifications required to make the data discrimination-free can be calculated by using the following formula:

$$M = \frac{(s \times \overline{s} \wedge +) - (\overline{s} \times s \wedge +)}{s + \overline{s}}$$

It means, we will change the class labels of M *victims* and M *profiters* in the training data. The modified impartial training data will then be used to train a classifier. In our experiments, we again use a Naive Bayesian classifier as a future classification model.

B. Running Example

We will now continue the running example of last section. Recall that the task at hand is to learn, based on the training data in Table I, a classifier that does no longer discriminates the people with *Age* equal to *Young*.

We start with learning a *Naive Bayesian* classification model as a ranker based on the discriminatory data of Table I. We follow the steps mentioned above. In Table II, the probabilities assigned by this ranker to the different tuples of being in class *Good* have been given.

TABLE II
DATA OBJECTS WITH THE PROBABILITY OF BEING IN CLASS *Good*.

Per_status	Age	House	Cred_hist	Class	Prob
Fem:dv/m/s	Young	own	Ex-cred paid	*Good*	62%
Fem:dv/m/s	Young	rent	Ex-cred paid	*Bad*	6%
Fem:dv/m/s	Young	own	Critical	*Bad*	69%
Male:m/w	Young	own	Ex-cred paid	*Good*	76%
Male:m/w	Young	rent	Ex-cred paid	*Bad*	3%
Male:single	Aged	own	Critical	*Good*	76%
Male:single	Aged	rent	Ex-cred paid	*Good*	62%
Male:single	Aged	own	no credit	*Good*	63%
Male:m/w	Aged	own	Critical	*Good*	81%
Male:m/w	Aged	rent	no credit	*Bad*	10%

In the second step, we arrange the data separately for *Young* people with class *Bad* in descending order and for *Aged* people with class *Good* in ascending order with respect to the probability of being in the *Good* credit class. The ordered promotion and demotion lists have been given in Table III and Table IV respectively.

Now, we consider the following three options for the removal of discrimination between *Aged* and *Young* with respect to C:

- We may assign *Young* people who are currently in *Bad* credit class and are more likely to qualify for *Good*

credit class to *Good* credit class. We will make this assessment on the basis of the corresponding positive class probability values.

- We may deny *Good* credit class to those *Aged* people who have the lowest positive class probability.
- We may use both above options simultaneously.

TABLE III
PROMOTION LIST: DATA OBJECTS WITH $SA = s$ (*Age=Young*) AND CLASSIFIED IN THE $-$ (*Bad*) CREDIT CLASS ARRANGED IN DESCENDING ORDER.

Psn_status	Age	House	Cred_hist	Class	Pos Prob
Fem:dv/m/s	Young	own	Critical	*Bad*	69%
Fem:dv/m/s	Young	rent	Ex-cred paid	*Bad*	6%
Male:m/w	Young	rent	Ex-cred paid	*Bad*	3%

TABLE IV
DEMOTION LIST: DATA OBJECTS WITH $SA \neq s$ (*Age=Aged*) AND CLASSIFIED IN THE $+$ (*Good*) CREDIT CLASS ARRANGED IN ASCENDING ORDER.

Psn_status	Age	House	Cred_hist	Class	Pos Prob
Male:single	Aged	rent	Ex-cred paid	*Good*	62%
Male:single	Aged	own	no credit	*Good*	63%
Male:single	Aged	own	Critical	*Good*	76%
Male:m/w	Aged	own	Critical	*Good*	81%

As discussed before, *CND* uses the third option for the modification of data in order to maintain the ratio between the *Good* and the *Bad* class. The number M of required modifications in this case is:

$$M = \frac{(\textit{Young} \times \textit{Aged} \wedge \textit{Good}) - (\textit{Aged} \times \textit{Young} \wedge \textit{Good})}{\textit{Young} + \textit{Aged}}$$
$$= \frac{(5 \times 4) - (5 \times 2)}{5 + 5} = 1$$

So, we have to change 1 class label in each of both the promotion and the demotion list. We change the labels of these selected objects as shown in Table V. This dataset shows the discrimination-free data which will be used for future classifier learning. The resulting classifier will classify the future data objects with minimum discrimination.

Algorithm 1 *Classification with No Discrimination (CND)*

Input $(D, s, SA, +)$
Output Classifier CND learnt on D without discrimination

1: $(pr, dem) := Rank(D, SA, s, +)$
2: $existDisc := Disc(D, SA, s, +)$
3: Calculate M, the number of necessary modifications based on $existDisc$
4: **for** M times **do**
5:    Select the data object from the top of pr
6:    Change the class label of the selected object in D
7:    Select the data object from the top of dem
8:    Change the class label of the selected object in D
9:    Remove the top element both of pr and dem
10: **end for**
11: Train a classifier CND on the modified D
12: **return** CND

Algorithm 2 *Rank*

Input $(D, s, SA, +)$
Output (pr, dem): Two ordered lists of data objects on the basis of target class probability.

1: Learn a ranker R based on D
2: Calculate the class probabilities $R(x)$ for all $x \in D$
3: Add all x in D with $x.SA = s$ and $x.c \neq +$ into the list pr in descending order w.r.t. $R(x)$
4: Add all x in D with $x.SA \neq s$ and $x.c = +$ into the list dem in ascending order w.r.t. $R(x)$
5: **return** (pr, dem)

C. Algorithm

The pseudocode of our algorithm has been given in Algorithm 1 and 2. Algorithm 1 describes the approach of *CND* for data massaging and classifier learning, and Algorithm 2 describes the process of ordering the data objects separately for the promotion and demotion lists.

TABLE V
MODIFIED DATA.

Per_status	Age	House	Cred_hist	Class
Fem:dv/m/s	Young	own	Ex-cred paid	*Good*
Fem:dv/m/s	Young	rent	Ex-cred paid	*Bad*
Fem:dv/m/s	Young	own	Critical	**Good**
Male:m/w	Young	own	Ex-cred paid	*Good*
Male:m/w	Young	rent	Ex-cred paid	*Good*
Male:single	Aged	own	Critical	*Good*
Male:single	Aged	rent	Ex-cred paid	**Bad**
Male:single	Aged	own	no credit	*Good*
Male:m/w	Aged	own	Critical	**Bad**
Male:m/w	Aged	rent	no credit	*Good*

V. EXPERIMENTS

In our experiments, we will compare the following two approaches:

1) Our proposed approach; i.e., we will use *CND* for massaging the training data and to make it discrimination free. The ranking function will be based on a Naive Bayesian model learned on the raw data. Then we learn a Naive Bayesian classifier *CND* on the discrimination-free data.
2) For reasons of comparison, we also learn a Naive Bayesian classifier directly on the original data without massaging. We refer to this second approach as "*Classification without the Massaging*".

In all experiments, following commonly accepted evaluation strategies, the dataset will be split into a training and a test set. The test set will be used solely for the purpose of evaluating the performance, so no massaging will be applied on this dataset. Notice that this implies that the accuracy of *CND* will be evaluated on biased data.

As a sanity check, next to biased datasets we also use one unbiased dataset in our experiments, as, clearly, *CND* should reduce to *Classification without the Massaging* in this setting.

The experiments conducted support the following claims:

- The redlining effect discussed in this paper is indeed real: experiments with *Classification without the Massaging*

where the sensitive attribute is removed turn out not to be satisfactory w.r.t. the goal of learning a non-discriminatory classifier. The *Classification without Massaging* works fine with nondiscriminatory data but gives very biased results with discriminatory data. In some examples, the *Classification without Massaging* even further suppresses the discrimination-affected community.

- *CND* classifies non-discriminatory data and discriminatory data with high accuracy and low discrimination. Sometimes, it may even favor the deprived community. The drop in accuracy is low in all our experiments.

Although we need to be cautious with generalizing our preliminary experimental results, the results at least show that our proposed technique has potential and deserves further exploration.

A. The German Credit dataset

We use the German Credit Dataset available in the UCI ML-repository [2] for our experiments. The dataset has 1000 instances which classify the bank account holders into credit class *Good* or *Bad*. Each data object is described by 20 attributes which include 13 categorical and 7 numerical attributes. In our experiments we consider the following setting for the sensitive attribute and value:

- *Age* as sensitive attribute. Since this attribute is not binary, we have to discretize it into *Young* and *Aged* account holders at some certain cut-off value. We considered different cut-off values. After discretization, the domain of this *SA* becomes { *Young*, *Aged* } and *Young* is selected as the sensitive attribute value s.
- In some experiments, we also used *Foreign Worker* as the sensitive attribute. In contrast to *Age*, *Foreign Worker* is already a binary attribute, so no discretization was required here. *Foreign Worker* equals *True* was used as the sensitive attribute value.

The other attributes in the datasets include: existing checking account status, duration of loan, credit history, purpose of loan, savings status, property, type of housing: own, rented or free; credit amount, installment plans, existing credits, employment status, employment since, number of dependents, telephone, personal status: depends upon gender and marital status, resident since and foreign worker.

B. Proposed Solution

We test the accuracy and the discriminatory attitude of the proposed classification model on *Age* as *SA*. We split the data into *Young* and *Aged* account holders at age 25. At this threshold, 190 account holders are categorized as *Young* and 810 account holders as *Aged*. We use the discrimination identifier function for the calculation of discrimination:

$$Disc(D, Age, Young, Good)$$

where D is the German Credit Dataset under experiment. This discrimination turns out to be as high as

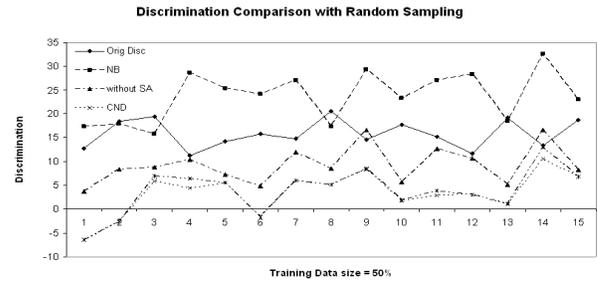$$conf(Aged \rightarrow Good) - conf(Young \rightarrow Good)$$
$$= 72.83 - 57.89 ,$$

Fig. 1. Accuracy comparison of different methods at different discrimination levels with respect to Age

which shows that *Aged* people are more likely to be classified into the *Good* credit class than *Young* people. We implement our proposed solution *CND* and compare it with the *Classification without Massaging*. We further divide each classification method into two techniques; in the first technique we classify the data with the *SA* while in the second, we omit the *SA* for future classification. Whereas for learning the ranking function the presence or absence of the sensitive attribute does not matter, it might have an influence for learning the final classifier *CND* in our proposed approach, and surely has a large influence for the *Classification without Massaging* approach.

In Figure 1, the results of 15 experiments with random sampling have been shown on the the X-axis while the resultant discrimination has been given on the Y-axis. For each of the 15 experiments the discrimination in the training data is given, as well as the discrimination attained by the *Classification without Massaging* (label NB), the discrimination of *Classification without Massaging* when we do not include the sensitive attribute in the training data (label without SA), and our approach with and without the sensitive attribute (labels CND and CND without SA). We find that *CND* classifies the future data with minimum discrimination as compared to the *Classification without Massaging*. Though the discriminatory behavior of the classification models is affected by the change of discrimination level in the data, *CND* always shows more impartiality as compared to the *Classification without Massaging*. We further elaborate on the effect of the discrimination level of the dataset on the discriminatory attitude of the classification model. For this purpose, we repeat our experiments on datasets having different discrimination levels. We observe that if we decrease the *Age* threshold from 52 to 25, the discrimination level also changes from 0 to 15, i.e., 0 represents no discrimination while 15 represents high discrimination. We find that *CND* and the *Classification without Massaging* work in a similar way on the dataset with no discrimination but as the discrimination level increases in the dataset, the *Classification without Massaging* classifies the future data with more discrimination while *CND* continues to classify the future data at the minimum level of discrimination, even though the discrimination in the dataset increases. Figure 2 shows the results of these experiments. The discrimination in the dataset for each experiment is shown on the X-axis and the values on Y-axis show the resultant discrimination by the different classification models in our experiments.
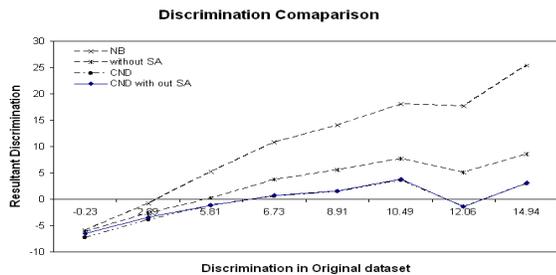
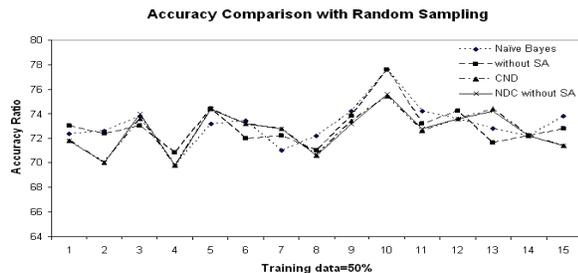Fig. 2. Discriminatory behavior of different methods at different discrimination levels with respect to Age

Fig. 3. Accuracy comparison of different methods at different discrimination levels with respect to Age

C. Accuracy Trade off

We find that the accuracy for *CND* drops to some extent but the difference in our experiments is so small that it can be ignored. Figure 3 shows the results of 15 experiments with random sampling. We find that *CND* classifies the data with high accuracy. It was feared that we would have to trade off accuracy for impartial future classification. We observe in Figure 3, however, that *CND* works with a very reasonable degree of accuracy.

Figure 4 shows the accuracy level when we apply both approaches on datasets having different levels of discrimination. The accuracy ratio is shown on the Y-axis while the X-axis shows different experiments which are conducted at different levels of discrimination.

VI. RELATED WORK

A detailed representation of Bayesian classification can be found in [3]. There are only few papers which propose classification with no discrimination. To the best of our knowledge, this paper is the second one on the classification with no discrimination in data mining. The work of [1] has similar motivation towards the solution of the discrimination problem. They have used similar concepts of discriminatory attribute
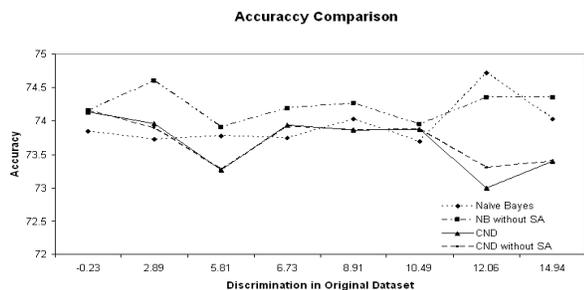
values. The authors of [1] concentrate on identifying discriminatory rules that are present in a dataset, hence they learn potential discriminatory guidelines that have been followed in the decision procedure. A central notion in their work is that of the context of the discrimination. That is, specific regions in the data are identified in which the discrimination is particularly high. They also use background knowledge for the identification of discriminatory guidelines in the dataset, in the case the discriminatory attribute is not present in the dataset. *CND* on the other hand assumes that historical data containing discrimination is available and attempts to massage the data before learning a classifier. This data is then used for learning a classifier with no discrimination for future decisions. The work of [4] 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 k-anonymity [5]. Although the goal there is different, namely removing data that allows for the identification of individuals, the mechanism is the same: before the data is released for mining, it is sanitized and the altered dataset is released. Other interesting related work concerns the incorporation of background knowledge in rule mining [6], [7]. The similarity is, however, artificial, as the techniques of [6], [7] are stated in a completely different setting of pattern mining and no guarantees regarding the discriminatory behavior of the discovered rules can be given.

VII. CONCLUSION AND FUTURE WORK

The notion of discrimination is non trivial and poses ethical and legal issues as well as obstacles in practical applications. *CND* provides us with a simple yet powerful starting point for the solution of the discrimination problem. *CND* classifies the future data (both discriminatory and non discriminatory) with minimum discrimination and high accuracy. It also addresses the problem of redlining. In future, we will explore other classification models for discrimination-free classification. Furthermore, we plan to study the incorporation of numerical attributes and groups of attributes as sensitive attribute(s).

REFERENCES

[1] D. Pedreschi, S. Ruggieri, and F. Turini, "Discrimination-aware data mining," in *Proceedings of the Fourteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2008.

[2] D. Newman, S. Hettich, C. Blake, and C. Merz, "(uci) repository of machine learning databases," 1998.

[3] R. Duda, P. Hart, and D. Stork, *Pattern classification and scene analysis*. Wiley New York, 1973.

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

[5] 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*, 1998, pp. 1–19.

[6] S. Jaroszewicz and T. Scheffer, "Fast discovery of unexpected patterns in data, relative to a bayesian network," in *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2005, pp. 118–127.

[7] S. Jaroszewicz and D. A. Simovici, "Interestingness of frequent itemsets using bayesian networks as background knowledge," in *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2004, pp. 178–186.

Fig. 4. Accuracy comparison of different methods at different discrimination levels with respect to Age