Data Min Knowl Disc (2010) 21:277–292
DOI 10.1007/s10618-010-0190-x

Three naive Bayes approaches for discrimination-free classification

Toon Calders · Sicco Verwer

Received: 30 April 2010 / Accepted: 20 June 2010 / Published online: 27 July 2010

Abstract In this paper, we investigate how to modify the naive Bayes classifier in order to perform classification that is restricted to be independent with respect to a given sensitive attribute. Such independency restrictions occur naturally when the decision process leading to the labels in the data-set was biased; e.g., due to gender or racial discrimination. This setting is motivated by many cases in which there exist laws that disallow a decision that is partly based on discrimination. Naive application of machine learning techniques would result in huge fines for companies. We present three approaches for making the naive Bayes classifier discrimination-free: (i) modifying the probability of the decision being positive, (ii) training one model for every sensitive attribute value and balancing them, and (iii) adding a latent variable to the Bayesian model that represents the unbiased label and optimizing the model parameters for likelihood using expectation maximization. We present experiments for the three approaches on both artificial and real-life data.

Keywords Discrimination-aware classification · Naive Bayes · Expectation maximization

1 Introduction

The topic of Discrimination-Aware classification was first introduced in Kamiran and Calders (2009), Calders et al. (2009), and is motivated by the observation that often training data contains unwanted dependencies between the attributes. Given a labeled dataset and a sensitive attribute; e.g., ethnicity, the goal of our research is to learn a classifier for predicting the class label that does not discriminate w.r.t. the sensitive

Responsible editor: José L Balcázar, Francesco Bonchi, Aristides Gionis, Michèle Sebag.

T. Calders · S. Verwer (✉)
Eindhoven University of Technology, Eindhoven, Netherlands
e-mail: s.verwer@tue.nl

attribute; e.g., for every ethnic group the probability of being in the positive class should roughly be the same. We call such constraints *independency constraints*. The paper will be about different techniques of learning and adapting Bayesian classifiers to make them discrimination-aware.

Throughout the paper we will assume that a labeled dataset D is given, with a binary class attribute C which takes values $\{-, +\}$ and one binary sensitive attribute S which takes values $\{S_-, S_+\}$ that has an unwanted correlation with the class attribute. The goal now is to learn a classifier on this data that optimizes predictive accuracy and is subject to the condition that its predictions are non-discriminatory. Discrimination in this paper is measured by the *discrimination score*, which is defined as the difference $P(C = + \mid S_+) - P(C = + \mid S_-)$. We will concentrate on naive Bayes classifiers. We assume that the sensitive attribute is available for training as well as for prediction.

1.1 Contributions

Our contributions in the paper are as follows:

- The discrimination-aware classification problem is illustrated and motivated. We show that simply removing the sensitive attribute from the training dataset does not solve the problem, due to the so-called red-lining effect.
- We propose three approaches to tackle the problem of discrimination-aware classification with naive Bayes classifiers:
  - in a post-processing phase we modify the probability of the decision being positive by changing the probabilities in the model,
  - we train one model for every sensitive attribute value and balance them, and
  - we add a latent variable in the Bayesian model that represents an unbiased, discrimination-free label and optimize the model parameters for likelihood using expectation maximization.
- We present and discuss experiments for the three approaches on both artificial and real-life data.

The organization of the paper is as follows: in Sect. 2 we motivate the discrimination-aware classification problem with two examples and illustrate the red-lining effect on the census-income dataset. In Sect. 3 the three Bayesian approaches are introduced and in Sect. 4 the expectation maximization technique for the third method involving the latent variable is detailed. Section 5 presents the results of the experimental evaluation and Sect. 6 concludes the paper.

2 Motivation and problem illustration

2.1 Motivation

We motivate the *discrimination-aware classification problem* setting with an example of a bank wanting to partially automate their loan issuing system. Consider, e.g., a bank that wants to use historical information on personal loans to learn models for predicting for new loan applicants the probability that they will default their loan.

It could very well be that this data shows that members of certain ethnic groups are more likely to default their loan. Nevertheless, from an ethical and legal point of view it is unacceptable to use the ethnicity of a person to deny the loan to him or her, as this would constitute an infringement of the discrimination laws. In such cases, the ethnicity of a person is likely to be an *information carrier* rather than a distinguishing factor; people from a certain ethnic group are more likely to default their loan because, e.g., the average level of education in this group is lower. In such a situation it is in general perfectly acceptable to use level of education for selecting loan candidates, even though this would lead to favoring one ethnic group over another. The bank could legally decide to split up the group of loan applicants according to their education level, and learn more fine-grained models for each of these groups separately. A prerequisite for this grouping or *stratification* approach is of course that the attribute education level is present in the dataset.

The overall effect of stratification will be that one ethnic group may be favored over another. Nevertheless, in each of the groups separately, the model should give equal probability to both classes. In the different strata, however, there may still be a strong dependency between ethnicity and loan defaulting. The reasons for this dependency may not be present in the dataset and latent to the decision maker. For example, it could be that the age distribution is different for the ethnic groups (e.g., one group has much more very young people), but the age of the loan applicants is not present in the dataset. As such, even though the discrimination-freeness requirement does not apply to the general setting, it does apply to the modelling problems for the individual strata. A straightforward approach to avoid that the classifier's prediction will be based on the sensitive attribute would be to remove that attribute from the training dataset. This approach, however, does not work, as we will show in the next section. The reason that suppression of the sensitive attribute does not work is that there may be other attributes that are highly correlated with it. In such a situation the classifier will use these correlated attributes to indirectly discriminate. We call this the *red-lining effect*. In the banking example, e.g., postal code may be highly correlated with ethnicity. Removing ethnicity would not solve much, as postal code is an excellent predictor for this attribute. Obviously, one could decide to remove these highly correlated attributes from the dataset as well. Although this would resolve the discrimination problem, in the process useful information will get lost. For example, when giving a loan for renovating a house it may be quite important to know if the house is located in the city center or in one of the suburbs. Postal code can reveal racial information and yet at the same time, still give useful, non-discriminatory information on loan defaulting.

The main motivation for starting out this research topic stems from a recently started collaboration with WODC; a Dutch study center associated with the department of Justice. The goal of this agency is providing data and modelling demographic and crime data to support policy making. 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 on such data could lead to discriminatory recommendations to the decision makers. Removing the ethnicity attributes would not solve the problem due to the red-lining effect, but in contrast even aggravate it, as the discrimination still would be present, only it would be hidden better. In such situations our discrimination-aware data mining paradigm clearly applies; even though racial discrimination would improve the accuracy of our classifier, in the policy making context it is unacceptable.

2.2 Illustration

In order to illustrate the difficulties of the problem of discrimination-free classification, we give some examples from the census income data-set.[1] From this data set we try to learn a naive Bayes classifier that can be used to decide whether a new individual should be classified as having a high or a low income. Historically, this decision has been biased towards the male sex, as can be seen in the following contingency table (containing co-occurrence counts):

	Male	Female
High income	3256	590
Low income	7604	4831

This table shows the number of male and female individuals in the high and low income class. About 30% of all male individuals and only about 11% percent of all female individuals have a high income. In total about 24% of all individuals are classified with a high income. If one learns a naive Bayes classifier from this data, this difference in income will be learned as a rule, resulting in even more distinction between the male and female individuals:

	Male	Female
High income	4559	422
Low income	6301	4999

Learning this classifier results in about 42% of all males having a high income, and only 8% of all females. This is of course highly undesirable, and can even lead to big fines for banks if they were to implement a decision system (for instance for assigning loans) based on such a classifier.

The problem we are trying to solve is how to obtain a good classifier that makes no distinction between males and females (or any other sensitive attribute). A simple solution that comes to mind is to leave out the sex attribute from the training data. In such a setting, the classifier will not be able to infer this distinction as a rule. Unfortunately, this is false:

	Male	Female
High income	4134	567
Low income	6726	4854

Even with the sex attribute left out of the data, the male individuals are still clearly favored by the naive Bayes classifier (38% against 10%). Moreover, they are still more favored than in the training data itself! The reason for this is the so-called *red-lining* effect: the classifier uses features that correlate with the sex attribute in order to learn similar rules. In other words, it will discriminate indirectly. The main goal of this paper is to not only remove direct discrimination, but to remove this red-lining effect as well.

[1] http://archive.ics.uci.edu/ml/datasets/Census+Income.

2.3 Measuring discrimination

Unfortunately, it is still unclear how to test for discrimination. A simple solution is the *discrimination score*, which we define as the difference between the probability of a male and a female of being in the high-income class. For the data and two classifiers above we get:

1. **Data.** 0.30 − 0.11 = 0.19
2. **Naive Bayes.** 0.42 − 0.08 = 0.34
3. **Naive Bayes without sensitive attribute.** 0.38 − 0.10 = 0.28

In an ideal world, this value would be 0.0. However, we do not live in an ideal world. It is a good and difficult question whether a discrimination value of 0.0 is desirable. It may very well be the case that the attributes that are indicative of whether a person should get a loan correlate with the sex of that person, for instance that person's salary. Without any additional knowledge, it is impossible to distinguish between this correlation and the red-lining effect. That is why we test for discrimination by checking whether this value equals 0.0. A nice feature of this measure is that it also tells us the severity of the discrimination. For instance, a discrimination value of 0.05 can be acceptable in many domains due to the aforementioned reason.

3 Three naive Bayes approaches

We investigate three approaches for removing discrimination from a naive Bayes classifier.

3.1 Modifying naive Bayes

The most straightforward method for removing discrimination is to modify the probability distribution $P(S|C)$ of the sensitive attribute values S given the class values C. Simply keep adding probability to the discriminated sensitive values S_- given the positive class C_+, and removing probability from the favored sensitive values S_+ given the positive class. Unfortunately, this simple scheme has the unwanted side-effect of either always increasing or always decreasing the number of positive labels assigned by the classifier, depending on whether favored sensitive values occur less frequently or more frequently in the data-set. In many applications, keeping this number close to the number or positive labels in the data-set is highly favorable. For instance, in the setting of banks assigning loans to individuals, the bank does not suddenly want to assign less or more loans. We therefore change the naive Bayes model slightly by changing $P(S|C)$ into $P(C|S)$ (see the first graph in Fig. 1). The joint distribution over the class C, sensitive S, and all other $A_1 \dots A_n$ attributes becomes

$$P(C, S, A_1, \dots, A_n) = P(S)P(C|S)P(A_1|C) \dots P(A_n|C)$$

instead of

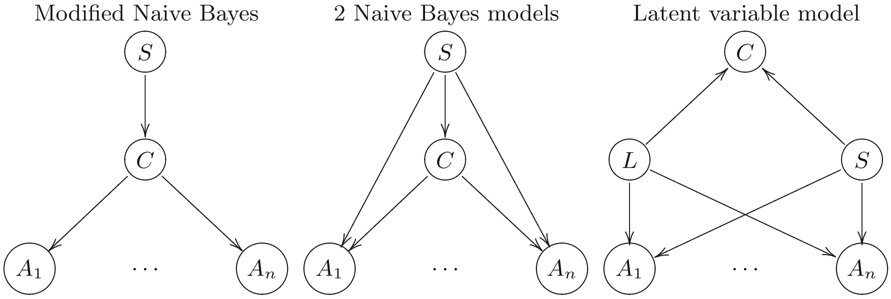

Fig. 1 Graphical models of the three naive Bayes approaches for discrimination-free classification

$$P(C, S, A_1, \ldots, A_n) = P(C)P(S|C)P(A_1|C)\ldots P(A_n|C)$$

We then modify $P(C|S)$ until there is no more discrimination in the labels assigned using the new model. We make these modifications in such a way that the number of assigned positive labels does not deviate much from the number of positive labels in the data-set. The exact method in which we perform this change is shown in Algorithm 1.

Algorithm 1 removes discrimination from a naive Bayes classifier, but it does not actively try to avoid the red-lining effect. Although the resulting decision is discrimination-free, the decision is not necessarily independent from the correlated attributes A_s.

Algorithm 1 Modifying naive Bayes

```
Require: a probabilistic classifier M that uses distribution $P(C|S)$ and a data-set D
Ensure: M is modified such that it is (almost) non-discriminating, and the number of positive labels
  assigned by M to items from D is (almost) equal to the number of positive items in D

  Calculate the discrimination disc in the labels assigned by M to D
  while disc > 0.0 do
    numpos is the number of positive labels assigned by M to D
    if numpos < the number of positive labels in D then
      $N(C_+, S_-) = N(C_+, S_-) + 0.01 \times N(C_-, S_+)$
      $N(C_-, S_-) = N(C_+, S_-) - 0.01 \times N(C_-, S_+)$
    else
      $N(C_-, S_+) = N(C_-, S_+) + 0.01 \times N(C_+, S_-)$
      $N(C_+, S_+) = N(C_-, S_+) - 0.01 \times N(C_+, S_-)$
    end if
    Update M using the modified occurrence counts N for C and S
    Calculate disc
  end while
```

3.2 Two naive Bayes models

Our second approach is to avoid this dependence on A_s by removing the correlation between S and A_s from the data-set used to train the naive Bayes classifier. This can for instance be achieved by removing A_s from the data-set altogether; the resulting classifier will be independent without modification. The price to pay, however, is a big loss in accuracy due to the reduction in number of attributes. Therefore, a more sophisticated method is not to remove the attributes A_s, but to remove the fact that they can be used to decide S. An easy way to achieve this is to split the data-set into two separate sets: one for S_+, and one for S_-. The model M_+ is learned using only the tuples from the data-set that have a favored sensitive value S_+. The model M_- uses only those that have a discriminated sensitive value S_-. The overall classifier chooses either M_+ or M_- depending on the value of S and uses that model's classification. Thus, in our banking example, we would learn two different models: one for males, and one for females. Such an approach is intuitively appealing since sex discrimination occurs when males and females are treated differently.

Overall, since M_+ and M_- share the same naive Bayes structure, this approach can be modeled by connecting S to all other attributes in this structure. This is shown by the second graph in Fig. 1. Since all probability distributions in the naive Bayes structure depend on S, this equals two different naive Bayes models. In this overall model, we remove discrimination by modifying the probability $P(C|S)$ using the same method as before (Algorithm 1).

3.3 A latent variable model

Our third and most complicated approach tries to model the discrimination process in order to discover the actual class labels that the data-set should have contained if it would have been discrimination free. Since they are not observed, these actual class labels are modeled using a latent (or hidden) variable L. How to include this latent variable in a naive Bayes like model depends crucially on our knowledge of this variable. Regarding this, we assume the following:

1. L is independent from S, i.e., the actual labels are discrimination-free;
2. C is determined by discriminating the L labels using S uniformly at random.

These are two strong assumption of the actual way in which discrimination occurs that simplify the way in which we deal with discrimination. The first assumption allows us to focus only on the overall discrimination, i.e., the difference between $P(C_+|S_+)$ and $P(C_+|S_-)$. Any other form of discrimination, such as discrimination dependent on an attribute A is neglected. Thus, in the resulting model $P(L_+|S_+, A)$ can be very different from $P(L_+|S_-, A)$. In such a case, we still call the model discrimination-free. The second assumption speaks for itself. Every tuple has an equal chance of being discriminated, again independent of attributes $A_1, \ldots, A_n$, and thus also independent of the probability of being assigned a positive label $P(L_+|A_1, \ldots, A_n)$.

These two assumptions might not correspond to how discrimination is being applied in practice, but because they result in a simple model, they do allow us to study the problem of discrimination-free classification in detail. The resulting model is the third graph in Fig. 1. In this model, we again remove the fact that an attribute A can be used to decide S by splitting $P(A|L)$ into $P(A|L, S_+)$ and $P(A|L, S-)$. We now show how to find the values of the latent variable in this model, and then we discuss some results we obtained with all three approaches.

4 Finding the latent values

The parameters of the first two models described in the previous section can be trivially observed from the data-set. Discrimination can then be removed from the models by applying Algorithm 1. Estimating the third model from a data-set is more difficult. In order to do so, we need to find good values to assign to the latent attribute in every tuple from the data-set. Essentially, this problem comes down to finding two groups (or clusters) of tuples: the ones that should have gotten a positive label L_+, and those that should have gotten a negative label L_-. We now describe the standard approach of expectation maximization that we use to find these two clusters.

4.1 Expectation maximization

Given a model M with a latent attribute L, the goal of this algorithm is to set the parameters of M such that they maximize the likelihood of the data-set D. Unfortunately, since L is unobserved, the parameters involving L can be set in many different ways. Exhaustively searching all of these settings is a hopeless task. Instead, expectation maximization iteratively optimizes these settings given D (the M-step), then calculates the expected values of the L attribute given those settings (the E-step), and incorporates these into D. This is a greedy procedure that converges to a local optimum of the likelihood function. Typically, random restarts are applied (randomizing the initial values of L in D) in order to find better optima.

4.2 Using prior knowledge

For the problem of finding the actual discrimination-free class labels L_+ and L_- we can do a lot better than simply running EM and hoping that the found solution corresponds to discrimination-free labels. For starters, it makes no sense to modify the labels of tuples with favored sensitive values S_+ and negative class labels C_-. The same holds for tuples with discriminated sensitive values S_- and positive class labels C_+. Modifying these can only result in more discrimination, so we fix the latent values of these tuples to be identical to the class labels in the data-set. We exclude these values from the E-step of the EM algorithm.

Another improvement over blindly applying EM is to incorporate prior knowledge of the distribution of C given L and S, i.e., $P(C|L, S)$. In fact, since the ultimate goal is to achieve zero discrimination, we can pre-compute this entire distribution. We show how to do this using an example.

Example 1 Suppose we have a data-set consisting of 100 tuples, distributed according to the following frequency counts:

	S_+	S_-
C_+	40	20
C_-	10	30

Clearly, there is some discrimination: the ratio of tuples with S_+ that have a positive class label C_+ $\left(\frac{4}{5}\right)$ is much larger than the ratio of tuples with S_- that have the positive class $\left(\frac{2}{5}\right)$. Initially, we set the distribution over L to be equivalent to the distribution over C, keeping the discrimination intact:

	S_+			S_-	
	L_+	L_-		L_+	L_-
C_+	40	0	C_+	20	0
C_-	0	10	C_-	0	30

Next, we rectify this situation by subtracting n from L_+ with S_+, and compensate by adding n to L_- with S_+. We also subtract n from L_- with S_-, and add n to L_+ with S_-. Since we want the number of tuples with actual positive labels L_+ to be equal to the number of tuples with positive labels in the data S_+, there is a unique and easy to compute setting to n that achieves zero discrimination. In this case, it is 10, resulting in the following distribution:

	S_+			S_-	
	L_+	L_-		L_+	L_-
C_+	30	10	C_+	20	0
C_-	0	10	C_-	10	20

We use these counts to pre-compute the probability table $P(C|L, S)$ in the latent variable model.

5 Experiments

In order to test the three naive Bayes approaches for discrimination-free classification, we performed tests on both artificial and real-world data. Here we make use of our model for discrimination (the third model from Fig. 1) to generate the artificial datasets. A big advantage of this artificial data is that we can also generate the actual class labels that would have been assigned to the tuples as if there was no discrimination. These labels are used to test the accuracy of the classifiers. In the real-world data, we do not have this luxury of a discrimination-free test-set. Therefore we can only test the accuracy on real-world data using (the sometimes incorrect) discriminated class labels. Clearly, this is an inferior method to measure classifier performance. On the other hand, a problem of using our discrimination model is that it is based on assumptions that might not always hold in practice.

5.1 Tests on artificial data

We first generate data using our latent variable model M, and then test our three approaches on this data. We first describe how we performed this data generation, afterwards we discuss the results.

5.1.1 Generating data

We require a random initialization of the parameters of M. However, this initialization should not be completely random; it is unlikely that the joint distribution of an attribute A and the latent class L is completely different for tuples with $S = 1$ than for those with $S = 0$. For example, suppose that we initialize the probability distribution $P(A|L, S)$ in this way:

$P(A = 1\|L, S)$	L	S
0.8	0	0
0.4	0	1
0.1	1	0
0.6	1	1

This would imply that for $S = 0$, $A = 1$ is a strong indication for $L = 0$, whereas for $S = 1$, all of a sudden this relation turns around. Not only is such a switch unnatural, but it would make our artificial problem trivial. Therefore the probabilities are drawn in such a way that the joint distribution of A and L for $S = 0$ is roughly (not completely of course) the same as for $S = 1$. Similarly for $L = 0$ and $L = 1$ not too large differences in probability are allowed as this would make getting a high accuracy too easy.

Finally, the conditional distribution of C w.r.t. L and S is fixed as follows:

$P(C = 1\|L, S)$	L	S
0.1	0	0
0.2	0	1
0.8	1	0
0.9	1	1

Hence, a subject with $L = 0$ has a small chance of getting class label $C = 1$ in the dataset anyway. This chance is slightly higher for subjects with $S = 1$. On the other hand, a subject with latent true class $L = 1$ has a very high chance of getting class label $C = 1$. This chance is higher for subjects with $S = 1$. These two inequalities together cause the discrimination in the dataset.

Even with these constraints, the problem of discrimination-free classification can become very easy simply because the probability of attribute A can be very high for tuples with a positive actual class label L_+, and very low for those with a negative actual class label L_-. If this occurs for multiple attributes, then these attributes effectively divide the feature-space into two clearly distinguishable clusters. We want to bound this effect. This is done by constraining the allowed difference. If two probabilities (drawn uniformly at random) do not meet this constraint, we simply draw them again. In addition, we bound the differences between the probabilities of an attribute A with a favored or discriminated sensitive value S_+ or S_- in the same way.

Given a random but constrained model M, we use it to generate data. We generate two separate sets: the first containing tuples with the discriminated class label C, the second containing tuples with the actual class label L. We learn a model from the first set, and test it using the second set. The sets each contain 10 000 tuples, 20 boolean attributes in addition to C, S, and L, and they are constrained using maximal differences of 0.2, 0.4, and 0.8. For every combination of these differences, we generate 5 different models, resulting in 5 different pairs of sets. On these sets, we test our three approaches for discrimination-free classification. We test the expectation maximization method both with and without using prior information.

5.1.2 Results

The results of these tests are shown in Figs. 2 and 3. These plots show some interesting behavior, especially of the expectation maximization method. Initially, we would have expected this method to work best. However, it ties with the 2 naive Bayes models approach in terms of accuracy on the non-discriminating test-set. In terms of discrimination it is even outperformed by both the simple modified naive Bayes model and the 2 NB model approaches. In terms of accuracy, the naive Bayes model scores clearly less than the 2 NB model. The difference between these two methods increases when we increase the constraint on the maximum of the difference $|P(A|S_+) - P(A|S_-)|$. This makes sense since the naive Bayes model does not model the dependence of A on S, while the 2 NB model does.

Of all results, the results of the expectation maximization methods are the most surprising. We expected that since we model the latent variable L in such a way that there is no dependence between L and S, the maximum likelihood assignment

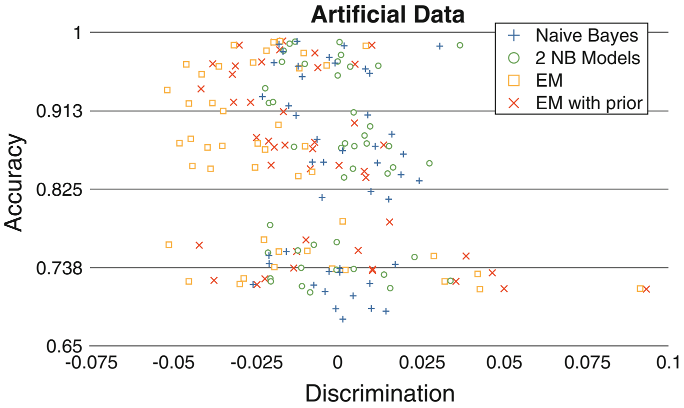

Fig. 2 The resulting discrimination and accuracy values of the trained classifiers on the discrimination-free test-set

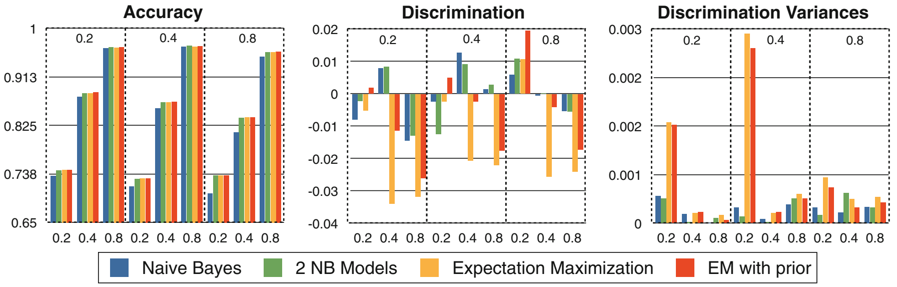

Fig. 3 The results of Fig. 2 (accuracy, discrimination, and discrimination variance) grouped per maximal difference value. The charts show the average values achieved by all methods for all combinations of the maximum bound values 0.2, 0.4, and 0.8. The values on the x-axis are the maximum bounds on $|P(A|L_+) - P(A|L_-)|$, the values in the x-axis boxes (at the top) are the maximum bounds on $|P(A|S_+) - P(A|S_-)|$

(or expectations) of L would also be without such dependence. Especially since we generate data from the exact same model, i.e., one where there is no such dependence. Unfortunately, this turned out to be false. Expectation maximization on average converges to approx. -0.025 discrimination, which gives the discriminated class 2.5% more chance to be assigned a positive latent label L. In addition, the exact discrimination value that EM converges to varies a lot, in one case it even reaches 0.09. This high variance seems not to depend on EM being a greedy algorithm that can converge to a local optimum: running it multiple times yields the same results. It is also not caused by the fact that we fix some of the latent values. We believe it might be the case that the maximum likelihood assignment does not correspond to a zero discrimination assignment. Investigating this behavior is left as future work.

5.2 Tests on census income

Census income is a data-set containing both numerical and categorical attributes that can be used to decide whether a new individual should be classified as having a high or a low income. We want to do so with zero discrimination with respect to the gender attribute. Our methods, however, are based on simple implementations and do not (yet) contain methods to deal with numerical attributes.[2] We therefore first discretize these attributes into 4 bins. The boundaries of these bins are the boundaries of the interquartile ranges. In addition, we remove low frequency counts (which may lead to problems for EM) by pooling any bin that occurs less than 50 times (out of a total of about 16 000). Thus, all infrequent attribute values are replaced by a unique (more frequent) ‘pool’ value. On this modified data-set we tested our algorithms.

5.2.1 Results

Figure 4 shows runs of the algorithms on the discretized and pooled census income data-set. The consecutive accuracy-discrimination values reached by the algorithms are connected by lines. These values are based on the data-set itself, not on a separate test-set. There are some interesting observations we can make from these plots. First of all, both the modified naive Bayes and 2 naive Bayes models seem to perform very well: the drop in accuracy is much smaller than the drop in discrimination. Logically, both of these plots stop at zero discrimination. This does not hold for the expectation maximization methods. After reaching zero discrimination, these methods are not yet converged to a local optimum of the likelihood function. Unfortunately, they converge to a point that is considerably worse in terms of accuracy and discrimination than the first time they reached zero discrimination. Interestingly, the first expectation maximization method reaches this point after just one iteration.

Another interesting observation is that, although the EM methods do not perform well in the end, they do start out good. Of all methods, they have the smallest

[2] We want to study the effect of discrimination and do not focus on maximizing the accuracy scores of our classifiers. The resulting accuracy values can therefore be a little lower than expected.

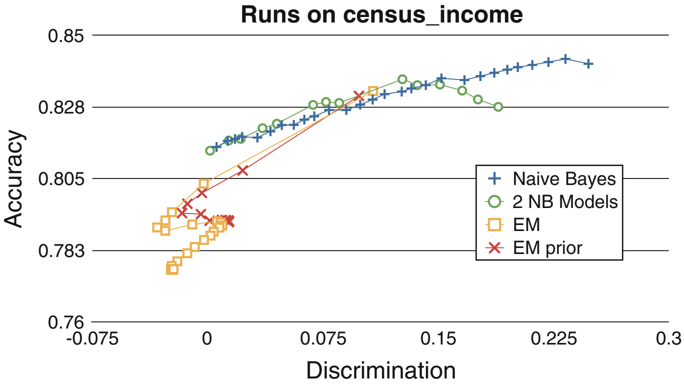

Fig. 4 Lines showing the the consecutive values reached by the runs of each of our algorithms. The accuracy and discrimination values are determined using the data-set

discrimination and a respectable accuracy. These points are determined by fixing the latent values for females with a positive class label and males with a negative class label, randomizing the other latent values, and using this set to estimate the latent variable model. Thus the latent variable model itself seems to perform good with respect to the other two approaches, only EM does not converge to the point we want it to converge to.

We tested all methods using 10-fold cross-validation on census income. The results are shown in the left part of Table 1. In addition to the four methods already discussed, we included the accuracy and dependency values of the EM methods if they would be stopped after the iteration in which they reached less than 0.01 discrimination. As expected, these perform better than the EM methods that were left to converge to a local optimum. The main conclusion we draw from the left part of Table 1 is that both naive Bayes modifications perform really well. However, they are still outperformed by previous methods that use decision trees Calders et al. (2010). This is not very surprising since decision trees in general work better than naive Bayes methods on

Table 1 Discrimination and accuracy values resulting from of 10-fold cross-validation of all methods with and without marginalizing over S on census income

	S included		Marginalizing over S	
	discrimination	Accuracy	discrimination	Accuracy
NB	−0.003	0.813	0.286	0.818
2 NB Models	−0.003	0.812	0.047	0.807
EM	0.000	0.773	0.081	0.739
EM prior	0.013	0.790	0.077	0.765
EM stopped	−0.006	0.797	0.061	0.792
EM prior stopped	−0.001	0.801	0.063	0.793

this data-set. We can also conclude that using prior information in the EM method improves its performance.

On the right part of Table 1, we show the performance of each of the methods when S is unknown during the testing phase. This shows the dependence of the classifiers on the S attribute. Ideally, we would like this dependence to be zero. In this case, it really does not matter what the value of the S attribute is, and hence there is no discrimination whatsoever. In our case, however, we modify the models using the S attribute, so there will be some dependence. We want this dependence to be as small as possible. The scores of the classifiers have been computed by marginalizing over S.

The first interesting observation we make from these scores is that the modified naive Bayes method suddenly obtains a very high discrimination. This is not surprising since this model is almost identical to the standard naive Bayes model when S is removed from the training data-set. The second observation is that for the (non-stopped) EM methods, the accuracy drops significantly. Thus there is a very high dependence on the S attribute. It is unclear why this is the case, but we hope it will provide some hints for investigating the convergence behavior of EM, which is planned for future research. The 2 naive Bayes models method has the lowest dependence on S, resulting in only about 5% discrimination if S is removed. This is somewhat surprising since this model uses S to split the data and then learn two separate models. Apparently, these two separate models are good at estimating S from the other attributes $A_1, \ldots, A_n$.

The overall conclusion of these experiments is that the 2 naive Bayes models method performs best: it achieves high accuracy scores with zero discrimination, and has the smallest dependency on S.

6 Related work

In a series of recent papers (Pedreschi et al. 2008, 2009; Kamiran and Calders 2009; Calders et al. 2009), the topic of discrimination in data mining received quite some attention. In Pedreschi et al. (2008, 2009), concepts of undesired dependency due to discrimination were introduced. These works, however, concentrate mainly on identifying the discriminatory rules that are present in a dataset, and the specific subset of the data where they hold, rather than on learning a classifier with independency constraints for future predictions. Discrimination-aware classification and its extension to independence constraints, were first introduced in Kamiran and Calders (2009), Calders et al. (2009) where the problem of undesired dependencies is handled by "cleaning away" the dependency from the dataset before applying the traditional classification algorithms. Other existing approaches include resampling (Kamiran and Calders 2010), and the construction of discrimination-aware decision trees (Calders et al. 2010).

There are many relations with more traditional techniques in classification as well. Due to space restrictions, we only discuss the most relevant links. Despite the abundance of related works, none of them satisfactory solves the classification with independency constraints problem. In Constraint-Based Classification, next to a training dataset also some constraints on the model have been given. Only those models that satisfy the constraints are considered in model selection. For example, when learning a decision tree, an upper bound on the number of nodes in the tree can be imposed (Nijssen and Fromont 2007). Our proposed classification problem with independency constraints clearly fits into this framework. Most existing works on constraint based classification, however, impose purely syntactic constraints limiting, e.g., model complexity, or explicitly enforcing the predicted class for certain examples. One noteworthy exception is monotonic classification (Kotlowski et al. 2007; Duivesteijn and Feelders 2008), where the aim is to find a classification that is monotone in a given attribute. Of all existing techniques in classification, monotone classification is therefore probably the closest to our proposal. In Cost-Sensitive and Utility-Bases learning (Chan and Stolfo 1998; Elkan 2001; Margineantu and Dietterich 1999), it is assumed that not all types of prediction errors are equal and not all examples are as important. The type of error (false positive versus false negative) determines the cost. Sometimes costs can also depend on individual examples. Nevertheless it is unclear how this can be generalized to independency.

7 Discussion and future work

We studied three Bayesian methods for discrimination-aware classification. In the first method, we change the observed probabilities in a naive Bayes model in such a way that its predictions become discrimination-free. The second method involved learning two different models; one for $S = 0$ and one for $S = 1$, and balancing these models afterwards. In the third and most involved method we introduced a latent variable L reflecting the latent "true" class of an object without discrimination. The probabilities in the model are then learned with the expectation maximization technique. All three methods were evaluated experimentally on an artificial and a real-life dataset. Surprisingly, the two-models approach outperformed the model with the latent variable. We plan to extend this work in the following ways:

- Right now our definition of discrimination is quite brute force. No discrimination at all is allowed. We want to extend the notion of discrimination to that of conditional discrimination; e.g., instead of requiring that there is no discrimination at all, we could weaken this condition to no discrimination *unless it can be explained by other attributes.* Another extension we plan to consider are numerical attributes (e.g., income) as sensitive attribute.
- There are many other graphical models possible. We could consider turning the arrows towards S, reflecting the idea that we can derive quite some information about the sensitive attribute S from the attributes A_i, but the attribute L should not help us any further for deriving S; i.e., S is conditionally independent of L given the attributes A_i. One obvious drawback of such a method is that the number of parameters to describe the distribution of S is exponential in the number of attributes A_i. Therefore it would be beneficial to consider other models that could be "inserted" into the Bayesian model to replace the probability table, such as, e.g., a decision tree.

- Obviously we want to further explore why the convergence of EM was relatively poor, even for the synthetic datasets where all conditions for a successful convergence were satisfied.

Acknowledgements This research was partially funded by the Netherlands Organisation for Scientific Research (NWO) through the *Responsible Innovation* project grant *Data Mining Without Discrimination* (KMVI-08-29).

References

Calders T, Kamiran F, Pechenizkiy M (2009) Building classifiers with independency constraints. In: IEEE ICDM workshop on domain driven data mining. IEEE press

Calders T, Kamiran F, Pechenizkiy M (2010) Constructing decision trees under independency constraints. Technical report, TU Eindhoven

Chan PK, Stolfo SJ (1998) Toward scalable learning with non-uniform class and cost distributions: a case study in credit card fraud detection. In: Proceedings of ACM SIGKDD, pp 164–168

Duivesteijn W, Feelders AJ (2008) Nearest neighbour classification with monotonicity constraints. In: Proceedings of ECML/PKDD'08. Springer, Berlin, pp 301–316

Elkan C (2001) The foundations of cost-sensitive learning. In: Proceedings of IJCAI'01, pp 973–978

Kamiran F, Calders T (2009) Classifying without discriminating. In: Proceedings of IC409. IEEE press

Kamiran F, Calders T (2010) Classification with no discrimination by preferential sampling. In: Proc. Benelearn

Kotlowski W, Dembczynski K, Greco S, Slowinski R (2007) Statistical model for rough set approach to multicriteria classification. In: Proceedings of ECML/PKDD'07. Springer, Berlin

Margineantu DD, Dietterich TG (1999) Learning decision trees for loss minimization in multi-class problems. Technical report, Department Computer Science, Oregon State University

Nijssen S, Fromont E (2007) Mining optimal decision trees from itemset lattices. In: Proceedings of ACM SIGKDD

Pedreschi D, Ruggieri S, Turini F (2008) Discrimination-aware data mining. In: Proceedings of ACM SIGKDD

Pedreschi D, Ruggieri S, Turini F (2009) Measuring discrimination in socially-sensitive decision records. In: Proceedings of SIAM DM