Three naive Bayes approaches for discrimination-free classification

Toon Calders · Sicco Verwer - 2010, Data Mining and Knowledge Discovery 21, no. 2

Overview

- How to modify naive Bayes classifier to avoid discrimination
- Labeled data is biased (both direct and indirect discrimination)
- Three approaches
  - Modifying probability of the decision being positive
  - Training one model for every sensitive attribute value and balancing them
  - Adding a latent variable to the Bayesian model that represents the unbiased label

Performance of Naive Bayes without discrimination awareness

Case:

- Classifying individual as high-income or low-income
- Focus on gender discrimination
- Census Income Data set
  (UCI Machine Learning Repository)
- Highly discriminatory labeling
  - About 30% of all male individuals and only about 11% percent of all female individuals have a high income

Performance of Naive Bayes without discrimination awareness

	Male	Female
High income	3256	590
Low income	7604	4831

(training data)

	Male	Female
High income	4559	422
Low income	6301	4999

(classification result)

“Learning this classifier results in about 42% of all males having a high income, and only 8% of all females”

Performance of Naive Bayes: Removal of discriminatory feature

	Male	Female
High income	4134	567
Low income	6726	4854

- Slight improvement of results
  - (38% male classified positively, against 10% of female)
- Still more discriminatory than labeled data
  - Redlining with features that correlate with gender.

Measuring discrimination

Discrimination score: difference between the probability of a male and a female of being in the high-income class

Data. $0.30 - 0.11 = 0.19$
Naive Bayes. $0.42 - 0.08 = 0.34$
Naive Bayes without sensitive attribute. $0.38 - 0.10 = 0.28$

Zero would be ideal, assumes probability of positive classification should be the same

Measuring discrimination

First note: Discrimination score measure seems to be simplistic!

- Measures both direct and indirect discrimination together.
- Assumes sample is representative of whole population
  - This case is ok, since it’s a census
  - Counter example - Credit rating - hypothetic:
    - Only successful females access credit
    - Most men access credit
    - Discrimination score zero is still discriminatory to women

Solution 1: Modifying probabilities

(aka: cheating)

Main idea

$$p(C_k|\mathbf{x}) = \frac{p(C_k)\ p(\mathbf{x}|C_k)}{p(\mathbf{x})}.$$

Removing discrimination by modifying the probability distribution P(S|C) of the sensitive attribute values S given the class values C.

eg.

Increase P(S♀|C+) and reduce P(S♂|C+), if we want to positively discriminate ♀, or vice-versa

First problem

If data is not symmetric (more males than females in the dataset, for eg.), total number of positive labels won't be constant.

By positively discriminating the smaller group, number of positive results will increase, and vice versa.

First problem

We change the naive Bayes model slightly by changing P(S|C) into P(C|S) - simple switch via bayes theorem.
The joint probability becomes

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

becomes

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

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

Summary of solution 1

- Removes discrimination from a naive Bayes classifier
- 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 As

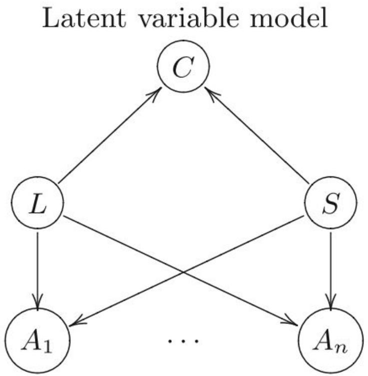
Solution 2: Two naive Bayes models

Main idea

- Previous solution does not remove indirect discrimination due to correlation of other variables
- Divide the model into 2 different models
- Train a model for $S_{♀}$ and another for $S_{♂}$

Solution 3 - Latent Variable

General Idea



Our data has discriminated labels, what if we can try to find how the training data should have looked like?

Add a Latent variable (L) to the model

L is independent of the sensitive parameter

C is determined by discriminating the L labels (using S uniformly at random)

Calculating latent variable

Expectation maximization

- Randomly initializing the L labels and iterating to optimize probability of dataset.

With prior information

- Starting from all the positively discriminated sensitive values
- P(C|L, S) - pre-compute distribution (we know we want to achieve zero discrimination)

Calculating latent variable

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

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

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+

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

Experiments

Results on artificial data

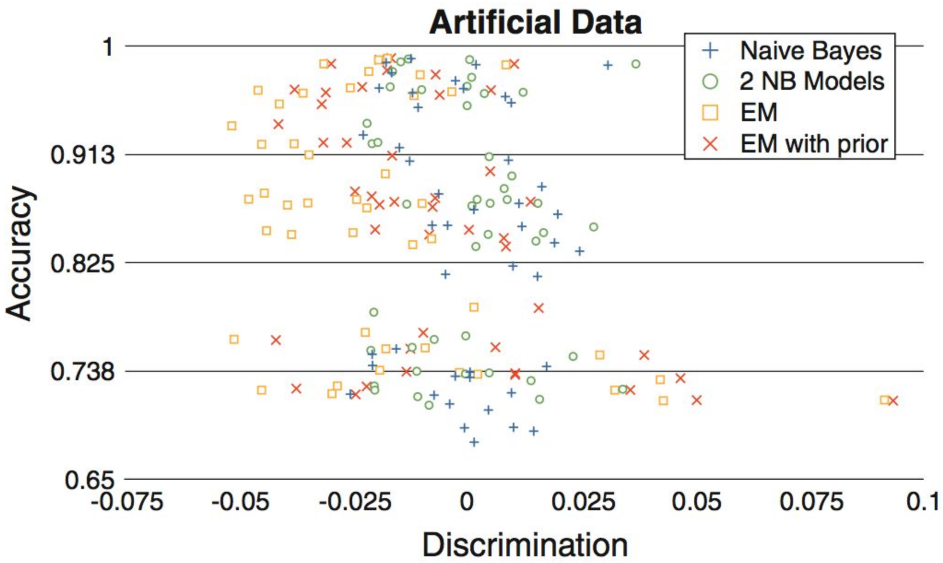


Fig. 2 The resulting discrimination and accuracy values of the trained classifiers on the discrimination-free test-set

Results on artificial data

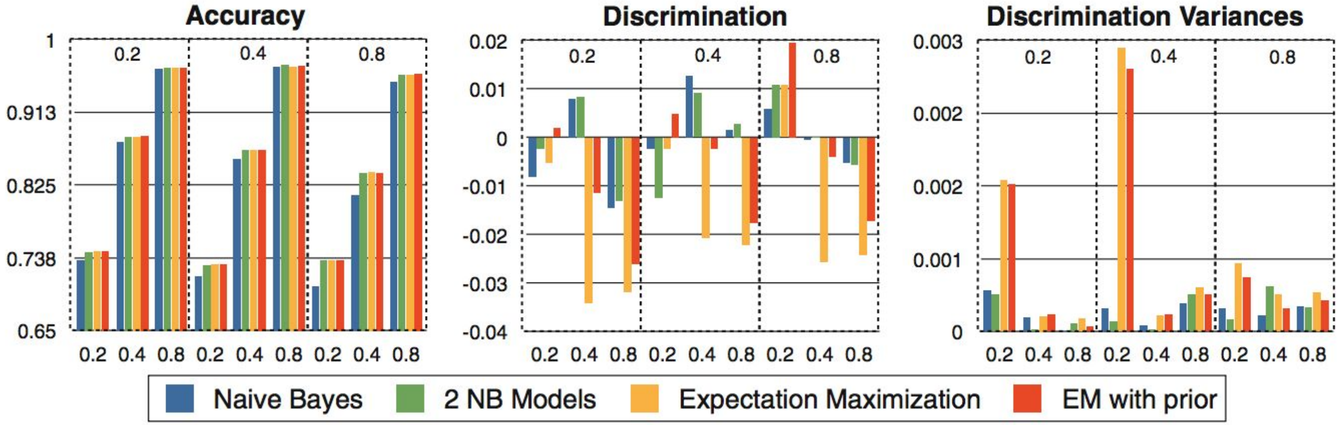


Fig. 3 The results of Fig. 2 (accuracy, discrimination, and discrimination variance) grouped per maximal difference value. The charts show the average values achieved by all methods for all combinations of the maximum bound values 0.2, 0.4, and 0.8. The values on the x-axis are the maximum bounds on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

Results on real data (census)

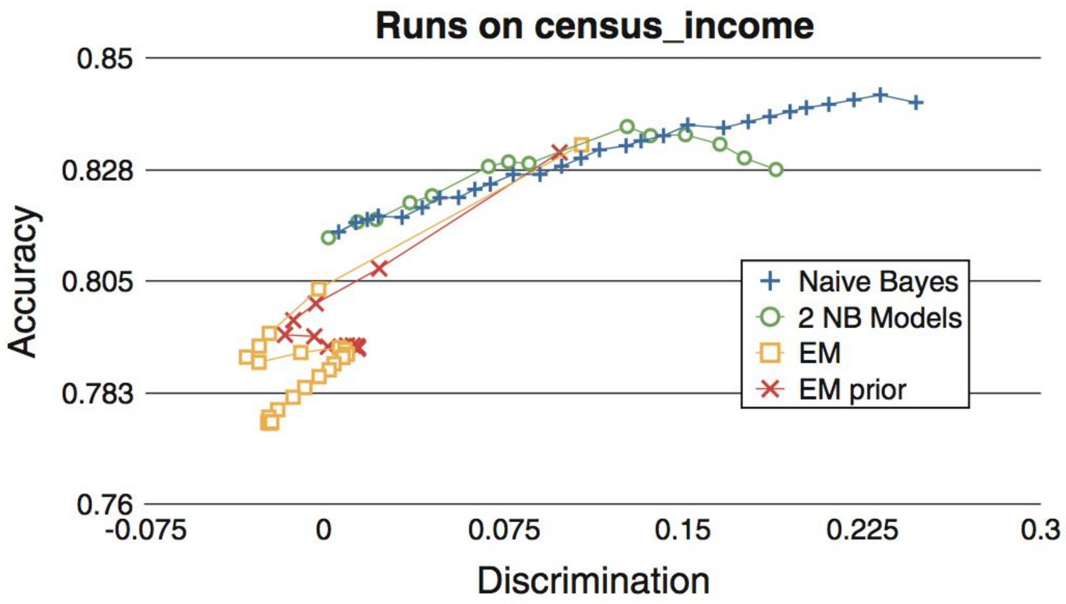


Fig. 4 Lines showing the the consecutive values reached by the runs of each of our algorithms. The accuracy and discrimination values are determined using the data-set

Results on real data (census)

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

Thank you