Three Naive Bayes Approached For Discrimination Free Classification

論文情報

- Toon Calders (Eindhoven University of Technology)
- Sicco Verwer (Eindhoven University of Technology)
- Data Mining and Knowledge Discovery, 2010
- https://link.springer.com/article/10.1007/s10618-010-0190-x

選んだ理由

- アノテーションバイアスの除去に関係ありそうな論文であったから.

Abstract

与えられたセンシティブ属性が独立であるという条件のもと，分類を行えるように単純ベイズ分類モデルを修正する方法を調べた．このような独立性制約は，何かを決めるために使うデータセットのラベルに例えば，男女や人種の差別などのバイアスが入っているときに起こる．この問題設定は，一部に差別が関わっている決定を許さないような既存の法律などさまざま状況に基づいて動機付けられた．通常の機械学習技術を素直に利用すると，会社に多額の罰金をもたらしてしまう．我々は差別のない単純ベイズ分類を作る3つの方法を提案する. (i) 決定の確率をポジティブに修正する．(ii) 全てのセンシティブな属性を用いて1つのモデルを学習し，それらを調整する方法.(iii) バイアスのないラベルとする潜在変数をモデルの中に追加して，期待値最大の尤度を用いて最適化する方法．人工データとリアルデータの２種類で実験を行なった.

Introduction

Discrimination-Aware classificationが初めて登場したのが，2009年にKamiranとCaldersの論文である．(あとで入れる)ここでは，データないの属性が，あって欲しくない関係を持つことがあることが観察されている．我々の研究ではセンシティブ属性を含んでいるデータセットを用いて，差別のない分類を行うことを目標としている．ここでいう差別のないとは，異なるセンシティブ属性でもラベルの比率が一定になるようなことをさす．またこの差別のないを我々は独立性制約という.

ここで，説明に使う表記を下にまとめる.

表記	説明
D	ラベル付きデータセット
C={+,-}	２種類のクラス属性
S={+,-}	ある一つのセンシティブな属性. クラス属性とは独立であってほしい.
A	S以外の属性．離散でも連続値でも何でもいい.

また差別をはかる指標として，ここではdiscrimination scoreを次のように定義する.

```
P(C=+|S_+) − P(C=+|S_−)
```

このdiscrimination scoreを0にしていくことが目標である． この論文の貢献は次の通り．

1. 差別の疑いがある分類問題の事例を紹介し，問題設定を作った．また単純にセンシティブな属性を取り除くだけではred-lining効果のせいで意味がないことを示した．
2. 公平なモデルを作る方法として3つ紹介する．
    1. データの後処理でモデルの確率を変更することで，ポジティブラベルを出すように調整する．
    2. すべてのセンシティブ属性を用いて，モデルを作り，バランスをとる．
    3. バイアスのない属性を潜在変数として追加し，モデルパラメータを尤度をEMアルゴリズムで最適化していく．

Preliminary

差別のない分類の大変さを説明するために，国勢調査のデータでの例を示す．

	Male	**Female**
High income	3256	590
Low income	7604	4831

男性のうち，High incomeの人の割合は30%,女性は10%である．また全体としては24%がHigh income. このまま単純ベイズで学習すると以下のようになる．

	Male	**Female**
High income	4559	422
Low income	6301	4999

男性のHigh incomeは42%,女性は8%になった．これは明らかに女性が差別されている． 単純に**性別**というセンシティブ属性を取り除いた状態で単純ベイズを行うと下のようになる．

	Male	**Female**
High income	4134	567
Low income	6726	4854

これでも男性のHigh incomeは38%, 女性10%という結果になってしまった．性別の情報を取り除いたのに，差別のある結果になってしまったのは，他の属性と男女の属性が独立でなく，間接的に影響を与えてしまっている．これがred-lining効果というやつ．我々の目標は直接の差別を取り除くだけでなく，このred-lining効果を取り除きたい．

差別指標として, discrimination scoreを用いる．今回の例でいうと．

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

である．理想的な差別のない状態であるなら，このスコアは0になる．

Proposal

Modifying naive Bayes

差別を除く方法として単純なものとでは，P(S|C)の確率分布を修正することである． この同時確率は次のようになる．

```
P(C,S,A_1,...,A_n) = P(S)P(C|S)P(A_1|C)...P(A_n|C)
```

後処理として,調整するアルゴリズムは下のようになる．

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

if-elseのところで後処理を行なっている．この方法では，差別は取り除けているが，red-lining効果を取り除けていない．

Two naive Bayes models

red-lining効果を取り除く手法を説明する．データセットをS_-に使う分とS_+に使う分で二等分する．モデルM_+はS_+を好むようなデータのみを使い，M_-jはS_-を好むようなデータのみを使う．全体のモデルは，SによってM_+とM_-のどっちを使うかを決め，分類に使う．この手法はM_+とM_-が同じBayes structureを用いているため同時分布は下のようにかける．

```
P(C,S,A_1,...,A_n) = P(S)P(C|S)P(A_1|S,C)...P(A_n|S,C)
```

これにAlgorithm1でP(C|S)を調整することで，差別を取り除くことができる．

A latent variable model

データセットの本物クラスラベル(バイアスがかかっていない)を見つけるための手法を提案する．この本物のクラスは直接観測されるものではないので，潜在変数Lとしてモデルに追加する．この潜在変数には次のことを仮定する．

1. LはSと独立である．つまり本物のラベルは，差別がない状態である．
2. Cは，Sをランダムに使ってLのラベルによって決定される? (C is determined by discriminating the L labels using S uniformly at random.)

一つ目の仮定によって，我々はP(C_+|S_+)とP(C_+|S_-)の違いによる差別のみを気にするだけでいい．二つ目の仮定は，すべてのデータ平等に差別されていることがあり，A_1,...,A_nは違いに独立である．つまりP(L+|A_1,...,A_n)で決定される．

このモデルは単純であるが，仮定がきびしいため，実世界に近くない．しかし，このモデルによって，詳細に差別のない分類の問題を見ることができる．同時分布は次のようになる．

```
P(A_1,...,A_n,S,C,L) = P(L)P(S)P(C|L,S)P(A_1|L,S)...P(A_n|L,S)
```

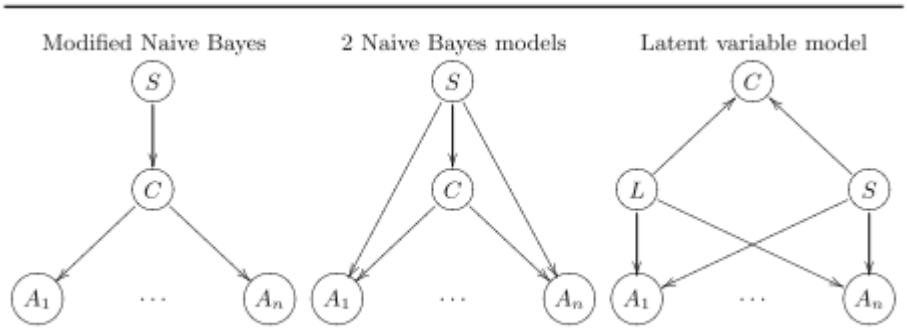

Fig. 1 Graphical models of the three naive Bayes approaches for discrimination-free classification

Finding the latent values

潜在変数を含むモデルの最適化として，EMアルゴリズムを用いる．

本物のクラスラベルL_+とL_-を見つけるために，ただ単にEMアルゴリズムを適用するよりもうちょっといい方がある．はじめの方では(S_+,C_-)や(S_-,C_+)のデータを修正するのはあまり意味をなさない．これをただ単に修正するともっと差別を産むことがある．だからこれらの潜在変数が同一である中で変更を行う．そしてこの変更されたデータはEステップでは使わない．二つ目として，P(C|L,S)の事前分布を用いる方法がある．前もって全体の分布を計算することは可能であるので，この方法は可能である．

(ちょっと省略...)

Experiments

潜在変数を用いたモデルによって人工データを生成する．人工データで実験の行う利点として，差別のないデータとして実験が行える点である．潜在変数Lを正しいラベルとしてAccuracyを計算することができる．実世界のデータでは，差別のないデータセットがないため，差別のある状態のクラスラベルを用いてAccuracyを計算していくことにする．

Tests on artificail data

潜在変数モデルMで生成するデータの検証をする． (省略していく)

Results

DiscriminatinとAccuracyはトレードオフの関係である．目標はdicriminationが小さく，Accuracyをあげること．2単純ベイズモデルと単純ベイズモデルがいい成績を出せた．

EMアルゴリズムの結果が以外にも悪かった．これは，EMアルゴリズムの最適解が必ずしも差別がない状態になるのと一致しないかもしれないからである. (Future workになる．)

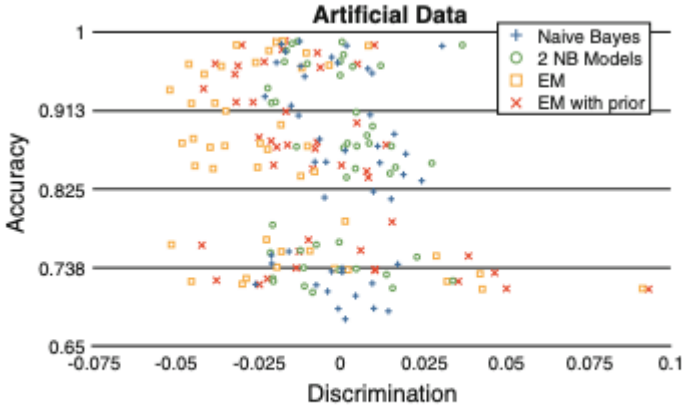

Fig. 2 The resulting discrimination and accuracy values of the trained classifiers on the discrimination-free test-set

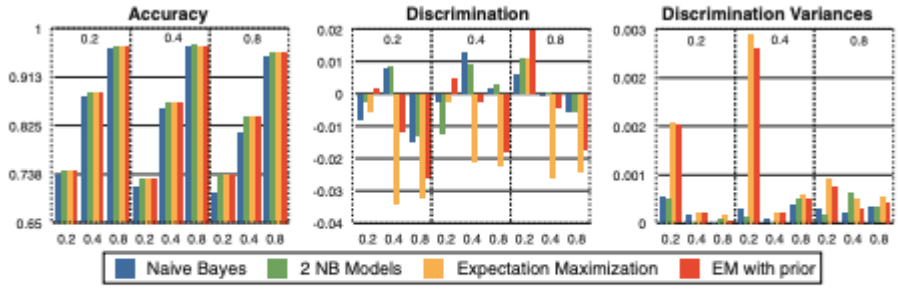

Fig. 3 The results of Fig. 2 (accuracy, discrimination, and discrimination variance) grouped per maximal difference value. The charts show the average values achieved by all methods for all combinations of the maximum bound values 0.2, 0.4, and 0.8. The values on the x-axis are the maximum bounds on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

Tests on census income

国勢調査の結果を用いて実験を行なった．各属性は離散化したデータにした．2単純ベイズと単純ベイズモデルはAccuracyの落ち具合がdiscriminationの落ち具合より小さくて，良い結果になった．EMアルゴリズムでははじめは

よかったが，最終的には結果が良くなくなってしまった．

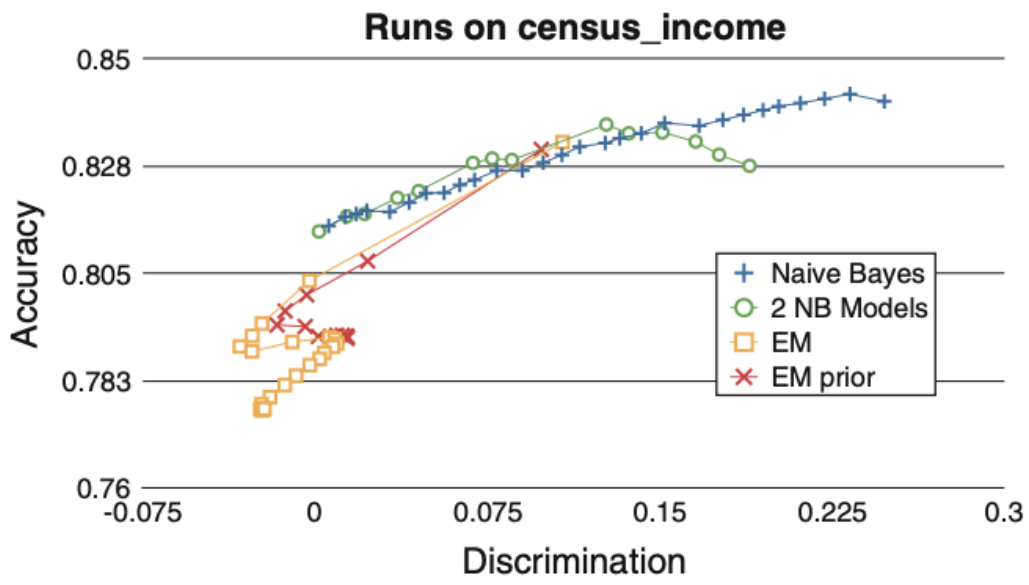

Fig. 4 Lines showing the the consecutive values reached by the runs of each of our algorithms. The accuracy and discrimination values are determined using the data-set

結論としては，2単純ベイズモデルが良い！

Discussion and future work

3つの差別を考慮したベイズモデルについて調査した．潜在変数モデルよりも，2単純ベイズモデルの方が性能がよかったのが意外だった．Future workとしては次の通り．

- 今回はセンシティブ属性は離散値のみを扱っているが，収入などの連続値も扱えるようにしていきたい．
- 今回の3つ以外のグラフィカルモデルがたくさんある．ベイズモデルだけでなく，決定木とかでもできたらいいのではないのか？

所感

- sensitive scoreを0にしていくことで差別がないと判断するのはデータによるのではないのかな？と思ってしまった．今回で言ったら．男女の高所得低所得の比が同じになるべきであるが，現実世界では本当に同じとは言えない? Accuracyとdiscriminationのトレードオフはどっちを取るべきか難しそう．
- NB,2NBでどうしてセンシティブ属性が取り除かれるかあまりわからない． 後処理でデータを操作しているところが取り除くのに影響していると思われるが，あまり実感できない．
- 潜在変数モデルの方が現実に近い気がするけど，悪い結果になったのは仮定が強すぎるから？
- 時間かけすぎた，反省．