The Independence of Fairness-Aware Classifiers

Toshihiro Kamishima[*], Shotaro Akaho[*], Hideki Asoh[*], and Jun Sakuma[**]

[*]*National Institute of Advanced Industrial Science and Technology (AIST), Japan*
[**]*University of Tsukuba, Japan; and Japan Science and Technology Agency*

IEEE International Workshop on Privacy Aspects of Data Mining (PADM)

@ Dallas, USA, Dec. 7, 2013

START

I'm Toshihiro Kamishima.
Today, we would like to talk about fairness-aware classification problem.

Introduction

[Romei+ 13]

Fairness-aware Data Mining
data analysis taking into account potential issues of fairness, discrimination, neutrality, or independence

Fairness-aware Classification

- One of a major task of fairness-aware data mining
- A problem of learning a classifier that predicts a class as accurately as possible under the fairness constraints from potentially unfair data

✤ We use the term “fairness-aware” instead of “discrimination-aware,” because the word “discrimination” means classification in the ML context, and this technique applicable to tasks other than avoiding discriminative decisions

Fairness-aware data mining is a data analysis taking into account potential issues of fairness, discrimination, neutrality, or independence.
In this talk, we focus on a fairness-aware classification task, which is one of major task of fairness-aware data mining.
This is a problem of learning a classifier that predicts a class as accurately as possible under the fairness constraints from potentially unfair data.

Introduction

Calders & Verwer's 2-naive-Bayes (CV2NB) is very simple, but highly effective

Other fairness-aware classifiers are equally accurate, but less fairer

↓

Why the CV2NB method performed better:

- model bias
- deterministic decision rule

↓

Based on our findings, we discuss how to improve our method

A fairness-aware classifier, Calders and Verwer's 2-naive-Bayes (CV2NB) method, is very simple, but highly effective.
We show the reasons why the CV2NB method performed better: the influences of a model bias and a deterministic decision rule.
Based on our findings, we discuss how to improve our method.

Outline

Applications of Fairness-aware Data Mining

- prevention of unfairness, information-neutral recommendation, ignoring uninteresting information

Fairness-aware Classification

- basic notations, fairness in data mining, fairness-aware classification, Connection with PPDM, Calders & Verwer's 2-naive-Bayes

Hypothetical Fair-factorization Naive Bayes

- hypothetical fair-factorization naive Bayes (HFFNB), Connection with other methods, experimental results

Why Did the HFFNB Method Fail?

- model bias, deterministic decision rule

How to Modify the HFFNB Method

- actual fair-factorization method, experimental results

Conclusion

This is an outline our talk.
After showing an applications of fairness-aware data mining, we introduce a problem of fairness-aware classification.
We propose our a simple method and compared it with Calders & Verwer's naive Bayes method.
We analyze why our simple method failed, and discuss how to modify the method.

Outline

Applications of Fairness-aware Data Mining
- prevention of unfairness, information-neutral recommendation, ignoring uninteresting information

Fairness-aware Classification
- basic notations, fairness in data mining, fairness-aware classification, connection with PPDM, & Verwer's 2-naive-Bayes

Hypothetical Fair-factorization Naive Bayes
- hypothetical fair-factorization naive Bayes (HFFNB), Connection with other methods, experimental results

Why Did the HFFNB Method Fail?
- model bias, deterministic decision rule

How to Modify the HFFNB Method
- actual fair-factorization method, experimental results

Conclusion

We begin by showing applications of fairness-aware data mining.

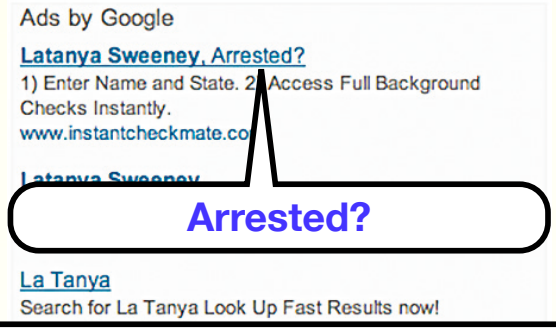
Prevention of Unfairness

[Sweeney 13]

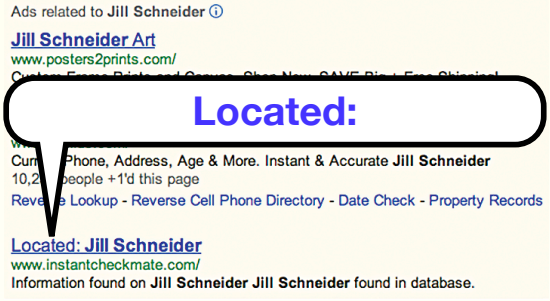
An suspicious placement keyword-matching advertisement

Advertisements indicating arrest records were more frequently displayed for names that are more popular among individuals of African descent than those of European descent

African descent names



European descent names



- This situation is simply due to the optimization of click-through rate, and no information about users' race was used
- Such unfair decisions can be prevented by FADM techniques

The first application is the prevention of unfairness.
The is an example of an suspicious placement keyword-matching advertisement.
Advertisements indicating arrest records were more frequently displayed for names that are more popular among individuals of African descent than those of European descent.
This situation is simply due to the optimization of click-through rate, and no information about users' race was used.
Such unfair decisions can be prevented by FADM techniques

Recommendation Neutrality

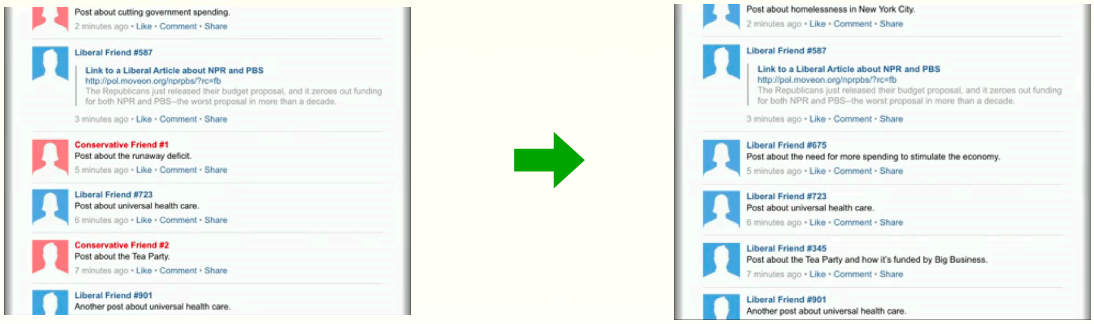
[TED Talk by Eli Pariser, http://www.filterbubble.com/]

The Filter Bubble Problem

Pariser posed a concern that personalization technologies narrow and bias the topics of information provided to people

Friend recommendation list in the Facebook

To fit for Pariser's preference, conservative people are eliminated from his recommendation list, while this fact is not noticed to him



- FADM technologies are useful for providing neutral information

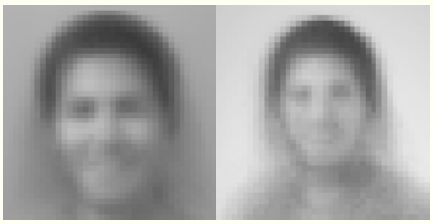
The second application is related to the filter bubble problem, which is a concern that personalization technologies narrow and bias the topics of information provided to people.
Pariser shows an example of a friend recommendation list in the Facebook.
To fit for his preference, conservative people are eliminated form his recommendation list, while this fact is not notified to him.
FADM technologies are useful for providing neutral information.

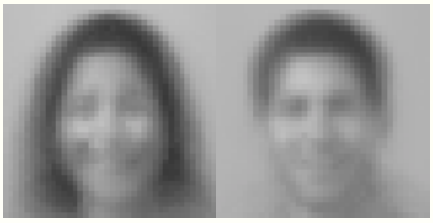
Ignoring Uninteresting Information

[Gondek+ 04]

non-redundant clustering : find clusters that are as independent from a given uninteresting partition as possible

a conditional information bottleneck method,
which is a variant of an information bottleneck method





clustering facial images

- Simple clustering methods find two clusters: one contains only faces, and the other contains faces with shoulders
- Data analysts consider this clustering is useless and uninteresting
- By ignoring this uninteresting information, more useful male and female clusters could be obtained

- Uninteresting information can be excluded by FADM techniques

The third application is ignoring uninteresting information.
The goal of the non-redundant clustering is to find clusters that are as independent from a given uninteresting partition as possible.
This is an example of clustering facial images:
Simple clustering methods find two clusters: one contains only faces, and the other contains faces with shoulders.
Data analysts consider this clustering is useless and uninteresting.
By ignoring this uninteresting information, more useful male and female clusters could be obtained.

Outline

Applications of Fairness-aware Data Mining
- prevention of unfairness, information-neutral recommendation, ignoring uninteresting information

Fairness-aware Classification
- basic notations, fairness in data mining, fairness-aware classification, connection with PPDM, & Verwer's 2-naive-Bayes

Hypothetical Fair-factorization Naive Bayes
- hypothetical fair-factorization naive Bayes (HFFNB), Connection with other methods, experimental results

Why Did the HFFNB Method Fail?
- model bias, deterministic decision rule

How to Modify the HFFNB Method
- actual fair-factorization method, experimental results

Conclusion

We then introduce a problem of fairness-aware classification.

Basic Notations

Y **objective variable**

- **a result of serious decision**
  ex., whether or not to allow credit

S **sensitive feature**

- **socially sensitive information**
  ex., gender or religion

$\mathbf{X}$ **non-sensitive feature vector**

- **all features other than a sensitive feature**
- non-sensitive, but may correlate with S

We begin by showing basic notations:
An objective variable Y represents a result of serious decision.
A sensitive feature S represents socially sensitive information.
All the other features consist of non-sensitive feature vector X.

Fairness in Data Mining

Fairness in Data Mining
Sensitive information does not influence the target value

↓

red-lining effect: Everyone inclines to eliminate a sensitive feature from calculations, but this action is insufficient

↑

Non-sensitive features that correlate with sensitive features also contains sensitive information

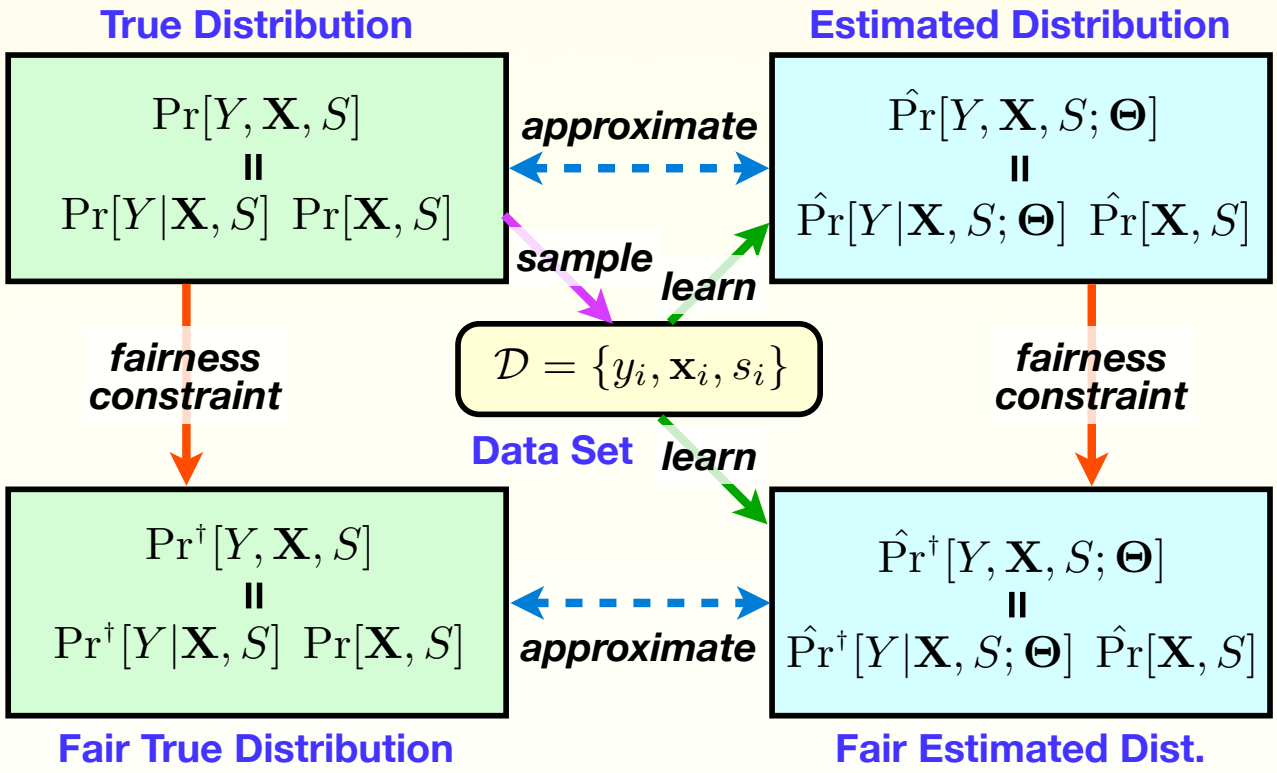
Y and S are merely conditionally independent: $Y \perp\!\!\!\perp S \mid \mathbf{X}$

↓

A sensitive feature and a target variable must be unconditionally independent $Y \perp\!\!\!\perp S$

To make a data mining process fair, sensitive information does not influence the target value.
For this purpose, everyone inclines to eliminate a sensitive feature from calculations, but this action is insufficient, because non-sensitive features that correlate with sensitive features also contains sensitive information.
Therefore, sensitive features and target variables must unconditionally independent.

Fairness-aware Classification

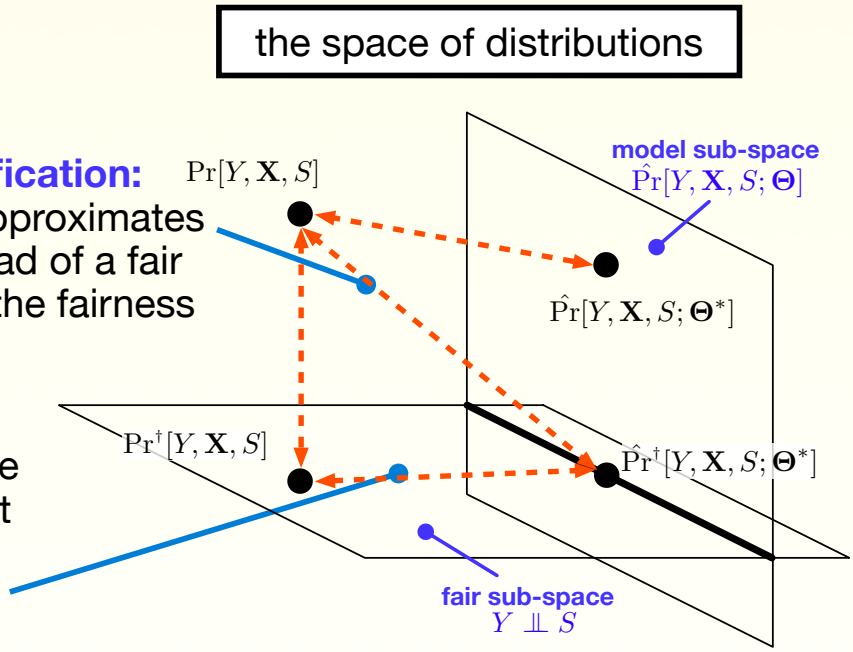


Fairness-aware classification is a variant of a standard classification.
In a case of a standard classification task, training data is sampled from a unknown true distribution.
A goal of a standard classification task is to estimate a model approximating the true distribution.
In a case of a fairness-aware classification task, we assume a fair true distribution that satisfies the fairness constraint.
A goal of fairness-aware classification task is to estimate a model approximating this fair true distribution.

Fairness-aware Classification

fairness-aware classification: find a fair model that approximates a true distribution instead of a fair true distribution under the fairness constraints

We want to approximate fair true distribution, but samples from this distribution cannot be obtained, because samples from real world are potentially unfair



We want to approximate fair true distribution, but samples from this distribution cannot be obtained, because samples from real world are potentially unfair.
Therefore, in fairness-aware classification, we have to find a fair model that approximates a true distribution instead of a fair true distribution under the fairness constraints.

Privacy-Preserving Data Mining

Fairness in Data Mining
the independence between an objective Y and a sensitive feature S

↓

from the information theoretic perspective,
mutual information between Y and S is zero

↓

from the viewpoint of privacy-preservation,
protection of sensitive information when an objective variable is exposed

Different points from PPDM

- introducing randomness is occasionally inappropriate for severe decisions, such as job application
- disclosure of identity isn't problematic in FADM, generally

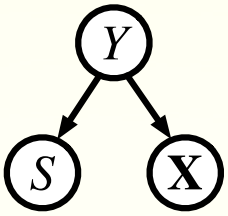
We here point out the connection with PPDM.
The fairness in data mining refers the independence between Y and S.
From information theoretic perspective, this means that mutual information between Y and S is zero.
From the viewpoint of privacy-preservation, this is interpreted as the protection of sensitive information when an objective variable is exposed.
However, there are some different points from PPDM.
introducing randomness is occasionally inappropriate for severe decisions. For example, if my job application is rejected at random, I will complain the decision and immediately consult with lawyers.
Further, disclosure of identity isn't problematic in FADM, generally.

Calders-Verwer's 2-Naive-Bayes

[Calders+ 10]

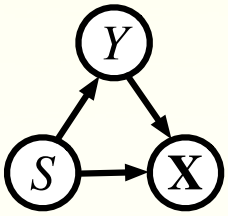
Unfair decisions are modeled by introducing the dependence of $\mathbf{X}$ on S as well as on Y

Naive Bayes



- S and $\mathbf{X}$ are conditionally independent given Y

Calders-Verwer Two Naive Bayes (CV2NB)



- non-sensitive features in $\mathbf{X}$ are conditionally independent given Y and S

✤ It is as if two naive Bayes classifiers are learned depending on each value of the sensitive feature; that is why this method was named by the 2-naive-Bayes

We compared our method shown in later with Calders-Verwer's two-naïve-Bayes method.
Unfair decisions are modeled by introducing the dependence of X on S as well as on Y.
As a result, non-sensitive features in X are conditionally independent given Y and S.

Calders-Verwer's 2-Naive-Bayes

[Calders+ 10]

parameters are initialized by the corresponding sample distributions

$$\hat{\Pr}[Y, X, S] = \hat{\Pr}[Y, S] \prod_i \hat{\Pr}[X_i | Y, S]$$

$\hat{\Pr}[Y, S]$ is modified so as to improve the fairness

estimated model $\hat{\Pr}[Y, S]$ **fair** → fair estimated model $\hat{\Pr}^{\dagger}[Y, S]$

keep the updated marginal distribution close to the $\hat{\Pr}[Y]$

```
while Pr[Y=1 | S=1] - Pr[Y=1 | S=0] > 0
    if # of data classified as "1" < # of "1" samples in original data then
        increase Pr[Y=1, S=0], decrease Pr[Y=0, S=0]
    else
        increase Pr[Y=0, S=1], decrease Pr[Y=1, S=1]
    reclassify samples using updated model Pr[Y, S]
```

update the joint distribution so that its fairness is enhanced

After parameters are initialized by the corresponding sample distributions, joint distribution of Y and S is modified by this algorithm.
This algorithm updates the joint distribution so that its fairness increases while keeping the updated marginal distribution close to the distribution of Y.

Outline

Applications of Fairness-aware Data Mining

- prevention of unfairness, information-neutral recommendation, ignoring uninteresting information

Fairness-aware Classification

- basic notations, fairness in data mining, fairness-aware classification, connection with PPDM, Calders & Verwer's 2-naive-Bayes

Hypothetical Fair-factorization Naive Bayes

- hypothetical fair-factorization naive Bayes (HFFNB), Connection with other methods, experimental results

Why Did the HFFNB Method Fail?

- model bias, deterministic decision rule

How to Modify the HFFNB Method

- actual fair-factorization method, experimental results

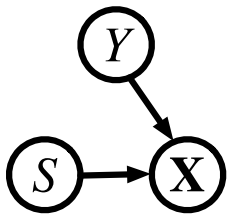
Conclusion

We propose a simple alternative method for fairness-aware classification.

Hypothetical Fair-factorization

Hypothetical Fair-factorization

A modeling technique to make a classifier fair



- In a classification model, a sensitive feature and a target variable are decoupled
- A graphical model of a fair-factorized classifier
- **By this technique, a sensitive feature and a target variable become statistically independent**

Hypothetical Fair-factorization naive Bayes (HFFNB)

A fair-factorization technique is applied to a naive Bayes model

$$\hat{\Pr}^{\dagger}[Y, \mathbf{X}, S] = \hat{\Pr}^{\dagger}[Y]\hat{\Pr}^{\dagger}[S]\prod_k \hat{\Pr}^{\dagger}[X^{(k)}|Y, S]$$

Under the ML or MAP principle, model parameters can be derived by simply counting training examples

Hypothetical fair-factorization is a modeling technique to make a classifier fair.
In a classification model, a sensitive feature and a target variable are decoupled.
By this technique, a sensitive feature and a target variable become statistically independent.
HFFNB is obtained by applying a fair-factorization technique to a naive Bayes model.

Connection with the ROC Decision Rule

[Kamiran+ 12]

In a non-fairized case, a new object $(\mathbf{x}, s)$ is classified into class 1, if

$$\hat{\Pr}[Y{=}1|\mathbf{x}, s] \geq 1/2 \equiv p$$

↓

Kamiran et al.'s ROC decision rule

A fair classifier is built by changing the decision boundary, p, according as the value of sensitive feature

↓

The HFFNB method is equivalent to changing decision boundary, p, to

$$p' = \frac{\hat{\Pr}[Y|S](1 - \hat{\Pr}[Y])}{\hat{\Pr}[Y] + \hat{\Pr}[Y|S] - 2\hat{\Pr}[Y]\hat{\Pr}[Y|S]}$$

(Elkan's theorem regarding cost-sensitive learning)

The HFFNB can be considered as a special case of the ROC method

We note the connection of the HFFNB method with the ROC decision rule.
In a non-fairized case, a new object is classified into class 1, if this conditional probability is larger then one half.
In Kamiran's ROC decision rule, a fair classifier is built by changing the decision boundary, p, according as the value of sensitive feature.
The HFFNB method is equivalent to changing decision boundary to this; hence, the HFFNB can be considered as a special case of the ROC method.

CV2NB vs HFFNB

The CV2NB and HFFNB methods are compared
in their accuracy and fairness

Accuracy
The larger value indicates
more accurate prediction

Unfairness
(Normalized Prejudice Index)
The larger value indicates
unfairer prediction

	Accuracy	Unfairness
HFFNB	0.828	1.52×10^{-2}
CV2NB	0.828	6.89×10^{-6}

The HFFNB method is **equally accurate** as the CV2NB method,
but it made **much unfairer prediction**

WHY?

The CV2NB and HFFNB methods are compared in their accuracy and fairness.
The HFFNB method is equally accurate as the CV2NB method, but it made much unfairer prediction.

Outline

Applications of Fairness-aware Data Mining

- prevention of unfairness, information-neutral recommendation, ignoring uninteresting information

Fairness-aware Classification

- basic notations, fairness in data mining, fairness-aware classification, connection with PPDM, Calders & Verwer's 2-naive-Bayes

Hypothetical Fair-factorization Naive Bayes

- hypothetical fair-factorization naive Bayes (HFFNB), Connection with other methods, experimental results

Why Did the HFFNB Method Fail?

- model bias, deterministic decision rule

How to Modify the HFFNB Method

- actual fair-factorization method, experimental results

Conclusion

Hereafter, we analyze why did the HFFNB method fail?

Why Did the HFFNB Method Fail?

HFFNB method

explicitly imposed the fairness constraint to the model

CV2NB method

The fairness of the model is enhanced by the post processing

Though both models are designed so as to enhance the fairness, the CV2NB method constantly learns much fairer model

Two reasons why these modeled independences are damaged

- **Model Bias:** the difference between model and true distributions
- **Deterministic Decision Rule:** class labels are not probabilistically generated, but are deterministically chosen by the decision rule

Though both models are designed so as to enhance the fairness, the CV2NB method constantly learns much fairer model.
We hypothesize two reasons why the modeled independences are damaged.
The one is a model bias, which widens the difference between model and true distributions.
The other is a deterministic decision rule; class labels are not probabilistically generated, but are deterministically chosen by the decision rule.

Model Bias

In a hypothetically Fair-factorized model, data are assumed to be generated according to the estimated distribution

$$\hat{\Pr}[Y]\hat{\Pr}[S]\hat{\Pr}[\mathbf{X}|Y,S]$$

Input objects are firstly generated from a true distribution, then the object is labeled according to the estimated distribution

estimated $\hat{\Pr}[Y|\mathbf{X},S]\,\Pr[\mathbf{X},S]$ **true**

These two distributions are diverged, especially if model bias is high

We first discuss a model bias.
In a hypothetically Fair-factorized model, data are assumed to be generated according to the estimated distribution.
However, actually, input objects are firstly generated from a true distribution, then the object is labeled according to the estimated distribution.
These two distributions are diverged, especially if model bias is high.

Model Bias

Synthetic Data

- Data are truly generated from a naive Bayes model
- Model bias is controlled by the number of features

Changes of the NPI (fairness)

HFFNB	1.02×10^{-1}	1.10×10^{-1}	1.28×10^{-1}
CV2NB	5.68×10^{-4}	9.60×10^{-2}	1.28×10^{-1}

high bias ⟷ **low bias**

As the decrease of model biases,
the differences between two methods in their fairness decreases

↓

The divergence between estimated and true distributions due to a model bias damages the fairness in classification

We tested on synthetic data.
As the decrease of model biases, the differences between two methods in their fairness decreases.
From this result, it can be concluded that the divergence between estimated and true distributions due to a model bias damages the fairness in classification.

Deterministic Decision Rule

In a hypothetically Fair-factorized model, labels are assumed to be generated probabilistically according to the distribution:

$$\hat{\Pr}[Y|\mathbf{X}, S]$$

Predicted labels are generated by this deterministic decision rule:

$$y^* = \arg \max_{y \in \mathrm{Dom}(Y)} \hat{\Pr}[Y|\mathbf{X}, S]$$

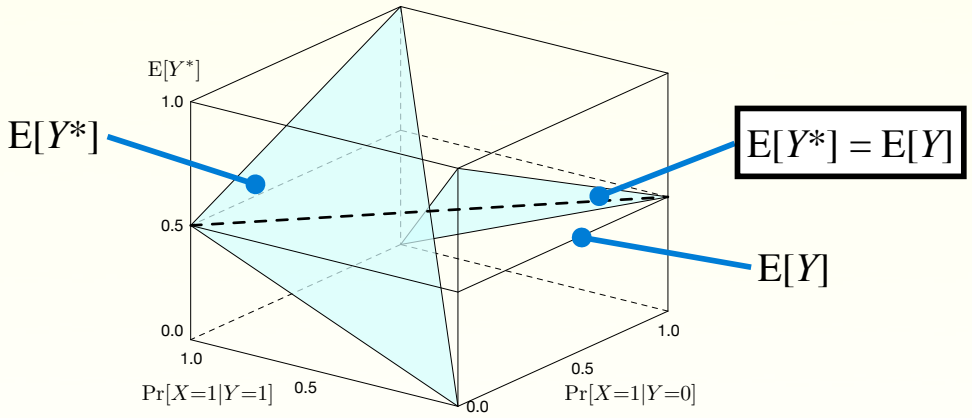
Labels generated by these two processes do not agree generally

We move on to a deterministic decision rule.
In a hypothetically Fair-factorized model, labels are assumed to be generated probabilistically according to the distribution.
However, actually, predicted labels are generated by this deterministic decision rule.
Labels generated by these two processes do not agree generally

Deterministic Decision Rule

simple classification model: one binary label and one binary feature

- class distribution is uniform: $\hat{\Pr}[Y=1] = 0.5$
- Y^* is deterministically determined: $Y^* = \arg\max \Pr[Y|X]$
- changing parameters: $\Pr[X=1|Y=1]$ and $\Pr[X=1|Y=0]$



E[Y^*] and E[Y] do not agree generally

We analyze a simple classification model, like this.
In this Figure, the expectations of class variable and the expectations of deterministically decided class agree with only on this line.
These two expectations do not agree generally.

Outline

Applications of Fairness-aware Data Mining

- prevention of unfairness, information-neutral recommendation, ignoring uninteresting information

Fairness-aware Classification

- basic notations, fairness in data mining, fairness-aware classification, connection with PPDM, Calders & Verwer's 2-naive-Bayes

Hypothetical Fair-factorization Naive Bayes

- hypothetical fair-factorization naive Bayes (HFFNB), Connection with other methods, experimental results

Why Did the HFFNB Method Fail?

- model bias, deterministic decision rule

How to Modify the HFFNB Method

- actual fair-factorization method, experimental results

Conclusion

Based on our findings, we finally discuss how to modify the HFFNB method.

Actual Fair-factorization

The reason why the HFFNB failed is the ignorance of the influence of a model bias and a deterministic decision rule

Actual Fair-factorization

A class and a sensitive features are decoupled not over the estimated distribution, but over the actual distribution

- As the hypothetical fair-factorization, a class label and a sensitive feature are made statistically independent
- We consider not the estimated distribution, $\hat{\Pr}[Y, \mathbf{X}, S]$, but the actual distribution, $\hat{\Pr}[Y|\mathbf{X}, S]\,\Pr[\mathbf{X}, S]$
- As a class label, we adopt a deterministically decided labels

The reason why the HFFNB failed is the ignorance of the influence of a model bias and a deterministic decision rule.
Therefore, a class and a sensitive features are decoupled not over the estimated distribution, but over the actual distribution.
We call this modified technique an actual fair-factorization.

Actual Fair-factorization naive Bayes

Actual Fair-factorization naive Bayes (AFFNB)

An actual fair-factorization technique is applied to a naive Bayes model

model bias

The multiplication of a true distribution, $\hat{\Pr}[Y|\mathbf{X}, S]\Pr[\mathbf{X}, S]$, is approximated by a sample mean, $(1/|\mathcal{D}|)\sum_{(\mathbf{x},s)\in\mathcal{D}}\hat{\Pr}[Y|\mathbf{X}{=}\mathbf{x}, S{=}s]$

deterministic decision rule

Instead of using a distribution of class labels, we count up the number of deterministically decided class labels

Y^* and S are made independent

under the constraint that the marginal distribution of Y* and S equal to the corresponding sample distribution

By applying an actual fair-factorization technique to a naive Bayes model, actual fair-factorization naive Bayes method is obtained.
To fixing the influence of a model bias, the multiplication of a true distribution is approximated by a sample mean.
To fixing the influence of a deterministic decision rule, Instead of using a distribution of class labels, we count up the number of deterministically decided class labels.
A deterministic class and a sensitive feature are made independent under the constraint that the marginal distribution of a deterministic class and a sensitive feature equal to the corresponding sample distribution

CV2NB and AFFNB

The CV2NB and AFFNB methods are compared in their accuracy and fairness

	Accuracy	Unfairness
AFFNB	0.828	5.43×10^{-6}
CV2NB	0.828	6.89×10^{-6}

CV2NB and AFFNB are equally accurate as well as equally fair

The superiority of the CV2NB method is considering the independence not over the estimated distribution, but over the actual distribution of a class label and a sensitive feature

Finally, the CV2NB and AFFNB methods are compared in their accuracy and fairness.
The performance is drastically improved; The CV2NB and the AFFNB methods are equally accurate as well as equally fair.
From this result, it can be concluded that the superiority of the CV2NB method is considering the independence not over the estimated distribution, but over the actual distribution of a class label and a sensitive feature.

Conclusion

Contributions

- After reviewing a fairness-aware classification task, we focus on why the CV2NB method can attain fairer results than other methods
- We theoretically and empirically show the reason by comparing a simple alternative naive-Bayes modified by a hypothetical fair-factorization technique
- Based on our findings, we developed a modified version, an actual fair-factorization technique, and show that this technique drastically improved the performance

Future Work

- We plan to apply our actual fair-factorization technique in order to modify other classification methods, such as logistic regression or a support vector machine

Our contributions are as follows.
We plan to apply our actual fair-factorization technique in order to modify other classification methods, such as logistic regression or SVM.

Program Codes and Data Sets

Fairness-aware Data Mining

http://www.kamishima.net/fadm

Information-neutral Recommender System

http://www.kamishima.net/inrs

Acknowledgements

- This work is supported by MEXT/JSPS KAKENHI Grant Number 16700157, 21500154, 23240043, 24500194, and 25540094.

Program codes and data sets are available at these sites.
That's all I have to say. Thank you for your attention.