Fairness-aware Classifier with Prejudice Remover Regularizer

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

ECMLPKDD 2012 (22nd ECML and 15th PKDD Conference)

@ Bristol, United Kingdom, Sep. 24-28, 2012

START

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

Overview

Fairness-aware Data Mining

data analysis taking into account potential issues of fairness, discrimination, neutrality, or independence

- fairness-aware classification, regression, or clustering
- detection of unfair events from databases
- fairness-aware data publication

Examples of Fairness-aware Data Mining Applications

- exclude the influence of sensitive information, such as gender or race, from decisions or predictions
- recommendations enhancing their neutrality
- data analysis while excluding the influence of uninteresting information

Fairness-aware data mining is a data analysis taking into account potential issues of fairness, discrimination, neutrality, or independence.
There are several kinds of tasks: fairness-aware classification, detection of unfair events, and fairness-aware data publication.
These are examples of fairness-aware data mining applications.

Outline

Applications
- fairness-aware data mining applications

Difficulty in Fairness-aware Data Mining
- Calders-Verwer's discrimination score, red-lining effect

Fairness-aware Classification
- fairness-aware classification, three types of prejudices

Methods
- prejudice remover regularizer, Calders-Verwer's 2-naïve-Bayes

Experiments
- experimental results on Calders&Verwer's data and synthetic data

Related Work
- privacy-preserving data mining, detection of unfair decisions, explainability, fairness-aware data publication

Conclusion

This is an outline of our talk.
We first give examples of fairness-aware data mining applications.
After showing difficulty in fairness-aware data mining, we define a fairness-aware classification task.
We then propose our method for this task, and empirically compare it with an existing method.
Finally, we review this new research area, and conclude our talk.

Outline

Applications
- applications of fairness-aware data mining

Difficulty in Fairness-aware Data Mining
- Calders-Verwer's discrimination score, red-lining effect

Fairness-aware Classification
- fairness-aware classification, three types of prejudices

Methods
- prejudice remover regularizer, Calders-Verwer's 2-naïve-Bayes

Experiments
- experimental results on Calders&Verwer's data and synthetic data

Related Work
- privacy-preserving data mining, detection of unfair decisions, explainability, situation testing, fairness-aware data publication

Conclusion

Let me start by giving examples of applications.

Elimination of Discrimination

Due to the spread of data mining technologies...

- **Data mining is being increasingly applied for serious decisions**
  ex. credit scoring, insurance rating, employment application
- **Accumulation of massive data enables to reveal personal information**
  ex. demographics can be inferred from the behavior on the Web

excluding the influence of these socially sensitive information from serious decisions

- **information considered from the viewpoint of social fairness**
  ex. gender, religion, race, ethnicity, handicap, political conviction
- **information restricted by law or contracts**
  ex. insider information, customers' private information

The first application is the elimination of discrimination.
Due to the spread of data mining technologies, data mining is being increasingly applied for serious decisions. For example, credit scoring, insurance rating, employment application, and so on.
Additionally, accumulation of massive data enables to reveal personal information.
To cope with these situation, fairness-aware data mining techniques are used for excluding the influence of socially sensitive information from serious decisions: information considered from the viewpoint of social fairness, and information restricted by law or contracts.

Filter Bubble

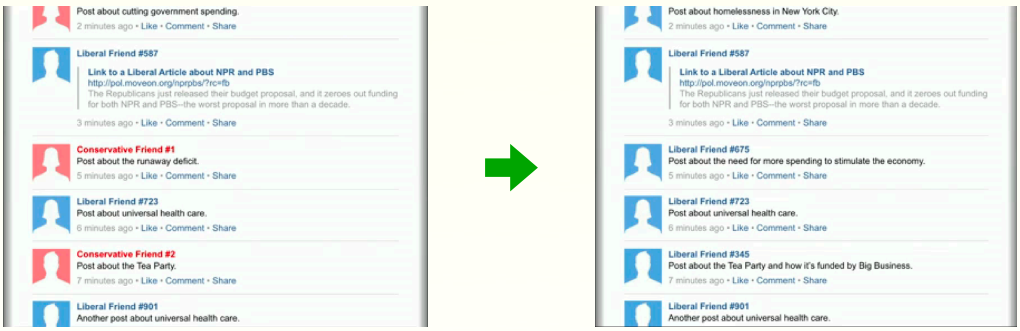
[TED Talk by Eli Pariser]

The Filter Bubble Problem

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

http://www.thefilterbubble.com/

Friend Recommendation List in Facebook



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

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 Facebook.
To fit for his preference, conservative people are eliminated form his recommendation list, while this fact is not notified to him.

Information Neutral Recommender System

[Kamishima+ 12]

The Filter Bubble Problem

↑

Information Neutral Recommender System

enhancing the neutrality from a viewpoint specified by a user and other viewpoints are not considered

↑

fairness-aware data mining techniques

ex. A system enhances the neutrality in terms of whether conservative or progressive, but it is allowed to make biased recommendations in terms of other viewpoints, for example, the birthplace or age of friends

To cope with this filter bubble problem, an information neutral recommender system enhances the neutrality from a viewpoint specified by a user and other viewpoints are not considered.
For this purpose, fairness-aware data mining techniques are applied.
In the case of Pariser's Facebook example, a system enhances the neutrality in terms of whether conservative or progressive, but it is allowed to make biased recommendations in terms of other viewpoints, for example, the birthplace or age of friends.

Non-Redundant Clustering

[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
- A non-redundant clustering method derives more useful male and female clusters, which are independent of the above clusters

The third application is excluding uninteresting information.

Outline

Applications
- applications of fairness-aware data mining

Difficulty in Fairness-aware Data Mining
- Calders-Verwer's discrimination score, red-lining effect

Fairness-aware Classification
- fairness-aware classification, three types of prejudices

Methods
- prejudice remover regularizer, Calders-Verwer's 2-naïve-Bayes

Experiments
- experimental results on Calders&Verwer's data and synthetic data

Related Work
- privacy-preserving data mining, detection of unfair decisions, explainability, situation testing, fairness-aware data publication

Conclusion

We then show the difficulty in fairness-aware data mining.

Difficulty in Fairness-aware Data Mining

[Calders+ 10]

US Census Data : predict whether their income is high or low

	Male	Female
High-Income	3,256	590
Low-income	7,604	4,831

fewer

Females are minority in the high-income class

- # of High-Male data is 5.5 times # of High-Female data
- While 30% of Male data are High income, only 11% of Females are

Occam's Razor : Mining techniques prefer simple hypothesis

↓

Minor patterns are frequently ignored
and thus minorities tend to be treated unfairly

This is Caldars & Verwer's example to show why fairness-aware data mining is needed.
This is a task to predict whether their income is high or low.
In this US census data, females are minority in the high-income class.
Because mining techniques prefer simple hypothesis, minor patterns are frequently ignored; and thus minorities tend to be treated unfairly.

Red-Lining Effect

[Calders+ 10]

Calders-Verwer discrimination score (CV score)

Pr[High-income | Male] - Pr[High-income | Female]

the difference between conditional probabilities of advantageous decisions for non-protected and protected members. The larger score indicates the unfairer decision.

- US Census Data samples
  - ➡ The baseline CV score is 0.19
- Incomes are predicted by a naïve-Bayes classifier trained from data containing all sensitive and non-sensitive features
  - ➡ The CV score increases to 0.34, indicating unfair treatments
- Even if a feature, gender, is excluded in the training of a classifier
  - ➡ improved to 0.28, but still being unfairer than its baseline

red-lining effect : Ignoring sensitive features is ineffective against the exclusion of their indirect influence

Caldars & Verwer proposed a score to quantify the degree of unfairness.
This is defined as the difference between conditional probabilities of High-income decisions for males and females. The larger score indicates the unfairer decision.
The baseline CV score for the US Census Data Set is 0.19.
If objective variables are predicted by a naïve Bayes classifier, the CV score increases to 0.34, indicating unfair treatments.
Even if a feature, gender, is excluded, the CV score is improved to 0.28, but still being unfairer than its baseline.
Consequently, ignoring sensitive features is ineffective against the exclusion of their indirect influence. This red-lining effect makes fairness-aware data mining difficult.

Outline

Applications
- applications of fairness-aware data mining

Difficulty in Fairness-aware Data Mining
- Calders-Verwer's discrimination score, red-lining effect

Fairness-aware Classification
- fairness-aware classification, three types of prejudices

Methods
- prejudice remover regularizer, Calders-Verwer's 2-naïve-Bayes

Experiments
- experimental results on Calders&Verwer's data and synthetic data

Related Work
- privacy-preserving data mining, detection of unfair decisions, explainability, situation testing, fairness-aware data publication

Conclusion

We next define a task of fairness-aware classification

Variables

Y

objective variable

- a binary class variable {0, 1}
- **a result of serious decision**
  ex., whether or not to allow credit

S

sensitive feature

- a binary feature variable {0, 1}
- **socially sensitive information**
  ex., gender or religion

$\mathbf{X}$

non-sensitive features

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

We introduce three types of variables.
An objective variable Y represents a result of serious decision.
A sensitive feature S represents socially sensitive information.
The other variables are non-sensitive features X.

Prejudice

Prejudice : the statistical dependences of an objective variable or non-sensitive features on a sensitive feature

Direct Prejudice $Y \not\perp\!\!\!\perp S \mid \mathbf{X}$

- a clearly unfair state that a prediction model directly depends on a sensitive feature
- implying the conditional dependence between Y and S given $\mathbf{X}$

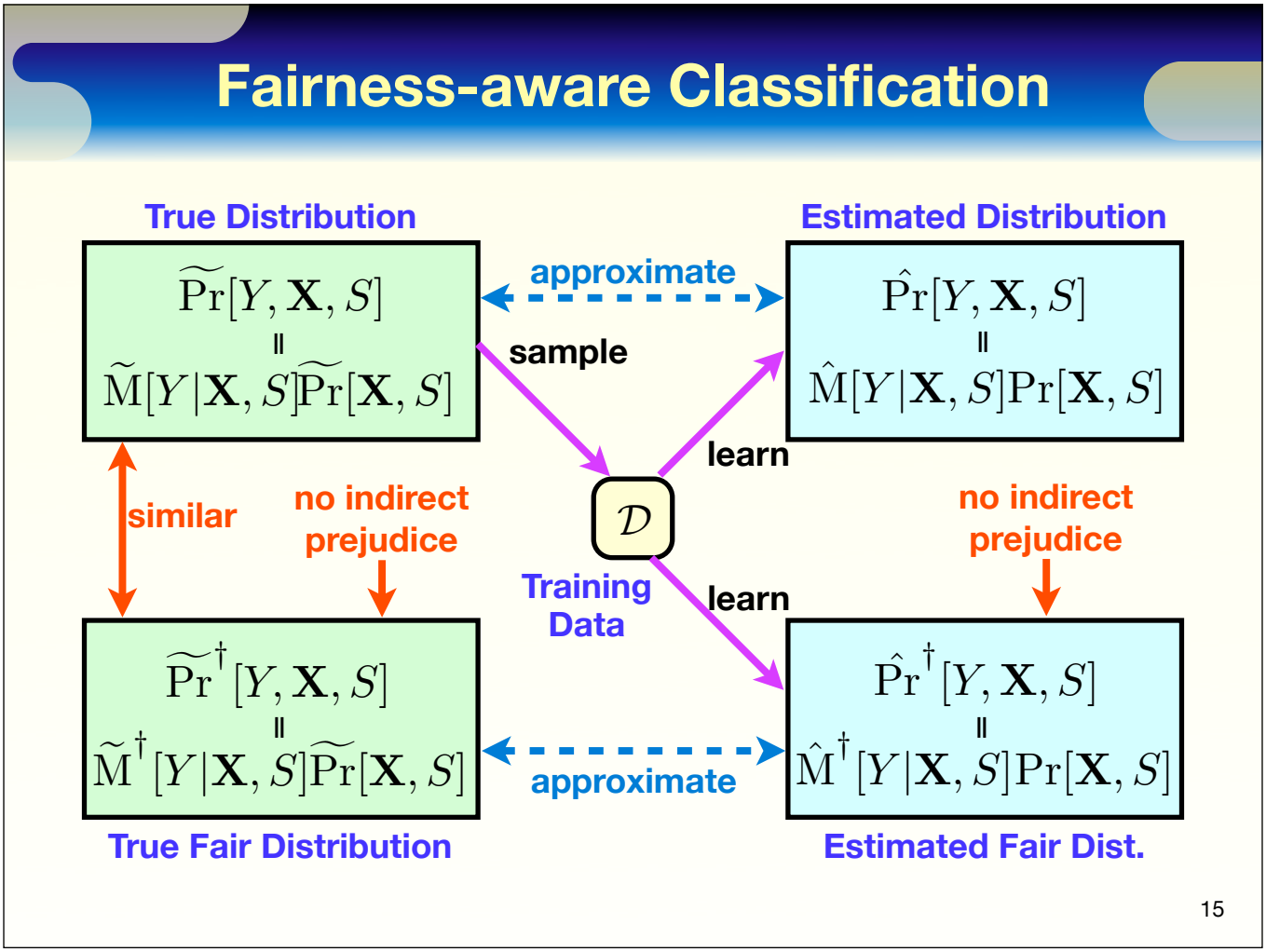
Indirect Prejudice $Y \not\perp\!\!\!\perp S \mid \phi$

- statistical dependence of an objective variable on a sensitive feature
- bringing red-lining effect

Latent Prejudice $\mathbf{X} \not\perp\!\!\!\perp S \mid \phi$

- statistical dependence of non-sensitive features on a sensitive feature
- completely excluding sensitive information

A prejudice is one of causes of unfairness, which is defined as the statistical dependences of an objective variable or non-sensitive features on a sensitive feature.
Prejudices is classified into three types:
Direct prejudice is a clearly unfair state that a prediction model directly depends on a sensitive feature.
Indirect prejudice is the statistical dependence of an objective variable on a sensitive feature.
Latent prejudice is the statistical dependence of non-sensitive features on a sensitive feature.
We here focus on removing this indirect prejudice.



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 this task is to learn a model approximating the true distribution.
In a case of a fairness-aware classification task, we assume a true fair distribution that satisfies two constraints.
This distribution should be similar to the true distribution, and should satisfy no indirect prejudice condition.
A goal of this task is to learn a model approximating this true fair distribution from training data sampled from the true distribution.

Outline

Applications
- applications of fairness-aware data mining

Difficulty in Fairness-aware Data Mining
- Calders-Verwer's discrimination score, red-lining effect

Fairness-aware Classification
- fairness-aware classification, three types of prejudices

Methods
- prejudice remover regularizer, Calders-Verwer's 2-naïve-Bayes

Experiments
- experimental results on Calders&Verwer's data and synthetic data

Related Work
- privacy-preserving data mining, detection of unfair decisions, explainability, situation testing, fairness-aware data publication

Conclusion

We show our logistic regression with prejudice remover regularizer and Calders-Verwer's 2-naïve-Bayes method.

Logistic Regression with Prejudice Remover Regularizer

modifications of a logistic regression model

- add ability to adjust distribution of Y depending on the value of S
- add a constraint of a no-indirect-prejudice condition

add ability to adjust distribution of Y depending on the value of S

$$\Pr[Y=1|\mathbf{x}, S=0] = \mathrm{sigmoid}(\mathbf{x}^\top \mathbf{w}_0)$$
$$\Pr[Y=1|\mathbf{x}, S=1] = \mathrm{sigmoid}(\mathbf{x}^\top \mathbf{w}_1)$$

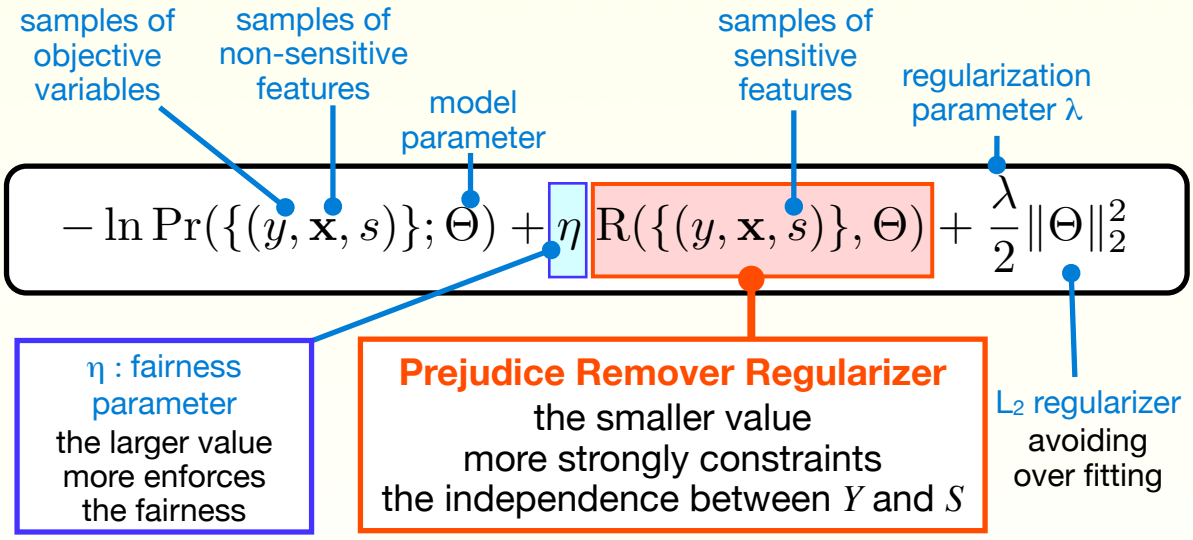
weights depending on the value of S

- Multiple logistic regression models are built separately, and each of these models corresponds to each value of a sensitive feature
- When predicting scores, a model is selected according to the value of a sensitive feature

We propose logistic regression with prejudice remover regularizer.
We change an original logistic regression model in these two points.
First, we add ability to adjust distribution of Y depending on the value of S.
For this purpose, multiple logistic regression models are built separately, and each of these models corresponds to each value of a sensitive feature.

No Indirect Prejudice Conditon

add a constraint of a no-indirect-prejudice condition



Second, we add a constraint of a no-indirect-prejudice condition, which is a main idea of this presentation.
For this purpose, we add this term to an original objective function.
This prejudice remover regularizer is designed so that the smaller value more strongly constraints the independence between Y and S.
Eta is a fairness parameter. The lager value more enforces the fairness.

Prejudice Remover Regularizer

no-indirect-prejudice condition = independence between Y and S

Prejudice Remover Regularizer

mutual information between Y (objective variable) and S (sensitive feature)

$$\sum_{Y\in\{0,1\}} \sum_{\mathbf{X},S} \Pr[\mathbf{X},S]\, \hat{\mathrm{M}}[Y|\mathbf{X},S] \ln \frac{\hat{\Pr}[Y,S]}{\hat{\Pr}[S]\hat{\Pr}[Y]}$$

$$\sum_{Y\in\{0,1\}} \sum_{(\mathbf{x},s)} \hat{\mathrm{M}}[y|\mathbf{x},s;\boldsymbol{\Theta}] \ln \frac{\hat{\Pr}[y|s;\boldsymbol{\Theta}]}{\sum_s \hat{\Pr}[y|s;\boldsymbol{\Theta}]}$$

expectation over $\mathbf{X}$ and S is replaced with the summation over samples

As this prejudice remover regularizer, we adopt mutual information between S and Y so as to enforce the independence between Y and S.
This term can be computed by replacing with the summation over samples.
But the computation of this term is rather complicated.

Computing Mutual Information

$$\sum_{Y\in\{0,1\}}\sum_{(\mathbf{x},s)}\hat{\mathrm{M}}[y|\mathbf{x},s;\boldsymbol{\Theta}]\ln\frac{\hat{\Pr}[y|s;\boldsymbol{\Theta}]}{\sum_s\hat{\Pr}[y|s;\boldsymbol{\Theta}]}$$

This distribution can be derived by marginalizing over $\mathbf{X}$

$$\int_{\mathrm{Dom}(\mathbf{x})}\hat{\mathrm{M}}[y|\mathbf{x},s;\Theta]\hat{\Pr}[\mathbf{x}]d\mathbf{x}$$

But this is computationally heavy...

approximate by the sample mean over X for each pair of y and s

$$\sum_{\mathbf{x}\in\mathcal{D}}\hat{\mathrm{M}}[y|\mathbf{x},s;\Theta]$$

Limitation: This technique is applicable only if both Y and S are discrete

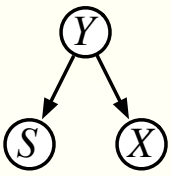
This distribution can be derived by marginalizing over X, but this is computationally heavy. We hence approximate by the sample mean over x for each pair of y and s. Unfortunately, this technique is applicable only if both Y and S are discrete.

Calders-Verwer's 2 Naïve Bayes

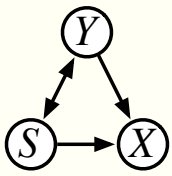
[Calders+ 10]

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

Naïve Bayes → **Calders-Verwer Two Naïve Bayes (CV2NB)**



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



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

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

We compared our method with Calders-Verwer's two-naïve-Bayes method.
Unfair decisions are modeled by introducing of the dependence of X on S as well as on Y.

Calders-Verwer's Two Naïve Bayes

[Calders+ 10]

parameters are initialized by the corresponding empirical distributions

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

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

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

keep the updated distribution similar to the empirical distribution

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

update the joint distribution so that its CV score decreases

After parameters are estimated by the corresponding empirical distributions, joint distribution of Y and S is modified by this algorithm.
This algorithm updates the joint distribution so that its CV score decreases while keeping the updated distribution similar to the empirical distribution.

Outline

Applications
- applications of fairness-aware data mining

Difficulty in Fairness-aware Data Mining
- Calders-Verwer's discrimination score, red-lining effect

Fairness-aware Classification
- fairness-aware classification, three types of prejudices

Methods
- prejudice remover regularizer, Calders-Verwer's 2-naïve-Bayes

Experiments
- experimental results on Calders&Verwer's data and synthetic data

Related Work
- privacy-preserving data mining, detection of unfair decisions, explainability, situation testing, fairness-aware data publication

Conclusion

We next show our experimental results.

Experimental Conditions

Calders & Verwer's Test Data

- *Adult* / *Census Income* @ UCI Repository
- Y : a class representing whether subject's income is *High* or *Low*
- S : a sensitive feature representing whether subject's gender
- $\mathbf{X}$: non-sensitive features, all features are discretized, and 1-of-K representation is used for logistic regression
- # of samples : 16281

Methods

- **LRns** : logistic regression without sensitive features
- **NBns** : naïve Bayes without sensitive features
- **PR** : our logistic regression with prejudice remover regularizer
- **CV2NB** : Calders-Verwer's two-naïve-Bayes

Other Conditions

- L_2 regularization parameter $\lambda = 1$
- five-fold cross validation

We first test these four methods on Calders and Vewer's Test Data.
LRns and NBs are respectively pure logistic regression and naïve Bayes without sensitive features.
PR is our logistic regression with prejudice remover regularizer.
CV2NB is Calders and Verwer's two-naïve-Bayes method.

Evaluation Measure

Accuracy

- **How correct are predicted classes?**
- the ratio of correctly classified sample

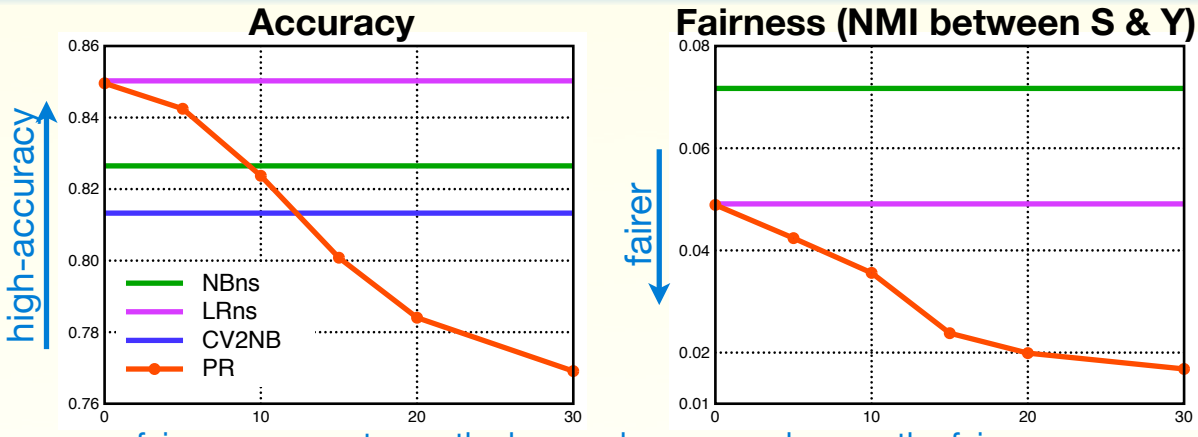
NMI (normalized mutual information)

- **How fair are predicted classes?**
- mutual information between a predicted class and a sensitive feature, and it is normalized into the range [0, 1]

$$\mathrm{NMI} = \frac{\mathrm{I}(Y;S)}{\sqrt{\mathrm{H}(Y)\mathrm{H}(S)}}$$

We use two types of evaluation measures.
Accuracy measures how correct predicted classes are.
Normalized mutual information measures how fair predicted classes are.

Experimental Results



fairness parameter η : the lager value more enhances the fairness

fairness parameter η → accuracy ↓ NMI ↑

- Our method **PR** (Prejudice Remover) could make fairer decisions than pure **LRns** (logistic regression) and **NBns** (naïve Bayes)
- **PR** could make more accurate prediction than **NBns** or **CV2NB**

CV2NB achieved near-zero NMI, but PR could NOT achieve it

These are experimental results.
X-axes correspond to fairness parameters, the lager value more enhances the fairness.
This chart (left) shows the change of prediction accuracy.
This chart (right) shows the change of the degree of fairness.
As the increase of a fairness parameter η, prediction accuracy is worsened and the fairness is enforced.
Our metod could make fairer decisions than pure logistic regression and naïve Bayes.
Additionally, our method could make more accurate prediction than naïve Bayes and CV2NB.
Unfortunately, CV2NB achieved near-zero NMI, but our method could not achieve it.

Synthetic Data

Why did our prejudice remover fail to
make a fairer prediction than that made by CV2NB?

sensitive feature

$$s_i \in \{0, 1\} \sim \mathrm{DiscreteUniform}$$

non-sensitive feature

$$x_i = \epsilon_i$$

independent from s_i

$$w_i = \begin{cases} 1 + \epsilon_i, \text{if } s_i = 1 \\ -1 + \epsilon_i, \text{otherwise} \end{cases}$$

depend on s_i

$$\epsilon_i \sim \mathcal{N}(0, 1)$$

objective variable

$$y_i = \begin{cases} 1, \text{if } x_i + w_i < 0 \\ 0, \text{otherwise} \end{cases}$$

both x_i and w_i equally influence an objective variable

27

To investigate why our method failed to make a fairer prediction than that made by CV2NB, we tested our method on synthetic data.
Each sample composed of one sensitive feature, two non-sensitive features, and one objective variable.
Non-sensitive feature X is independent from S, but W depends on S.
And, both X and W equally influence an objective variable.

Analysis on Synthetic Data Results

	M[Y | **X**, S=0]			M[Y | **X**, S=0]		
	X	W	bias	X	W	bias
η=0	11.3	11.3	-0.0257	11.3	11.4	0.0595
η=150	55.3	-53.0	-53.6	56.1	54.1	53.6

Prejudice Remover

- When η=0, PR regularizer doesn't affect weights, and these weights for X and W are almost equal
- When η=150, absolutes of weights for X are larger than those for W

➡ PR ignores features depending on S if the influences to Y are equal

CV2NB CV2NB treats all features equally

Prejudice Remover can consider the differences among individual features, but CV2NB can improve fairness more drastically

A prejudice remover regularizer doesn't affect weights if η=0, but it heavily does if η=150.
When η=0, weights for x and w are almost equal.
When η=150, absolutes of weights for X are larger than those for W.
This indicates that our prejudice remover regularizer ignores features depending on s if the influences to y are equal.
On the other hand, CV2NB method treats all features equally.
Consequently, our prejudice remover can consider the differences among individual features, but Calders-Verwer's 2-naïve-Bayes can improve fairness more drastically.

Outline

Applications
- applications of fairness-aware data mining

Difficulty in Fairness-aware Data Mining
- Calders-Verwer's discrimination score, red-lining effect

Fairness-aware Classification
- fairness-aware classification, three types of prejudices

Methods
- prejudice remover regularizer, Calders-Verwer's 2-naïve-Bayes

Experiments
- experimental results on Calders&Verwer's data and synthetic data

Related Work
- privacy-preserving data mining, detection of unfair decisions, explainability, situation testing, fairness-aware data publication

Conclusion

We finally review this new research area.

Relation to Privacy-Preserving Data Mining

indirect prejudice
the dependency between a objective Y and a sensitive feature S

↓

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

↓

from the viewpoint of privacy-preservation,
leakage of sensitive information when an objective variable is known

differences from PPDM

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

Here, we'd like to point out the relation between an indirect prejudice and PPDM.
From information theoretic perspective, an indirect prejudice implies that mutual information between Y and S is non-zero.
From the viewpoint of privacy-preservation, this can be interpreted as the leakage of sensitive information when an objective variable is known.
On the other hand, there are some differences from PPDM like this.

Finding Unfair Association Rules

[Pedreschi+ 08, Ruggieri+ 10]

ex: association rules extracted from German Credit Data Set

(a) **city=NYC ⇒ class=bad** (conf=0.25)
0.25 of NY residents are denied their credit application

(b) **city=NYC & race=African ⇒ class=bad** (conf=0.75)
0.75 of NY residents whose race is African are denied their credit application

extended lift (elift)

$$\text{elift} = \frac{\text{conf}(A \wedge B \Rightarrow C)}{\text{conf}(A \Rightarrow C)}$$

the ratio of the confidence of a rule with additional condition
to the confidence of a base rule

***a*-protection :** considered as unfair if there exists association rules whose elift is larger than *a*
ex: (b) isn't *a*-protected if a = 2, because elift = conf(b) / conf(a) = 3

They proposed an algorithm to enumerate rules that are not *a*-protected

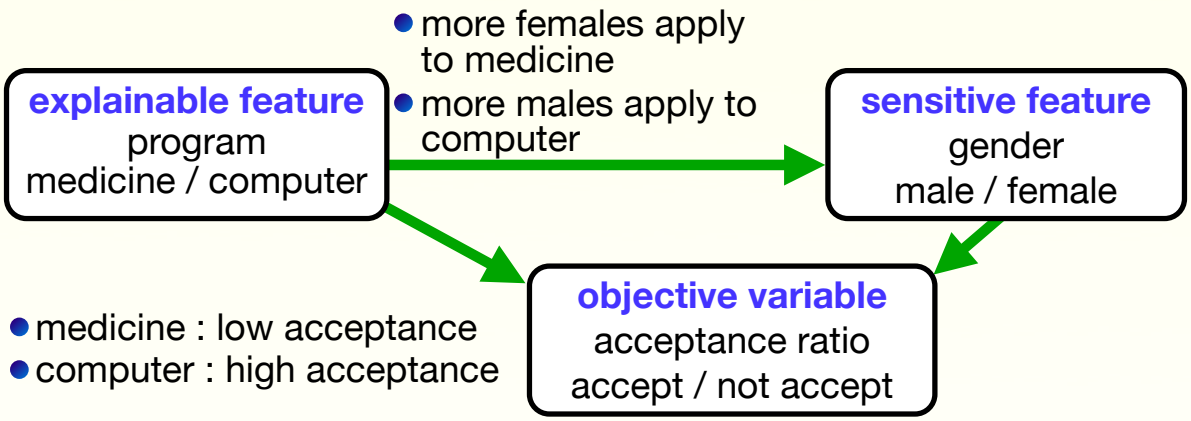
Pedreschi et al. firstly addressed the problem of discrimination in data mining.
Their algorithm enumerates unfair association rules.

Explainability

[Zliobaite+ 11]

conditional discrimination : there are non-discriminatory cases, even if distributions of an objective variable depends on a sensitive feature

ex : admission to a university



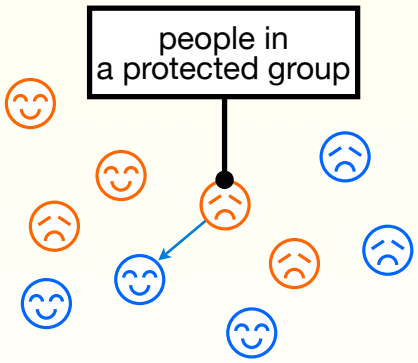
Because females tend to apply to a more competitive program, females are more frequently rejected
Such difference is explainable and is considered as non-discriminatory

Zliobaite et al. proposed a notion of explainability.

Situation Testing

[Luong+ 11]

Situation Testing : When all the conditions are same other than a sensitive condition, people in a protected group are considered as unfairly treated if they received unfavorable decision



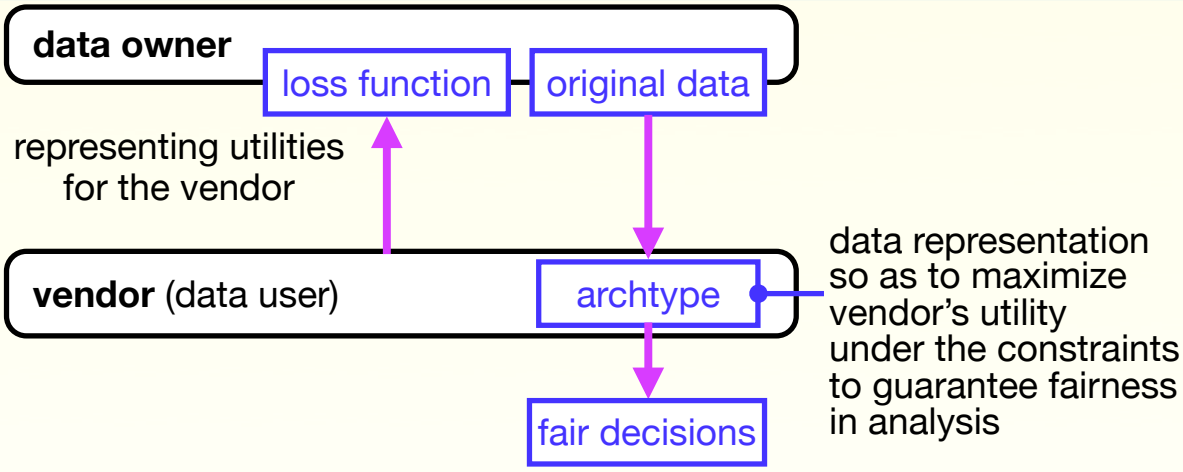
k-nearest neighbors

- They proposed a method for finding unfair treatments by checking the statistics of decisions in *k*-nearest neighbors of data points in a protected group
- Condition of situation testing is $Pr[Y \mid X, S=a] = Pr[Y \mid X, S=b] \; \forall X$ This implies the independence between S and Y

Luong et al. proposed another notion, situation testing.

Fairness-aware Data Publishing

[Dwork+ 11]



- They show the conditions that these archtypes should satisfy
- This condition implies that the probability of receiving favorable decision is irrelevant to belonging to a protected group

Dwork et al. proposed a framework for fairness-aware data publishing.

Conclusion

Contributions

- the unfairness in data mining is formalized based on independence
- a prejudice remover regularizer, which enforces a classifier's independence from sensitive information
- experimental results of logistic regressions with our prejudice remover

Future Work

- improve the computation of our prejudice remover regularizer
- fairness-aware classification method for generative models
- another type of fairness-aware mining task

Socially Responsible Mining

- Methods of data exploitation that do not damage people's lives, such as fairness-aware data mining, PPDM, or adversarial learning, together comprise the notion of **socially responsible mining**, which it should become an important concept in the near future.

Our contributions are as follows.
In future work, we have to improve computation of prejudice remover regularizer
Methods of data exploitation that do not damage people's lives, such as fairness-aware mining, PPDM, or adversarial learning, together comprise the notion of socially responsible mining, which it should become an important concept in the near future.

Program Codes and Data Sets

Fairness-Aware Data Mining

http://www.kamishima.net/fadm

Information Neutral Recommender System

http://www.kamishima.net/inrs

Acknowledgements

- We wish to thank Dr. Sicco Verwer for providing detail information about their work and program codes
- We thank the PADM2011 workshop organizers and anonymous reviewers for their valuable comments
- This work is supported by MEXT/JSPS KAKENHI Grant Number 16700157, 21500154, 22500142, 23240043, and 24500194, and JST PRESTO 09152492

Program codes and data sets are available at these URLs.
We finally wish to thank Dr. Sicco Verwer for providing detail information about their work and program codes.
That's all I have to say. Thank you for your attention.