Fairness-aware Learning through Regularization Approach

Toshihiro Kamishima†, Shotaro Akaho†, and Jun Sakuma‡
† National Institute of Advanced Industrial Science and Technology (AIST)
‡ University of Tsukuba & Japan Science and Technology Agency
http://www.kamishima.net/

IEEE Int'l Workshop on Privacy Aspects of Data Mining (PADM 2011)
Dec. 11, 2011 @ Vancouver, Canada, co-located with ICDM2011

START

I'm Toshihiro Kamishima, and this is joint work with Shotaro Akaho and Jun Sakuma.
Today, we would like to talk about fairness-aware data mining.

Introduction

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**

Fairness-aware / Discrimination-aware Mining

taking care of the following sensitive information in its decisions or results:

- **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

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.
Furthermore, accumulation of massive data enables to reveal personal information.
To cope with these situation, several researchers have begun to develop fairness-aware mining techniques, which take care the sensitive information in their decisions.
First, information considered from the viewpoint of social fairness. For example, gender, religion, race, ethnicity, handicap, political conviction, and so on.
Second, information restricted by law or contracts. For example, insider information, customers' private information.

Outline

Backgrounds

- an example to explain why fairness-aware mining is needed

Causes of Unfairness

- prejudice, underestimation, negative legacy

Methods and Experiments

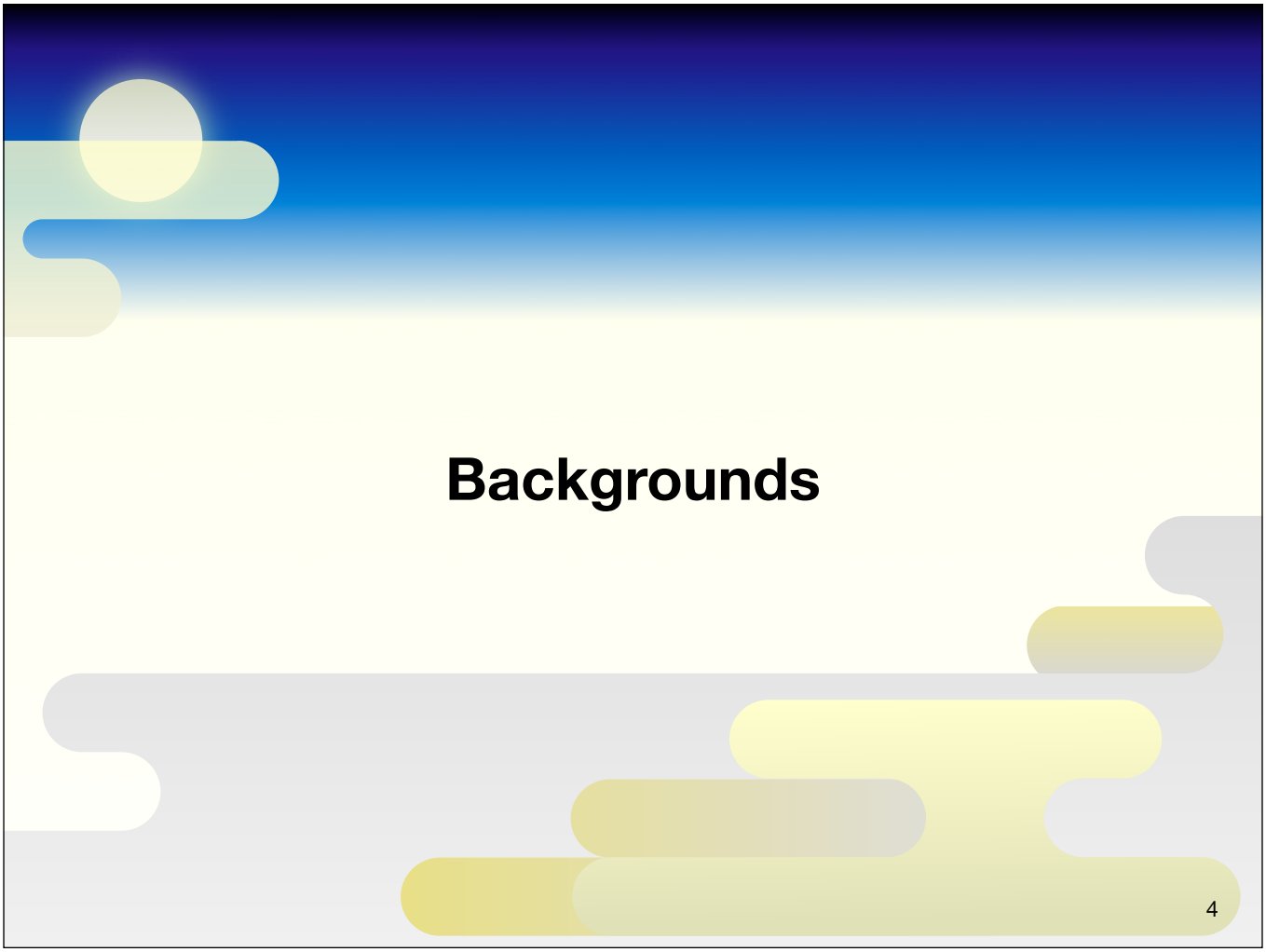
- our prejudice removal technique, experimental results

Related Work

- finding unfair association rules, situation testing, fairness-aware data publishing

Conclusion

This is an outline of our talk.
First, we show an example to explain why fairness-aware mining is needed.
Second, we discuss three causes of unfairness.
Third, we show our prejudice removal technique.
Finally, we review this new research area



Let’s turn to backgrounds of fairness-aware mining.

Why Fairness-aware 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

Occum'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 mining is needed.
Consider 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.

Calders-Verwer Discrimination Score

[Calders+ 10]

Calders-Verwer discrimination score (CV score)

Pr[Y=High-income | S=Male] − Pr[Y=High-income | S=Female]

Y: objective variable, S: sensitive feature

The conditional probability of the preferred decision
given a sensitive value subtracted that given a non-sensitive value

- As the values of Y, the values in sample data are used
  - ➡ The baseline CV score is 0.19
- Objective variables, Y, are predicted by a naive-Bayes classifier trained from data containing all sensitive and non-sensitive features
  - ➡ The CV score increases to 0.34, indicating unfair treatments
- Even if sensitive features are excluded in the training of a classifier
  - ➡ improved to 0.28, but still being unfairer than its baseline

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

Caldars & Verwer proposed a score to quantify the degree of unfairness.
This is defined as the difference between conditional probabilities of High-income cases given Male and given Female. The larger score indicates the unfairer decision.
The baseline CV score is 0.19.
If objective variables are predicted by a naive Bayes classifier, the CV score increases to 0.34, indicating unfair treatments.
Even if sensitive features are excluded, the score is improved to 0.28, but still being unfairer than its baseline.
Ignoring sensitive features is ineffective against the exclusion of their indirect influence. This is called by a red-lining effect. More affirmative actions are required.

Social Backgrounds

Equality Laws

- Many of international laws prohibit discrimination in socially-sensitive decision making tasks [Pedreschi+ 09]

Circulation of Private Information

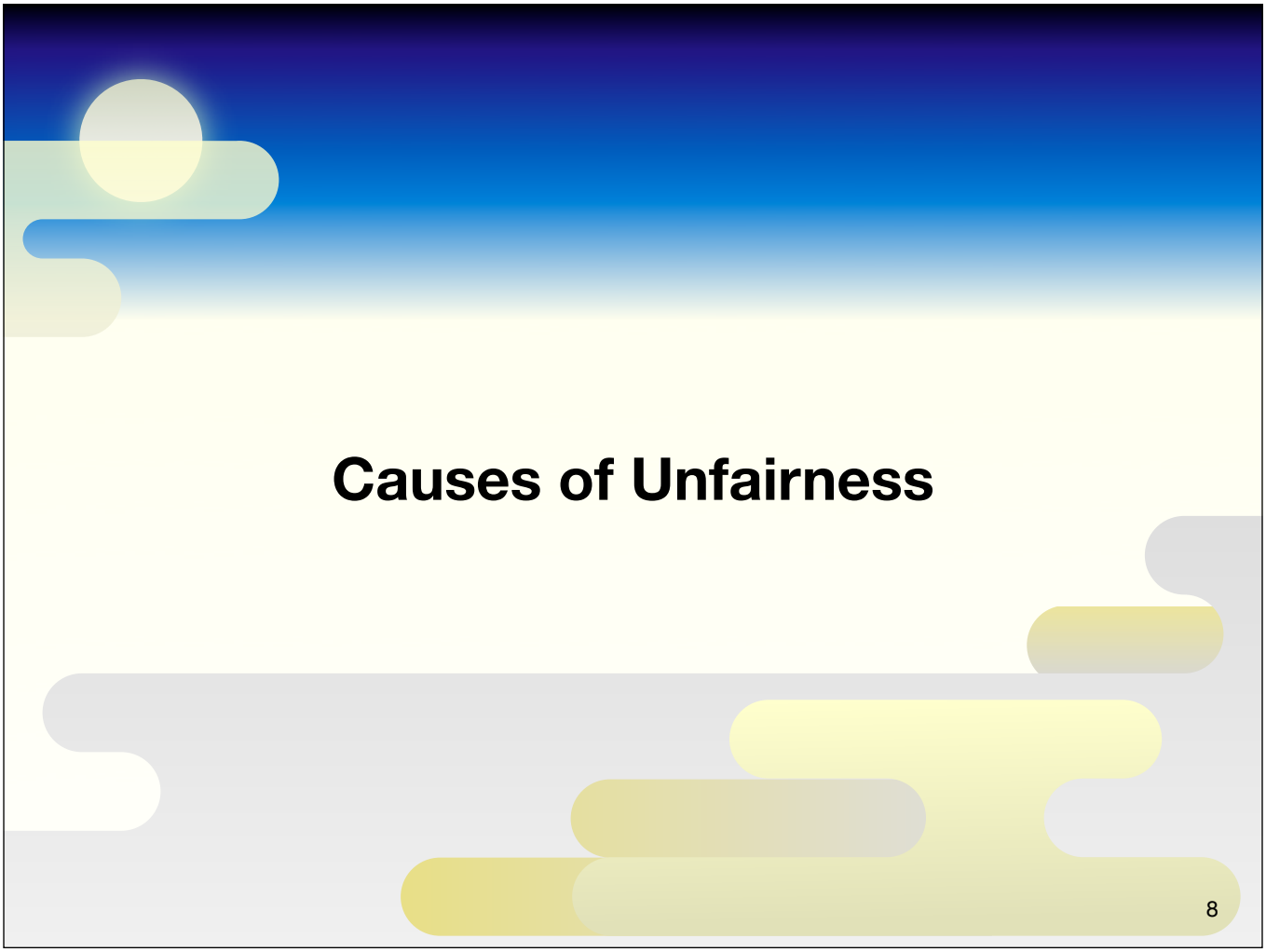
- Apparently irrelevant information helps to reveal private information ex. Users' demographics are predicted from query logs [Jones 09]

Contracts with Customers

- Customers' information must be used for the purpose within the scope of privacy policies

We need sophisticated techniques for data analysis whose outputs are neutral to specific information

Furthermore, fairness-aware mining are required due to social backgrounds, such as equality laws, the circulation of private information, contracts with customers, and so on.
Because these prohibit to use of specific information, we need sophisticated techniques for data analysis whose outputs are neutral to the information.



We next discuss causes of unfairness.

Three Causes of Unfairness

There are at least three causes of unfairness in data analysis

Prejudice

- the statistical dependency of sensitive features on an objective variable or non-sensitive features

Underestimation

- incompletely converged predictors due to the training from the finite size of samples

Negative Legacy

- training data used for building predictors are unfairly sampled or labeled

We consider that there are at least three causes of unfairness in data analysis: prejudice, underestimation, and negative legacy.
We sequentially show these causes.

Prejudice: Prediction Model

variables

- **objective variable Y** : a binary class representing a result of social decision, e.g., whether or not to allow credit
- **sensitive feature S** : a discrete type and represents socially sensitive information, such as gender or religion
- **non-sensitive feature X** : a continuous type and corresponds to all features other than a sensitive feature

prediction model

Classification model representing Pr[Y | X, S]

$M[Y \mid X, S]$

Joint distribution derived by multiplying a sample distribution

$Pr[Y, X, S] = M[Y \mid X, S] Pr[X, S]$

considering the independence between these variables over this joint distribution

Before talking abut prejudice, we'd like to give some notations.
There are three types of variables: an objective variable Y, sensitive features S, and non-sensitive features X.
We introduce a classification model representing the conditional probability of Y given X and S.
Using this model, joint distribution is predicted by multiplying a sample distribution.
We consider the independence between these variables over this joint distribution.

Prejudice

Prejudice : the statistical dependency of sensitive features on an objective variable or non-sensitive features

Direct Prejudice $Y \not\perp\!\!\!\perp S \mid X$

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

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

- a sensitive feature depends on a objective variable
- bringing red-lining effect (unfair treatment caused by information which is non-sensitive, but depending on sensitive information)

Latent Prejudice $X \not\perp\!\!\!\perp S \mid \phi$

- a sensitive feature depends on a non-sensitive feature
- completely excluding sensitive information

The first cause of unfairness is a prejudice, which is defined as the statistical dependency of sensitive features on an objective variable or non-sensitive features.
We further classify this prejudice into three types: direct, indirect, and latent.
Direct prejudice is a clearly unfair state that a prediction model directly depends on a sensitive feature. This implies the conditional dependence between Y and S given X.
Indirect prejudice is a state that a sensitive feature depends on a objective variable. This brings a red-lining effect.
Latent prejudice is the state that a sensitive feature depends on a non-sensitive feature.

Relation to PPDM

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

different conditions 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 indirect prejudice and PPDM.
Indirect prejudice refers the dependency between Y and S.
From information theoretic perspective, this means that mutual information between Y and S is non-zero.
From the viewpoint of privacy-preservation, this is interpreted as the leakage of sensitive information when an objective variable is known.
On the other hand, there are different conditions 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. disclosure of identity isn't problematic generally.

Underestimation

Underestimation : incompletely converged predictors due to the training from the finite size of samples

If the number of training samples is finite, the learned classifier may lead to more unfair decisions than that observed in training samples.

↓

Though such decisions are not intentional,
they might awake suspicions of unfair treatment

↓

- Notion of asymptotic convergence is **mathematically rationale**
- Unfavorable decisions for minorities due to the shortage of training samples **might not be socially accepted**

Techniques for an anytime algorithm or a class imbalance problem might help to alleviate this underestimation

Now, we turn back to the second cause of unfairness.
Underestimation is caused by incompletely converged predictors due to the training from the finite If the number of training samples is finite, the learned classifier may lead to more unfair decisions than that observed in the training samples.
Though such decisions are not intentional, they might awake suspicions of unfair treatment.
Notion of asymptotic convergence is mathematically rationale.
However, unfavorable decisions for minorities due to the shortage of training samples might not be socially accepted

Negative Legacy

Negative Legacy : training data used for building predictors are unfairly sampled or labeled

Unfair Sampling
If people in a protected group have been refused without investigation, those people are less frequently sampled

This problem is considered as a kind of a sample selection bias problem, but it is difficult to detect the existence of the sampling bias

Unfair Labeling
If the people in a protected group that should favorably accepted have been unfavorably rejected, labels of training samples become unfair

Transfer learning might help to address this problem, if additional information, such as the small number of fairly labeled samples, is available

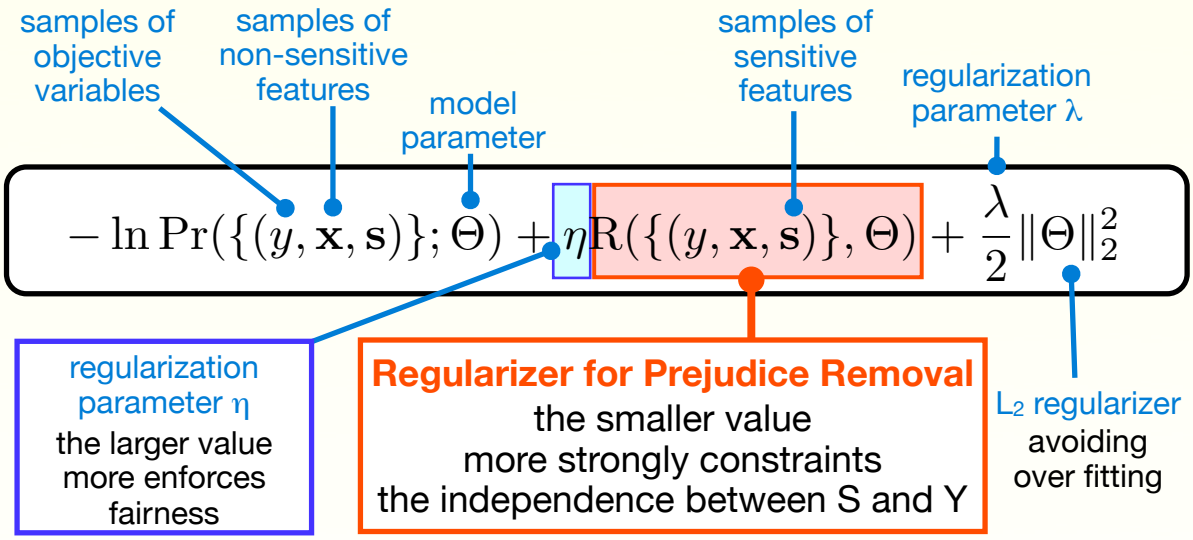
The third cause of unfairness is a negative legacy, which refers that training data used for building predictors are unfairly sampled or labeled.
Unfair sampling occurs when, for example, if people in a protected group have been refused without investigation, those people are less frequently sampled.
Unfair labeling occurs when, for example, if the people in a protected group that should favorably accepted have been unfavorably rejected, labels of training samples become unfair.

Among these causes of unfairness, we focus on the removal indirect prejudice.
We next talk about our methods and experimental results.

Logistic Regression

Logistic Regression : discriminative classification model

A method to remove indirect prejudice from decisions made by logistic regression



We propose a method to remove indirect prejudice from decisions made by logistic regression.
For this purpose, we add this term to an original objective function of logistic regression.
Regularizer for prejudice removal is designed so that the smaller value more strongly constraints the independence between S and Y.
Eta is a regularization parameter. The lager value more enforces fairness.

Prejudice Remover

Prejudice Remover Regularizer
mutual information between S and Y
so as to make Y independent from S

$$\sum_{Y,X,S} \mathcal{M}[Y|X,S]\Pr[X,S]\ln\frac{\Pr[Y,S]}{\Pr[S]\Pr[Y]}$$

$$\sum_{Y\in\{0,1\}}\sum_{(\mathbf{x},\mathbf{s})} \mathcal{M}[y|\mathbf{x},\mathbf{s};\Theta]\ln\frac{\Pr[y|\bar{\mathbf{x}}_{\mathbf{s}},\mathbf{s};\Theta]}{\sum_{\mathbf{s}}\Pr[y|\bar{\mathbf{x}}_{\mathbf{s}},\mathbf{s};\Theta]}$$

true distribution is replaced with sample distribution

approximate by the mutual info at means of X instead of marginalizing X

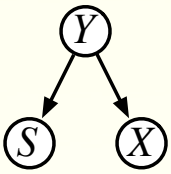
We are currently improving a computation method

To remove indirect prejudice, we develop a prejudice remover regularizer.
This regularizer is mutual information between S and Y so as to make Y independent from S.
This is calculated by this formula. Rather brute-force approximation is adopted.

Calders-Verwer Two Naive Bayes

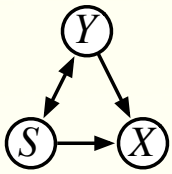
[Calders+ 10]

Naive Bayes



- S and X are conditionally independent given Y

Calders-Verwer Two Naive Bayes (CV2NB)

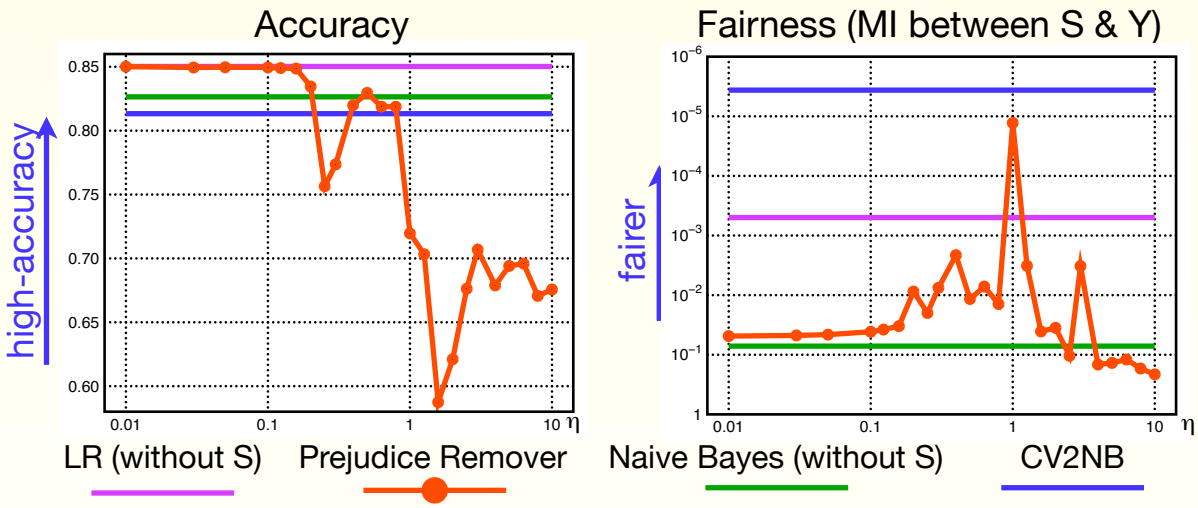


- non-sensitive features X are mutually conditionally independent given Y and S

- Unfair decisions are modeled by introducing of the dependency of S on X, as well as that of Y on X
- A model for representing joint distribution Y and S is built so as to enforce fairness in decisions

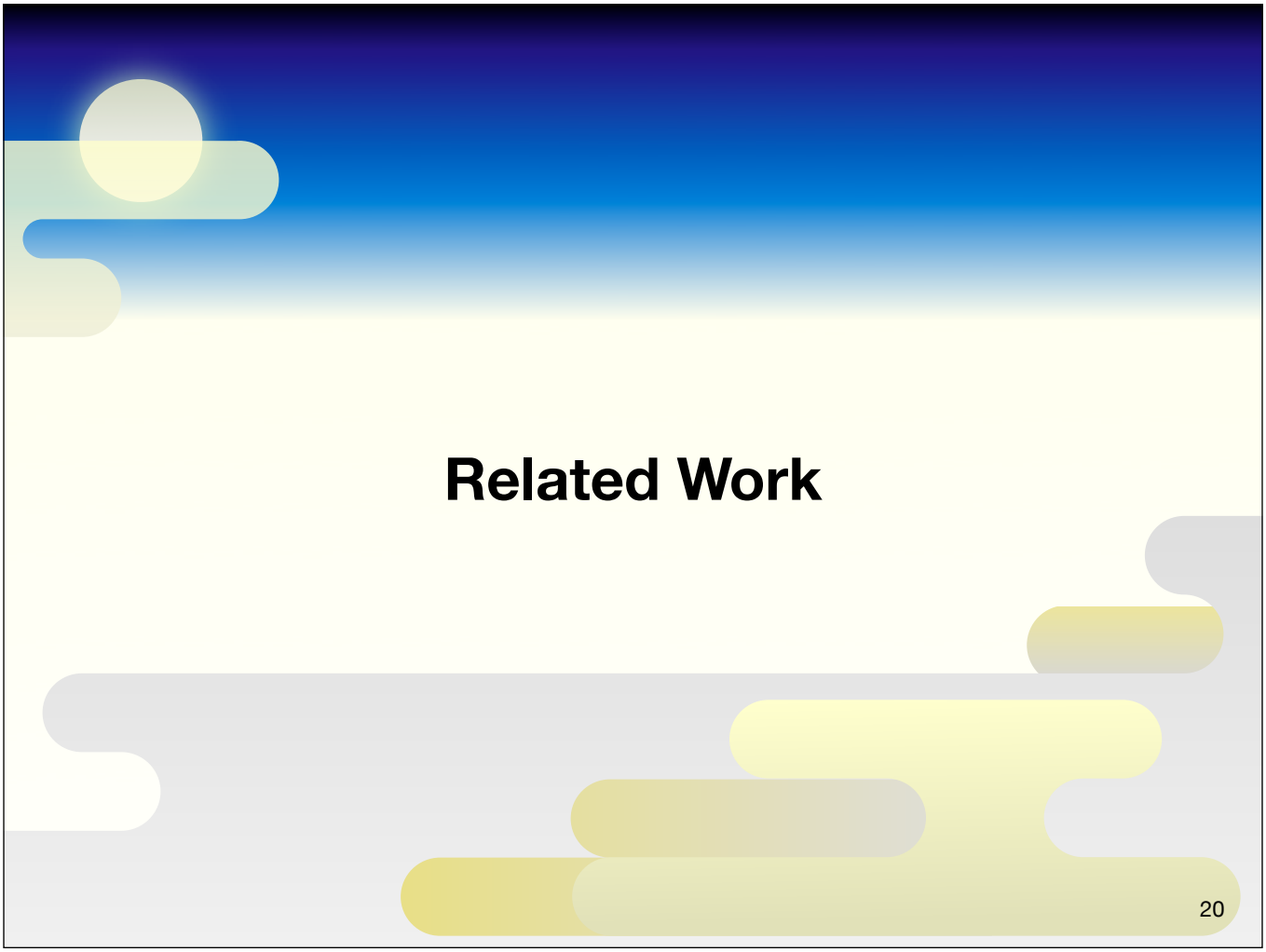
We compared our logistic regression with Calders-Verwer two-naive-Bayes classifier.
Unfair decisions are modeled by introducing of the dependency of X on S as well as on Y.
A model for representing joint distribution Y and S is built so as to enforce fairness in decisions.

Experimental Results



- Both CV2NB and PR made fairer decisions than their corresponding baseline methods
- For the larger η, the accuracy of PR tends to be declined, but no clear trends are found in terms of fairness
- The unstablity of PR would be due to the influence of approximation or the non-convexity of objective function

This is a summary of our experimental results.
Our method balances accuracy and fairness by tuning η parameter.
A two-naive-Bayes classifier is very powerful. We could win in accuracy, but lost in fairness.
Both CV2NB and PR made fairer decisions than their baseline methods.
For the larger η, the accuracy of PR tends to be declined, but no clear trends are found in terms of fairness
The unstablity of PR would be due to the influence of brute-force approximation or the non-convexity of objective function. We plan to investigate this point.



Fairness-aware mining is a relatively new problem, so we briefly review related work.

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

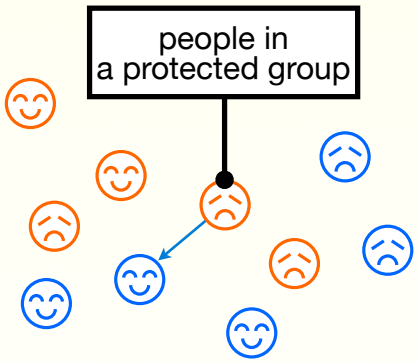
They proposed an algorithm to enumerate rules that are not *a*-protected

To our knowledge, Pedreschi, Ruggieri, and Turini firstly addressed the problem of fairness in data mining.
They proposed a measure to quantify the unfairness of association rules.
ELIFT is defined as the ratio of the confidence of an association rule with additional condition to the confidence of a base rule.
They proposed a notion of a-protection. Rules are considered as unfair if there exists association rules whose extended lift is larger than a.
They proposed an algorithm to enumerate rules that are not a-protected

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



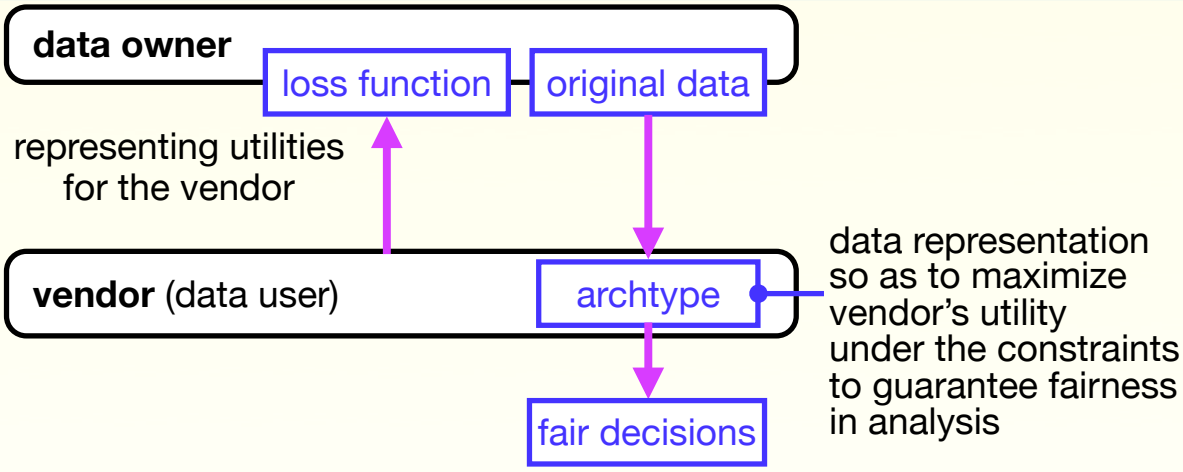
k-nearest neighbors

- 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

This group proposed another notion, 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.
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.
This situation testing has relation to our indirect prejudice.

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 and colleagues proposed a framework for fairness-aware data publishing.
Vendor send a loss function, which represents utilities for the vendor.
Using this function, data owner transforms original data into archtypes.
Archtype is a data representation so as to maximize vendor's utility under the constraints to guarantee fair results.
They show the conditions that these archtypes should satisfy, and this condition implies that the probability of receiving favorable decision is irrelevant to belonging to a protected group.
This further implies the independence between S and Y, and has relation to our indirect prejudice.

Conclusion

Contributions

- three causes of unfairness: prejudice, underestimation, and negative legacy
- a prejudice remover regularizer, which enforces a classifier's independence from sensitive information
- experimental results of logistic regressions with our prejudice remover

Future Work

- Computation of prejudice remover has to be improved

Socially Responsible Mining

- 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.

Our contributions are as follows.
In future work, computation of prejudice remover has to be improved.
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.
That's all we have to say. Thank you for your attention.