Considerations on Fairness-aware Data Mining

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 ICDM International Workshop on Discrimination and Privacy-Aware Data Mining

@ Brussels, Belgium, Dec. 10, 2012

START

I'm Toshihiro Kamishima.
Today, we would like to talk about fairness-aware data mining and its connections with other techniques for data analysis.

Overview

Fairness-aware Data Mining

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

Introduction of the following topics to give an overview of fairness-aware data mining

- Applications of fairness-aware data mining for the other purposes besides avoiding discrimination
- An integrated view of existing fairness measures based on the statistical independence
- Connections of fairness-aware data mining with other techniques for data analysis

Fairness-aware data mining is a data analysis taking into account potential issues of fairness, discrimination, neutrality, or independence.
In this talk, we introduce these topics to give an overview of fairness-aware data mining:
First, we show applications of fairness-aware data mining in addition to avoiding discrimination.
Second, we demonstrate an integrated view of existing fairness measures based on the statistical independence.
Third, we discuss the connections of fairness-aware data mining to other techniques for data analysis.

Applications of Fairness-Aware Data Mining

We first show applications of the fairness-aware data mining techniques in addition to avoiding discrimination.

Discrimination-aware Data Mining

[Pedreschi+ 08]

Discrimination-aware Data Mining
Detection or elimination of the influence of socially sensitive information to 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

⬇

This framework for detecting or eliminating the influence of a specific information to the analysis results can be used for other purposes besides avoiding discrimination

- enhancing the neutrality of the prediction
- excluding uninteresting information from analysis results

✳This is the reason why we use the term "fairness-aware" instead of "discrimination-aware"

Fairness-aware data mining techniques were originally developed for avoiding social discrimination.
The goal of discrimination-aware data mining is to detect or to eliminate the influence of socially sensitive information to serious decisions, such as job application or credit scoring.
Socially sensitive information includes information considered from the viewpoint of social fairness or information restricted by law or contracts.
This framework for detecting or eliminating the influence of a specific information to the analysis results can be used for other purposes besides avoiding discrimination.
We then show these two types of applications.

Filter Bubble

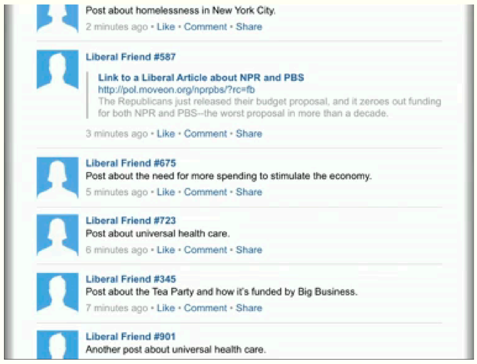
[TED Talk by Eli Pariser]

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 first 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 are used for removing the influence of the specified information

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.
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.
Fairness-aware data mining techniques are used for removing the influence of the specified information.

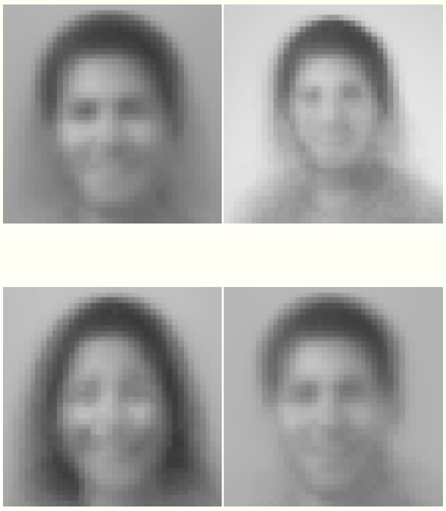
Ignoring Uninteresting Information

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

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

An Integrated View of Existing Fairness Measures

We next demonstrate an integrated view of existing fairness measures.

Notations of Variables

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

$\mathbf{X}^{(E)}$ **explainable non-sensitive feature vector**

$\mathbf{X}^{(U)}$ **unexplainable non-sensitive feature vector**

- Non-sensitive features can be further classified explainable and unexplainable (we will introduce later)

We begin by introducing some 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.
Non-sensitive features can be further classified explainable and unexplainable.

Prejudice

[Kamishima+ 12]

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

$Y \not\perp\!\!\!\perp S \mid \emptyset$

- the dependence of an objective variable Y on a sensitive feature S
- bringing a red-lining effect

Indirect Prejudice with Explainable Features

$Y \not\perp\!\!\!\perp S \mid \mathbf{X}^{(E)}$

- the conditional dependence between Y and S given explainable non-sensitive features, if a part of non-sensitive features are explainable

We next introduce our notion of prejudice, which is one of causes of unfairness
This is defined as the statistical dependences of an objective variable or non-sensitive features on a sensitive feature.
There are several types of prejudices:
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.
If a part of non-sensitive features are explainable, indirect prejudice becomes the conditional dependence between Y and S given explainable non-sensitive features.

Prejudice

The degree of prejudices can be evaluated:

- computing statistics for the independence (among observations)
  ex. mutual information, χ^2-statistics
- applying tests of independence (among sampled population)
  χ^2-tests

Relations between this prejudice and other existing measures for unfairness

- elift (extended lift)
- CV score (Calders-Verwer's discrimination score)
- CV score with explainable features

The degree of prejudices can be evaluated by computing statistics for independence or by applying tests of independence.
We then show the relations between this prejudice and other existing measures for unfairness: elift, CV score, and CV score with explainable features.

elift (extended lift)

[Pedreschi+ 08, Ruggieri+ 10]

elift (extended lift) = $\dfrac{\mathrm{conf}(X=\mathbf{x}, \boxed{S{=}-} \Rightarrow Y{=}-)}{\mathrm{conf}(X=\mathbf{x} \Rightarrow Y{=}-)}$

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

The condition elift = 1 means that no unfair treatments, and it implies

$$\Pr[Y{=}-, S{=}- \mid X{=}\mathbf{x}] = \Pr[Y{=}- \mid X{=}\mathbf{x}]$$

If this condition is the case for any $\mathbf{x} \in \mathrm{Dom}(X)$
(the condition of no direct discrimination in their definition)
and S and Y are binary variables:

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

This is equivalent to the condition of no direct prejudice

Similarly, their condition of no-indirect-discrimination
is equivalent to that of no-indirect-prejudice condition

elift is a measure to quantify the unfairness of association rules, and is defined as the ratio of the confidence of an association rule with additional condition to the confidence of a base rule.
The condition elift equals to one means no unfair treatments, and it implies this equation.
If S and Y are binary, this equation can be derived.
This is equivalent to the condition of no direct prejudice.

CV Score

(Calders-Verwer's Discrimination Score)

[Calders+ 10]

Calders-Verwer discrimination score (CV score)

$$\Pr[Y{=}+\,|\,S{=}+] - \Pr[Y{=}+\,|\,S{=}-]$$

The conditional probability of the advantageous decision for non-protected members subtracted by that for protected members

⬇

The condition CV score = 0 means
that neither unfair or affirmative treatments

⬇

$$\Pr[Y\,|\,S] = \Pr[Y]$$

This is equivalent to the condition of no indirect prejudice

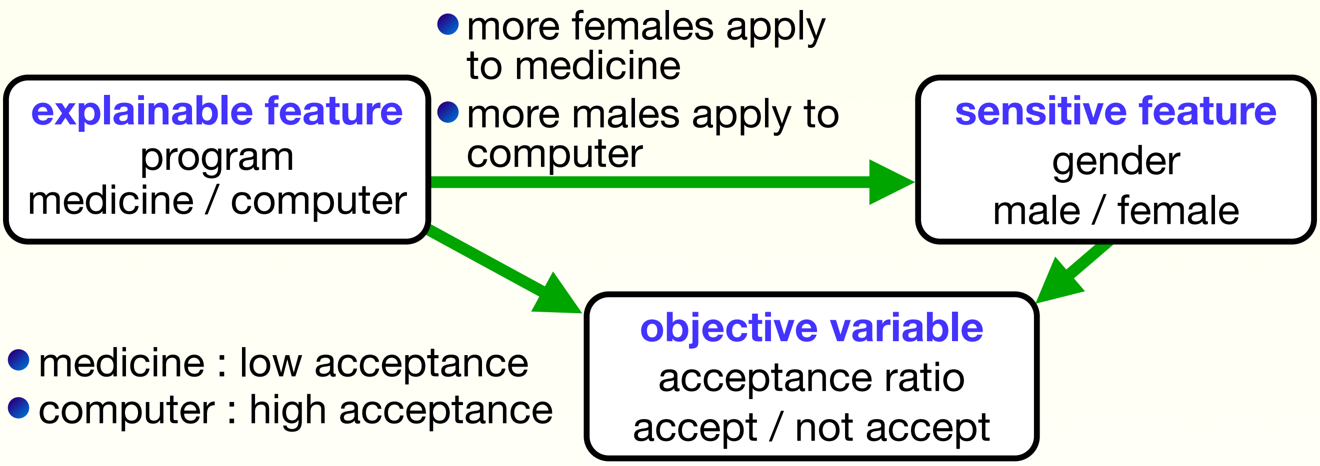
Calders and Verwer proposed another discrimination score.
We here call this a CV score, which is the conditional probability of the advantageous decision for non-protected members subtracted by that for protected members.
The condition CV score = 0 means that neither unfair or affirmative treatments, and it implies this equation.
This is equivalent to the condition of no indirect prejudice

Explainability

[Zliobaite+ 11]

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. showed non-discriminatory cases, even if distributions of an objective variable depends on a sensitive feature.
In this example, 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.

Explainability

To remove unfair treatments under such a condition, the probability of receiving an advantageous decision, $\Pr^{\dagger}[Y{=}+\,|X^{(E)}]$, is chosen so as to satisfy the condition:

$$\Pr^{\dagger}[Y{=}+\,|X^{(E)}] =$$
$$\Pr^{\dagger}[Y{=}+\,|X^{(E)}, S{=}+] = \Pr^{\dagger}[Y{=}+\,|X^{(E)}, S{=}-]$$

This condition implies

$$\Pr[Y|X^{(E)}, S] = \Pr[Y|X^{(E)}]$$

This is equivalent to
the condition of no indirect prejudice with explainable features

✳Notions that is similar to this explainability are proposed as a *legally-grounded feature* [Luong+ 11] and are considered in the rule generalization procedure [Haijan+ 12].

To remove unfairness under such a condition, the probability of receiving an advantageous decision is chosen so as to satisfy the condition, and it implies this equation, if both S and Y are binary.
This is equivalent to the condition of no indirect prejudice with explainable features.

Connections with Other Techniques for Data Analysis

We next discuss connections of FADM with other techniques for data analysis.

Privacy-Preserving Data Mining

indirect prejudice
the dependence between an 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 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 first point out the connection with PPDM.
Indirect prejudice refers the dependence 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 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.
Disclosure of identity isn't problematic in FADM, generally.

Causal Inference

[Pearl 2009]

Causal inference

a general technique for dealing with probabilistic causal relations

- Both FADM and causal inference can cope with the influence of sensitive information to an objective, but their premises are quite different
- While causal inference demands more information about causality among variables, it can handle the facts that were not occurred

example of the connection between FADM and causal inference

- A person of S=+ actually receives Y=+
- If S was changed to −, how was the probability of receiving Y=−?
- Under the condition called exogenous, the range of this probability is

$$\max[0, \underline{\Pr[Y{=}+|S{=}+] - \Pr[Y{=}+|S{=}-]}] \leq \mathrm{PNS} \leq \min[\Pr[Y{=}+, X{=}+], \Pr[Y{=}-, X{=}-]]$$

This lower bound coincides with a CV score

Causal inference is a general technique for dealing with probabilistic causal relations.
Both FADM and causal inference can cope with the influence of sensitive information to an objective, but their premises are quite different.
This is an example of the connection between FADM and causal inference.
A person in a non-protected group actually receives advantageous decision.
If he or she was in a protected group, how was the probability of receiving a disadvantageous decision.
Under the condition called exogenous, the range of this probability is represented by this formula.
This lower bund coincides with a CV score.

Cost-Sensitive Learning

[Elkan 2001]

Cost-Sensitive Learning: learning classifiers so as to optimize classification costs, instead of maximizing prediction accuracies

⬇

FADM can be regarded as a kind of cost-sensitive learning that pays the costs for taking fairness into consideration

Cost matrix $C(i \mid j)$: cost if a true class j is predicted as class i

Total cost to minimize is formally defined as (if class Y=+ or −):

$$\mathcal{L}(\mathbf{x}, i) = \sum_j \Pr[j \mid \mathbf{x}] C(i \mid j)$$

An object $\mathbf{x}$ is classified into the class i whose cost is minimized

The goal of cost-Sensitive Learning to obtain classifiers so as to optimize classification costs, instead of maximizing prediction accuracies
Formally, an object x is classified into the class i whose cost is minimized.
Broadly speaking, FADM can be regarded as a kind of cost-sensitive learning that pays the costs for taking fairness into consideration.

Cost-Sensitive Learning

[Elkan 2001]

Theorem 1 in [Elkan 2001]

If negative examples in a data set is over-sampled by the factor of

$$\frac{C(+|-)}{C(-|+)}$$

and a classifier is learned from this samples, a classifier to optimize specified costs is obtained

↓

In a FADM case, an over-sampling technique is used for avoiding unfair treatments

↓

A corresponding cost matrix can be computed by this theorem, which connects a cost matrix and the class ratio in training data

✳Note that this over-sampling technique is simple and effective for avoiding unfair decisions, but its weak point that it completely ignores non-sensitive features

Elkan proposed a method to learn a cost-sensitive classifier by over-sampling training data based on this theorem,
In a FADM case, an over-sampling technique is used for avoiding unfair treatments.
A corresponding cost matrix can be computed by this theorem, which connects a cost matrix and the class ratio in training data.

Other Connected Techniques

Legitimacy

- Data mining models can be deployed in the real world

Independent Component Analysis

- Transformation while maintaining the independence between features

Delegate Data

- To perform statistical tests, specific information is removed from data sets

Dummy Query

- Dummy queries are inputted for protecting users' demographics into search engines or recommender systems

Visual Anonymization

- To protect identities of persons in images, faces or other information is blurred

Many other techniques and notions are connectng with FADM: legitimacy, ICA, delegate data, dummy query, visual anonymization.

Conclusion

Contributions

We introduced the following topics to give an overview of fairness-aware data mining

- We showed other FADM applications besides avoiding discrimination: enhancing neutrality and ignoring uninteresting information.
- We discussed the relations between our notion of prejudice and other existing measures of unfairness.
- We showed the connections of FADM with privacy-preserving data mining, causal inference, cost-sensitive learning, and so on.

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.
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

- 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 sites.
That’s all I have to say. Thank you for your attention.