バイアス考慮型分類器の安定性に関する予備調査

神嶌 敏弘[1]，赤穂 昭太郎[1]，馬場 雪乃[2]，鹿島 久嗣[3]

[1]産業技術総合研究所，[2]筑波大学，[3]京都大学

2021年度人工知能学会全国大会（第35回）@ オンライン，2021-06-10

https://www.kamishima.net

Overview

- Stability : a new desirable property for bias-aware classifier
- Dataset to examine the stability
- Bias-aware models
- Discussion about the reasons why the model was unstable

Notations

Y **target variable / object variable**

An objective of decision making, or what to predict

ex., loan approval, university admission, what to recommend

- Y: observed / true, $\hat{Y}$: predicted

S **sensitive feature**

To ignore the influence to the sensitive feature from a target

ex., socially sensitive information (gender, race), items' brand

- Specified by a user or an analyst depending on his/her purpose
- It may depend on a target or other features

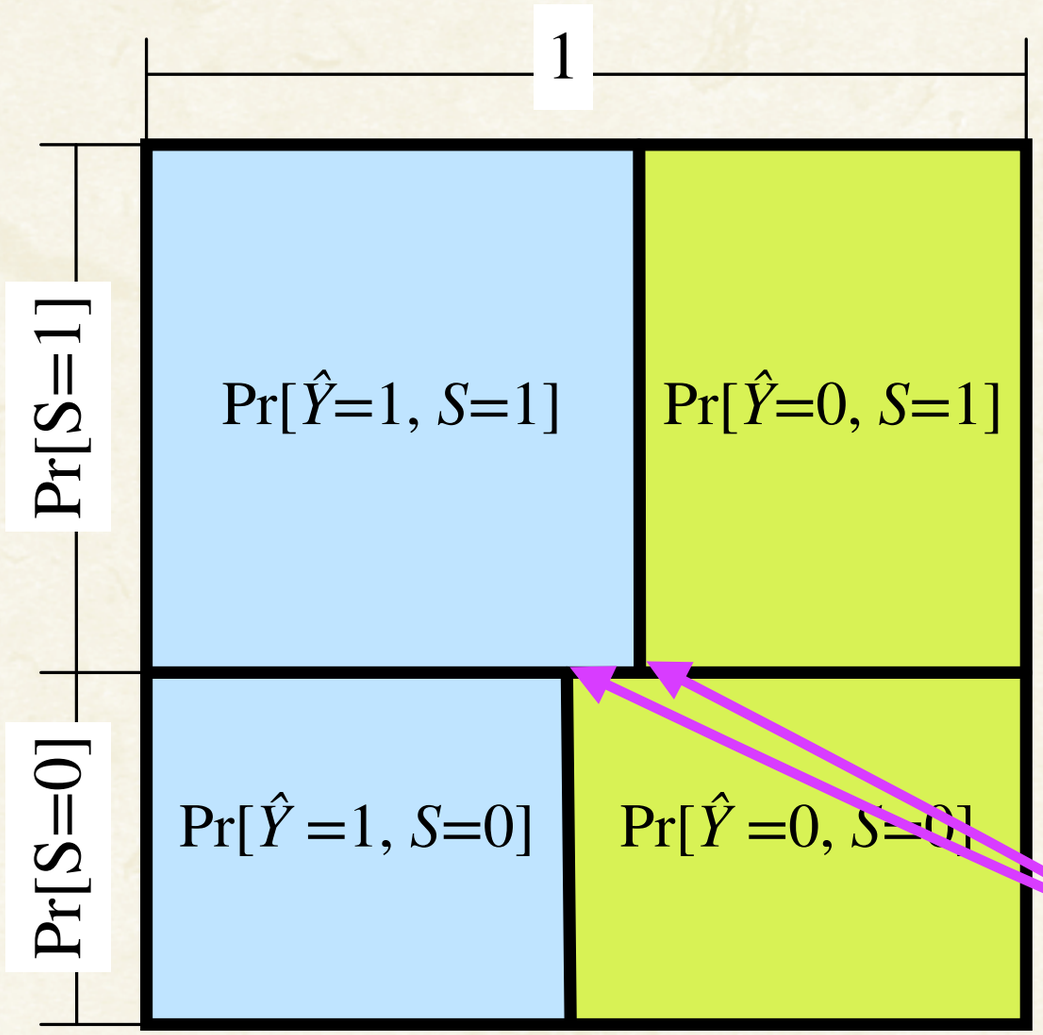
$\mathbf{X}$ **non-sensitive feature vector**

All features other than a sensitive feature

Independence / Statistical Parity

[Calders+ 10, Dwork+ 12]

Remove data bias ➡ Independence / Statistical Parity: $\hat{Y} \perp\!\!\!\perp S$



$$\hat{Y} \perp\!\!\!\perp S$$

$$\Leftrightarrow$$

$$\Pr[\hat{Y}, S] = \Pr[\hat{Y}]\Pr[S]$$

Ratios between positives and negatives in prediction are matched among all sensitive values

Bias-aware Classifier

Bias-aware classifier : The most accurate predictor while satisfying a fairness constraint

loss prediction predictor fairness constraint

$$\min \ell(\hat{Y} = f(\mathbf{X}, S); \mathcal{D}) \text{ s.t. } S \perp\!\!\!\perp \hat{Y}$$

The loss is evaluated on **a biased dataset**,
because an unbiased dataset is unavailable
However, the loss should be evaluated on **an unbiased dataset**

Other desirable property for bias-aware classifier

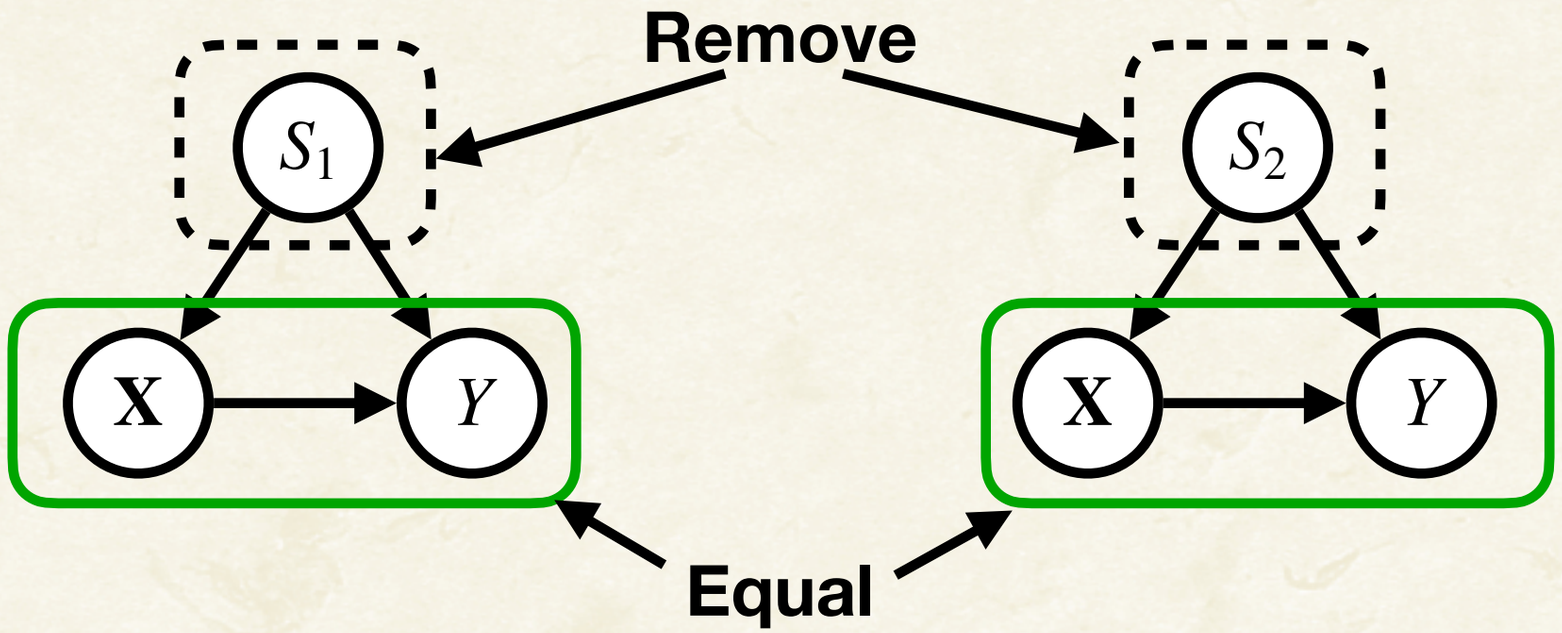
Stability of Bias-aware Classifier

Bias-aware Classifier : Removing the influence of S on Y

Stability of Bias-aware Classifier

Two datasets consist of the same information
except for the information represented by S

A bias-aware classifier should learn the same models



Dataset to Examine the Stability

Paired Comparison Task

The subjects choose the more preferred item from a pair of items

fatty tuna | tuna | shrimp | salmon roe | sea eel | sea urchin | tuna roll | squid | egg | cucumber roll

Popular ← → Unpopular

a proxy of a sensitive feature

Cognitive Bias : the factors of a user interface that influences the subjects' choices

The same models are learned from datasets influenced by different kinds of cognitive biases

↓

The bias-aware classifier is considered stable

Cognitive Biases: Positional Effect

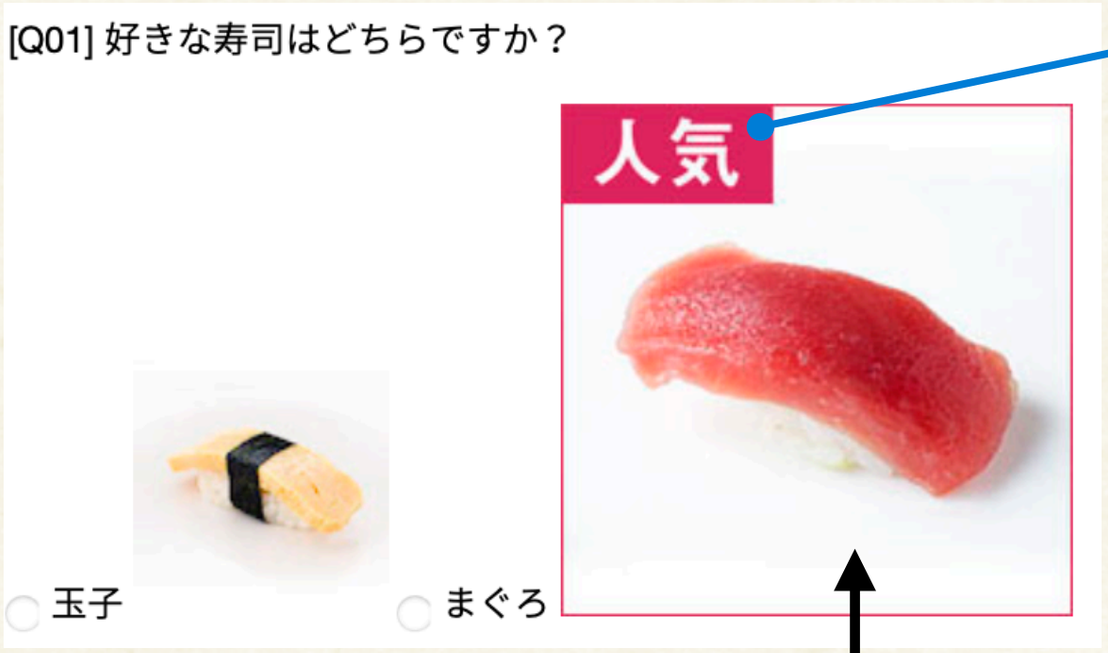
Positional Effect : the item displayed near the upper-left corner of the interface screen is more frequently chosen.

frequently chosen

- **Baseline** : the positions of items are randomly assigned, ideal RCT
- **Fixed** : the positions of items are affected by the popularity of items

Cognitive Biases: Bandwagon Effect

Bandwagon Effect : the item indicated that other people prefer is more frequently chosen

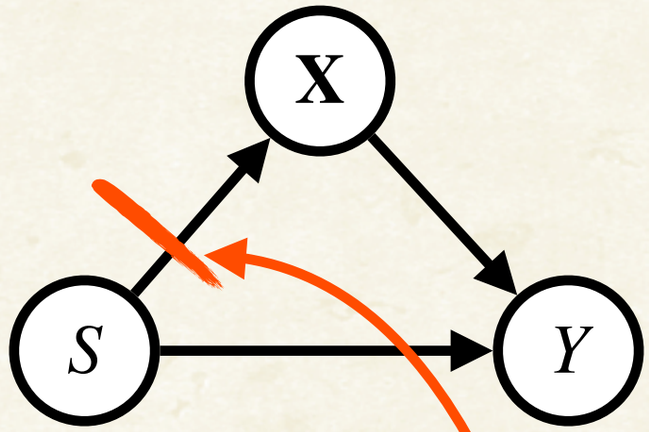


- **Bandwagon** : the items indicated as popular are affected by the popularity of items as in the same procedure as the Fixed procedure

Measurement of Cognitive Bias

a pair of candidate items **confounder**

cognitive bias
intervention



subject's choice
effect

Candidate items are randomly selected

Causal effect : $\Pr[Y = 1 \mid S = 0] - \Pr[Y = 1 \mid S = 1]$

As we designed, our datasets are influenced by cognitive biases

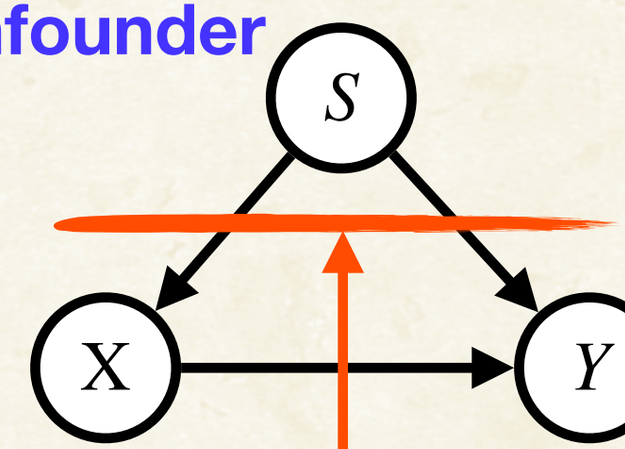
Random	Fixed	Bandwagon
0.0229	0.0077	0.3451

* Positive value indicates the influence of cognitive biases

Bias-aware Model

Model for estimating the influence $\mathbf{X}$ on Y
while removing the influence of S

cognitive bias **confounder**



a pair of candidate items
intervention

subject's choice
effect

The Influence of S can be removed by **stratification**

$$\Pr[Y \mid \mathbf{X}] = \sum_s \Pr[S = s] \Pr[Y \mid \mathbf{X}, S = s]$$

total effect of $\mathbf{X}$ on Y

size of stratum

effect of $\mathbf{X}$ on Y per stratum

Bias-aware Model

This stratification technique has connection with
a generative type of a fairness-aware model

biased generative classifier

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

fairness constraint

$$= \Pr[S]\Pr[Y]\Pr[\mathbf{X}|Y, S] \quad (\leftarrow Y \perp\!\!\!\perp S)$$

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

size of stratum

total effect of Y on $\mathbf{X}$

effect of Y on $\mathbf{X}$ per stratum

Removing Biases by Stratification

Hypothesis : If a stratification technique is stable as a bias-aware classifier, similar models would be learned from three types of datasets: Random, Fixed, and Bandwagon

For each dataset, we get the probabilities $\Pr[Y = 1 \mid \mathbf{X} = \mathbf{x}]$ for all $\mathbf{x}$

If stratification is stable, these probabilities would be similar

Similarities (Frobenius norm) between a pair of probability matrices

Random vs Fixed	Random vs Bandwagon
0.0521	0.127

Contrary to our hypothesis,
this stratification techniques are not stable

Failure of Random Assignment

select item pairs

randomly assign subjects

Popular

enhance a **more popular** item

Popular

enhance a **less popular** item

merge subjects' responses

Cognitive biases may not be fully removed by the stratification

- The scheme of data processing might change the causal structure
  - ➡ S may behave as a mediator as well as a confounder
- We will explore other types of causal structure

Next Steps

enhance a **more popular** item

honest

enhance a **less popular** item

dishonest

Confounders other than the controlled cognitive bias might exist

- We showed “Popular” marks irrelevant to its real popularity
  - ➡ This might cause another type of cognitive bias
- We plan to collect data influenced by other cognitive bias

memory effect

show one item

ask subjects to choose

Conclusions

- The notion of stability of bias-aware techniques
- A dataset to examine the stability of bias-aware techniques
- The relationship between bias-aware techniques and causal inference
- Preliminary experimental results, showing the instability of our techniques
- Next steps for collecting new datasets or for developing new models