Learning Individually Fair Classifier with Path-Specific Causal-Effect Constraint

Yoichi Chikahara[1,3], Shinsaku Sakaue[2], Akinori Fujino[1], Hisashi Kashima[3]

[1] NTT [2] University of Tokyo [3] Kyoto University

Problem: Learning fair classifier with causal graph

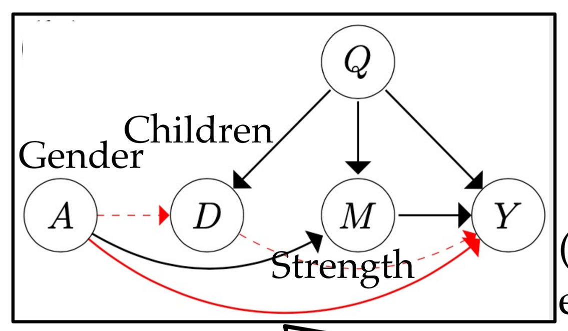
Input

Training data

A Sensitive	Q	D	M	Y
Female	B	0	B	Accept
Male	A	1	B	Reject
Male	C	2	C	Reject

$\boldsymbol{X} = \{A, Q, D, M\}$: Features of each individual

Causal graph



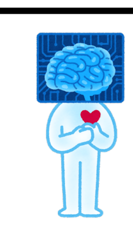
(Given by experts or estimated from data)

Minimize loss L_θ + penalty on unfairness G_θ

$$\min_\theta \frac{1}{n}\sum_{i=1}^{n} L_\theta(\boldsymbol{x}_i, y_i) + \lambda G_\theta(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$$

Output

Fair binary classifier $h_{\hat{\theta}}(\boldsymbol{X})$



Avoid imposing unnecessary fairness constraints using causal graph that expresses *what is unfair*

Example 1: Hiring decisions for physically-demanding jobs

Following reasons for rejection is **unfair**:
1. female ($A \to Y$) 2. female, has no child ($A \to D \to Y$)
while following is **fair**:
3. female, has little physical strength ($A \to M \to Y$)

To formulate G_θ based on unfair pathways $\pi = \{A \to Y, A \to D \to Y\}$, we measure the unfairness as **path-specific causal effects (PSEs)** [1].

Weaknesses of existing methods

Table 1: Comparison with existing methods

Method	Individually fair	Functional assumptions
Our method	Yes	Unnecessary
PSCF	Yes	Necessary
FIO	No	Unnecessary

According to Wu et al. [2], a classifier achieves **(path-specific) individual-level fairness** if the **conditional expected value of PSEs is zero** for any input $\boldsymbol{x}$:

$$\mathbb{E}_{Y_{A\Leftarrow 0}, Y_{A\Leftarrow 1\|\pi}}[Y_{A\Leftarrow 1\|\pi} - Y_{A\Leftarrow 0} | \boldsymbol{X} = \boldsymbol{x}] = 0$$

PSE: difference of two predictions (i.e., $Y_{A\Leftarrow 0}$ and $Y_{A\Leftarrow 1\|\pi}$), obtained by modifying input feature attributes $\boldsymbol{x}$. In **Example 1**, for each woman ($A = 0$),
$Y_{A\Leftarrow 0}$ is made by directly taking her observed feature attributes as input
$Y_{A\Leftarrow 1\|\pi}$ is made with *counterfactual attributes*, observed if she were male ($A = 1$)

How can we learn individually fair classifier without restrictive functional assumptions?

Proposed method

Main idea: Make $Y_{A\Leftarrow 0} = Y_{A\Leftarrow 1\|\pi} = 0$ or $Y_{A\Leftarrow 0} = Y_{A\Leftarrow 1\|\pi} = 1$ for all individuals (i.e., regardless of input feature value $\boldsymbol{x}$)

1. Penalty by upper bound on PIU

To achieve this goal, we **force probability of individual unfairness (PIU) to be zero**, whose upper bound can be derived as

$$\underbrace{\mathrm{P}(Y_{A\Leftarrow 0} \neq Y_{A\Leftarrow 1\|\pi})}_{\text{PIU}} \leq \underbrace{2\,\mathrm{P}^I(Y_{A\Leftarrow 0} \neq Y_{A\Leftarrow 1\|\pi})}_{\text{upper bound on PIU}}.$$

$\mathrm{P}^I(Y_{A\Leftarrow 0}, Y_{A\Leftarrow 1\|\pi}) = \mathrm{P}(Y_{A\Leftarrow 0})\,\mathrm{P}(Y_{A\Leftarrow 1\|\pi})$
is an *independent joint distribution*, which **can be inferred from data without any restrictive functional assumptions**

To make the upper bound value close to zero, we use the estimator of $\mathrm{P}^I(Y_{A\Leftarrow 0} \neq Y_{A\Leftarrow 1\|\pi})$ as penalty function, which is formulated as

$$G_\theta(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = \hat{p}_\theta^{A\Leftarrow 1\|\pi}(1 - \hat{p}_\theta^{A\Leftarrow 0}) + (1 - \hat{p}_\theta^{A\Leftarrow 1\|\pi})\hat{p}_\theta^{A\Leftarrow 0},$$

where $\hat{p}_\theta^{A\Leftarrow 0}$ and $\hat{p}_\theta^{A\Leftarrow 1\|\pi}$ are estimator of $\mathrm{P}(Y_{A\Leftarrow 0} = 1)$ and $\mathrm{P}(Y_{A\Leftarrow 1\|\pi} = 1)$.
In **Example 1**, they are given as weighted averages of $c_\theta(\boldsymbol{X}) = \mathrm{P}(Y = 1|\boldsymbol{X})$:

$$\hat{p}_\theta^{A\Leftarrow 0} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}(a_i = 0)\hat{w}_i c_\theta(a_i, q_i, d_i, m_i) \qquad \hat{p}_\theta^{A\Leftarrow 1\|\pi} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}(a_i = 1)\hat{w}'_i c_\theta(a_i, q_i, d_i, m_i)$$

2. Comparison with existing fairness constraint

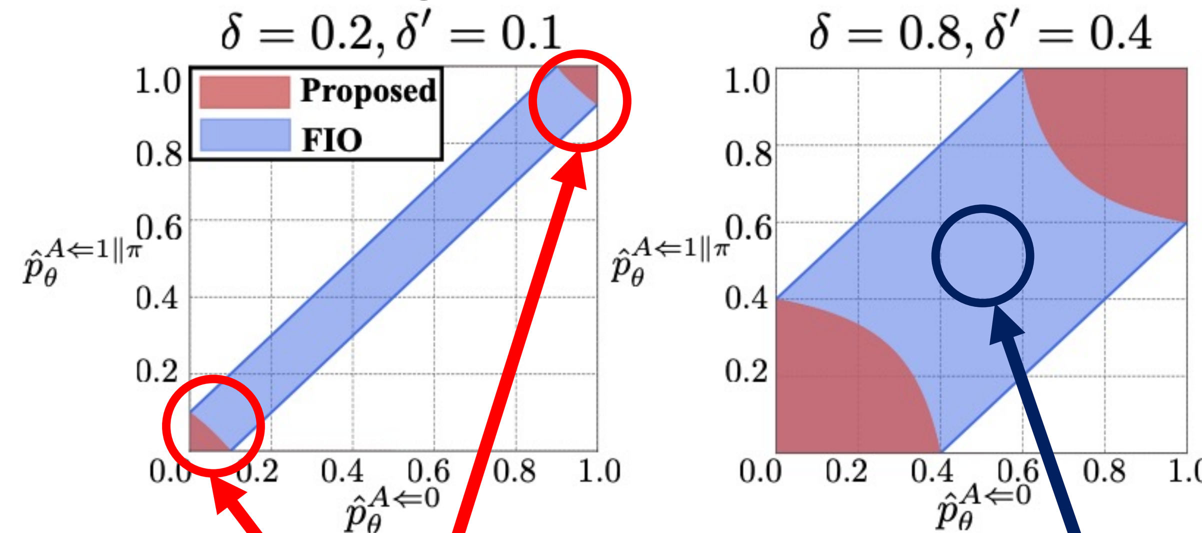
Our method aims to satisfy the following condition:

$$\hat{p}_\theta^{A\Leftarrow 1\|\pi}(1 - \hat{p}_\theta^{A\Leftarrow 0}) + (1 - \hat{p}_\theta^{A\Leftarrow 1\|\pi})\hat{p}_\theta^{A\Leftarrow 0} \leq \delta.$$

By contrast, the existing FIO method [3] imposes the following one:

$$-\delta' \leq \hat{p}_\theta^{A\Leftarrow 1\|\pi} - \hat{p}_\theta^{A\Leftarrow 0} \leq \delta'$$

Figure 1: Feasible regions of our constraint (red) and FIO (blue)



$Y_{A\Leftarrow 0} = Y_{A\Leftarrow 1\|\pi} = 0$ or $Y_{A\Leftarrow 0} = Y_{A\Leftarrow 1\|\pi} = 1$ holds with high probability.

It is uncertain whether $Y_{A\Leftarrow 0}$ and $Y_{A\Leftarrow 1\|\pi}$ take the same value.

3. Extension for addressing latent confounders

Marginal probabilities $\hat{p}_\theta^{A\Leftarrow 0}$ and $\hat{p}_\theta^{A\Leftarrow 1\|\pi}$ are difficult to estimate when there are unobserved variables called latent confounders.

Nevertheless, if their lower and upper bounds are available, we can achieve individual-level fairness using the following penalty:

$$G_\theta(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = \hat{u}_\theta^{A\Leftarrow 1\|\pi}(1 - \hat{l}_\theta^{A\Leftarrow 0}) + (1 - \hat{l}_\theta^{A\Leftarrow 1\|\pi})\hat{u}_\theta^{A\Leftarrow 0}$$

Experimental results

We compared our method with the following four baselines:

1. **FIO** [3]: constrains the expected value of PSEs
2. **PSCF** [4]: aims to reduce the conditional expected value of PSEs
3. **Unconstrained**: imposes no fairness constraint or penalty
4. **Remove** [5]: not use any features that are affected by sensitive feature

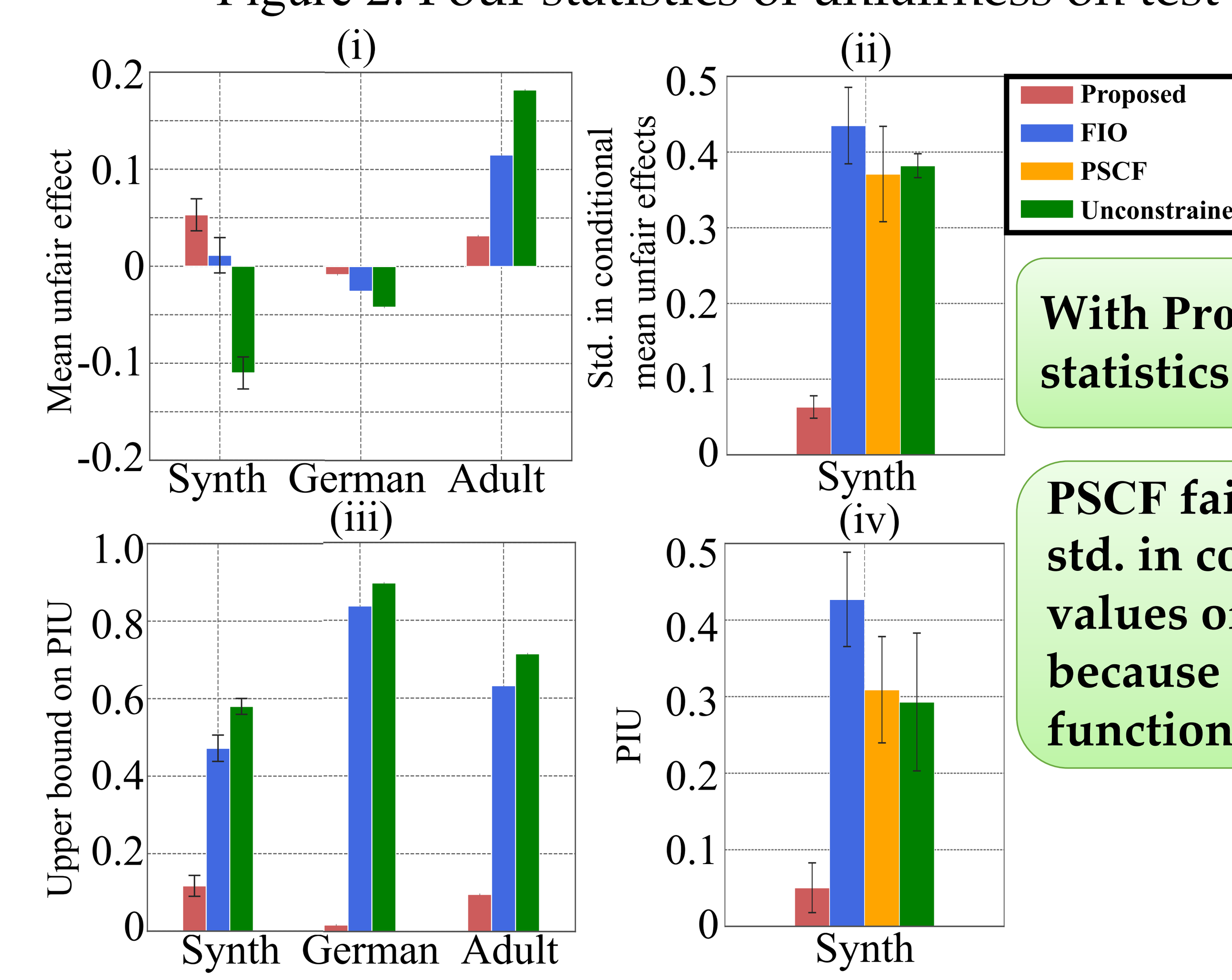
Table 2 and Figure 2 shows the test accuracy and the four statistics of unfairness: (i) **the expected value of PSEs**, (ii) **the std. in conditional expected values of PSEs**, (iii) **Upper bound on PIU**, and (iv) **PIU**.

Table 2: Test accuracy (%) on each dataset

Method	Synth	German	Adult
Proposed	80.0 ± 0.9	75.0	75.2
FIO	84.8 ± 0.6	78.0	81.2
PSCF	74.8 ± 1.6	76.0	73.4
Unconstrained	88.2 ± 0.9	81.0	83.2
Remove	76.9 ± 1.3	73.0	74.7

Proposed achieved comparable accuracy to PSCF.

Figure 2: Four statistics of unfairness on test data



With Proposed, all unfairness statistics were close to zero.

PSCF failed to reduce the std. in conditional expected values of PSEs (i.e., (ii)) because the data violates the functional assumptions.

References

[1] Chen Avin, Ilya Shpitser, and Judea Pearl. "**Identifiability of path-specific effects.**" In IJCAI, pages 357–363, 2005.

[2] Yongkai Wu, Lu Zhang, Xintao Wu, and Hanghang Tong. "**PC-fairness: A unified framework for measuring causality-based fairness.**" In NeurIPS, pages 3399–3409, 2019.

[3] Razieh Nabi and Ilya Shpitser. "**Fair inference on outcomes.**" In AAAI, pages 1931–1940, 2018.

[4] Silvia Chiappa and Thomas PS Gillam. "**Path-specific counterfactual fairness.**" In AAAI, pages 7801–7808, 2019.

[5] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. "**Counterfactual fairness.**" In NeurIPS, pages 4066–4076, 2017.

For more details, please check out our paper!

https://arxiv.org/abs/2002.06746