02

Learning individually fair classifier based on causality

Ask me how to make a fair decision for everyone

Abstract

Machine learning predictions are increasingly used to make critical decisions that severely affect people's lives, including loan approvals, hiring, and recidivism prediction. For this purpose, we developed a novel machine learning technology that makes predictions that are accurate and fair with respect to sensitive features such as gender, race, religion, and sexual orientation. To achieve high prediction accuracy, we utilize prior information about societal demands for each decision-making scenario, e.g., "rejecting applicants based on physical strength is fair if the job requires physical strength." Although existing methods cannot ensure fairness when the data are not generated by a restricted class of functions, our proposed method can use various data to guarantee fairness. Thus, admitting that "what is fair" depends on a particular sense of societal values, we create innovative machine learning technologies that can more flexibly respond to societal demands by bridging the gap between technical limitations and societal needs. In this way, we hope to mold a society that can make automatic decisions while ensuring that nobody will suffer detrimental treatment.

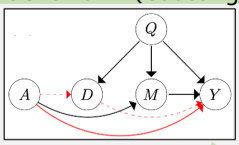
Problem: How can we build an AI that makes fair predictions (decisions) for individuals?

Input — Training data

A Gender	Q Qualification	D Department	M Physical strength	Y Decision
Female	A	Economics	B	Accept
Male	B	Literature	B	Accept
Male	C	Science	C	Reject

Sensitive feature

Prior information about "**What is fair**" (Causal graph)



Arrows provided by experts or estimated from data

Learn

Output: **Classifier (AI) that makes fair predictions**

Examples
- Hiring
- Loan approval
- Child abuse screening
- Recidivism prediction

1. Achieve high prediction accuracy

Using prior knowledge about societal demands for fairness **to learn an AI without imposing unnecessary constraints** and **to achieve high prediction accuracy**

E.g., hiring decisions for physically demanding jobs

1. Making decision Y based on gender A is **unfair** (A→Y)
2. Making decision Y based on department D is also **unfair** (A→D→Y)
3. Making decision Y based on physical strength M (which is necessary) is **fair** (A→M→Y)

Although **3. yields gender difference in rejection rates, it is unnecessary to impose a constraint on it**, which only decreases prediction accuracy.

2. Guarantee individual-level fairness using various data

Building an AI that makes **individually fair** predictions **regardless of what functional model generates data**

Reduce unfair differences in decision outcomes to zero for each individual

Original data (C.V.)

A Gender	Q Qualification	D Department	M Physical strength	Y Decision
Female	A	Economics	B	Accept

Reduce differences in acceptance probabilities to zero

Hypothetical data (Modified C.V.)

A Gender	Q Qualification	D Department	M Physical strength	Y Decision
Male	A	**Science**	**B**	Accept

Change gender to male — **Likely department for the male applicant** — **Leave unchanged since physical strength is irrelevant to discrimination**

Difficulty

To modify C.V., we must express true data-generating processes, which are impossible to approximate if data are not generated from simple functional models

Proposed method

- Propose unfairness measure that can be computed regardless of data-generating processes
- We can make an unfair difference in decision outcomes zero by forcing this unfairness measure to be zero

References

[1] Y. Chikahara, S. Sakaue, A. Fujino, H. Kashima, "Learning Individually Fair Classifier with Path-Specific Causal-Effect Constraint," in *Proc. the 24-th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2021.

Contact

Yoichi Chikahara / Learning and Intelligent Systems Research Group, Innovative Communication Laboratory
Email: cs-openhouse-ml@hco.ntt.co.jp