Learning individually fair classifier based on causality

02 因果関係に基づく一人一人にとって公平な分類器の学習

みんなにとって公平な決め方、教えます

どんな研究	融資承認や人材採用、囚人の累犯予測など、人間に対する重大な意思決定を機械学習予測によって行うための研究です。この実現に向け、性別・人種・宗教・性的指向など、人間が持つセンシティブな特徴に関して公平で、かつ高精度な予測を行うための技術を考案しました。
どこが凄い	例えば「体力を要する職種なので体力が理由の不採用は差別的でない」といった要請に基づき、不要な制約なしに学習することで予測精度を高めます。従来法ではデータが特定のモデルから生じている場合のみ公平性を保証可能でしたが、提案法では様々なデータで公平性を保証できます。
めざす未来	「何が公平か」は人や社会の価値観とともに移り変わるものです。社会的な要請と技術的な限界に向き合いながら、より柔軟に社会的要請に応じられる機械学習技術を実現し、一人一人が不利益を被ることなく、同時に、効率的な意思決定を行える社会に貢献できればと考えています。

問題：人間に対して公平な予測（意思決定）を行うAIを学習

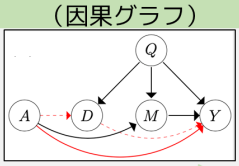
入力

訓練データ

A 性別	Q 資格	D 専攻	M 体力	Y 採否
女性	A	経済	B	採用
男性	B	文学	B	採用
男性	C	理工	C	不採用

センシティブな特徴

「何が公平か」を表す事前知識（因果グラフ）



矢印の有無・方向は専門家が与える or データから推定

学習

出力

公平な予測を行う分類器（AI）

応用例
- 人材採用
- 融資の承認
- 児童の虐待検知
- 囚人の累犯予測

ポイント１：高精度な予測を実現

個々の意思決定に求められる要請を
事前知識として活用することで
不要な制約を課すことなく学習し
高精度な予測を実現

例：体力を要する職種における人材採用の場合

1. 性別Aによって採否Yを決定するのは**差別的**（A→Y）
2. 専攻Dによって採否Yを決定して採用率に性差が生じるのは**差別的**（A→D→Y）
3. （体力は必要なので）体力Mによって採否Yを決定するのは**差別的でない**（A→M→Y）
   1., 2.によって生じる採用率の性差には制約を課す必要があるが
   3.によって生じる採用率の性差に制約を課すのは不要（予測精度を下げてしまうだけ）

ポイント2：様々なデータに対し個人公平性を保証可能

どのような関数形（モデル）から生じたデータでも
各個人一人一人に対し予測が公平になるように学習できる

各個人ごとに、差別による決定結果のちがいをゼロにする

元のデータ（履歴書）

A 性別	Q 資格	D 専攻	M 体力
女性	A	経済	B

採用確率の差がゼロになるように学習

仮想的なデータ（書き換えた履歴書）

A 性別	Q 資格	D 専攻	M 体力
男性	A	理工	B

性別を男性に書き換える　尤もらしい専攻に書き換える　体力は差別と関係ないのでそのまま

難しさ

履歴書を適切に書き換えるためにはデータの生成式を正しく表す必要があり、扱いやすい関数形（モデル）から生じたデータでなければ近似不可能☹

提案法

- データの生成式によらず推定できる差別度を新たに提案
- この差別度をゼロにすれば差別による決定結果のちがいをゼロにできる

関連文献

[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.

連絡先

近原 鷹一 （Yoichi Chikahara） 協創情報研究部 知能創発環境研究グループ
Email: cs-openhouse-ml@hco.ntt.co.jp