＜研究レポート＞

一人一人にとって公平な予測を因果効果制約を用いた機械学習

京都大学大学院・情報学研究科
（NTTコミュニケーション科学基礎研究所・協創情報研究部）
近原　鷹一

融資承認や人材採用など，個人に対する意思決定を機械学習で行う際，その予測が各個人の性別や人種について公平であることは重要である．本研究では因果効果制約を用いて，各個人に公平な予測を行うための学習技術を提案した．

公平かつ高精度なAIをめざして

近年のAI・機械学習技術の目覚ましい発展に伴い，融資の承認や罪人の釈放，企業における人材採用など，個人に対する重要な意思決定の問題を機械学習によって行う事例が増えつつある．例えば，みずほ銀行とソフトバンク社による共同出資で設立された株式会社J. scoreでは，個人向け融資のためのスコアリングを機械学習で行うサービスを2017年に開始している．機械学習を活用することで高精度に個人に対する意思決定を行うことが期待されるが，予測精度を優先するあまり，予測が人種・性別・障害・性的指向など、センシティブな特徴に関して差別的になるという問題が指摘されている[1]．これは，学習に用いる訓練データ(つまり人間による過去の決定結果の履歴であるが)に偏りがあり，そのようなデータに即して予測を行うことで，予測結果が差別的になってしまうためである．本研究では，各個人一人一人にとって公平で，かつ高精度な予測を行うため，予測の不公平性を経路特異的因果効果[2]という量で定量し，これに制約を課しながら予測精度を最大化する学習技術を提案した．

因果効果による不公平性の定量

初めに，予測結果の不公平性を効率的に測る尺度である経路特異的因果効果[2]について述べる．これは，個人の各特徴と予測結果の間の因果関係を矢印で表した，因果グラフに基づいて定義される．例えば，体力を要する職(消防士など)において採用/不採用を決定するケースを考えよう．このケースでは，性別によって採否を決定するのは性差別的であるが，体力を要する職であることから，性別に大きく影響される体力によって採否を決定することは性差別的でないとされる場合がある．各応募者の性別，資格，体力，採否結果を表す変数を$A, Q, M, \hat{Y}$とおくと，この場合の因果グラフは，例えば図１のように与えられ，上述の「性別によって採否が決定される場合のみ差別的である」という人間の事前知識は，因果関係$A \to \hat{Y}$を”不公平な経路”とみなすことで表現できる．経路特異的因果効果は，このような特定の因果関係の経路がどの程度強い影響をもたらすかを各個人に対して定義するものである．これは２つの”反実仮想的な状況”における予測結果の差として定義される．例えば，体力を要する職の採否予測の場合，「性別を女性に性転換した場合の予測結果」と「男性に性転換しつつも体力は女性に性転換した際のものを持つ場合の予測結果」の違いとして定義される．一般に，このような量をデータから推定することは非常に困難であるが，いくらかの仮定の下で，個人の集団における平均値を推定することは可能である．

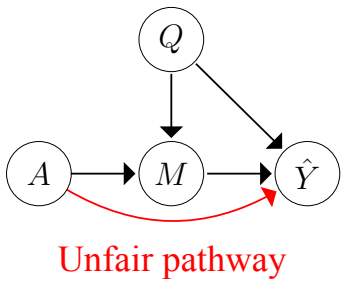

図1: 体力を要する職において採用予測を考える際の因果グラフ

既存技術の問題点

このため，既存文献[3]では因果効果の平均値に制約を課しながら予測精度を最大化する学習技術を提案している．しかしこのような学習技術では，集団全体でみれば平均的に差別は無いものの，一部の個人にとっては著しく差別的な予測をする場合があるという問題がある．　本研究では，この問題を解決し，各個人一人一人に対して公平で，かつ高精度な予測を実現するための，新たな学習のフレームワークを提案する．

提案技術の概要

提案技術[4]では，与えられた因果グラフに基づいて，各個人に対する因果効果がゼロになるように学習を行う．ここで，因果グラフは人手で与える，もしくは既存技術を用いてデータから推定するなどして事前に用意する必要がある[5]．提案技術では，因果効果の平均値に代わる量として，個人に対する因果効果がゼロにならない確率（PIU）を考え，PIUがゼロになるよう制約を課しながら学習を行う．PIUはデータから推定できない量であるが，本研究ではPIUの上界，すなわちPIUより常に大きな値をとる量であるが，これをデータから推定できる形で導出した．このPIUの上界がゼロになるように制約を課せば，データから推定できないPIUの値がゼロになるように制約を課すことが可能になり，結果として，各個人一人一人に対して公平な予測を行うよう，学習することを可能にした．

評価実験

提案技術の有効性を示唆する評価実験結果として，COMPASデータ[1]とAdultデータ[6]を用いた実験結果について述べる．前者は，アメリカ合衆国のいくつかの州で用いられている，囚人の累犯(犯罪を繰り返すか否か)を予測するスコアリングシステムCOMPASに関するデータであり，人種A，過去の犯罪歴M,年齢などその他の情報$\boldsymbol{C}$などのデータが含まれる．一方後者は，アメリカ合衆国における国勢調査のデータであり，性別A，結婚歴M，学歴L，勤務時間など職業に関する情報$\boldsymbol{R}$，性別と国籍$\boldsymbol{C}$などのデータが含まれる．

表1: 実データを用いた評価実験結果

データセット	手法	予測精度(%)	平均因果効果	PIUの上界値
COMPASデータ	提案技術	65.2	3.09×10^{-5}	$\mathbf{1.29\times10^{-3}}$
	既存技術[3]	65.5	$\mathbf{-7.40\times10^{-6}}$	0.492
	公平性制約なし	**66.3**	3.02×10^{-2}	0.844
Adultデータ	提案技術	76.6	$\mathbf{5.95\times10^{-4}}$	$\mathbf{1.11\times10^{-4}}$
	既存技術[3]	78.1	2.41×10^{-2}	0.111
	公平性制約なし	**80.0**	0.204	0.966

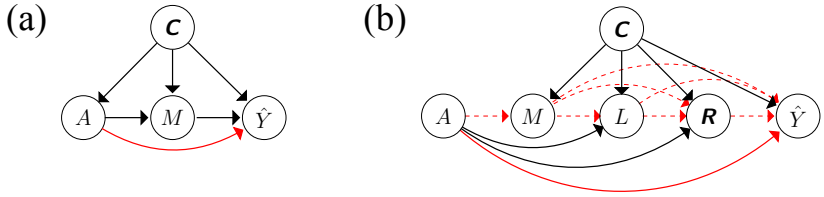

図2: (a)COMPASデータで囚人の累犯予測を行う際の因果グラフ (b)Adultデータで個人の年収予測を行う際の因果グラフ

COMPASデータにおいて累犯するか否か, Adultデータにおいて年収が50,000ドル以上か否かを予測した際の結果を$\hat{Y}$とするとき， 因果グラフをそれぞれ図２(a), (b)のように与えた．ここで，COMPASデータでは人種Aをセンシティブ特徴として因果関係$A\rightarrow\hat{Y}$を不公平な経路とし，Adultデータでは性別Aをセンシティブ特徴として，因果関係$A\rightarrow\hat{Y}$のみならず，Aから$\hat{Y}$に至る経路のうち，結婚歴Mを経由するものを全て不公平とした．これらの実データを用いて，提案技術の予測精度と平均因果効果とPIUの上界値を，既存技術[3]および何も公平性に関する制約を用いず予測精度を優先した場合と比較したところ，表１のようになった．既存技術[3]では平均因果効果がゼロになるように学習するが，これではPIUの上界値が大きくなることがわかり，したがって各個人に対して予測の公平性を保証できていないことがわかる．一方，提案技術では，平均因果効果だけでなくPIUの上界値も限りなくゼロに近い値になっており，また公平性制約なしに学習した場合と比べて僅かな精度の減少に留まっていることがわかる．これらは，提案技術が各個人に対して公平，かつ高精度な予測を行っていることを示唆している．

おわりに

本研究では，公平かつ高精度な予測を行うための因果効果制約を用いた学習技術を提案した．因果効果は，予測結果の不公平性を，人間の事前知識に基づいて効果的に定量することを可能にするが，一方でその推定は困難を極める．今後，さらに幅広い状況で公平かつ高精度な予測を行うための検討を実施する．

＜参考文献＞

[1] J. Angwin, J. Larson, S. Mattu, L. Kirchner; Machine Bias, 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[2] C. Avin, I. Shpitser, J. Pearl; Identifiability of path-specific effects. IJCAI, pages 357–363, 2005

[3] R. Nabi, I. Shpitser; Fair inference on outcomes. AAAI, pages 1931–1940, 2018.

[4] Y. Chikahara, S. Sakaue, A. Fujino, H. Kashima; Learning individually fair classifier with path-specific causal-effect constraint. arXiv, 2002.06746, 2020.

[5] C. Glymour, K. Zhang, P. Spirtes; Review of causal discovery methods based on graphical models. Frontiers in Genetics, 10, 2019.

[6] K. Bache, M. Lichman; UCI machine learning repository Datasets. http://archive.ics.uci.edu/ml/datasets, 2013.