2023/1/23

データ倫理の規範形成

アルゴリズムの公平性について

福家 佑亮　立命館大学　非常勤講師
fukuya.yusuke.86n@gmail.com

本発表の概要

- 本発表が扱うアルゴリズムの公平性の問題を明確化する
- アルゴリズムの公平性の中心問題が公平性の不可能性定理であることを確認する
- 不可能性定理に対する既存の対応に検討を加える
- 不可能性定理に対する悲観的な反応を退ける

本発表の目次

- アルゴリズムの公平性という問題
- 公平性基準とその両立不可能性
- 不可能性定理への応答
- まとめ

アルゴリズムの公平性という問題

本発表で扱う問題

- データに基づいた予測アルゴリズムpredictive algorithmの普及

➡金融、教育、医療、司法の領域で予測アルゴリズムが浸透。

- 差別的広告やCOMPASなど実際にアルゴリズムの公平性は大きな問題となっている

➡ACM(Association for Computing Machinery)では2018年以降アルゴリズムの公平性を扱ったカンファレンスが毎年開かれているなど社会的関心が高まっている。

本発表で扱う問題

- アルゴリズムにかかわる様々なバイアス

(1)データのラベリングに関わるバイアス(Annotator bias)

(2)データのサンプリングに関わるバイアス(Sampling bias)

(3)人種差別や性差別などのデータに内在するバイアス(Historical bias)

他にMeasurement biasやInherited biasなど…(Hellström et al. [2020])

本発表で扱う問題

- 本発表で問題にしたいバイアス

➡帰納バイアス(inductive bias・learning bias)。データから学習した予測モデルによる推論が不適切(e.g. 集団毎で偽陽性と偽陰性の割合が異なるなど…)。

- 帰納バイアスと他のバイアスの関係

➡本発表では扱わない(Hellman [2020] や Lippert-Rasmussen[2022] 等を参照)

本発表で扱う問題

- アルゴリズムの予測の公平性に焦点をあてる

➡予測と、予測に基づいた行動や決定に関係する問題は区別可能(Hellman [2020]; Hedden [2021]; Beigang [2022])。

- 予測に基づいてどのように行動・決定すべきかは文脈も重要

➡民事裁判と刑事裁判の立証基準の違い(Hellman [2020])。

本発表で扱う問題

- 帰納バイアスを取り除くための具体的なアルゴリズムや対処案についても考察の対象外

- 本発表で扱う問題

➡帰納バイアス(の予測)の公平性

公平性基準とその両立不可能性

具体例：COMPAS

- COMPAS(Correctional Offender Management Profiling for Alternative Sanctions)とは

➡再犯リスクを10段階で評価する予測アルゴリズム。

- NPOの報道機関であるProPublicaによる告発

➡ COMPASは白人に比べて黒人に不利な仕方で再犯リスクの評価を行っているとProPublicaは批判。

具体例：COMPAS

偽陽性：誤ってリスクが高いと評価

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

偽陰性：誤ってリスクが低いと評価

- **偽陽性の割合が黒人の方が高く、偽陰性の割合が白人の方が高い**

➡黒人に対して不利であるというProPublicaの主張(Angwin et al. [2016])。

具体例：COMPAS

- ProPublicaの主張

➡偽陽性と偽陰性の割合が不平等であるのは公平に反する(Angwin et al. [2016])。

- Northpointの主張

➡予測の全体的な正確性は人種間で違いがないのでCOMPASは公平。実際、再犯予測の合致率は人種に関わらず約60%であった(Dietrich et al. [2016]; Flores et al. [2016])。

アルゴリズムの公平性基準について

- この両者の対立をどう理解するか？

➡両者の主張とも間違っていない。両者はEqualized oddsとCalibrationという異なる公平性基準に訴えかけている

- 一般にアルゴリズムの公平性には複数の基準が存在する

➡しかも複数の基準を同時に満たすことは不可能であることが数理的に証明されている(Hunter and Schmidt [1976]; Kleinberg et al. [2016]; Corbett-Davies and Goel [2018])。

アルゴリズムの公平性基準について

- アルゴリズムの公平性基準はいくつ存在するか？

➡約10～20個提案されている（Verma and Lubin [2018] ; Hedden [2021] ）

取り上げられることが多い5つの基準（Kamishima[2022]）

(1)Fairness through unawareness

Fairness through awareness
- (2)Individual fairness（Counterfactual fairness）
- Group Fairness
  - (3)Calibration（Sufficiency）
  - (4)Equalized odds （Separation）
  - (5)Statistical parity（Demographic parity）

アルゴリズムの公平性基準について

- **Equalized odds**

$$P(\hat{y}|y\&a) = P(\hat{y}|y\&\acute{a})$$
$$P(\hat{y}|\sim y\&a) = P(\hat{y}|\sim y\&\acute{a})$$

➡目的変数が与えられた場合、保護された属性にかかわらず予測値が等しい

A: 保護された属性・センシティブ情報$(a, \acute{a} \in A)$

y:目的変数

$\hat{y}$:アルゴリズムによる予測値

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Equalized odds を満たしてない

アルゴリズムの公平性基準について

- **Calibration**

$$P(y|\hat{y}\&a) = P(y|\hat{y}\&\acute{a})$$
$$P(y| \sim \hat{y}\&a) = P(y| \sim \hat{y}\&\acute{a})$$

➡予測値が与えられた場合、保護された属性にかかわらず、目的変数が等しい

- 具体例：COMPAS

Higher risk(予測値)と推定された人は、人種(保護された属性)にかかわらず、実際に再犯する確率(目的変数)が60%と等しい

公平性の不可能性定理

Equalized odds

$$P(\hat{y}|y\&a) = P(\hat{y}|y\&\acute{a})$$
$$P(\hat{y}|\sim y\&a) = P(\hat{y}|\sim y\&\acute{a})$$

Calibration

$$P(y|\hat{y}\&a) = P(y|\hat{y}\&\acute{a})$$
$$P(y|\sim \hat{y}\&a) = P(y|\sim \hat{y}\&\acute{a})$$

特殊な状況を除いて、Equalized oddsとCalibrationを同時に満たすことは不可能(他の基準間でも両立不可能性が証明されている)
➡公平性の不可能性定理the impossibility theorems of fairness

不可能性定理への応答

不可能性定理に対する悲観的な反応

‘any assignment of risk scores can in principle be subject to natural criticisms on the grounds of bias.’

（Kleinberg et al. [2016] ）

‘The implications of the impossibility results are huge…… The goal of complete race or gender neutrality is unachievable.’ （Berk et al. [2018] p.20 ）

‘Race neutrality is not attainable,’ （Mayson [2019] p.2238 ）

不可能性定理に対する応答

①実は公平性概念は対立していない

(Holm [2022] ; Castro et al. [2022])

②両立不可能な公平性基準の中から最も妥当な公平性を決定する(Hedden [2021] ; Long [2021] ; Loi and Heitz [2022] ; Eva [2022])

不可能性定理に対する応答①

- Broomeのfairness概念を援用して公平性概念が対立していないことを示す（Holm [2022] ）。

‘What, the, does fairness require? It requires, I suggest, that *claims should be satisfied in proportion to their strength*.’
（Broome [1990] p. 90 ）

➡fairnessとは請求の比例的充足

Broomeの公平性概念

- アルゴリズムの公平性概念の文献においてもある程度支持を受けているが・・・(Holm [2022] ; Grote and Keeling [2022] p. 91)

<u>問題点</u>

Broomeのfairnessは個人をベースとした議論だが、 CalibrationやEqualized oddsといったGroup-fairnessの議論にBroomeの議論が適用可能か疑わしい(Castro et al. [2022])

不可能性定理に対する応答②

- もう1つの応答は、両立不可能な公平性基準の中から最も妥当な公平性を選ぶこと（Hedden [2021] ; Long [2021] ; Loi and Heitz [2022] ; Eva [2022] ）

- 議論の状況

➡Calibration、Equalized odds、 Statistical parity、いずれも公平性の基準として必要十分ではない。

不可能性定理に対する応答②

- Equalized oddsを例に考えてみる

➡Equalized oddsが必要になるのはbase rates(基準率)の違いがあるから

- 犯罪率の男女比の偏りは人種間のそれよりも大きい(Long [2021])

➡base ratesの違いから男性の方が偽陽性率が高く出ているはずだが、男女間でEqualized oddsを達成することは本当に公平？

不可能性定理に対する応答②

- CalibrationやStatistical parityについても、思考実験等を通じて同様の指摘が行われている

- 以上のことは結局、悲観的な反応が正しいことを示しているか？

➡必ずしもそうではない

公平性の意味

- 1つのあり得る解釈

➡アルゴリズムの公平性と道徳的な意味での公平性を分けるべき

- アルゴリズムの「公平性」の「公平性」の意味

➡アルゴリズムの公平性が問題となるとき、「公平性」という言葉は非常に形式的な意味しかもっていない(Castro [2022])。

公平性の意味

- アルゴリズムの公平性と道徳的な意味での公平性は場合によっては重なるときもあるが基本的には区別可能

➡ただアルゴリズムの(不)公平性は道徳的な意味での(不)公平性の何らかの証拠evidenceにはなりうる(Fleisher [2021])

- だとすると、公平性の不可能性定理は、道徳的にそれほど憂慮すべき事態でない？

残りの論点

- アルゴリズムの公平性と差別の規範理論(前田 [2021])や統計的差別statistical discriminationとの関係

- COMPASをモデルケースとする功罪

➡医療分野などの他の領域では別のバイアスやアルゴリズムの公平性が問題となる可能性

まとめ

- アルゴリズムの公平性基準は複数

➡両立不可能な基準の存在が数理的に証明

- アルゴリズムの公平性が問題となるとき、道徳的な意味での公平性とは異なる可能性

➡アルゴリズムの公平性を満たせなくともそれほど悲観的になる必要はない？

参考文献

- Angwin, Julia; Larson, Jeff; Mattu, Surya; and Kirchner, Lauren. 2016. ‘Machine Bias.’ *ProPublica* : https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing （2023年1月19日最終アクセス）
- Berk, Richard, et al. "Fairness in criminal justice risk assessments: The state of the art." *Sociological Methods & Research* 50.1 (2021): 3-44.
- Beigang, Fabian. "On the Advantages of Distinguishing Between Predictive and Allocative Fairness in Algorithmic Decision-Making." Minds and Machines (2022): 1-28.
- Broome, John. "Fairness." *Proceedings of the Aristotelian Society* 91(1990): 87-101.
- Dietrich, William; Mendoza, Christina; and Brennan, Tim. 2016. ‘COMPAS Risk Scales:Demonstrating Accuracy Equity and Predictive Parity.’ Technical Report, Northpointe, July 2016. https://www.documentcloud.org/documents/2998391-ProPublica-Commentary-Final070616.html (2023年1月19日最終アクセス)

参考文献

- Fleisher, Will. "Evidence of Fairness: On the Uses and Limitations of Statistical Fairness Criteria." *Available at SSRN 3974963* (2021).
- Flores, Anthony W., Kristin Bechtel, and Christopher T. Lowenkamp. "False positives, false negatives, and false analyses: A rejoinder to machine bias: There's software used across the country to predict future criminals. and it's biased against blacks." *Federal Probation* 80 (2016): 38-46.
- Grote, Thomas, and Geoff Keeling. "On algorithmic fairness in medical practice." *Cambridge Quarterly of Healthcare Ethics* 31.1 (2022): 83-94.
- Hedden, Brian. "On statistical criteria of algorithmic fairness." *Philosophy and Public Affairs* 49.2 (2021): 209-231
- Hellman, Deborah. "Measuring algorithmic fairness." *Virginia Law Review* 106.4 (2020): 811-866.

参考文献

- Hellström, Thomas, Virginia Dignum, and Suna Bensch. "Bias in Machine Learning--What is it Good for?." *arXiv preprint arXiv:2004.00686* (2020).
- Holm, Sune. "The Fairness in Algorithmic Fairness." *Res Publica* (2022): 1-17.
- Hunter, John E., and Frank L. Schmidt. "Critical analysis of the statistical and ethical implications of various definitions of test bias." *Psychological Bulletin* 83.6 (1976): 1053-1071
- Kamishima, Toshihiro. Fairness-Aware Machine Learning and Data Mining: https://www.kamishima.net/jp/kaisetsu/(2023年1月19日最終アクセス)
- Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan. "Inherent trade-offs in the fair determination of risk scores." *arXiv preprint arXiv:1609.05807* (2016).

参考文献

- Lippert-Rasmussen, Kasper. "Using (Un) Fair Algorithms in an Unjust World." *Res Publica* (2022): 1-20.
- Loi, Michele, and Christoph Heitz. "Is calibration a fairness requirement? An argument from the point of view of moral philosophy and decision theory." *arXiv preprint arXiv:2205.05512* (2022).
- Long, Robert. "Fairness in machine learning: against false positive rate equality as a measure of fairness." *Journal of Moral Philosophy* 19.1 (2021): 49-78.
- Mayson, Sandra G. "Bias in, bias out." *The Yale Law Journal* 128.8 (2019): 2218-2300.
- Verma, Sahil, and Julia Rubin. "Fairness definitions explained." 2018 IEEE/ACM international workshop on software fairness (fairware). IEEE, 2018.
- 前田春香. "アルゴリズムの判断はいつ差別になるのか: COMPAS 事例を参照して." 応用倫理 12 (2021): 3-21.