アルゴリズムによる「差別」をなくすには？: 公平かつ正確なアルゴリズムの設計に向けて Liang et al. (2022)

奥村恭平 *

December 6, 2022

1 アルゴリズムの公平性に対する需要の高まり

アルゴリズムによる意思決定がますます盛んになってきている。動画やニュースなどのコンテンツ配信、ソーシャルメディアや EC における広告配信はもちろんのこと、金融、裁判、監視にいたるまで、アルゴリズムによる予測や分類を用いた意思決定が爆発的に広がっている。

そのような状況の中、アルゴリズムが一部の集団に属する人々に対し「不公平」な決定をしている可能性が指摘され始め、政府・企業・大学など (以下、「為政者」と呼ぶ) は対応を求められている。2016 年には、保釈を認めるか否かの決定の指針としてアメリカの一部の州で用いられていた再犯予測アルゴリズムにおいて、黒人の被告人の偽陽性率（将来の犯罪のリスクが高いと誤って分類すること）が白人の被告人の 2 倍であることが指摘され議論を呼んだ (Angwin et al., 2016)。

しかし、為政者にとって公平性だけが重要な指標である訳ではない。アルゴリズムの公平性を追求しようとすると、アルゴリズムの正確性などの他の指標との間にトレードオフが生まれることが予想される。アルゴリズムの正確性と公平性の間のトレードオフはどのような関係にあるのだろうか。どのような場合に正確性を犠牲にして公平性を追求することが正当化されるのだろうか。また、現実では公平性担保を理由に「特定の特徴量 (例えば人種) の使用を禁止する」という政策がしばしば観察されるが、これは果たして妥当な政策なのだろうか。

2 設定

為政者は意思決定を下すためのアルゴリズムを設計したい。アルゴリズムは、ある個人の観察可能な特徴量 X を入力として、行動 $A \in \{0,1\}$ を出力する関数 f として表せる。[*1] 各個人は (事前には観察不可能な) 特徴量 Y を持っており、行動 A と特徴量 Y が決まると損失 $\ell(A,Y)$ が定まる。アルゴリズムが正確であるほど、損失は小さくなる。例として、患者に対しある治療法を試すか否かを決めるアルゴリズムを設計することを考える。この場合、X は通院歴や血液検査結果などの患者に関する情報、$A \in \{1,0\}$ は実際に治療を受けたか否か、$Y \in \{1,0\}$ は治療が必要か否かとなる。損失関数 ℓ としては、例えば誤分類の数が考えられる。[*2]

さらに、各個人は特徴量 X に加えその個人が属する集団を表す特徴量 $G \in \{r,b\}$ を持っており [*3] (先述の例では、所得の高低などが G の例として考えられる)、各個人の特徴量 (X,Y,G) は同一の分布に独立に従っているとする。すると、集団 $g \in \{r,b\}$ のアルゴリズム f における損失 $e_g(f)$ が次のように定義される:

$$e_g(f) := \mathbb{E}\left[\ell(f(X),Y) \mid G = g\right].$$

為政者が自由にアルゴリズム f を選べるとき実現可能な損失の組の集合を $\mathcal{E}(X)$ で表す。[*4] アルゴリズムを

* kyohei.okumura@gmail.com

*1 アルゴリズムが確率的に行動を選ぶことを許容している。また、行動は二種類しかないと仮定している。

*2 この場合、損失関数は $\ell(a,y) := \mathbb{1}\{a \neq y\}$ となる。ただし、$\mathbb{1}$ は指示関数を表す。

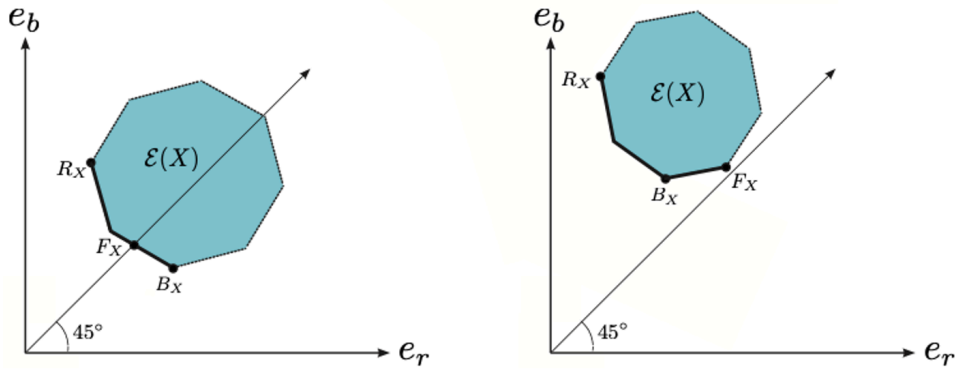
*3 r は"red", b は"blue" の略である。差別に関する論文では、political correctness への配慮からか、black/white, rich/poor などといった表現ではなく、より「価値中立的」な表現が好まれる傾向がある。過去に筆者が見た論文では"○,□" (丸、四角) で集団を表しているものもあった。

*4 $\mathcal{E}(X) := \left\{(e_r(f), e_b(f)) \in \mathbb{R}^2 : f \in (\Delta(\mathcal{A}))^{\mathcal{X}}\right\}$, ただし、$\mathcal{X}$ は確率変数 X の値域. $\mathcal{E}(X)$ は閉凸多角形となることが示せる。

適切に選択・設計することで、為政者は $\mathcal{E}(X)$ 内の損失の組 (e_r, e_b) を一つ実現することができる。為政者は各集団の損失の組に関する選好を有しており、為政者が公平性をどの程度重視するかによって、どのような損失の組を好むかが変わりうる。例えば、功利主義的な為政者は社会全体の損失の和 $p_r e_r + p_b e_b$ (ただし、p_g は集団 g が人口に占める割合) の最小化、平等主義的な為政者は集団間の損失の差 $|e_r - e_b|$ の最小化を望む。[*5]

3 集団平衡性とその含意

ここで一つ、データの持つ統計的性質に関する重要な概念を導入しよう。各集団 g について、集団 g の損失を可能な限り小さくするようにアルゴリズムを設計したとき g の損失が他の集団 g' の損失より小さくなるとき、特徴量 X は**集団平衡**であると言うことにする。また、そうならない時、つまり、ある集団 g について、g の損失を可能な限り小さくするようにアルゴリズムを設計したとしても g の損失が他の集団 g' の損失より大きくなるとき、特徴量 X は g'**-非平衡**であると言う。



(a) 集団平衡 **(b) r-非平衡**

Fig. 1: 集団平衡性 (Liang et al. (2022), Figure 3 より)

図 1 (a) は X が集団平衡であるときの、図 1 (b) は X が r-非平衡であるときの例になっている。図中の R_X は集団 r の損失が最小になる損失の組、B_X は集団 b の損失が最小になる損失の組、F_X は最も公平な損失の組を、太線はパレートフロンティアをそれぞれ表している。ただし、ここでのパレート効率性は、正確性と公平性という二つの基準を元に定められている。[*6] つまり、為政者の公平性に対する選好がどのようなものであれ、妥当な選好を持つ為政者であれば皆、(1) 両集団について損失がより少ない (より正確)、(2) 集団間の損失の差がより少ない (より公平) という二条件が同時に満たされるアルゴリズムをより好むはずだと考えており、そのような為政者は必ず太線内のどれか一点を実現するアルゴリズムを選ぶ。

Liang et al. (2022) は、パレートフロンティアの形状が集団平衡性と密接に関連していることを主張する次の定理を示した:

定理 1. パレートフロンティアは、特徴量 X が集団平衡であるとき R_X と B_X を結ぶ直線より下の部分に、特徴量 X が r-非平衡であるとき R_X と F_X を結ぶ直線より下の部分になる。

この定理によれば、アルゴリズムの正確性と公平性に関するトレードオフについては、実は図 1 (a), (b) のような二種類の状況しか存在しないことになる。このことから特に、公平性を高めるために両集団に対する正確性をどちらも犠牲にすることは、特徴量 X が集団非平衡である場合 (F_X がフロンティア上のある点よりも右上に位置しうる場合) にしか正当化されないことがわかる。

特に重要なケースとして、集団に関する特徴量 G がアルゴリズムの入力として使用可能な場合が考えられる。次の命題は、為政者の選好がどんなものであれ、集団に関するデータを追加的に使用することで、X の下で不利な集団の損失を小さくできることを主張している。[*7]

(Liang et al., 2022, Lemma B.1.)

[*5] ここで与えられた公平性の定義は、既存研究で提唱されてきた複数の公平性の尺度を一般化したものになっている。詳細は、Liang et al. (2022) の Appendix を参照されたい。

[*6] (e_r, e_b) が (e'_r, e'_b) を**パレート支配**するとは、(1) $e_r \leq e'_r, e_b \leq e'_b$, (2) $|e_r - e_b| \leq |e'_r - e'_b|$, かつ、少なくとも一つの不等号が厳密に成立することとして定義される。

[*7] 実は Liang et al. (2022) ではもう少し強い主張が示されている。詳細は元論文の Corollary 3 を参照されたい。

命題 1. 為政者の選好が妥当であり、[*8] かつ、X が g'-非平衡であるとする。このとき、(X, G) が入力として使える場合に為政者が選ぶアルゴリズムにおける集団 g の損失は、X のみが入力として使える場合に為政者が選ぶアルゴリズムにおける集団 g の損失より小さくなる。

3.1 為政者とアルゴリズム設計者が異なる場合: 公開データのデザイン

Liang et al. (2022) では、為政者 (データを収集・公開する主体, 例えば政府) とアルゴリズム設計者 (例えば企業) が異なる場合についても考察している。アルゴリズム設計者は正確性のみを気にしている一方で為政者は公平性も重んじており、為政者が元のデータ (X, G) を「加工」した上で公開することができるとしたとき、為政者はどのようにしてデータを公開すべきだろうか。[*9]

紙面の都合上モデルの詳細は省くが、ある仮定のもとで、データの加工の仕方をうまく工夫すると、為政者は自分でアルゴリズムを設計する場合に自身が選択する損失の組と同じものをアルゴリズム設計者に選ばせることができることが示せる。また、「為政者はある特定の特徴量に関するデータを加工前のデータから消去すべきか?」という問いに対して、次のような結果を得ている:

命題 2. 特徴量 X が集団平衡であるならば、どんな為政者も集団特徴量 G を加工前データに含むことを好む。

命題 3. (X, G) が加工前データに含まれている状況で、X の中に含まれる一部の特徴量 X' を加工前データから消去するか否かを考える。このとき、弱い条件のもとで、どんな為政者も X' を保持することを好む。

2021 年に、カリフォルニア州は州立大学入試において特定の人種や裕福な学生に有利であるという理由から、SAT や ACT などの筆記試験の点数を合否判断に使用しないことを決定した。これは果たして妥当な政策決定だろうか。[*10] アメリカの多くの州では、人種に関するデータの収集が許されている。命題 3によれば、この場合、いくら筆記試験の点数のデータについて人種ごとに偏りがあるとしても、為政者は点数のデータを保持することを好む。一方で、人種に関するデータが元からない場合 (カリフォルニア州はこのケースにあたる) 為政者の好みと特徴量の性質によっては、筆記試験の点数のデータを消去することが好まれるかもしれない。

なお、以上の分析ではあくまで「為政者が元のデータを加工した上でアルゴリズム設計者に公開する」という状況を考えており、無加工のままデータを公開する場合に集団に関するデータを消去して公開すべきか否かという議論ではないことに注意されたい。

3.2 集団平衡性とは何なのか

これまで、集団平衡性と呼ばれる性質がアルゴリズムの公平性と正確性のトレードオフを分析する上で重要な役割を果たすことを見てきた。ではそもそも、どういうときに特徴量 X は集団非平衡になるのだろうか。

Liang et al. (2022) は集団平衡性について「特徴量 X が一部の集団にとってより有益な情報を伝える場合、X は集団非平衡になる」と言及している。以下例を見てみよう。今、人々が高所得層と低所得層に分けられるとし、過去の通院回数 $X \in \{多, 少\}$ をもとに施術するか否かを決めるアルゴリズムを設計しているとする。低所得層は、仮に治療が必要 $(Y = 1)$ であったとしても、病院に行かないことが高所得層と比べて多いとすると、通院回数 X はどちらの層についても Y に関して同様に有益な情報であり、かつ、高所得者についてより有益な情報を与える。このようなとき、X は高所得層-非平衡な特徴量となることが示せる。

4 おわりに

Liang et al. (2022) は、損失を「負の効用」とも解釈可能な形で定義した上で、アルゴリズムの正確性と公平性についてのパレートフロンティアを求め、入力データの統計的性質との間の関係を分析するという手法を提案した。このことは公平性の尺度や為政者の選好の詳細に左右されずに公平性と正確性のトレードオフを一般的に議論することを可能にした。主な結果として、集団平衡性というデータの持つ統計的特性が公平性/正確性パレートフロンティアの形状と密接に関係している (定理 1) ことが示されており、また、「データから特徴量を排除することが良いことである」という主張を公平性を理由に擁護することは一般的には難しそうである (命題 1, 2, 3) ことが示唆されている。行動が二値より多くの値を取り得る場合や、行動に追加的な制約 (例えば、大学入試の場合は定員があるため、合格を出せる人数に制約がかかる) がある場合などのより複雑なケー

[*8] 為政者の選好が注*6で定義したパレート支配の関係と整合的であるという意味。

[*9] ここでの「加工」とは、数学的には garbling(例えばde Oliveira (2018) を参考にされたい) のことを指している。(1) ある特徴量を元のデータから消した上で公開する, (2) データの粒度を粗くする (例: cm 単位で測られていた身長のデータを、「高」「低」の二値にして公開する), (3) 元のデータにノイズを加えて公開する　などの操作は全て garbling の一例である。

[*10] "University of California Will No Longer Consider SAT and ACT Scores," *The New York Times*, May 15, 2021.

スにおいて本論文と同様の分析がどれだけ適用可能か検討することは、今後の研究課題の一つである。

References

Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner, "Machine bias," in "Ethics of Data and Analytics," Auerbach Publications, 2016, pp. 254–264.

de Oliveira, Henrique, "Blackwell's informativeness theorem using diagrams," *Games and Economic Behavior*, 2018, *109*, 126–131.

Liang, Annie, Jay Lu, and Xiaosheng Mu, "Algorithmic design: Fairness versus accuracy," in "Proceedings of the 23rd ACM Conference on Economics and Computation" 2022, pp. 58–59.