The 34th Annual Conference of the Japanese Society for Artificial Intelligence, 2020

分布変動に対する公平な機械学習アルゴリズムの頑健性

Robustness of Fair Machine Learning Algorithm against Distribution Shift

福地 一斗 *1*2
Kazuto Fukuchi

*1筑波大学 University of Tsukuba
*2理化学研究所 革新知能統合研究センター RIKEN AIP

Fairness in machine learning is a problem where the learned machine learning model outputs biased decisions against individuals' sensitive attributes, such as race and gender, and has been recognized as a crucial problem in the machine learning community. Many researchers hence have devoted to the development of fair machine learning algorithms. Basically, these algorithms are specifically designed for the situation where the sample distributions between training and test phases are equivalent. However, such an assumption may not hold in a practical situation. For example, suppose we build a fair machine learning model from the applicants' resumes from five years ago to predict hiring decisions. Then, we can easily imagine that the rule of the hiring decision may change from five years ago to now because of the change in social situations. When the sample distribution changes, a model learned with the training sample might be unfair in the test sample. In this paper, we assess the possibility that such a situation occurs even with a small change of sample distributions. To this end, we develop an algorithm that generates a test sample distribution in which the learned model would be unfair. Also, we demonstrate by empirical experiments that the developed algorithm can generate the unfair test sample distribution against the existing fair learning algorithms.

1. はじめに

深層学習の登場などによって機械学習の技術が広く使われるようになってきている一方で，機械学習が行う予測が性別，人種，信仰などに偏ってしまうことが問題視されている．例えば，Amazon.com Inc は AI リクルーティングツールとして，履歴書にからその人の能力をスコアとして予測するツールの開発を行っていた．このスコアが高いほど履歴書の人物の能力が高いことを表し，このようなスコアを採用者に提示することによって採用プロセスの負荷を軽減することができる．しかしながら，このツールを解析したところ，「女性の」という単語が入っていたり，2 つの女子大出身だったりするとスコアに罰則を加えることがわかった [Dastin 18]．採用が性別で偏ることは法的にも問題になる懸念があるため，Amazon はこれを修正しようと試行錯誤を行った．しかし，修正することができず，最終的に簡易的なツールを使う方針に切り替えた．このように機械学習の出力が個人のセンシティブな属性によって偏る問題を，機械学習の公平性と呼ぶ．この問題意識は機械学習のコミュニティに広く浸透しており，公平な学習アルゴリズムの開発に関する研究が盛んに行われている [Kamishima 12, Hardt 16, Cotter 19]．これらの研究によって，公平性を保証しつつ機械学習の運用を行うことができるようになりつつある．

現状，公平な機械学習アルゴリズムを直ちに実際のサービスに組み込めるわけではなく，様々な課題を抱えている．その一つとして分布変動に対する頑健性の問題が挙げられる．多くの機械学習のアルゴリズムはデータを解析して予測モデルを構築する学習段階と，予測モデルを新たなデータに適用する予測段階の 2 つに分けることができる．例えば AI リクルーティングツールの例における学習段階では，過去の応募者の履歴書とその採否の情報を解析することで履歴書に対してスコアリングを行うツールを構築する．その後予測段階において，現在来ている応募者の履歴書に対して学習段階で構築したツールを適用することで，これから採否を決める応募者に対するスコアリングを行う．多くの機械学習の技術は，過去の応募者の履歴書と現在の応募者の履歴書が同じ傾向を持っており，同じルールで正しいスコアリングをすることができるという仮定のもとで構築されている．従って，過去の履歴書と現在の履歴書の傾向が変わってしまった場合，予測段階で行ったスコアリングが正しくない可能性がある．公平性においても同様で，公平な機械学習のアルゴリズムは学習段階において過去の履歴書に対して公平なスコアリングルールを出力するため，予測段階において現在の応募者の履歴書に対して行ったスコアリングは公平でない可能性がある．このような問題を，機械学習の分野においてはデータの傾向をデータを生成する確率分布で表現することから，**分布変動による不公平**と呼ぶ．分布変動に対する頑健性とは，学習アルゴリズムが分布変動に対する不公平を引き起こさず，耐性を持つという性質のことを指す．

分布変動によって公平性に関する問題が発生することが予想できるが，実際に機械学習のアルゴリズムが分布変動による不公平を発生させることへの調査は十分に行われていなかった．そこで本稿では，この分布変動に対する不公平が発生するのか実際に調査を行った．調査のために，実際に不公平を誘発する分布変動を求めるためのアルゴリズムの開発を行った．このアルゴリズムで求める分布変動は，不公平を誘発する変動の中でもなるべく小さい変動を求める．分布変動が小さいほど現実に発生しやすいと予測できるため，小さな分布変動を見つけることによって分布変動による不公平の問題のリスクがどれぐらい大きいか評価する．

本稿の主な貢献は，開発したアルゴリズムを用いて分布変動による不公平の調査した結果，少しのデータの書き換えによって不公平が誘発されてしまうことを明らかにしたことである．具体的には，開発したアルゴリズムを用いて [Kamishima 17] で提案されている公平な学習手法を攻撃し，不公平が誘発され

連絡先: 福地 一斗，筑波大学 システム情報系，fukuchi@cs.tsukuba.ac.jp

るか Adult データセットを使った数値実験で評価した．実験の結果，1 つの特徴量を書き換える程度の分布変動によって不公平が誘発されてしまうことを示した．

2. 分布変動による不公平

2.1 分布変動が起こる原因

実際の現実的な問題として分布変動が起こりうる状況をいくつか例として挙げ，分布変動に対処する際の課題を整理する．はじめの分布変動が起こりうる例として，機械学習の予測モデルを適用する人の層が変化する状況が挙げられる．

例 1 (適用対象の変化) ある大学 A の数理系学科は，数理系の試験の点数をもとに男性，女性ともに応募者の半分が合格するように合否ルールを設計したとする．大学 A の運営の上ではこの合否ルールはうまくいっていたため，他の大学 B に対して合否ルールを提供し大学 B の理数系学科もこの合否ルールを運用したとする．大学 A での運用実績から，大学 B でも男性，女性ともに応募者の半分が合格することが期待できる．しかしながら，例えば大学 B の周辺地域に理数系が強い男子高校があるなどの状況があった場合は，女性よりも男性の応募者に対する合格割合が大きくなる可能性がある．この例では，大学 A への応募者と大学 B の応募者の傾向が変わる分布変動が起きており，これによって合否ルールが不公平になってしまっている．

この例ではルールの適用対象が，大学 A から大学 B に変化することによって分布変動が起こった．適用対象の変化は機械学習の技術の応用を考えると，現実的に簡単に起こりうると想像できる．例えば，自社のある Web サービスのデータを用いて学習を行い，そこで得られたルールを使って自社の別のサービスを提供したとする．サービスの提供する内容が，例えば，学習段階では車やバイクに関連した SNS であり，予測段階では化粧品の EC サイトであったとするとサービスを利用する人の層がまったく異なることが予想できる．その他にも，他社が公開している機械学習のアルゴリズムで得られた予測モデルなどを自社のサービスに組み込むなどといった状況では分布変動が起こると考えられる．

他の分布変動が起こりうる例として，社会情勢が変化する状況が挙げられる．

例 2 (社会情勢の変化) ある企業 A は過去に収集した履歴書のデータを機械学習のアルゴリズムを使って解析することで，履歴書にスコア付けを行い採用にかかる負担を減らすためのツールを導入している．過去のデータでは過去の社会情勢において形成された公平性の基準でもって公平な採択を行っており，それを反映したデータになっているとする．一方，現在では世間の公平性に対する認識が高まってきたこともあり，公平性の基準がより厳しくなっていたとする．すると，過去の基準で問題なかったルールで採否を決定した場合，現在の基準に照らし合わせると不公平になってしまう可能性がある．

社会情勢の変化による分布変動は，機械学習の技術を導入したシステムの管理者自身が分布変動を認識できていない可能性がある．従って，自動で分布変動に対処してくれるようなシステムの構築が必要になってくる．

ではこれらの問題に対して，技術的に解決できうる範囲はどこまでなのかが様々な制度設計などでは問題になってくる．次の節では関連研究をもとにどのような状況ならば対処可能であるか考察を行う．

2.2 実現可能な公平性の頑健性

分布変動の問題はすでに，機械学習や最適化の分野において分布頑健性 [Calafiore 06]，ドメイン適合 [Redko 19]，転移学習 [Yang 20] などと呼ばれ盛んに研究が行われている．分布頑健性では可能性のある分布変動の最悪ケースにおいて予測性能が良くなるように学習を行うことによって，分布変動に対する頑健性を実現する．一方で，ドメイン適合・転移学習では，予測時のデータを使って予測モデルを修正することによって分布変動に対する頑健性を実現する．これらの技術をもとに，公平性の分布変動に対する頑健性をもつ学習アルゴリズムも構築することができると考えられる．

一方で，分布頑健性，ドメイン適合，転移学習の知見を応用したとしても限界があり，最終的には公平性の分布変動に対する頑健性と予測性能はトレードオフの関係になる場合が出てくる．このような状況では，公平性への要求を強めるほど予測性能を犠牲にするため，あまり強い公平性基準を要求すると産業的価値を著しく損なうことにつながる．このバランスをとるためには，政治的，法的などといった様々な観点からの検討が必要不可欠である．

3. 攻撃アルゴリズム

ここでは実際に構築した攻撃アルゴリズムを説明するために，問題の定式化を行う．入力 $X \in \mathbb{R}^d$ からラベル $Y \in [k]$ を予測する分類問題において，個人のセンシティブな属性を $S \in [m]$ に対して公平性を保証することを考える [*1]．S に対して公平になるように学習した分類器を関数 $f : \mathbb{R}^d \to \mathbb{R}^k$ と表し，入力 X に対する予測ラベルは $F(X) = \arg\max_{y\in[k]} f_y(X)$ で決定する．ここで，f_y は f の出力の y 番目の要素を表す．

分布変動は，データに以下の式で表す線形変換を行うことによってシミュレートする．

$$g(X) = X + WX, \tag{1}$$

ここで，$W \in \mathbb{R}^{d\times d}$ はパラメータである．

アルゴリズムの目的は，不公平を誘導しかつ分布変動が少なくなるような g のパラメータ W を求めることである．不公平性の指標として以下の式で定義される demographic parity [Calders 10] と呼ばれる評価指標を用いる．

$$\mathrm{DP} = |\mathbb{P}\{Y = 1|S = 1\} - \mathbb{P}\{Y = 1|S = 0\}|. \tag{2}$$

ただしここでは，$k = 2$ かつ $m = 2$ としている．DP は F を用いており，$\arg\max$ が入っているためパラメータ W について微分不可能関数である．そこで，実際には F の $\arg\max$ を scale パラメータを掛けた softmax 関数で近似した関数 $\tilde{F}(X) = \mathrm{softmax}(\gamma f(X))$ を用いて評価した不公平指標 $\tilde{\mathrm{DP}}$ を最大化する，ここで，γ は scale パラメータを表しており，大きいほど近似の精度が良い．分布変動の大きさとして，以下の式で定義される指標を用いる [*2]．

$$\mathrm{WD} = \frac{1}{2}\mathbf{E}\big[\|X - g(X)\|_1\big]. \tag{3}$$

これは Adult のような X が $\{0, 1\}$ の値をとるようなデータセットの場合，変化した特徴量の個数を期待的に評価してい

*1 自然数 m について $[m] = \{1, ..., m\}$

*2 ベクトル $y \in \mathbb{R}^k$ について softmax の出力は $k-1$ 次元確率シンプレクス上の値をとり，その i 番目の要素は $\mathrm{softmax}_i(y) = \exp(y_i)/\sum_j \exp(y_j))$ である．

る．またこの指標は，元の分布と g の押し出しよる分布の間の 1-Wasserstein 距離の定数倍を評価していることと等価である．Wasserstein 距離は分布間の距離尺度として広く用いられているため，評価量は分布変動の評価量として妥当性を持つ．

上記の 2 つ指標を使って，アルゴリズムをトレードオフを制御するパラメータ $\eta > 0$ を導入した以下の最適問題で定式化する．

$$\max_{W} \ \tilde{\mathrm{DP}} + 2\eta \mathrm{WD}. \tag{4}$$

W に対して spectral normalization [Miyato 18] と呼ばれるテクニックを導入し，g が極端な関数になって勾配ベースの最適化アルゴリズムでの最適化に悪影響を与えないようにした，実験では，上記の最適化問題を最適化アルゴリズム Adam を使って最適化した．

4. 分布変動に対する頑健性調査

実験設定.

第 3. 節で構築したアルゴリズムを用いて，実際に公平な学習のアルゴリズムの分布変動に対する耐性の調査を行った．ここでは，[Kamishima 17] で提案されている ROCSVM-AI という公平な学習アルゴリズムに対して，提案アルゴリズムを用いて分布変動をシミュレーションし公平性と分布変動の大きさの関係性を調査した．調査は UCI リポジトリで提供される Adult データセットで実験を行う [Dua 17]．Adult は，32561 件の訓練データと 16281 件のテストデータを含むデータセットになっている．

この実験では提案アルゴリズムによって作成された分布変動における不公平性と分布変動の大きさを評価する．不公平性の評価量として式 (2) で示される DP，分布変動の大きさとして式 (3) で定義される指標を用いる．トレードオフを制御するパラメータを $\eta \in \{2^{-5}, 2^{-4}, ..., 2^{5}\}$ と変化させたときのそれぞれの指標を評価する．

ROCSVM-AI のハイパーパラメータとして正則化パラメータがあるが，これは 10 に設定した．また，Adam の学習率と 2 つのパラメータ β_1, β_2 はそれぞれ 0.01，0.90.999 に設定した．また，提案法の scale パラメータは $\gamma = 10^4$ に設定した．

実験結果.

図 1 に実験結果を示す．図 1 の左図は提案アルゴリズムによって得られた不公平性と分布変動の大きさの関係性を表している．図からわかるように，学習された予測モデルの元のデータセットでの公平性は DP ≈ 0.001 であることに対して，元の公平性が WD が 1 未満で DP を 0.1 付近まで上昇していることがわかる．ROCSVM-AI のベースとなる学習アルゴリズムである SVM における不公平性は DP ≈ 0.16 となるため，WD が 1 未満でこれに匹敵するほど不公平にすることに成功している．

図 1 の右図は，横軸に η の変化による公平性，変動分布の大きさの変化を表している．図からわかるようにパラメータ η によって 2 つの間のトレードオフを制御することができている．公平性指標が $\eta = 2^{-3}$ で一度下がっているが，提案アルゴリズムで解いている最適化問題が非凸であるため，その影響によるものであると考えられる．

5. まとめ

本稿では，分布変動による不公平の問題を取り扱い，実際に機械学習のアルゴリズムが分布変動による不公平を発生させることへの調査を行った．調査のために，不公平をなるべく少ない変動で誘発させる分布変動を求めるアルゴリズムを開発した．開発したアルゴリズムを用いて分布変動による不公平の調査した結果，少しのデータの書き換えによって [Kamishima 17] のアルゴリズムで学習した予測モデルに対して不公平が誘発されてしまうことを明らかにした．

今回はデータを書き換えによって不公平を誘発する分布を生成したが，この方法は最近機械学習コミュニティで機械学習アルゴリズムの脆弱性の一つとして問題視されている**敵対的事例** [Goodfellow 15] に関連する．敵対的事例は，データを人間が気づかれない程度書き換えることによって予測モデルの出力を変化させることができてしまう問題である．今回の提案方法は，この性質を活用して少量のデータの書き換えによって不公平の誘発に成功したと考えることができる．敵対的事例の防御方法に関する知見を分布変動への対処に導入することなども考えられる．

参考文献

[Calafiore 06] Calafiore, G. C. and Ghaoui, L. E.: On Distributionally Robust Chance-Constrained Linear Programs, *Journal of Optimization Theory and Applications*, Vol. 130, No. 1, pp. 1–22 (2006)

[Calders 10] Calders, T. and Verwer, S.: Three naive Bayes approaches for discrimination-free classification, *Data Mining and Knowledge Discovery*, Vol. 21, No. 2, pp. 277–292 (2010)

[Cotter 19] Cotter, A., Jiang, H., Gupta, M., Wang, S., Narayan, T., Mobility Seongnam-si, K., and Korea Karthik Sridharan, S.: Optimization with Non-Differentiable Constraints with Applications to Fairness, Recall, Churn, and Other Goals, Technical report (2019)

[Dastin 18] Dastin, J.: Amazon scraps secret AI recruiting tool that showed bias against women (2018)

[Dua 17] Dua, D. and Graff, C.: {UCI} Machine Learning Repository (2017)

[Goodfellow 15] Goodfellow, I. J., Shlens, J., and Szegedy, C.: Explaining and Harnessing Adversarial Examples, in *International Conference on Learning Representations* (2015)

[Hardt 16] Hardt, M., Price, E., and Srebro, N.: Equality of Opportunity in Supervised Learning, in Lee, D. D., Sugiyama, M., Luxburg, von U., Guyon, I., and Garnett, R. eds., *Advances in Neural Information Processing Systems 29*, pp. 3315–3323, Barcelona, Spain (2016)

[Kamishima 12] Kamishima, T., Akaho, S., Asoh, H., and Sakuma, J.: Fairness-Aware Classifier with Prejudice Remover Regularizer, in Flach, P. A., Bie, T. D., and Cristianini, N. eds., *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD, Part II*, Vol. 7524 of *Lecture Notes in Computer Science*, pp. 35–50, Bristol, UK (2012), Springer

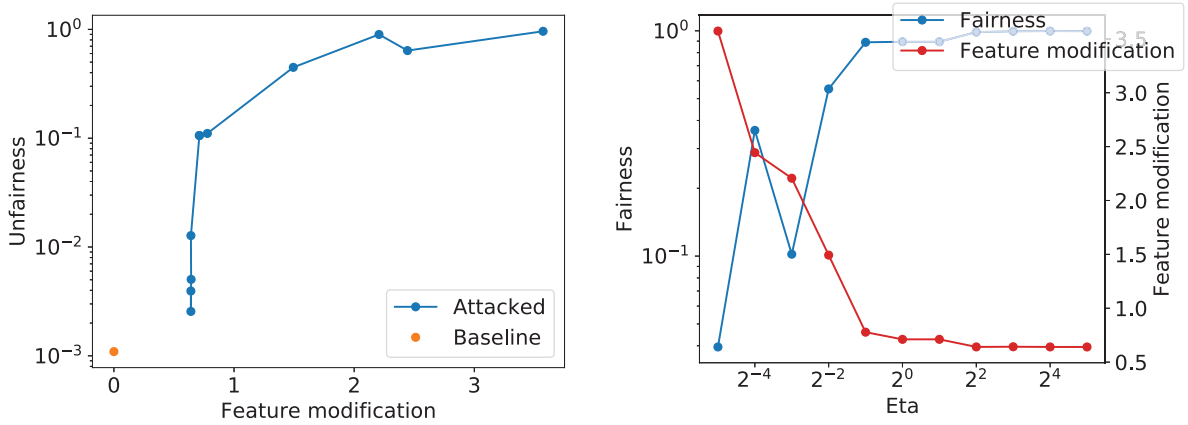

図 1: 実験結果．左図は縦軸が不公平性指標，横軸が分布変動の大きさを表している．オレンジの点で示す点でデータセットに変更を加えずに評価した結果を，青い線で提案法で得られた結果を示している．右図は，不公平性と分布変動の大きさを制御するためのパラメータ η を横軸に取ったときの公平性と分布変動の大きさを表している．青い線が公平性指標 $1 - \mathrm{DP}$ を表しており，赤い線が分布変動の大きさを表している．

[Kamishima 17] Kamishima, T., Akaho, S., Asoh, H., and Sakuma, J.: Model-based and actual independence for fairness-aware classification, *Data Mining and Knowledge Discovery* (2017)

[Miyato 18] Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y.: Spectral Normalization for Generative Adversarial Networks, in *6th International Conference on Learning Representations, {ICLR} 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, OpenReview.net (2018)

[Redko 19] Redko, I., Habrard, A., Morvant, E., Sebban, M., and Bennani, Y.: *Advances in Domain Adaption Theory*, Elsevier (2019)

[Yang 20] Yang, Q., Zhang, Y., Dai, W., and Pan, S. J.: *Transfer Learning*, Cambridge University Press (2020)