2C-01

公平なAI開発

髙橋 悠文†
アドソル日進株式会社 AI研究所‡

1. 序論

近年、AI 技術が発達し、社会のあらゆる場面で AI が利活用されることが増加している。それとともに、AI モデルが特定のグループに対して不公平な結果を出力してしまう事象が発生し、AI の公平性が社会的な課題となっている。産業技術総合研究所(以下産総研)は、AI を用いた製品・サービスの品質を管理するための指針となる機械学習品質マネジメントガイドライン(以下ガイドライン)を公開しており、その第 8 章では AI の公平性に関する品質マネジメントについて論じられている[1]。このように、AI 公平性に関する活動が活発に行われており、今後は実社会への適用が求められている。

本研究では、ガイドライン第 8 章を参考に公平性を考慮した AI 開発を行うことで、広くガイドラインユーザおよびガイドライン次版検討チームに対して、有益な課題・解決策の提供を目指す。

2. 開発概要

ガイドラインの 8.4 節では、AI 公平性品質確保のプロセスが示されている。このプロセスでは、正義、人権など抽象度の高いレベルにおける定性的な公平性要件の定義から開始し、AI 要素への要件などの下層の要件に具体化していく。それらの要件分析の後、学習に用いる具体的なデータが検討される段階で、定量的なメトリクス要件決定すなわち「定性→定量」へシフトし、その後、定められたメトリクスを満たすべく各種バイアス軽減アルゴリズムを活用する。本研究ではこのプロセスに従って公平性確保を目指す。

その際、より現実的な問題点を洗い出すため、リアルな架空のユースケースを想定する。今回作成する AI は、高額商品のレコメンドを行うかどうか決定するために、顧客アンケートの結果から年収が 5 万ドルを超えるか否か予想する AI であり、男女間の公平性実現を要件とする。よって要配慮属性は性別の属性である。

今回は AI 要素への定性的要件として以下の二つを定める。

1. 大人数に対する判断結果を集計した際に、不公平を感じさせるような結果にならないこと。
2. 男女の違いのみによって判断結果が変わらないこと。

さらに、ガイドラインに沿って実際のデータ内容も検討したうえで、1 の要件をもとに disparate impact(DI)、2 の要件をもとに different result rate(DRR)をメトリクスとして選択した。DI は、ぞれぞれの要配慮属性における良い結果を得る人の割合の比であり、DRR は、要配慮属のみ変更した際に出力が変わるデータの割合である。DRR は一般的なメトリクスではなく、今回独自に定義、実装する。識別モデルのアリゴリズムとしてはロジスティック回帰を用い、Reweighing[2]と Prejudice Remover[3]の二つのバイアス軽減アルゴリズムをそれぞれ適用する。Reweighing は学習用のデータに対して用いるアルゴリズムであり、Prejudice Remover は学習段階で用いるアルゴリズムである。独自定義の DRR 以外のメトリクス評価と、バイアス軽減アルゴリズム適用においては、IBM の AI Fairness 360 を用いる。バイアス軽減実施有無によるメトリクス値への影響を確認し、あらかじめ定義したメトリクス目標値などに従い最終的なモデル選択を行う。

3. 実施結果

バイアス軽減を行うことにより公平性メトリクスの値を改善することができた。ここでは例として、バイアス軽減を何も行わないモデルと Reweighing を適用したデータで学習したモデルのそれぞれにおける DRR と精度の値を図 1、図 2 に示す。

Development of fair AI
†Hisafumi Takahashi
‡Ad-Sol Nissin Corp. AI Research Institute

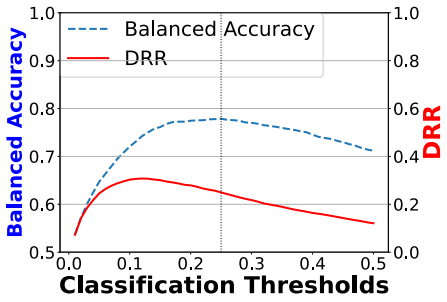

図 1　バイアス軽減を用いないモデル

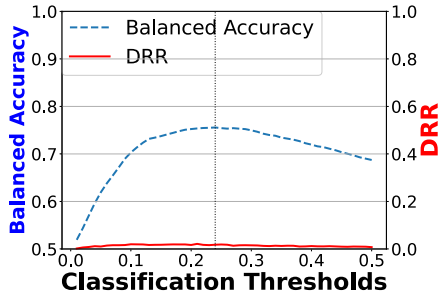

図 2　Reweighing を適用したモデル

Reweighing を用いた結果 DRR の値が低く望ましい値になっている。さらに、今回実装したすべてのモデルにおける精度および公平性メトリクス値について表 1 に示す。

表 1　全モデルの精度および公平性メトリクス

Bias Mitigator	Classier	bal_acc	DRR	1-min(DI, 1/DI)
	Logistic Regression	0.776	0.246	0.761
Reweighing	Logistic Regression	0.750	0.017	0.257
Prejudice Remover	Logistic Regression	0.722	0.171	0.128

バイアス軽減実施により、二つのメトリクスの値がどちらも改善された一方、精度(bal_acc)は下がっており、精度と公平性のトレードオフが確認できる。さらに、二つの公平性メトリクス間でもトレードオフが生じている。今回は次節で述べるメトリクス最低基準などを参考にし、Reweighing を適用したモデルを最終的に「公平な AI」として採用した。

4. 考察

本章では、開発を行ったうえで感じた課題とその解決策について述べる。まず、メトリクスの選択において、場合によっては今回の DRR のように自前のメトリクスの実装が必要になるという課題がある。この場合メトリクス算出のアルゴリズムの正確性は注意点である。それらしい数値が出力されているとミスには気付きづらいことが考えられるため、別途テストコードを用意するなどして正確性の確保に努めることが必要となる。今回の DRR においても、実装の前段階としてメトリクス算出の正確性を調べる手順を設けた。

さらに、最終的なモデル選択の際の課題について述べる。機械学習アルゴリズム、バイアス軽減アルゴリズム、公平性メトリクスの組み合わせは、多量の検証を行うほど多くなり、モデル選択の難易度が上がってしまう。そのため、あらかじめ明確なモデル選択のアルゴリズムを用意し、それに従って選択を行うことが重要と考えられる。例として、精度、メトリクスの目標だけでなく、最低基準を決めておくこと、メトリクスごとの重みを決め、最終的な評価指標を一つ算出することなどが挙げられる。各メトリクスでどの程度の数値が出力されるか分からない段階でこれらを行うことは難しいが、公平性を考慮しない場合のメトリクス算出を終えた後で、その値を参考に目標や最低基準を設定することが有用と考える。今回もその方法にて、二つのメトリクスおよび精度のそれぞれに対して目標の他に最低基準を設けた。

5. 結論

本研究では産総研の機械学習品質マネジメントガイドラインを参考に、公平性を考慮した AI モデルの開発を行った。バイアス軽減アルゴリズムによって公平性メトリクスの値を向上させ、精度と公平性のトレードオフや、異なるメトリクス間でのトレードオフを確認した。さらに考察として、開発における今後の課題について述べ、解決策を提案した。今後は、それらの解決策を検証し、その効果について調査する。

6. 参考文献

[1] 国立研究開発法人産業技術総合研究所, “機械学習品質マネジメントガイドライン”

[2] F. Kamiran and T. Calders, “Data Preprocessing Techniques for Classification without Discrimination,” Knowledge and Information Systems, 2012.

[3] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma, “Fairness-Aware Classifier with Prejudice Remover Regularizer,” Joint European Conference on Machine Learning and Knowledge Discovery in Databases, 2012.