The 35th Annual Conference of the Japanese Society for Artificial Intelligence, 2021

バイアス考慮型分類器の安定性に関する予備調査

Preliminary Investigation on the Stability of Bias-aware Classifiers

神嶌 敏弘 *1 Toshihiro Kamishima
赤穂 昭太郎 *1 Shotaro Akaho
馬場 雪乃 *2 Yukino Baba
鹿島 久嗣 *3 Hisashi Kashima

*1産業技術総合研究所 National Institute of Advanced Industrial Science and Technology (AIST)
*2筑波大学 University of Tsukuba
*3京都大学 Kyoto University

We here discuss a bias-aware classifier, which is designed to predict a class while removing the bias caused by the influence of specific information. Such type of classifiers are used for, say, taking fairness into account by removing socially sensitive information. We define the stability of such bias-aware classifiers as how similar predictions are made from the same information other than the bias-source information to be removed. In this paper, we collected the dataset for checking the stability through a crowdsourcing service.

1. はじめに

現状の公平性配慮型機械学習は，公平性の制約下でできるだけ正確な予測器を学習することを目標とする場合が多い．そして，公平性配慮型手法の開発では，予測精度と公平性の間のトレードオフを改善することに主眼が置かれている．しかしながら，予測精度は不公平なアノテーションが行われているデータ集合で評価しているため，公平な決定が行われる環境下での予測精度が評価できているかには疑問が残る．そこで，本研究では，公平性配慮方型の予測器に対する新たな望ましい性質として「安定性」を提案する．

本稿で公平性配慮型手法が安定であるとは，センシティブ情報以外の情報が全て同じであれば，同じモデルが学習されれることをさす．より詳細に，分類対象の特徴からクラスを予測する分類問題の場合を述べる．これらの特徴のうち，社会的公平性の観点から配慮が必要なセンシティブ情報を表すものをセンシティブ特徴とし，その他全ての特徴は非センシティブ特徴とする．このとき，公平性配慮型手法で学習したモデルではセンシティブ情報は除去されるため，非センシティブ情報が全て同じであれば，同じモデルが学習されるのが理想的といえよう．そうした理想的な分類器の状態をここでは安定的であるという．

この安定性を調べるために，統制されたデータ集合を作るのが今回の目的である．公平性配慮型分類器が扱うような，与信や入学といった問題は統制が困難なため，アイテムに関する嗜好判断を使う．ここでの目的では，公平性の確保だけでなく，任意の望ましくないバイアスを除去するという，より一般的な問題を扱っているので，嗜好判断を対象としても問題はない．よって，以後は公平性配慮方型手法を，より一般的な状況を扱うバイアス考慮型手法として参照する．

ここでは寿司に対する嗜好を扱い，クラウドソーシングを利用して被験者に一対比較法で嗜好データを収集した．さらに，センシティブ情報の代用として，認知バイアスを考える．そのために，被験者の選択が認知バイアスの影響を受けるように利用者インターフェースを設計した．今回は 2 種類の認知バイアス，位置効果 [Chandler 11] とバンドワゴン効果 [Eickhoff 18] を扱う．バイアス考慮型の手法は，この認知バイアスを除去するために用いる．そして，認知バイアスの種類が異なっていても，同じモデルが学習されるようであれば，そのバイアス考慮型手法は安定的であるとする．ここでは簡潔な手法を適用した予備的な実験結果を示し，今後の方向性を検討する．

本研究の貢献は次の通りである．

- バイアス考慮型手法の安定性を調べるための統制データを収集する．
- 簡潔な手法をこのデータに適用した予備実験結果を示す．

連絡先: ホームページ https://www.kamishima.net/

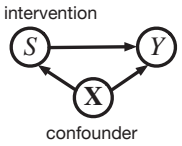
2. データ集合

データの生成モデルに続き，このデータの収集手続きに述べる．そして，このデータが認知バイアスの影響を受けていることを示す．

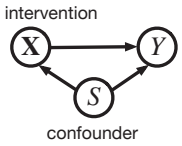
2.1 データの生成モデル

ここでは認知バイアスの分析で想定するモデルを示す．標準的な公平性配慮方型機械学習の枠組みと同様に，このモデルも S，$\mathbf{X}$，および Y の 3 種類の変数で構成される．S はセンシティブ特徴である．バイアス考慮型分類器で公平性に配慮する場合では，この変数は，性別や人種といった社会的にセンシティブな情報を表現する．しかしながら本研究では，認知バイアスをこのセンシティブ変数で表す．これは，本研究の目的では人間の決定に影響を与える任意の因子が実験に利用でき，また認知バイアスは統制が可能なためである．ここでは次の 2 種類の認知バイアスを検証した．一つは位置効果で，右上隅に近く表示されたアイテムがより頻繁に選択されやすいというものである [Chandler 11]．ここでは二つの寿司を水平方向に表示し，左側の寿司がより頻繁に選択されやすいと予測する．この場合，S は寿司が左右のどちらの枠に表示されたかを表す．もう一つは，バンドワゴン効果で，他の利用者がより好むことを示唆されたアイテムは選択されやすいというものである [Eickhoff 18]．ここでは一方の寿司に「人気」という印を付けて強調し，この強調された寿司がより頻繁に選択されやすいと予測する．この場合，S は寿司が強調されたかどうかを表す．

$\mathbf{X}$ は非センシティブ変数である．標準的なバイアス考慮型分類問題では，これらの変数は Y と S 以外の全ての変数に該当する．ここでは，$\mathbf{X}$ は具体的に X_1 と X_2 の二つの変数になる．これらの変数は被験者に表示される二つの寿司を表現する．これらは離散変数で，特定の寿司を指定するインデックス番号の値をとる．ここで，これらの変数以外にも，被験者が誰かや，



(a) 認知バイアスの効果を分析するためのモデル



(b) バイアス考慮型分類のためのモデル

図 1: 変数間の独立性のモデル化

(a) ランダム

(b) バンドワゴン

図 2: データ収集用の利用者インターフェース

NOTE: いずれのインターフェースでも，被験者に対して「どちらの寿司が好きか？」と尋ねた．「バンドワゴン」の場合では，「人気」と示すことで一方の寿司を強調した．

評価時刻など多様な因子がアイテムの選択に影響しているが，これらの因子は観測されていないことに留意されたい．

Y は，予測クラスを表す目的変数である．機械学習の公平性を扱う場合では，この変数は，入学，借款，雇用などの決定を表すが，ここでは，一方の寿司がもう一方より好まれるかどうかを示す．被験者が，寿司 X_1 より，寿司 X_2 を好んだ場合，この変数は 1 となり，そうでなければ 0 となる．

次に，認知バイアスの効果を測るためのモデルについて述べる．このモデルでの，変数間の依存性を表したのが図 1(a) である．すると因果推論の文脈では，S と Y はそれぞれ介入変数と効果変数に相当する．そして，被験者に提示するアイテムを表す $\mathbf{X}$ は，S と Y の交絡因子に相当する．

このモデルを用いて，もし介入を被験者に無作為に割り当てれば，認知バイアスの影響を測ることができる．具体的には，無作為に選択したアイテムを被験者に提示し，$S=0$ と $S=1$ の介入を被験者に無作為に割り当てる．この手続きは無作為化比較試験であるので，認知バイアスの影響は次式で測れる．

$$\mathrm{E}[Y|S{=}1] - \mathrm{E}[Y|S{=}0] = \Pr[Y{=}1|S{=}1] - \Pr[Y{=}1|S{=}0] \quad (1)$$

2.2 データ収集の手順

ここでは，一対比較法による嗜好データの収集手順について述べる．まず，過去の調査 [Kamishima 03] で用いた次の 10 種類の寿司から無作為に二つを選び，それを被験者に提示する．

トロ (FT) マグロ (Tn) エビ (Sh) イクラ (SR)　アナゴ (SE)
ウニ (SU) 鉄火巻 (TR) イカ (Sq) タマゴ (Eg) カッパ巻 (CR)

ただし，括弧内は寿司名の略号で，以下の実験結果の図表で用いる．過去の 5000 人を対象とした調査 [Kamishima 03] では，この一覧で示した順により好まれており，この順序を以下「人気順」として参照する．被験者のは，提示された二つの寿司を比較し，他方よりより好ましい方を選択する．

データの収集には，ベースラインとバンドワゴンの 2 種類の利用者インターフェースを用いた．ベースラインインターフェースでは，水平方向に並べて同じ大きさで二つの寿司を図 2(a) のように提示した．もう一方のバンドワゴンインターフェースでは，寿司をより大きく表示し，また「人気」というラベルを付けることで図 2(b) のように，一方の寿司が人気商品であるとして強調した．

表 1: 各データ集合の被験者数

ランダム	固定		バンドワゴン	
	左	右	人気	不人気
120	103	118	99	96

嗜好データは，次の 3 種類の手続きで収集した．一つ目は，ベースラインインターフェースを用いて各被験者に二つの寿司を提示し，また，どちらの寿司を左に表示するかは無作為に決める．すると，S は位置効果を示し，X_1 が示すアイテムを左に表示したとき $S{=}1$ となる．この手続きは理想的な無作為化比較試験に相当するため，位置効果の効果は厳密に式 (1) で測れる．この手続きを「ランダム」として参照する．

二つ目は，二つのグループの各被験者にベースラインインターフェースを用いて寿司を提示し，後に二つのグループのデータを併合する．一方のグループでは，人気順序で上位の寿司を常に左側に表示し，もう一方ののグループでは右側に表示する．比較する寿司の割当ては依然として無作為であり，グループへの被験者の割当ても無作為とみなせることから，この手続きも無作為化比較試験とみなせる．しかしながら，暗黙的な因子が完全に無作為化されているとは言い切れないので，完全なものとはいえない．例えば，逐次的に寿司を選択することで，被験者が以前の選択に影響される記憶効果や，左側がより選択されやすいことにより位置効果の効果がより強くなることなどがありうる．この手続きを，各被験者に対する寿司の配置は固定されていることから，「固定」として参照する．

三つ目も，被験者を二つのグループに分けるが，インターフェースにはバンドワゴンを用いる．一方のグループでは，人気順で上位の寿司を強調し，もう一方のグループでは下位の寿司を強調する．固定手続きと同様に，この手続きも不完全ではあるが無作為化比較試験とみなせる．この手続きを「バンドワゴン」として参照する．

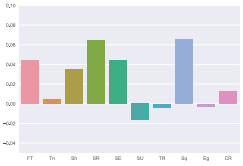
日本国内のクラウドソーシングを利用してデータを収集した．収集期間は 2020 年 1 月 31 日から 2 月 22 日までで，各被験者に 50 円を支払った．被験者あたりの質問数は 50 件である．そのうち二つは，右を選ぶよう要求する集中度テストで，この二つのテストを通過したデータのみを分析に用いた．各手続きで収集したデータ数を表 1 に示す．固定手続きの左右の列には，それぞれ人気寿司を左と右に表示した場合の数を示した．また，バンドワゴン手続きの左右の列には，それぞれ人気と不人気寿司を強調した場合の数を示した．

2.3 認知バイアスの評価

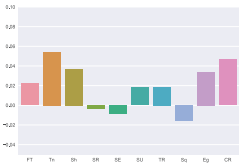
ここでは，式 (1) によって，認知バイアスの総因果効果を測る．総因果効果とは，介入変数から目的変数までの全ての因果パスの効果の総和のことである．なお，公平性配慮方型機械学習の文脈では，この効果は risk difference に相当する [Calders 10, Žliobaitė 17]．各手続きでの認知バイアスの効果を表 2 に示す．総因果効果が正の値になるのは，ランダムと固定手続きの場合は左側の寿司が選択されたことを，バンドワゴン手続きの場合では強調した寿司が選択されたことを示す．この表から寿司の選択が，認知バイアスに影響されていることがこ

表 2: 認知バイアスの総因果効果

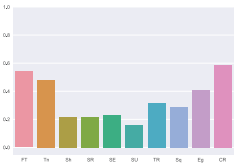
ランダム	固定	バンドワゴン
0.0229	0.0077	0.3451



(a) 「ランダム」手続き



(b) 「固定」手続き



(c) 「バンドワゴン」手続き

図 3: 各手続きでの認知バイアスの効果

NOTE: バンドワゴン手続きでの認知バイアスの効果の大きさが他の二つの手続と差があるため，Y 軸の尺度を変更している．

の表から確認できる．また，位置効果の効果はバンドワゴンのそれより強い．この結果は，文献 [Eickhoff 18] で報告されている，バンドワゴン効果が他の効果より強いという結果と整合している．

認知バイアスについてさらに調査するために，各寿司ごとの効果を評価する．各寿司 i について，この寿司 i が他の寿司より好まれるとき，目的変数 Y_i は 1 になり，それ以外では 0 になる．各寿司ごとの認知バイアスの効果を図 3 に示す．位置効果の場合では，特に明確な傾向は観察されない．一方で，バンドワゴン効果の場合では，特に人気がある，もしくは特に人気のない寿司が，より強く認知バイアスに影響されている．この観察結果も，位置効果よりバンドワゴン効果の方が強いという文献 [Eickhoff 18] の報告と整合している．

3. バイアス考慮型分類器の安定性

ここでは，因果推論とバイアス考慮型分類器との関係について論じたあと，認知バイアスの除去についての予備実験結果を示す．

3.1 因果推論の観点から見たバイアス考慮型分類器

ここでは，バイアス考慮型分類器と因果推論の関係について論じ，認知バイアスを除去する簡潔な技法について述べる．ここではグループ公平性や statistical parity [Calders 10, Žliobaitė 17, Dwork 12] と呼ばれる条件，形式的には Y と S の独立性，$Y \perp\!\!\!\perp S$ で定義される条件を対象とする．因果推論の文脈では，この条件は，S から Y への総因果効果を除去することに相当する [Zhang 18].

このバイアス考慮型分類器のモデルを因果推論の観点から解釈する．図 1(b) にあるように，今度は $\mathbf{X}$ ではなくセンシティブ変数 S を交絡因子，S ではなく非センシティブ変数 $\mathbf{X}$ を介入変数として扱う．ここで，$\mathbf{X}$ の Y への直接的な効果を保存したまま，S を通じたパスの効果を除去したい．因果推論では，交絡因子 S の値でデータ集合をまず層化しておくことで，この目的を達成できる [岩崎 15, 6 章]．このため，S の値が等しいデータを集めた層ごとに統計量を求める．そして，これらの統計量を，各層のデータ数に比例する重みで重み付けて和をとることで集約する．これは重み $\Pr[S]$ を用いた次の荷重和で表せる．

$$\Pr[Y|\mathbf{X}] = \sum_s \Pr[S{=}s] \Pr[Y|S{=}s, \mathbf{X}] \tag{2}$$

単純に $\mathbf{X}$ の全ての値に対してこの値を計算することで，S に含まれる認知バイアスの影響を除去できる．

ここでこの処理は，後処理型のバイアス考慮型分類と関連があることを指摘しておく [Kamiran 12, Hardt 16]．後処理型では，データ集合を S の値が等しいものごとに分割し，その分割したデータ集合ごとに分類器を学習する．そして，$\Pr[Y|\mathbf{X}, S{=}1]$ と $\Pr[Y|\mathbf{X}, S{=}0]$ の差を最小化するように，分類器の決定境界値を修正する．式 (2) より，本来の $\Pr[Y|\mathbf{X}]$ の式から，層化によって $\Pr[Y|\mathbf{X}, S{=}1]$ と $\Pr[Y|\mathbf{X}, S{=}0]$ とはともに $\Pr[Y|\mathbf{X}]$ に強制的に変更されている．よって，層化はバイアス考慮型の一手法とみなせる．

また文献 [Kamishima 18] の，バイアス考慮型の生成モデルとも関係がある．Y，$\mathbf{X}$，および S の同時分布に，statistical parity の条件 $S \perp\!\!\!\perp Y$ を加えると次式のように変形できる．

$$\begin{aligned}\Pr[Y, \mathbf{X}, S] &= \Pr[S] \Pr[Y|S] \Pr[\mathbf{X}|Y, S] \\ &= \Pr[S] \Pr[Y] \Pr[\mathbf{X}|Y, S] \quad (\leftarrow Y \perp\!\!\!\perp S) \\ &= \Pr[S] \Pr[Y|\mathbf{X}, S] \Pr[\mathbf{X}]\end{aligned} \tag{3}$$

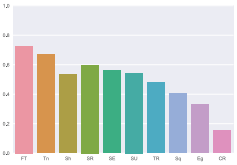
$\Pr[\mathbf{X}]$ は特定の入力に対して定数であるため，この式を S で周辺化すると，式 (2) を得る．このことから，statistical parity の条件を満たす同時分布を考えることと，S を交絡因子としてその効果を層別解析で除去することは等価とみなせる．

3.2 バイアス考慮型手法の安定性の実験結果

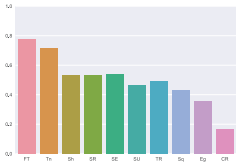
この節では，上記のバイアス除去手法を適用してモデルを学習し，学習したモデルの安定性を報告する．そのため，$\mathbf{X}$ の各値，すなわちアイテムのあらゆる対に対して，式 (2) を適用する．こうして，ランダム，固定，およびバンドワゴンの 3 種類のデータ集合から，それぞれ確率行列を得る．なおこの確率行列の要素は，行にある寿司より列にある寿司より好まれる確率となる．

認知バイアスは除去され，寿司の種類などの他の情報は変わらない．よって理論的には，バイアス除去技術が安定的であれば，これらの行列はデータ集合の収集手続きによらず等しくなるはずである．この仮説を検証するため，各手続きで得た確率行列の絶対距離を，ランダムと固定，およびランダムとバンドワゴンの間で求めた．前者の固定との間はやや小さく 0.0521 だが，後者のバンドワゴンとの間は大きく 0.127 であった．層化による手法が安定的であれば，これらいずれの距離も小さく，また同様の値になるはずだが，得られた値はかなりの差があった．

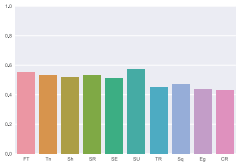
この結果をさらに詳細にしたものが図 4 である．この図には，各寿司が他のすしより好まれる確率を，層化によってバイアス除去した値を示してある．認知バイアスは除去してあるので，より人気のある寿司がより好まれるようになるはずである．しかしながら，グラフ 4(c) は，他の二つのグラフとは明らかに異なっている．



(a) 「ランダム」手続き



(b) 「固定」手続き



(c) 「バンドワゴン」手続き

図 4: 層化でバイアス除去した，各寿司が他の寿司より好まれる確率

以上のように，認知バイアスを除去することで同様の結果が得られることを期待していたが，実際に得られたデータではかなり乖離がある結果となった．

3.3 バイアス考慮型手法の安定性に関する議論

上述のように，仮説とは異なりバイアス除去後に得られたモデルにはかなりの相違があった．この点に関し，現状では次の二つの可能性を検討している．一つは，層化によって認知バイアスの効果は十分に除去されてない可能性である．もう一つは，認知バイアス以外の想定していないバイアスの影響が存在する可能性である．

まず，認知バイアスの効果が十分に除去されていない可能性について検討する．分析においては，例えばトロとイカの二つの寿司を比較したとき，一回の比較でトロを X_1 とするデータと，イカを X_1 とするデータの二つを作って分析している．そうしなければ，S が一方の値のデータしか得られなくなるためであるが，これが無作為割当ての条件を阻害している可能性がある．また，人気順に応じて表示位置や強調する寿司を固定していたが，これが $\mathbf{X}$ から S への依存性パスになってしまった可能性がある．これらの状況に対応した因果モデルを想定した分析を今後は検討する．

もう一つの想定していないバイアスの効果についても検討する．不人気アイテムを人気があるとして強調すると，被験者の主観と乖離がある結果が提示されることになり，何らかのバイアスの原因となる可能性がある．記憶効果など，他の種類の認知バイアスのデータを収集し，比較検討したい．

4. まとめ

本研究では，バイアス考慮型手法の安定性について論じた．この安定性を調査するために，認知バイアスの影響を受けた嗜好データを収集した．因果推論とバイアス考慮型手法の関係について論じた後，このバイアスを除去する簡潔な手法を示した．この手法を適用した結果，十分に安定的な結果はえられなかった．データと手法のいずれかに問題があると思われるが，これが今後はこれらの問題を検討してゆきたい．

謝辞：本研究は JSPS 科研費 JP24500194，JP15K00327，および JP18H03300 の助成を受けた．

参考文献

[Calders 10] Calders, T. and Verwer, S.: Three Naive Bayes Approaches for Discrimination-free Classification, *Data Mining and Knowledge Discovery*, Vol. 21, pp. 277–292 (2010)

[Chandler 11] Chandler, D. and Horton, J.: Labor Allocation in Paid Crowdsourcing: Experimental Evidence on Positioning, Nudges and Prices, in *AAAI Workshop: Human Computation*, pp. 14–19 (2011)

[Dwork 12] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R.: Fairness Through Awareness, in *Proc. of the 3rd Innovations in Theoretical Computer Science Conf.*, pp. 214–226 (2012)

[Eickhoff 18] Eickhoff, C.: Cognitive Biases in Crowdsourcing, in *Proc. of the 11th ACM Int'l Conf. on Web Search and Data Mining*, pp. 162–170 (2018)

[Hardt 16] Hardt, M., Price, E., and Srebro, N.: Equality of Opportunity in Supervised Learning, in *Advances in Neural Information Processing Systems 29* (2016)

[岩崎 15] 岩崎 学：統計的因果推論, 朝倉書店 (2015)

[Kamiran 12] Kamiran, F., Karim, A., and Zhang, X.: Decision Theory for Discrimination-aware Classification, in *Proc. of the 12th IEEE Int'l Conf. on Data Mining*, pp. 924–929 (2012)

[Kamishima 03] Kamishima, T.: Nantonac Collaborative Filtering: Recommendation Based on Order Responses, in *Proc. of The 9th Int'l Conf. on Knowledge Discovery and Data Mining*, pp. 583–588 (2003)

[Kamishima 18] Kamishima, T., Akaho, S., Asoh, H., and Sakuma, J.: Model-based and Actual Independence for Fairness-aware Classification, *Data Mining and Knowledge Discovery*, Vol. 32, pp. 258–286 (2018)

[Zhang 18] Zhang, L., Wu, Y., and Wu, X.: Anti-discrimination Learning: From Association to Causation, The 24th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining, Tutorial (2018)

[Žliobaitė 17] Žliobaitė, I.: Measuring Discrimination in Algorithmic Decision Making, *Data Mining and Knowledge Discovery* (2017)