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

独立性制約下の変換の認知バイアスの補正への適用

Application of the Transformation under the Independence Constraints for Canceling Cognitive Biases

神嶌 敏弘 *1　馬場 雪乃 *2　鹿島 久嗣 *3
Toshihiro Kamishima　Yukino Baba　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

The techniques of enhancing the independence between variables involved in a model have been exploited for removing a social bias. However, these techniques will be able to remove other kinds of biases. In this paper, we applied such techniques to remove a cognitive bias when eliciting preference data.

1. はじめに

機械学習の公平性の研究が 2016 年ごろから注目されてきた [神嶌 19]．これらの研究は，性別や人種などの社会的にセンシティブな情報に決定が影響されるバイアスが生じないようにする．そのために，予測モデルに含まれる変数間に独立性制約を導入してこれらのバイアスを補正してきた．例えば，クラス分類問題において予測クラス $\hat{Y}$ とセンシティブ情報 S を独立にする statistical parity，$\hat{Y} \perp\!\!\!\perp S$ などがある [Kamishima 12]．そして，この技術は社会的にセンシティブな情報によるバイアスだけでなく，その他の要因による様々なバイアスの補正にも利用できる．例えば，文献 [Adler 16] では，科学的化合物の特性予測にこうした独立性の検証技術を適用している．

そこで本研究では，嗜好データ収集に伴う認知バイアスの除去を目的として独立性制約を利用する．嗜好データとは，利用者のアイテムに対する嗜好の度合いを測るもので，5 段階尺度などを用いる採点法などを用いる．この嗜好データのように，利用者に質問してデータを収集するときには，認知バイアスが生じることが知られている．文献 [Cosley 03] は，アイテムの好みを尋ねるときに，予測値よりも高評価の値を入力時に示すと評価が上に偏ることなどを報告している．また，文献 [Eickhoff 18] は，クラウドソーシング環境下での文書の適合性判定において，多数派の決定に追随しやすいバンドワゴン効果などの認知バイアスが確認できることを報告している．これらの認知バイアスを削除するため，認知バイアスの原因と嗜好データの独立性を保つように変換する手法を検証する．センシティブ情報とデータを独立にしてデータからセンシティブな情報をを削除する試みは文献 [Pérez-Suay 17] などにもあるが，ここではこの文献のように次元削減などのデータ自体の変換は行わない．

ここでは，一対にアイテムを比較してどちらがより好みかを尋ねる一対比較法によって嗜好データを収集する．そして，文献 [Eickhoff 18] で影響の大きかったバンドワゴン効果対象にする．クラウドソーシングを利用し，100 人前後の規模で被験者データを収集した．以後，認知バイアスの影響を確認する検証結果を示したのち，その除去を独立性制約を用いて除去する実験結果を示す．

連絡先: ホームページ http://www.kamishima.net

(a) ベースライン入力画面

(b) バンドワゴン入力画面

図 1: 入力画面

2. 認知バイアスの確認実験

最初に，一対比較の手順について述べる．文献 [Kamishima 03] の実験で用いた次の 10 種類の寿司から，二つを被験者に提示し，より好きな方を選択させた．

トロ，マグロ，エビ，イクラ，アナゴ，
ウニ，テッカ巻，イカ，タマゴ，カッパ巻

この文献により，5000 人の調査では，この順番に人気があった．以後，実験データ中の寿司はこの順に並べて示し，グラフ中では次の略号で示す．

Tr, Mg, Eb, Ir, Ag, Un, Tk, Ik, Tm, Kp

これらの寿司のうち，二つを提示しいずれか一方の，より好きな寿司を被験者に選択させた．入力画面は，ベースラインと

バンドワゴンの 2 種類ある．ベースラインは図 1(a) のように二つの寿司を同じ大きさで左右に配置した．もう一つのバンドワゴンは図 1(b) で，一方の寿司を大きく，また人気があることを示すラベルを付加することで強調した．この人気があるラベルは，実際に上記の人気順に基づいて人気のある方にラベルをつけた場合と，実際には不人気である方にラベルを付けた場合との 2 通りを実験した．なお，左右の配置の影響を受けないように，左右は無作為に配置している．

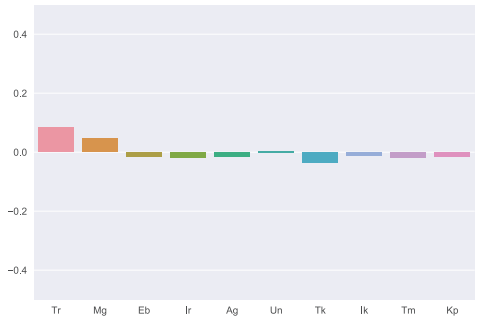
クラウドソーシング環境下で，この入力画面を用いて被験者実験を行った．データは 2020 年 1 月 31 日〜2 月 22 日の期間に収集し，1 人あたり 50 円の代価を支払った．1 人あたり 50 個の質問を行ったが，そのうち 2 個は「右はどちらですか」という集中度質問である．この 2 件の集中度質問に二つとも正解した被験者のみのデータを実験では用いた．ベースラインでは 120 人，バンドワゴン人気では 99 人，バンドワゴン不人気では 96 人の被験者から，それぞれ 48 件の一対比較結果を得た．

寿司 i と j を比較したときに，寿司 i が j より好まれた割合 $\Pr[i \succ j]$ を求めた．そしてこの割合を要素 x_{ij} とする 10×10 の行列を $\mathbf{X}$ とする．ベースライン，バンドワゴン人気，そしてバンドワゴン不人気のそれぞれの行列を $\mathbf{X}^{(b)}$，$\mathbf{X}^{(p)}$，および $\mathbf{X}^{(u)}$ のように上付き文字で示す．各寿司 i について，ベースラインに対するバンドワゴン効果量を次式で測る．

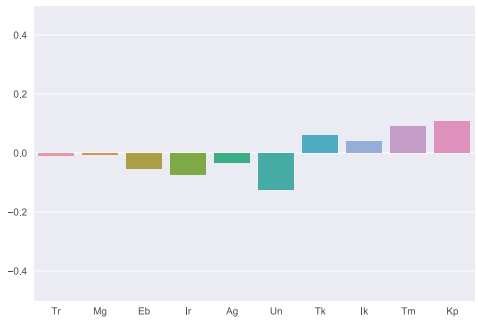
$$e_i = \sum_j \left(x_{ij}^{(a)} - x_{ij}^{(b)} \right), \; a \in \{p, u\} \tag{1}$$

バンドワゴン効果によりより頻繁に好まれるようになった場合には正に，好まれなくなったら負になる．各アイテムについてのこのバンドワゴン効果量を図 2 に示す．式 (1) で a が p と u の場合がそれぞれ図 2(a) と 2(b) にあたる．

図 2(a) は実際に人気のある寿司を強調したので，より上位（図中の左側に寿司）の方がより頻繁に強調され，より大きなバンドワゴン効果が生じていると推測される．そして実際に，この図では強調された上位 2 種の寿司はより好まれるようになっており，あまり強調されない下位の寿司ではその分好まれなくなっている．図 2(b) は実際に人気がない寿司を強調した．実際の人気とは逆に，人気のない下位の寿司がより強調されて，より大きなバンドワゴン効果生じると推察される．図 2(a) の場合よりも結果は顕著で，中位から下位の寿司はより頻繁に選ばれるようになっている．もう一方の上位の寿司に注目すると，本当に好まれている最上位の寿司は強調の影響を受けていないが，中位からやや上位の部分が負の効果を受けている．以上のことから，人気を強調することでバンドワゴン効果が生じることが確認できた．



(a) バンドワゴン人気



(b) バンドワゴン不人気

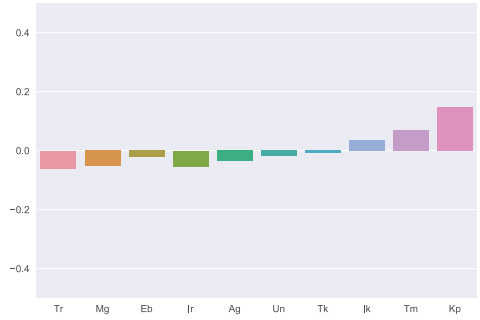
図 2: 補正前のバンドワゴン効果量

3. 認知バイアスの除去実験

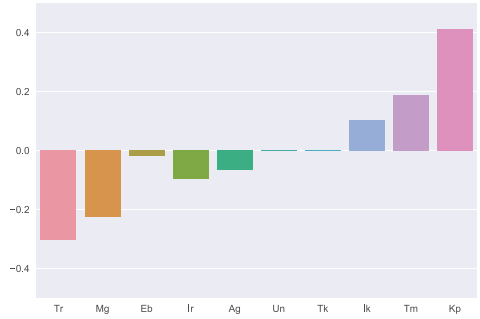
この認知バイアス，すなわちバンドワゴン効果が除去できるかを簡潔な方法で検証する．そのために，アイテムが強調されているかどうかを示す変数 S を導入する．S=0 では強調されており，S=1 では強調されていないものとする．そして，$\Pr[i \succ j | S = 0]$ をアイテム i が強調されていないときにアイテム j より i が好まれる割合，逆に $S = 1$ ならば強調されていないときの割合とする．

一対比較の結果がこの変数 S と独立であれば，すなわち $\Pr[i \succ j | S{=}0] = \Pr[i \succ j | S{=}0]$ となれば，強調の効果を除去できるという着想の実装を試みる．ここでは，一対比較結果がこの S に直接的に依存する次式で $\Pr[i \succ j]$ をモデル化する．

$$\Pr[i \succ j] = \Pr[i \succ j | S{=}0] \Pr[S{=}0] + \Pr[i \succ j | S{=}1] \Pr[S{=}1] \tag{2}$$



(a) バンドワゴン人気



(b) バンドワゴン不人気

図 3: 補正後のバンドワゴン効果量

ここで強調されるかどうかに偏りがあることでバイアスを生じているとして，これを一様にした確率を考える．

$$\tilde{\Pr}[i \succ j] = \frac{1}{2}\Pr[i \succ j | S{=}0] + \frac{1}{2}\Pr[i \succ j | S{=}1] \qquad (3)$$

条件付き確率 $\Pr[i \succ j | S]$ が常にこの式 3 の値であるように $\mathbf{X}$ を補正する．しかしながら，強調は人気寿司，もしくは不人気寿司に確定的に提示しているため，$\Pr[i \succ j | S{=}0]$ か $\Pr[i \succ j | S{=}1]$ のいずれか一方しか観測できないので式 3 は計算できない．そこで，最も簡潔な方法として，観測できなかった条件付き確率を一様な確率 0.5 を割り当てて補正を試みた．

図 3(a) は人気寿司に対するバンドワゴン効果の除去を試みたものである．図 2(a) と比較して，効果の向きが反転し，両端では効果量が拡大している．図 2(b) と図 3(b) の比較でも同様の問題を生じている．以上のことから補正の効果はみられなかった．

4. 考察

今回の手法では補正の効果は見られなかった．S から $\mathbf{X}$ へ直接的に効果があるモデルを想定していたが，このモデルが妥当だと仮定する．この場合は，観測されなかった反実仮想の場合の確率を一様な $1/2$ としたことが妥当ではなかったと推察される．何らかのモデルを用いて妥当な値を推定する方法が必要になるだおる．S から $\mathbf{X}$ へ直接的に効果があるモデル自体が妥当でない場合もありうる．この場合は強調の有無から，何らかの中間変数をえて一対比較結果 $\mathbf{X}$ が生成される新たなモデルを考察する必要が生じる．今後はこれらの問題点の修正に取り組みたい．

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

[Adler 16] Adler, P., Falk, C., Friedler, S., Rybeck, G., Schedegger, C., Smith, B., and Venkatasubramanian, S.: Auditing Black-box Models for Indirect Influence, in *Proc. of the 16th IEEE Int'l Conf. on Data Mining*, pp. 1–10 (2016)

[Cosley 03] Cosley, D., Lam, S. K., Albert, I., Konstan, J. A., and Riedl, J.: Is Seeing Believing? How Recommender Interfaces Affect Users' Opnions, in *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems*, pp. 585–592 (2003)

[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)

[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 12] Kamishima, T., Akaho, S., Asoh, H., and Sakuma, J.: Fairness-aware Classifier with Prejudice Remover Regularizer, in *Proc. of the ECML PKDD 2012, Part II*, pp. 35–50 (2012), [LNCS 7524]

[神嶌 19] 神嶌 敏弘, 小宮山 淳平：機械学習・データマイニングにおける公平性, 人工知能, Vol. 34, No. 2, pp. 196–204 (2019)

[Pérez-Suay 17] Pérez-Suay, A., Laparra, V., Mateo-García, G., Muños-Marí, J., Gómez-Chova, L., and Camps-Valls, G.: Fair Kernel Learning, in *Proc. of the ECML PKDD 2017, Part I*, pp. 339–355 (2017), [LNCS 10534]