公開

研 究 終 了 報 告 書

「行動経済学に基づく個人的・集団的評価の数理モデルの開発」

研究期間: 2019 年 10 月～2023 年 3 月
研 究 者: 馬場 雪乃

1. 研究のねらい

我々人間は，商品や人物，サービスなど様々なものを日常的に評価している．日常的行為であるにもかかわらず評価は簡単ではなく，一貫した評価をするのは難しい．たとえば，複数の入学候補者を面接し，それぞれを評価する場合を考える．候補者が多くなると，優れた候補の直後では評価が厳しくなることや，途中から評価基準が変化することは良くある．評価が一貫しないのは，限られた思考時間で評価を決めるために，直感を用いるからである．人間が直感を用いるとき，合理的・論理的ではない判断をしてしまうことが，行動経済学において実験的に示されている．このような判断のゆがみを認知バイアスという．認知バイアスにより，評価の実施順などの，評価対象の価値とは無関係の要素が影響し，評価がゆがんでしまう．

認知バイアスによる評価のゆがみは，個人による評価を集約し，集団としての評価を決める際にも生じる．集約過程において，利他性により他者の意見を優先したり多数派に追随したりと，評価対象の価値と無関係な要素が影響してしまう．そのため，評価の集約結果が，集団が本来もつ価値関数と一貫しないものとなる．

本研究では，個人の評価や集団の評価集約の過程を，認知バイアスを含めた数理モデルで表現する．数理モデルを用いて，個人や集団の判断や意思決定を支援する技術を開発する．個人と集団，そして，情報提示等による受動的な支援から，人間とのインタラクションを通じて積極的な介入を行う支援まで，様々な場面における意思決定支援の技術の開発を目指す．

2. 研究成果

(1) 概要

集団意思決定の支援を目的として，研究テーマ A「確証バイアスに対処する投票集約法」と研究テーマ B「他者意見の影響を利用した合意形成法」に取り組んだ．また，個人の意思決定の支援を目的として，「順序バイアスの影響を捉えた一対比較モデル」と研究テーマ D「公平な人物評価の機械教示」に取り組んだ．

研究テーマ A では，大量の候補の中から集団で採用案を決める際に，投票による選抜では各自の確証バイアスにより，多数派の価値観しか反映できないという問題に対処するため，投票結果を利用して多様な観点で有望な候補を選抜する手法 CrowDEA を開発した．研究テーマ B では，合意形成で広く用いられるデルファイ法では，集団がしばしば誤った解に収束するという問題に対処するため，提示する意見分布を操作する手法を検討した．意見分布の操作により，確信度が極端に高くない場合以外は，「多数派の意見」として提示されたものに追従する傾向があることを確認した．これにより，追従度合いを使って確信度を推定し，それを利用して正解を推定して，正解を多数意見として提示することで，正解への収束を促せる可能性があることがわかった．研究テーマ C では，人間の嗜好を推定する際によく用いられる一対比較では，評価順の影響を受けてしまうという問題に対処するため，評価順の影響を

捉えた数理モデルを開発した．自然言語処理でよく用いられる注意機構を採用して評価順の影響と比較相手の影響を捉えることで，精度の良いモデルが実現できることを示した．研究テーマ D では，公平性配慮型機械学習を用いることで，人間に公平な評価の仕方を教える方法を検討した．ユーザに人物評価をさせ，その結果に公平性配慮型機械学習を適用することで，ユーザの評価基準が公平になるように調整した「教師モデル」を獲得する．教師モデルを用いて，公平モデルの評価基準を学べるように，教材を学習者に提示する．被験者実験により，単純な手法と比較して学習効果が高いことを示した．

JST サイエンスインパクトラボの協力の下，研究テーマ A の社会実装を進めた．最初の取り組みとして，かえつ有明高校の高校 1 年生のクラスにおいて，クラス内の困り事の解決に CrowDEA を活用し，CrowDEA を用いることで少数派の意見にも配慮できるという結果を得ることができた．

(2) 詳細

■ 研究テーマ A「確証バイアスに対処する投票集約法」

例えば，キャラクターデザインを公募して，最終的に採用するデザインを決めるという状況を考える．候補が膨大な時，各自が投票し，得票数上位のものに絞った上で議論して決めるという方法が考えられる．しかし，人間は，確証バイアスにより，自分の価値観に合うものばかりに票を入れてしまう．結局，投票において選ばれるのは，多数派の価値観に合う，似通った候補となる．これでは，せっかく集めた候補の多様性が失われてしまう．

この問題に対処するため，投票を利用して，多様な有望候補を選抜する手法，CrowDEA を開発した．CrowDEA を用いることで，アイディアの多様性を失わずに有望なものを選抜し，最終案の議論の対象にすることができる．CrowDEA では，投票は一対比較で行うものとする．複数の評価者からの，複数のペアに対する一対比較の結果を用いて，図 1 のような Priority map を生成する．青くハイライトされたものが，各観点で有望な候補だとして選抜される．CrowDEA は，多次元ベクトルで表現された各候補の埋め込み表現及び最良観点と，各評価者の評価観点というパラメータを導入して，投票行動をモデル化し，一対比較結果からパラメータを推定することで，Priority map を出力する．

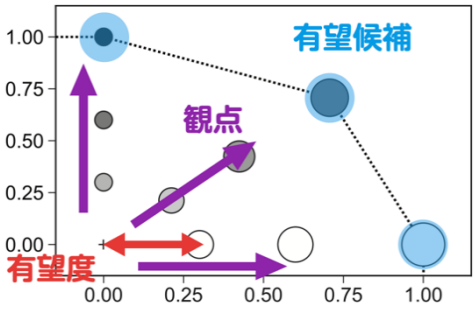

図 1 CrowDEA が出力する Priority map の例

オリンピックのエンブレムに対してクラウドソーシングで一対比較結果を収集した．単純な集約では，現代的な類似した候補が選抜されてしまうのに対し，CrowDEA を適用すると，CrowDEA では伝統的なものも含めて選抜することができることを確認した．また，「カンニング

の防止策」のアイディアについてCrowDEAを適用した例を表1に示す．罰則を与える，問題の配置を変える，といった，多様なアイディアが選抜できている．

表1 カンニング問題に対してCrowDEAが選抜したアイディア例

カンニングが判明した時の罰則を重いものにする．例えばその定期テスト全教科を全て0点，内申点0点など．
問題の順番を並び替えたものを2種類用意して，隣の列と同じ問題の順番にならないように配布する．
学生の後ろ側から監視する．テストを受けている身としては後ろに教師がいると思うと迂闊にカンニングはできない．
一目見ただけで回答を確認できてしまうような一問一答問題や選択問題の出題を極力避け，論述問題の出題を中心にすることでカンニング行為を無意味なものにする．

■ 研究テーマB「他者意見の影響を利用した合意形成法」

合意形成で用いられるデルファイ法は，他者の意見分布を見せ，各自の意見を再考させ，その手順を繰り返すことで，合意形成を促す手法だが，集団が誤った解に収束することがあることが知られている．提示する意見分布を操作することで，正しい解への収束を促す手法を検討するため，意見分布の操作により意見を変えることができるかどうかを調査した．絵画の価格予測，衣服の価格予測，映画の年代予測の三つの題材を採用した．実験参加者には，「あなたの他に 11 人が同時に同じ予測問題に取り組んでいる」と伝え，実際にはランダムに生成した意見分布を提示した．提示した意見分布に応じて，意見がどのように変化するかを調査した．その際，各選択肢に賭けさせることで，意見の裏にある予測分布を収集し，意見に対する確信度の分析に用いた．

実験により，確信度が高くない人は「ほとんどの人がこの意見です」と提示されると，それが元々の自分の意見をかけ離れていても，それに追従するという傾向があることがわかった．一方で，確信度が高い人は，多数意見には追従しにくいという傾向があった．この結果から，追従度合いを使って確信度を推定し，それを利用して正解を推定して，正解を多数意見として提示することで正解への収束を促せる可能性があると言える．

■ 研究テーマC「順序バイアスの影響を捉えた一対比較モデル」

人間の嗜好を推定する際に，一対比較はよく用いられるが，一対比較の結果は，評価の順番や，比較相手によるバイアスの影響を受けることが知られている．少ない評価回数でも，バイアスの影響を取り除いて，嗜好を推定する手法を開発した．自然言語処理で，単語間の影響を捉えるために用いられる，注意機構を用いて，過去の嗜好が今回の一対比較に与える影響を捉えた．また，比較相手に応じて決まるマスクを用いて，比較相手の影響を捉えた．モデルを図2に示す．

図形，キャラクター，映画の嗜好の推定を対象にして，クラウドソーシングにより一対比較結果を収集した．それぞれ 100 名の実験参加者に一対比較を 120 回実施させた．最初の 10 回，あるいは 20 回の比較結果を用いてモデルのパラメータを推定し，残りの比較結果を予測することでモデルの精度を評価した．バイアスの影響を考慮しない既存手法と比較して，提案モデルが高い予測精度を達成することを確認した．

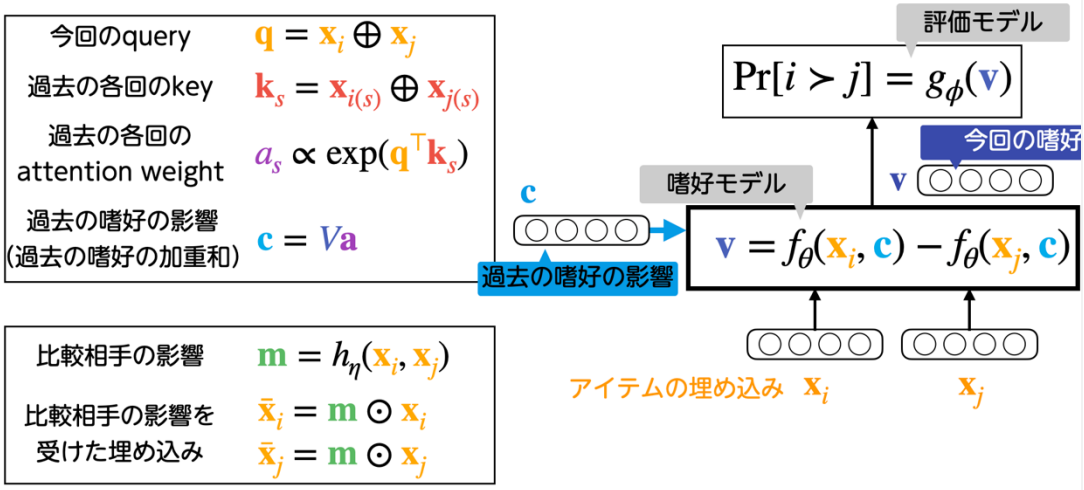

図 2　順序バイアスを捉えるモデル

■　研究テーマ D「公平な人物評価の機械教示」

人間が他者を評価するときに，相手の人種や性別の影響を受けて，不公平な評価をしてしまうことが知られている．公平なモデルを学習する手法である，公平性配慮型機械学習を用いることで，人間に公平な評価の仕方を教える方法を検討した．ユーザに人物評価をさせ，その結果に通常の機械学習を適用し，ユーザの評価基準を模倣した「不公平モデル」を獲得する．同時に，公平性配慮型機械学習によって，ユーザの評価基準が公平になるように調整した「公平モデル」を獲得する．ユーザが，公平モデルの評価基準を学べるように，教材を提示する．教材の例を図 3 に示す．不公平モデルの評価基準を左側，公平モデルの評価基準を右側に提示している．左側では，例えばこのユーザは，対象がアジア系でフィリピン出身の時に，不利な評価をしやすい，という傾向を提示している．右側の公平モデルにはそのような傾向はないことを示している．

約 100 名の被験者を対象に実験を行った．人物評価の事前テストを受けさせ，その後，教材提示を5セット行い，最後に事後テストを受けさせ，事前と事後の不公平度を比較した．ベースラインの手法は，例えば「あなたは，相手が白人の場合は 20%を採用，白人以外の場合は 15%を採用，と判断しています．両者の数字を近づけてください」という様にして，公平な判断を促す「ナッジ」である．今回の公平モデルを用いた手法が，あるタスクにおいては不公平度をベースラインよりもよく改善できていることが確認できた．

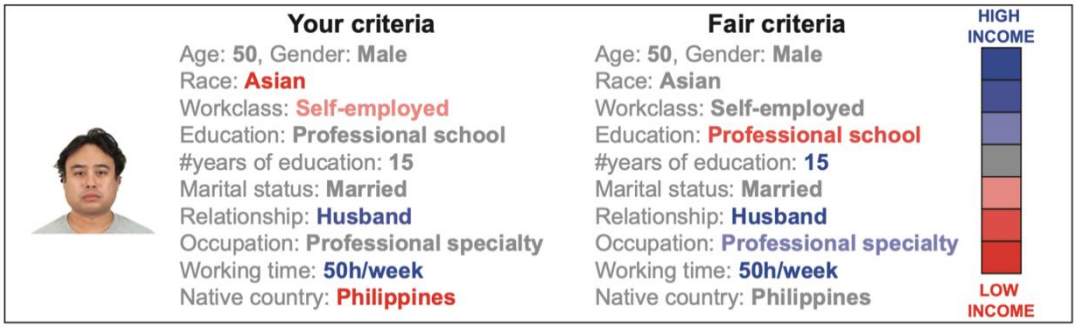

図 3　公平判断を教えるための教材例

3. 今後の展開

特に，研究テーマ A の社会実装を今後も進める予定である．CrowDEA をウェブプラットフォームとして実装することで，一般の人が本技術を利用できるようにする．同時に，教育，職場，街づくり，審査などの様々な場面において，ケーススタディを実施し，本技術を社会実装するためのニーズを洗い出して技術のブラッシュアップを進める．

4. 自己評価

■ 研究目的の達成状況

当初の研究目的では研究テーマ B, C のみを予定していたが，研究テーマ A, D という新たなテーマに取り組むことができ，当初目的以上の成果を上げることができた．また，JST サイエンスアゴラにおけるアウトリーチ活動や，JST サイエンスインパクトラボにおける社会実装・実証実験など，当初目的以上の活動を行うこともできた．

■ 研究実施体制

研究費は予定通り執行できている．研究実施体制は，複数の学生を研究補助として雇用することで複数のテーマを並列に実施することができた．

■ 研究成果の科学技術及び社会・経済への波及効果

特に研究テーマ A, D においては，人々の多様性に配慮した意思決定を支援する技術を開発することができた．私たちの社会が集団のポテンシャルを最大限に活かすためには，マイノリティへの配慮が喫緊の課題である．また，私たちが集団の知能を最大限に活かすためには，少数派だが有望なアイディアを見逃さないことが重要である．多様性に配慮できるように社会を変革するための，ベースとなる技術を開発できたと自負している．

5. 主な研究成果リスト

(1) 代表的な論文（原著論文）発表

研究期間累積件数：7件

1. Y. Baba, J. Li, H. Kashima. CrowDEA: Multi-view Idea Prioritization with Crowds. In

Proceedings of the 8th AAAI Conference on Human Computation and Crowdsourcing (HCOMP), pp.23--32, 2020.
集団で議論して，大量の候補の中から一つの採用案を決めるという場面において，まずは各自が投票し，得票数上位のものに絞った上で議論して決めるという方法がよく用いられる．しかし，投票において選ばれるのは，多数派の価値観に合う，似通った候補となる．これでは，せっかく集めた候補の多様性が失われてしまう．この問題に対処するため，投票を利用して，多様な有望候補を選抜する手法，CrowDEA を開発した．
2. S. Ito, Y. Baba, T. Isomura, H. Kashima. Synthetic Accessibility Assessment using Auxiliary Responses. Expert Systems with Applications, Vol. 145, 2020.
機械学習を用いた創薬技術の開発が進む一方で，機械学習モデルが提案した化合物の合成可能性の判定は，未だ人間の専門家によって行われている．合成可能性の判定の効率化のため，複数の準専門家の回答を統計的に統合する手法を提案した．特に，準専門家に，合成可能性のスコアだけではなく，化合物中の注意すべき構造をアノテーションさせ，それを利用して各自の判断バイアスを捉える手法を提案した．
3. K. Fujii, Y. Saito, S. Takamichi, Y. Baba, H. Saruwatari. HumanGAN: Generative Adversarial Network with Human-based Discriminator and its Evaluation in Speech Perception Modeling. In Proceedings of the 45th International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 6239—6243, 2020.
画像生成や音声合成において広く用いられる敵対的生成ネットワークは，生成器と識別器の二つの機械学習モデルを競わせることで，生成器の学習を行う．生成器に人間の嗜好を反映させるため，識別器を人間に置き換える HumanGAN を提案した．一対比較により，生成器が生成したサンプルと実際のサンプルの識別を人間に行わせる．Natural evolution strategy を用いることで，ブラックボックスである人間の判断を逆伝播して生成器の学習に活用することを実現した．

(2)特許出願

研究期間全出願件数:0 件(特許公開前のものも含む)

(3)その他の成果(主要な学会発表，受賞，著作物，プレスリリース等)

- [出展] サイエンスアゴラ 2022「集合知と人工知能 ～AI があなたの一票に命を宿す～」, 2022 年 11 月
- [受賞] 鈴木一史，馬場雪乃．第 28 回社会情報システム学シンポジウム優秀発表賞．個人の意見が他者の意見の多様性と多数意見との距離から受ける影響の分析．2022 年 1 月
- [解説記事] 馬場雪乃．集合知を生かすヒューマンコンピュテーション．人工知能学会誌．2022年 3 月