様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　４ 年　 ６ 月 ２１ 日現在

機関番号：８２６２６
研究種目：基盤研究(B)（一般）
研究期間：2018～2020
課題番号：１８Ｈ０３３００

研究課題名（和文）公平性配慮型データ変換技術の開発とそのクラウドソーシングによる効果検証

研究課題名（英文）The development of a fairness-aware data-transformation technique and the validation of its effectiveness through a cloudsoucing environment

研究代表者
神嶌　敏弘（Kamishima, Toshihiro）

国立研究開発法人産業技術総合研究所・情報・人間工学領域・主任研究員

研究者番号：５０３５６８２０

交付決定額（研究期間全体）：（直接経費）　　8,200,000円

研究成果の概要（和文）：本研究は，公平性を保証するために，社会的にセンシティブな性別や人種などの特徴と与信や採用などの判断とが統計的に独立となるようにする公平性配慮型機械学習を扱う．観測されたデータで予測精度を評価していたが，これは不公平な決定結果であり，真に公平な決定結果ではどうなっていたかは観測できない．センシティブな情報の代わりに統制実験が可能な認知バイアスを利用し，真に公平な決定を擬似的に作り出したデータをクラウドソーシングを利用して作り出した．認知バイアスを除去した後に，どれだけ真の決定の情報が得られているかを調べるために安定性の概念を開発した．

研究成果の学術的意義や社会的意義
機械学習の公平性については2011年から取り組んでいるが，2016年の米大統領選や，欧州のGDPR試行に伴い注目され，世界的に研究が拡大している研究分野である．しかしながら，本当はあるべき公平な決定というものが観測できない根本的な制限がある．この制限に対して，センシティブ情報の代用として認知バイアスを利用して，人工的にデータと収集するという手段で挑んだのが本研究である．

研究成果の概要（英文）：We tackle with one task related to fairness-aware machine learning, in which the predictor tries to satisfy some fairness constraint, such as statistical stability. In the previous literature of fairness-aware machine learning, the prediction accuracy is evaluated on the observed dataset whose decision labels are supposed to be unfair. This is due to the restriction that truly fair labels cannot be observed. We tried to evaluate the precision on the fair labels as proxy by using preference data influenced by cognitive biases. We collected such data through a crowdsourcing service. Then, to evaluate how much information of fair decisions are extracted from these observations, we developed the notion of stability and a method to quantify the stability.

研究分野：機械学習

キーワード：公平性　クラウドソーシング

１．研究開始当初の背景

本研究は，公平性配慮型データマイニングという機械学習分野の中でも新しい分野の研究である．この研究分野が生じた背景について述べておく．データマイニング技術の普及に伴い，与信・採用・入試など，個人の生活に大きな影響を及ぼす分野に適用されるようになった．それに伴い，人種や性別といった情報が決定に影響してしまう事例が散見されるようになった．例えば，2016 には米ウィスコンシン州の上級裁判所では保釈の決定に統計的手法による再犯リスクスコア COMPAS の利用を決めた．しかし，このスコアが，アフリカ系の人に対して再犯リスクをヨーロッパ系の人より高く推定することを ProPublica が指摘し，裁判所はそれに反論するなどの動きがあった．一方で EU おいても，機械学習の予測結果に対する正しい説明を求める General Data Protection Regulation が 2016 年に制定された．これらの動きから，機械学習分野において公平性は喫緊の問題となっている．

この公平性問題の解決を，特定の情報を排除しつつ予測する問題ととらえているのが公平性配慮型機械学習分野である．

２．研究の目的

公平性配慮型機械学習では，人種や性別など社会的にセンシティブな情報に対して公平性を保ちつつ予測できる予測器をデータから学習する．代表的な公平性規準の定式化として知られる統計的均一性は，センシティブな情報と，与信や採用といった決定変数とが統計的に独立であるというものであり，倫理学などでいう配分の公正に該当する規準である．

この公平性配慮型機械学習では，公平性と精度のトレードオフの関係にある．公平性を確保するために特定の情報を排除すると，利用できる情報は確実に減少するので，予測精度は確実に低下する．よって，公平性をある水準にしたとき，できるだけ高い予測精度を達成することが目標となる．多くの公平性配慮型機械学習手法では，この公平性と精度のトレードオフを調整するパラメータが存在するが，実際に適用するときに，このパラメータをどうせ一定すれば，現実世界で妥当な公平性を達成しつつ十分な予測精度を達成できるのかはオープンな問題であった．

この問題に取り組むべく提案書では，人間が主観的に認知できる公平性の限界を探るようなデータを収集する実験を計画していた．しかしながら，研究を進める上で公平性配慮型機械学習での精度とは，いったいどういうことなのかという疑問が生じた．どういう疑問かを述べると，予測精度は観測されたデータ集合に対して計測しているが，このデータに含まれる決定の情報は，公平で理想的な判断を人間ができないため不公平な決定を含んでいるものである．ところが，公平な決定を含んだデータは観測できないので，この不公平な決定に対して予測精度を測っている．よって，公平な決定についての予測精度をどうにかして求めつつ，公平性との相対的な関係を考える必要があることに気付いた．

そこで，公平な決定の情報が取得できるような統制環境を作りだし，クラウドソーシングを使ってデータを収集するようにして，「正しい」といえる予測精度を測る方法を構築するようにした．

３．研究の方法

実際にセンシティブな情報に対して公平な結果を用いると，今までと同様にセンシティブな情報に依存した決定を人間はしてしまうため，公平なデータを生成することはできない．そこで，センシティブな情報の代わりに，こちらから環境を統制できる認知バイアスを用いる着想を得た．決定変数には，利用者の嗜好，具体的には寿司の好き嫌いの情報を使った．そして，これらのデータをクラウドソーシングを通じて収集し，そのデータを評価する．以下，これらの段階をより詳細に述べる．

まず，決定タスクである．これは，二つの寿司を利用者に提示し，どちらが好きかを問う一対比較法と呼ばれるタスクを実施した．クラウドソーシングのワーカには 50 件ほどのタスクを割当てることで，10 種類の寿司の嗜好全体の順序を求めることができる．なお，このタスクを選んだのは，この一対比較法の数理モデルの一つに Bradly-Terry モデルがあり，このモデルは，我々が過去に作った公平性配慮型機械学習アルゴリズムでも採用していたロジスティック回帰モデルと同一になるためである．

次に，センシティブな情報の代わりに認知バイアスを採用した．ここでは 2 種類の認知バイアス，位置バイアスとバンドワゴン効果である．位置バイアスとは，普段，人間がものを並べる方向で上位のものを並べる上方や左方にあるものをよく選んでしまうというものである．もう一方のバンドワゴン効果とは，他の人がこちらが好きであるといった人気情報を与えられると，多くの人がそれに従ってしまうというものである．

これら二つのそれぞれの認知バイアスについて，クラウドソーシング実験環境下で，特定の方向にバイアスがかかるように統制実験を行った．ここで重要なのは，どちらの方向にバイアスをかけてワーカに質問したかの情報があるため，本当のセンシティブな情報とは違って，交絡因子を除去する層別分析によって認知バイアスの影響を除去できる点にある．

以上の手順で収集したデータを用いて分析を行った．

４．研究成果

分析は，2 種類の認知バイアスの影響を受けているものに加えて，位置バイアスの影響をも除去するように無作為割り付けをして認知バイアスの影響を受けにくくしたものとの三種類のデータを，いろいろな実験計画で収集した．

認知バイアスを層別分析で除去すると，行っている質問は全て，同じアイテム集合に対する一対比較の質問である．よって，センシティブな情報の代用として用いた認知バイアスの情報が十分に取り除かれているのであれば，それを除外した一対比較の決定の情報だけが残っているはずであり，データの収集時に与えた認知バイアスとは無関係に，嗜好にたいする決定情報はデータ集合間で類似するはずである．この類似性が高い状態を安定性（stability）と定義し，決定変数がどれだけバイアスの影響を受けているのかを評価できるようにした．従来は観測できないセンシティブ情報をあつかっていたため，こうしたことはできなかったので，本研究の貢献の一つといえる．

具体的にこの安定性を，バイアスが異っているデータ集合について，バイアスを取り除いたときに，決定が同じであればあるほど高いというように定式化することで，この安定性を定量的に測れるようにした．実際に計測してみたところ，無作為と位置バイアスの間では，データ集合は比較的安定的といえたが，影響の大きな認知バイアスであるバンドワゴン効果はかなり不安定であった．

この不安定性の原因の究明にあたったが，研究期間の終了までにはその原因を特定することはできなかった．この究明にあたっては，今後とも研究を継続してゆきたい．

５．主な発表論文等

〔雑誌論文〕　計9件（うち査読付論文　2件／うち国際共著　5件／うちオープンアクセス　2件）

１．著者名 T. Kamishima, S. Akaho, Y. Baba, and H. Kashima	４．巻
２．論文標題 Preliminary Experiemnts on the Stability of Bias-aware Techniques	５．発行年 2021年
３．雑誌名 2nd Int'l Workshop on Algorithmic Bias in Search and Recommendation {Bias 2021}	６．最初と最後の頁 25-35
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/978-3-030-78818-6_4	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 M. D. Ekstrand, P.-N. Schwab, J. Garcia-Gathright, T. Kamishima, and N. Sonboli	４．巻
２．論文標題 3nd FAccTRec Workshop: Responsible Recommendation	５．発行年 2020年
３．雑誌名 Proc. of the 14th ACM Conf. on Recommender Systems	６．最初と最後の頁 607-608
掲載論文のDOI（デジタルオブジェクト識別子） 10.1145/3383313.3411538	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

１．著者名 神嶌 敏弘	４．巻 37
２．論文標題 私のブックマーク「人工知能と公平性」	５．発行年 2022年
３．雑誌名 人工知能	６．最初と最後の頁 230-233
掲載論文のDOI（デジタルオブジェクト識別子） 10.11517/jjsai.37.2_230	査読の有無 無
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Abdollahpouri Himan、Adomavicius Gediminas、Burke Robin、Guy Ido、Jannach Dietmar、Kamishima Toshihiro、Krasnodebski Jan、Pizzato Luiz	４．巻 30
２．論文標題 Multistakeholder recommendation: Survey and research directions	５．発行年 2020年
３．雑誌名 User Modeling and User-Adapted Interaction	６．最初と最後の頁 127～158
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/s11257-019-09256-1	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

1．著者名 神嶌 敏弘、鹿島 久嗣	4．巻 34
2．論文標題 機械学習分野の俯瞰と展望	5．発行年 2019年
3．雑誌名 人工知能	6．最初と最後の頁 905-915
掲載論文のDOI（デジタルオブジェクト識別子） 10.11517/jjsai.34.6_905	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 神嶌 敏弘	4．巻 59
2．論文標題 サービスの公平性に配慮したデータ分析技術	5．発行年 2018年
3．雑誌名 情報処理	6．最初と最後の頁 433-436
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 神嶌 敏弘	4．巻 74
2．論文標題 変わりゆく機械学習と変わらない機械学習	5．発行年 2019年
3．雑誌名 日本物理学会誌	6．最初と最後の頁 5-13
掲載論文のDOI（デジタルオブジェクト識別子） 10.11316/butsuri.74.1_5	査読の有無 無
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

1．著者名 神嶌 敏弘, 小宮山 淳平	4．巻 34
2．論文標題 機械学習・データマイニングにおける公平性	5．発行年 2019年
3．雑誌名 人工知能	6．最初と最後の頁 196-204
掲載論文のDOI（デジタルオブジェクト識別子） 10.11517/jjsai.34.2_196	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

1．著者名 T. Kamishima, P.-N. Schwab, M. D. Ekstrand	4．巻 なし
2．論文標題 2nd FATREC Workshop: Responsible Recommendation	5．発行年 2019年
3．雑誌名 Proc. of the 12th ACM Conf. on Recommender Systems	6．最初と最後の頁 516
掲載論文のDOI（デジタルオブジェクト識別子） 10.1145/3240323.3240335	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

〔学会発表〕　計7件（うち招待講演　2件／うち国際学会　0件）

1．発表者名 神嶌 敏弘，馬場 雪乃，鹿島 久嗣
2．発表標題 独立性制約下の変換の認知バイアスの補正への適用
3．学会等名 人工知能学会全国大会（第34回）
4．発表年 2020年

1．発表者名 神嶌 敏弘，赤穂 昭太郎，馬場 雪乃，鹿島 久嗣
2．発表標題 バイアス考慮型分類器の安定性に関する予備調査
3．学会等名 人工知能学会全国大会（第35回）
4．発表年 2021年

1．発表者名 神嶌 敏弘
2．発表標題 機械学習の公平性への取り組み －Fairness-aware data miningを中心に－
3．学会等名 人工知能学会全国大会（第33回）
4．発表年 2019年

１．発表者名
神嶌 敏弘

２．発表標題
機械学習における公平性の概要

３．学会等名
第38回産総研AIセミナー（招待講演）

４．発表年
2019年

１．発表者名
機械学習における公平性の概要

２．発表標題
機械学習と公平性

３．学会等名
機械学習と公平性に関するシンポジウム（招待講演）

４．発表年
2020年

１．発表者名
神嶌 敏弘, 赤穂 昭太郎, 麻生 英樹, 佐久間 淳

２．発表標題
公平ロジスティック回帰での確定的決定則の影響

３．学会等名
人工知能学会全国大会（第32回）

４．発表年
2018年

１．発表者名
T. Kamishima

２．発表標題
Formal Fairness in Machine Learning

３．学会等名
Cybersecurity Cooperation between France and Japan / Intermediate Workshop

４．発表年
2019年

〔図書〕　計2件

1．著者名 ペドロ・ドミンゴス、神嶌 敏弘	4．発行年 2021年
2．出版社 講談社	5．総ページ数 522
3．書名 マスターアルゴリズム　世界を再構築する「究極の機械学習」	

1．著者名 神嶌 敏弘	4．発行年 2019年
2．出版社 近代科学社	5．総ページ数 1
3．書名 「機械学習の動向と深層学習の位置づけ」（AI事典第3版）	

〔産業財産権〕

〔その他〕

Fairness-Aware Machine Learning and Data Mining
http://www.kamishima.net/faml/

6．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	馬場　雪乃 (Baba Yukino) (40711453)	筑波大学・システム情報系・准教授 (12102)	

６．研究組織（つづき）

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	鹿島　久嗣 (Kashima Hisashi) (80545583)	京都大学・情報学研究科・教授 (14301)	

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関