研 究 終 了 報 告 書

「安全かつ透明な個別化のためのプライバシ保護データマイニング」

研究期間： 2017 年 10 月～2021 年 3 月
研 究 者： 荒井 ひろみ

1． 研究のねらい

人工知能技術の負の側面として，自動処理に対するユーザーの理解や制御が及ばなくなる可能性が指摘されており，それによりユーザーには倫理的な懸念や不安が生じている．人工知能の社会への浸透においてそれらのリスクを検証し，危険因子を取り除くことは技術の社会受容において重要な課題である．そのため人工知能の倫理的な懸念を明らかにし，問題がないことを示す手法の必要性が高まっている．特にパーソナルデータを利用したサービスにおいて提供データのプライバシー保護や，サービスに用いられるデータが適切に収集，分析されているかといった安全性，サービスの公平性などの倫理的な観点について近年関心が高まっており，これらの倫理面を向上させる技術や，ユーザーに対してサービスの倫理性を説明する透明性の確保が重要になってきた．

本研究課題では，ユーザーからパーソナルデータを収集，利用しユーザーにサービスを提供する事業者を想定し，事業者のデータの収集分析におけるプライバシー保護，安全性，透明性に関わる課題に取り組む．具体的にはデータ提供者のプライバシーを保護しつつデータの学習に悪影響を及ぼすような異常データを検知するためのプライバシー保護データマイニング技術，機械学習モデルの説明における不適切な説明の可能性の検証，及びプライバシーポリシーの記述の正確性やユーザーの説明の理解についての検証及び正確で効果的な説明方法の開発を実施した．また AI の倫理面において問題化した事例の分析や提言，事業者との意見交換や作成ツールの公開等を通じ，これらの技術の応用可能性を探る．

2． 研究成果

（1）概要

本研究課題では，ユーザーからパーソナルデータを収集，利用しユーザーにサービスを提供する事業者を想定し，そのプライバシー保護，安全性，透明性にまたがる課題の発見・解決を目指すものである．データの収集分析におけるプライバシー保護と安全性の両立のための技術の開発および，ユーザーにプライバシー保護や意思決定についての説明を行う際のリスクの分析や適切な説明方法の開発に取り組んだ．具体的には(A)データ提供者のプライバシーを保護しつつ異常データを検知するための暗号を用いたフレームワークの提案，(B)サービスの意思決定に用いられる機械学習モデルの説明のための技術における偽装可能性の指摘，(C)プライバシーポリシーにおける問題点の分析及び説明性向上の検証，を実施した．

（2）詳細

(A)データ提供者のプライバシーを保護しつつ異常データを検知するための暗号を用いたフレームワークの提案

個人情報を含むデータの収集およびその分析を行う上で，如何にして異常データの提供者のみを特定し，それ以外のデータ提供者のプライバシーを保護するかは重要な課題である．しかし何をもって異常データとするのかという規則をデータ収集前に決定することは困難であることから，データ形式や異常値判定規則によらず汎用的に利用可能なプライバシー保護手法であることが望ましい．そこで任意の異常検知手法に適用可能な暗号を用いたプライバシー保護プロトコルを開発し実データを用いた検証を行った[5(3)-4]．具体的にはメッセージ依存開示可能グループ署名（Group signatures with message-dependent opening, GS-MDO）及び非対話開示可能公開鍵暗号（Public key encryption with non-interactive opening, PKENO）からのプライバシー保護異常検知フレームワークの一般的構成を提案した．提案フレームワークにおいて，データ提供者は異常データを提供しない限り匿名性が担保される．またデータと提供者との紐付けを行う管理者を権限によって適切に分離することで(図 1)，データ提供者を単独で特定可能な，いわゆるビッグブラザーが存在しない．実装の結果，提案フレームワークのオーバーヘッドが高々数 10 ミリ秒程に収まることを確認した．さらに実データを用い，提供データのプライバシーを保護しつつ実際に異常検知が行えることを検証した[5(1)-2]．

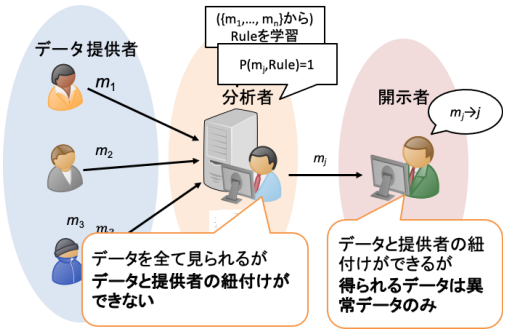

図 1：提案モデルの概要

(B)機械学習モデルの説明技術における偽装可能性の指摘

大勢のパーソナルデータを集めたデータベースから学習をする際，ときに学習モデルは人間の理解が及ばない複雑さになる．それを人間に理解可能な形で説明する主要な方法の一つにモデルの予測プロセスについて単純化や近似を行う方法がある．このような単純化や近似にはプライバシー保護の意味合いもある．一方，近年機械学習モデルに不公平性やバイアスなどの側面があることが問題視されてきている．例えば人種や性別の違いが学習モデルの判断に影響し差別的な扱いをするなどである．そこで，複雑なモデルを単純化したり特徴を抽出して説明を生成する際に，上辺だけ公平性に配慮していると装う(Fairwashing)ことが可能であると考え，そのリスクについて攻撃方法を提案し検証を行った．具体的には学習した複雑な学習モデルを単純なルールリストで近似して説明を行う場合を想定し，説明モデルの生成に元のモデルとの近さおよび近似モデルの公平さについての多目的最適化問題とし準最適解を列挙し，偽装に都合のよいモデルを選択する攻撃方法を提案した(図 2)．実際に提案攻撃方法を用いると説明において人種や性別の情報を用いずに公正なプロセスで予測を行っていると装うこと(rationalization)が可能であることを示した．本研究では学習モデルとしては大規模なルールリスト，説明方法としては短縮したルー

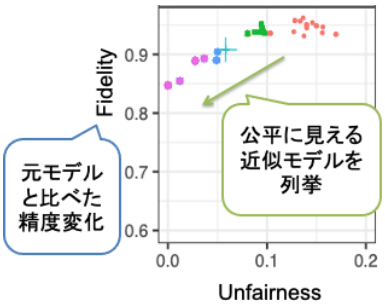

図 2：説明モデル列挙の例
(Adult income データの場合)

ルリストを用いたが，この方式は別の学習モデルにも拡張可能である[5(1)-2，5(3)-4]．報告時点では偽装の可能性の指摘までの研究となったが，当初の最終目標であった偽装の検出や，検出におけるプライバシー保護について引き続き取り組みたい．

(C) プライバシーポリシーにおける問題点の分析及び説明性向上の検証

まずパーソナルデータの利用における倫理的課題について整理した．広告やオンラインニュースなどでのパーソナルデータを用いたプロファイリングのユーザーへの影響の大きさを鑑みその効用とリスクについて主にプライバシーリスクや公平性について論点を整理し，プロファイリングを健全に実装するための自主的取組に関するチェックリストの作成に参加した[5(3)-1]．また採用におけるプロファイリングサービスを題材にプライバシーポリシーの同意における課題について整理した[5(3)-2]．人事や採用活動においてパーソナルデータを基にしたプロファイリングや分析結果をもとにしたスコアリングが普及しつつあるなか，いわゆる「リクナビDMP フォロー事件」が 2019 年 8 月に社会問題化したのを受け，この経緯を時系列で整理し法的・倫理的な論点を概観した．特に個人が識別可能な情報を第三者に譲渡するにもかかわらず「特定の個人を識別できない」と説明していたこと，サービス移行時に個人データの扱いを変更したがプライバシーポリシーに不備を生じさせ一部のユーザーから有効な同意を得ていなかったことが個人情報保護法，職業安定法に関し問題であったと分析された．

プライバシーポリシーの難解さや不正確さの問題を改善するため，情報の流れを正確にユーザーに読みやすく記述する方法を開発している．Nissenbaum の文脈整合性の理論に基づき，実際の日本語のプライバシーポリシーにおける情報の流れについて，流れを明確化するアノテーションを実施し流れの記述についての分析を行った[5(3)-6]．文脈整合性において情報の流れを構成するエンティティである sender，recipient，subject，情報を説明する attribute，情報の伝達条件 transmission principle をアノテーションの実施にあたり，文脈整合性の検証の既存研究が英語によるガイドラインであること，ガイドラインが抽象的でアノテーションの揺れが大きくなることから．日本語対応およびガイドラインの精密化を行った．さらに3つのプライバシーポリシーの抜粋に対し上述のガイドラインに基づき複数人のアノテーターによるアノテーションを実施した．アノテーション結果から日本語におけるプライバシーポリシー記述における特徴や問題点として，日本語では subject の欠損や sender，recipient，subject の数が少ない傾向があることがわかった．また曖昧な語が特定のパラメータにおいて頻繁に利用されるケースが見られた．一方で曖昧表現の有害性は場合により，例えばパーソナルデータの第三者提供の条件や提供先などが曖昧になるような場合は有害と考察される．これらの結果から，現状のプライバシーポリシーにおける曖昧性や欠損について改善することはユーザーのプライバシーポリシーの理解の向上につながると期待される．COVID-19 等の影響により実施が遅れているが，研究の最終目的である典型的な情報の流れについての正確な記述方法の指針やユーザーに説明する方法の開発を行っている．

3. 今後の展開

これまでの研究において説明における偽装可能性を見出したが，リスクの指摘にとどまっている．現在このような偽装の検出や防止のための方法の開発を進めている．偽装検出におけるプライバシー保護技術の適用も検討する，説明の適切さ，あり方について，調査研究も

実施し，機械学習の安全な説明方法についての指針をまとめたい．
また，プライバシーポリシーについての研究は現在典型的な情報の流れについての正確な記述方法のガイドラインの作成およびユーザーにわかりやすく説明する描画方法の開発，及びユーザー調査により実際のユーザーの理解につながっているか検証を行う計画である．さらに最終的には提案する記述方法の指針やツールのプライバシーポリシーの作成や検証への利用を目指す．より利用を促進するために，ワークショップや事業者，法務関係者等との意見交換を通じ関連事業者や研究者との連携を試みる．

4. 自己評価

本研究は，パーソナルデータを利用した人工知能等のサービスにおいて，プライバシー保護，安全性，透明性にまたがる課題について取り組むものである．研究分野の状況の変化や研究内容の検討を受けた計画変更はあったものの，ある程度の学術的な成果を得，トップカンファレンスへの論文採択や国内学会での受賞に至った．また，とくに透明性の研究については，説明において公平性などのある倫理的側面についてある程度操作できる可能性があるというリスクを指摘した AI の信頼性を高める上でも重要な視点を世界にさきがけて得ることができた．また，領域会議における助言や研究者間の交流の影響もあり，当初の計画にはなかったが自然言語での説明性の向上に関わる研究に取り組むことができ，国内会議の受賞等研究コミュニティ内でも評価を得ている．
研究成果については，セミナーやシンポジウムでの発表や提言書，解説論文による研究成果の普及活動などを実施した．また，当該研究の実施およびその成果についての事業者や人文社会学系の研究者との意見交換を通じ，パーソナルデータの利用におけるプライバシーの懸念や，パーソナルデータを利用したユーザー向けの予測などのサービスにおける倫理的問題，それらの解消のための説明技術における課題等の実社会利用時の課題を明確化した．これらの研究成果およびその普及は，パーソナルデータの利用における様々な障壁を解消しパーソナルデータをより信頼される形で利活用するシステムや制度の設計に寄与すると期待される．

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：3件

1. Ulrich Aïvodji, Hiromi Arai, Olivier Fortineau, Sébastien Gambs, Satoshi Hara, Alain Tapp, Fairwashing: the risk of rationalization, 2019, Proceedings of the 36th International Conference on Machine Learning (ICML 2019).
モデルを説明のために情報を落とし単純化する際に，上辺だけの公正性に配慮していると装う(Fairwashing)リスクについて検証を行った．具体的には学習した複雑な学習モデルを単純なルールリストで近似して説明を行う場合に，説明において人種や性別の情報を用いずに公正なプロセスで予測を行っていると装うこと(rationalization)が可能であることを示した．
2. Hiromi Arai, Keita Emura, Takuya Hayashi, Privacy-Preserving Data Analysis: Providing Traceability without Big Brother, 2021, IEICE transaction, Vol.E104-A,No.1,pp. XX.

データ提供者のプライバシーを保護しつつ異常データを検知するための，メッセージ依存開示可能グループ署名及び非対話開示可能公開鍵暗号からのプライバシー保護異常検知フレームワークの一般的構成を提案した．提案フレームワークにおいて，データ提供者は異常データを提供しない限り匿名性が担保される．またデータと提供者との紐付けを行う管理者を権限によって適切に分離することで，データ提供者を単独で特定可能な，いわゆるビッグブラザーが存在しない．

（2）特許出願

研究期間累積件数：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）

1.（著作物） パーソナルデータ+α研究会，「プロファイリングに関する提言案」、および同付属 中間報告書」の公表

2.（学会発表） 工藤郁子，荒井ひろみ，江間有沙，採用におけるプロファイリング・サービスの倫理的課題，2020 年度人工知能学会全国大会，2020.

3.（原著論文） Nenad Tomasev, Julien Cornebise, Frank Hutter, Shakir Mohamed, Angela Picciariello, Bec Connelly, Danielle CM Belgrave, Daphne Ezer, Fanny Cachat van der Haert, Frank Mugisha, Gerald Abila, Hiromi Arai, Hisham Almiraat, Julia Proskurnia, Kyle Snyder, Mihoko Otake-Matsuura, Mustafa Othman, Tobias Glasmachers, Wilfried de Wever, Yee Whye Teh, Mohammad Emtiyaz Khan, Ruben De Winne, Tom Schaul, Claudia Clopath, AI for social good: unlocking the opportunity for positive impact, Nature Communications, Vol. 11, Issue 1, pp. 1-6, 2020.

4.（学会発表・受賞） PWS 優秀論文賞 （荒井ひろみ，江村恵太，林卓也．プライバシー保護異常検知フレームワーク，コンピュータセキュリティシンポジウム 2017 論文集, vol. 2017, no.2, 2017 に対して.）

5.（学会発表・受賞） 全国大会優秀賞 （荒井 ひろみ，Ulrich Aïvodji，Olivier Fortineau，Sébastien Gambs，原 聡，Alain Tapp，機械学習の説明における公正さの偽装，2020 年度人工知能学会全国大会，2020 に対して）

6.（学会発表・受賞） CSS 優秀論文賞 （荒井 ひろみ，仲宗根 勝仁 ，濱本 鴻志，日本語のプライバシーポリシーにおける文脈完全性に基づいた情報抽出の一検討，コンピュータセキュリティシンポジウム 2020 論文集，に対して）