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　５ 年　 ６ 月 １２ 日現在

機関番号：８２４０１
研究種目：若手研究(B)
研究期間：2017～2022
課題番号：１７Ｋ１２６９６

研究課題名（和文）人工知能利用に伴うプライバシリスクの評価手法の開発

研究課題名（英文）The development of privacy risk evaluation methods associated with the use of artificial intelligence

研究代表者
荒井　ひろみ（Arai, Hiromi）

国立研究開発法人理化学研究所・革新知能統合研究センター・ユニットリーダー

研究者番号：２０６３１７８２

交付決定額（研究期間全体）：（直接経費）　　3,100,000円

研究成果の概要（和文）：画像などのデータ形式におけるプライバシー保護手法とそのリスクについて検証した。また、利活用におけるデータの流れを考慮したプライバシーリスクについての検討も行った。パーソナルデータ加工におけるプライバシーリスクについて整理し、説明責任のあり方を議論した。また、情報の流れについて整理・記述する方法を用い調査研究を実施した。また、人口集団についてのデータの偏りがある場合の学習結果やその評価への影響について検討した.

研究成果の学術的意義や社会的意義
様々な人工知能技術において、顔画像や行動履歴などのパーソナルデータを収集、加工し、共有したり開示したりする場面が想定される。そのような場面におけるプライバシーのリスクの評価、整理は、リスクを考慮したデータ利活用に資すると期待される。我々はいくつかのデータ利活用場面を想定したプライバシー保護およびリスク評価、プライバシーリスクの整理方法、データの偏りがある場合の学習や評価への検証を実施した。

研究成果の概要（英文）：We explored privacy preserving technologies and privacy risks in the use of them in data formats such as images. We also examined privacy risks in consideration of data flow in utilization. We organized privacy risks in personal data processing and discussed accountability and transparency. We conducted research using methods to organize and describe the flow of information. We also examined the impact of demographic bias in data on study results and their evaluation.

研究分野：プライバシー

キーワード：プライバシー保護

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

個人に関する秘匿性の高いデータの取得及び利用においてデータのプライバシ保護は不可欠である。一方でパーソナルデータは個別化したデータサービスや医療などのための人工知能技術に利用されることも多く、プライバシの問題でデータの利用を過度に妨げないことが望ましい。パーソナルデータの利用におけるプライバシリスクを評価できれば、高リスクなデータ利用の際にはアクセス制限を行う等の工夫をすることができる。

パーソナルデータをプライバシーを保護しつつ利活用するための方法の一つにプライバシー保護技術の適用がある。特に匿名化レコードや機械学習モデル、統計データなど、レコードを陽に含まないようなデータ加工はプライバシーを保護したデータ開示の 1 つの方法である。しかしこのようなデータ加工によるプライバシ保護において、開示データから元のデータを推定されるリスクが指摘されている。

また、個人に関するセンシティブな情報が他の情報から推定されるようなプロファイリングによるプライバシ侵害が懸念されている。様々な予測のための知識を得て人工知能として利用することが可能になってきているため，一見関連性の低いデータからでもセンシティブ情報が推定される可能性がでてきた。

パーソナルデータやそれを加工したものは一般的に複数のエンティティ間で流通し、様々な目的で利用される。こういった様々なテータ利用の局面において、これらのリスクを評価することができれば適切なデータ利用の助けになると期待される。

２．研究の目的

パーソナルデータを機械学習などを応用した人工知能技術に用いるような場合には、パーソナルデータの収集、加工、共有、学習モデルの公開などを通じ、エンティティ間の流通やデータ加工のプロセスが存在する。特にデータ共有については、匿名化や統計化などの加工をしてプライバシ保護に配慮したデータが共有されることが多い。このようにプライバシー保護に配慮したデータにおいても、元データが推定されるリスクが存在する。また意図せぬプロファイリングによるセンシティブ情報推定のリスクがある。

これまで元データの推定に関しては、特定のデータや方式に特化した方式が様々提案されてきていた。本研究では高次元で構造を持つデータに対するデータを対象としたデータ共有方法とリスクを扱う。

プロファイリングについては当初はデータ間の関係に関する知識を扱うことを想定していたが、それでは近年の急速なデータ知識の蓄積を考慮するとリスク評価が困難であったため、予測アプリケーションによってプロファイリングを行われるリスクを扱った。

また、実際のパーソナルデータの流通は複雑であり多くのエンティティが関わるため、当初は２つのエンティティを想定していたが、多くのエンティティがいる状況を整理しプライバシーリスクを検討するようにした。

３．研究の方法

パーソナルデータ利用において、データの共有や開示、利用におけるプライバシーリスクの評価のための技術開発を行う。特に画像やトレースなどの高次元で構造を持つデータは保護の難易度が高い。そのようなデータの保護技術を開発し、またその保護技術を適用したデータを開示する際のプライバシーリスクを評価する。

さらに、パーソナルデータやそれを加工したものは様々な用途、様々なエンティティによって扱われている。場合によってはユーザーが望まないプロファイリングやデータ利用を行われる可能性もある。しかしパーソナルデータの利用は時として複雑でわかりにくい。そのためパーソナルデータやその加工データの流れや利用を整理、評価する方法を提案する。

また、画像データのアプリケーションにおけるプライバシーリスクについて考察するために、顔識別における認識・誤認識の評価を行った。顔認証技術は、例えば犯罪者リストとの照合など人

生に重大な影響を与える用途におけるプロファイリングを行うことも可能である。それ以外にも意図せぬプロファイリング用途に用いられる可能性がある。また、データに人口集団内についての不均等がある場合に、それはデータを利用したシステムの振る舞いやリスクなどの評価に影響を与えると考えられる。このようなデータバイアスの機械学習を用いた顔認証システムへの影響の評価を行う。

４．研究成果

まず以下のような元データを推定されるリスクに関する研究を実施した。
機械学習モデルは陽にパーソナルデータのレコードを含まないことが多いが、そのような場合のプライバシーリスクの評価については、ある知識や計算能力をもった“攻撃者”によるレコード再特定や属性推定などの攻撃を想定し、その攻撃の成功確率を評価することが一般的である。そこで学習モデルの利用における学習データのプライバシーリスクなどについての既存研究についての整理を行った。
また、顔画像の匿名化方法の提案とその評価を行った。顔画像データは「高次元」及び「顔の構造が含まれる」という二つの特徴を持つデータである。そのため、顔画像データの匿名化では高次元データに対する計算量の観点と顔の構造維持という有用性の観点が重要となる。そこでまず、k-匿名化アルゴリズムである Mondrian と次元削減及び特徴抽出の手法として知られる非負値行列因子分解を組み合わせ、上記の要点を満たす顔画像データの匿名化手法を提案した。さらにその k-匿名性を確認し、さらに最大知識攻撃者モデルによる再識別攻撃によってプライバシーリスクを評価した。公開されている顔画像データセットを使用し、コスト及び匿名化されたデータの有用性について、提案手法と Mondrian 単体での匿名化を比較し、提案手法の優位性を確認した。
法制度上では、十分な匿名化が達成されれば、そのようなデータはパーソナルデータでなくなり、規制の対象外となる。他方でそのような非パーソナルデータ化が達成されるためにはどの程度の匿名化処理が行われればよいか、非パーソナルデータ化が達成されたことを示すために誰がどのような責任を果たすべきなのかの所在は研究時点で明らかでなかった。そこでパーソナルデータの保護制度における論点を整理し、技術面における具体的要件や実装について検討を行った。プライバシーに関する規制やガイドラインから各ステークホルダーに求められる要件を整理し、さらにこれまでのプライバシー保護技術研究を非パーソナルデータ化の視点から整理した。

また、パーソナルデータの利活用における複雑な情報の流れを整理し適切性を判断する方法の改良を試みた。このような方法の一つに文脈完全性という考え方がある。これを利用し、データの流れを抽出し、利用目的やデータの譲渡先の実態について調査を行った。サービス提供者がユーザーのパーソナルデータを収集、利用する際に、サービス提供者が利用規約等の説明をユーザーに提示し同意を取得することが一般的に行われている。パーソナルデータの扱い方に関してはプライバシーポリシーとしてユーザーに提示される。文脈完全性の理論に基づき，日本語のプライバシーポリシーに対してアノテーションを実施し情報の流れを明確化するパラメータの組を抽出し、ータ共有のパターンや利用目的についての考察を試みた。さらに利用目的や加工についても整理するようにパラメータを拡張したアノテーション方法を検討した。

また顔認証におけるプロファイリングやプライバシーについての考察の一側面として、人口統計学的データについての偏りについて取り上げた。ある特定のセンシティブ情報をもつグループがネガティブな特徴を持つと誤認識されやすいような状況は不公平であると言える。このような不公平さについて特に表現学習に基づく汎用的な顔認証システムを取り上げ、その公平性評価方法について検討、評価を行った。学習データセットや評価用のデータセットにおいて、ある特定のセンシティブ情報を持つサブグループで分けることを考える。サブグループ間のサンプル数の割合に偏りがあり、特定のサブグループのサンプル数が小さいような場合を考える。また、公平性評価値として、サブグループごとの認証精度などの評価値の差を導入した。MORPH データセットを用いた実験により、偏りのある学習セットにより作成された顔特徴抽出器を用いた場合に、認証精度の偏りが生じることや、偏りのあるテストセットを用いることで意図的に高い公平性評価値を算出可能である場合があることを示した。このような公平性評価を不当に操作しうるような場合の対応として、属性ごとのしきい値を設定することで認証精度への影響を軽減できることなどを示した。

５．主な発表論文等

〔雑誌論文〕　計1件（うち査読付論文　0件／うち国際共著　0件／うちオープンアクセス　0件）

１．著者名 荒井ひろみ	４．巻 32
２．論文標題 私のブックマーク「機械学習のプライバシーとセキュリティ」	５．発行年 2017年
３．雑誌名 人工知能	６．最初と最後の頁 804-807
掲載論文のDOI（デジタルオブジェクト識別子） 10.11517/jjsai.32.5_804	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 －

〔学会発表〕　計5件（うち招待講演　0件／うち国際学会　0件）

１．発表者名 大木哲史，荒井ひろみ
２．発表標題 顔認証における公平性評価の一検討
３．学会等名 コンピュータセキュリティシンポジウム2021
４．発表年 2021年

１．発表者名 荒井ひろみ,仲宗根勝仁,濵本鴻志
２．発表標題 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
３．学会等名 コンピュータセキュリティシンポジウム2020
４．発表年 2020年

１．発表者名 荒井ひろみ
２．発表標題 機械学習のセキュリティーとプライバシー
３．学会等名 computer security symposium 2018
４．発表年 2019年

１．発表者名 荒井ひろみ，加藤尚徳
２．発表標題 非パーソナルデータ化の実装についての一検討
３．学会等名 研究報告電子化知的財産・社会基盤（EIP）
４．発表年 2017年

１．発表者名 出町彰啓，荒井 ひろみ，中川裕志
２．発表標題 顔画像データにおける匿名化手法
３．学会等名 コンピュータセキュリティシンポジウム2017
４．発表年 2017年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

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

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