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　　元 年　　６ 月 １０ 日現在

機関番号：３４５０４
研究種目：挑戦的萌芽研究
研究期間：2016～2018
課題番号：１６Ｋ１２５６４

研究課題名（和文）教育データ匿名化ツールの開発

研究課題名（英文）The Development of Anonymization Tool for Educational Data

研究代表者
武田　俊之（TAKEDA, Toshiyuki）

関西学院大学・高等教育推進センター・教育技術主事

研究者番号：７０２２７０３１

交付決定額（研究期間全体）：（直接経費）　　2,400,000円

研究成果の概要（和文）：本研究の目的は、教育分野のデータについて、データの有用性とプライバシー保護のバランスの取れた匿名化ツールを開発することである。本研究の研究成果は以下の通りである。
(1)教育におけるプライバシー保護の法制や倫理的な要求について整理、検討、考察をおこなった。(2)教育研究のレビューをおこない、教育データ（変数)の匿名化上の性質の検討をおこなった。モデリングと測定の方法としてアナリシスパターンの適用を試みた。(3)匿名化を実現するPythonパッケージを既存ツールのラッパーとして開発した。既存の匿名化ツールについては、ドキュメントの日本語化をおこなった。

研究成果の学術的意義や社会的意義
本研究の教育研究のデータ取扱の倫理に関して整理検討をおこない、医療分野以上に教育研究はプライバシー上の課題が多いことを示すことができた。
本研究で開発したツールを活用することによって、データ主体のプライバシーを保護しながら、情報損失を評価した上での分析が可能になるであろう。これによって、データ主体のデータ二次利用への懸念が低減されて、無理な同意を求めることなく教育研究におけるデータ共有が促進されることが期待される。また、本研究で開発した匿名化ツールは他分野でも有用であり、諸外国にくらべて遅れがちな匿名化とデータ共有の促進も期待される。

研究成果の概要（英文）：The purpose of this study is to develop an anonymization tool with balancing disclosure risks and usefulness in education field. The research results of this research are as follows: (1) We reviewed legal and ethical requirements for privacy protection in education. (2) We created process models of anonymizing educational data and explored the nature of modeling, systems, measurement and variables based on analysis pattern. (3) We developed a Python package for anonymization. This tool is a wrapper library of existing tools.

研究分野：教育工学

キーワード：データ匿名化　ラーニング・アナリティクス　プライバシー保護

様　式　C－１９、F－１９－１、Z－１９、CK－１９（共通）

１．研究開始当初の背景

学習履歴データの活用によって、教育活動の方法や環境と、成績などの成果の関連を見出して、改善に役立てるための研究領域「ラーニング・アナリティクス」が期待されている。実践から得られたデータの共有・流通と、それにもとづく多面的な二次分析によって、教育実践の理解と教育改善のための強固な証拠がもたらされるであろう。

一方で、データの利用とプライバシー保護にはトレードオフの関係がある。プライバシー侵害の可能性は、利用者や研究者などのデータ提供者の不安につながり、データの利活用を阻害する要因となりうる。

データの有用性を損ねることなく、プライバシーを保護する方法の一つとして、匿名化手法(統計的開示制御: Statistical Disclosure Control やプライバシー保護データマイニング）の研究開発が進んでいる。

一方で、データを生みだす実践領域と、生成されるデータの性質はさまざまであり、データの領域固有な性質に適した匿名化の方法を開発する必要がある。たとえば、医療分野では実践的なリスク管理手法が開発・実施されている（El Emam、2014）が教育分野においては、このようなリスク管理の枠組はまだ確立していない。

２．研究の目的

本研究の目的は、教育データの保有者（研究者、教育機関、サービス組織）自身が匿名化作業（あるいは、その一部）をおこなうことができるようツールを作成することである。そのためには以下を達成しなければならない。

(1) 匿名化に関連した教育データの性質を明らかにすること。その際、教育実践および研究の側面だけではなく、法律、社会的な側面と倫理についても検討をおこなうこと。匿名化研究の成果を実践的に適用している医療分野では、収集されるデータの種類と性質が整理されている。

(2) 教育データに適した匿名化手法と指標を特定すること。実際のデータを匿名化するときには、データの性質を踏まえた上で、プライバシー保護とデータ利活用のバランスがとれた手法と指標を採用する必要がある。

(3) 匿名化ツールをデータ提供者（研究者、教育機関、サービス組織）にとって、理解可能な手続きと操作になるよう設計すること。データの所有者（組織）にとって、匿名化のリスクが十分に低いことを理解できなければ、利活用のためのデータ提供には不安が大きいであろう。

３．研究の方法

以下の 3 点を並行して研究を進めた。

(1) 教育データに関するプライバシー保護および倫理的側面に関する情報の収集、整理、検討をおこなった。特に施行された個人情報保護法と EU の GDPR（General Data Protection Regulation）の情報と、機微情報を取り扱う医療情報分野、欧米における教育データ取り扱いの動向について整理検討した。教育データの倫理的な取り扱いについては、研究、教育実践、法律、制度の各側面からレビューをおこなった。また、欧米の教育に関連する事例および日本の医療情報の事例の収集と、教育で生成、収集されるデータの種類の分類をおこない、倫理的な問題が生じるパターンについて検討をおこなった。

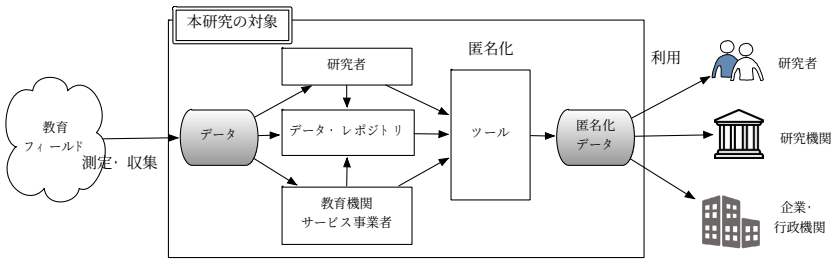
(2) 教育研究のレビューをおこない、教育データ(変数、属性)の匿名化上の性質の検討をおこなった。モデリングと測定の方法としてアナリシスパターンの適用を試みた。

(3) 匿名化ツールの開発。

匿名化を実現するための Python パッケージを開発した。そのために、まず、既存の匿名化ツールの調査、選択をおこない、ドキュメントの日本語化をおこなった。デザイン上の選択として有用なツールのラッパーとして開発をおこなった。

① まず、関連するツールのレビューをおこない、機能等で有用な ARX（Java / GUI)、sdcMicro（R / GUI)、Google Data Loss Prevention（DLC=オンラインライブラリ）について詳細な調査をおこない、日本語のドキュメントを作成した。

② これらのツールの機能を利用する Python のラッパーライブラリを開発することとした。デザイン原則としては、データ分析プロセスの一部としての匿名化（下図）を支援するために、複数の設定を変えながら匿名化モデルとその評価をおこなえること、新しい匿名化手法をアドオンできる仕組みとすることである。



４．研究成果

(1) 教育データとプライバシー保護および倫理の整理、検討

教育データは個人情報を含む情報の集合物であって個人情報保護法や GDPR の対象である。国や公共団体は個人情報保護法とは別の法律の上で義務を負うが法理としては同様と考えられる。また、研究倫理、実践倫理としてのプライバシー保護は法を包含する範囲の広いものと考えられる。

個人情報保護法では個人情報の要件として「特定の個人を識別することができるもの」とする。個人は特定できないがひとりひとりを識別できることを識別化という。氏名を記号で置き換える仮名化（pseudonymization）は非特定化であるが非識別化ではない。しかし、他の情報と連結することができる仮名化は匿名性があるとはいえないと一般的に考えられている。教育データも仮名化ではなく非識別化が必要と思われる。

プライバシーは文脈によって内実が異なる。たとえばあるクラス内で成績が開示されていることはプライバシー上問題がないが、他の学校の生徒に開示されているならプライバシーが侵害されている可能性が高くなる。グループ、組織等を横断してデータが共有されるときにはプライバシーが侵害される可能性が高くなるので、データは非識別化されることが望ましい。

プライバシー・バイ・デザインは事業の計画の前にプライバシーに関する事項を埋め込んでいくことである。研究の組織的な倫理審査（IRBs）はプライバシー・バイ・デザインの一種であると考えられる。日本の教育研究においても IRBs を通していない研究の査読は受けつけないなどが望ましいかもしれない。一方、教育実践のデータ（テスト、アンケート、入試成績、ポートフォリオ）などの取得、連結について、倫理審査の仕組みはほぼ皆無である。プライバシー・バイ・デザインのアプローチなど非識別化を含めた共有のためのガイドラインの設定が必要と思われる。

Drachsler ら（2016）は、ラーニング・アナリティクスの倫理とプライバシーについてレビューをおこなった上で、ステークホルダーが安心できるような実践のための DELICATE チェックリストを開発した。内容は、DETERMINATION（決定）、EXPLAIN（説明）、LEGITIMATE（遵法）、INVOLVE（関与）、CONSENT（同意）、ANONYMISE（匿名）、TECHNICAL（技術）、EXTERNAL（外部連携）である。本研究は、LEGITIMATE、CONSENT、ANONYMISE、EXTERNAL に関連している。

(2) 教育データの特徴

教育方法の研究（介入、観察のいずれの場合も）においては、ある教育方法をフィールドに適用した結果をデータとして得る。データを共有する場合のパターンと、データの比較、連結において重要となる、それぞれの研究におけるモデル、変数とその測定について考察、整理をおこなった。

① データの共有と二次利用のパターン

教育データの連結、統合、二次利用する主なパターンは以下の 3 つである。

横断的データは、別の学校や教室で同じ方法を用いられた場合、統合した横断的データ（cross-sectional data）を作成することが可能である。分析において学習者の属性は有用であるが、二次利用など分析の範囲が広がる場合には、個人情報やプライバシーの保護に留意しなければならない。

縦断的データは、教育の効果は長期的にあらわれることも少なくない。ある特定の個人のデータを連結した縦断的データ（longitudinal data）から得られる知見は少なくないであろう。教育における縦断的な効果の期間は、医療等と比較しても極めて長くなる場合がある。たとえば、幼少期の教育結果が 30 歳の職業的学習と結びつけられるようなことがある。

縦断的データにおいては、個体の識別性を確保してデータを蓄積、連結する必要がある。横断的データにくらべて、個人情報、プライバシー上の問題が生じやすい。研究上の取扱としては、分析者の限定、安全管理の徹底、事前の理論的考察と最小限の分析、分析に必要な変数以外の秘匿、最小限の数にサンプリングなどが考えられる。また、グループ化による分析が可能であるなら匿名化したデータを用いることができる。

メタ分析（meta-analysis）は、複数の統計的研究の結果を比較、統合する手法である。多くの場合、メタ分析の結果は措置とアウトカムの効果量を用いて表現される。教育分野では John Hattie（2008）の Visible Learning の achievement を達成した教育方法の比較が有名である。メタ分析は論文に記述されている内容のみで分析が可能であるが、詳細な検証のために個別のデータに下りて再分析をおこなうことは有用であろう。

外部ツール	ツールの所在	利用の要件
ARX	ローカル。要インストール。	py4j パッケージ（python）
sdcMicro	ローカル。 要インストール。	RPy2 パッケージ（python） R 言語
DLC	クラウド。要登録。	Google ライブラリ（Python）

② データ変数の性質の検討

どの変数が機微であるか、どのような変数の種別、属性について匿名化をすべきか検討をおこなった（武田、2016; 武田、2019）。機微情報以外の匿名化変数を自動的に識別することは困難であるが、（結合時を含めた）データ変数の組み合わせの個人識別性は匿名化ツールで分析可能になった。検討した学習活動に関するデータのソースには以下のようなものがある。

学習の種類	フォーマル（formal）/ ノンフォーマル（non-formal） / インフォーマル（informal）
データ・ソースのレイヤ	個人 / インタラクション / 授業 / カリキュラム / 学部 / 大学 / 行政
システム運用の形態	オンプレミス / リモート / クラウド
システムの種類	LMS / SIS（Student Information System） / e ポートフォリオ / 評価 / その他内部 / ソーシャルメディアなど
データの観測方法	仮説にもとづく測定（measured）/ システムが生成したデータの収集（collected）/ 外部からのインポート（imported） / 分析・推定値（inferred or computed）
データの種類	成果（成績等）/ アウトプット（テスト、レポート、作品等）/ アンケート / 行動履歴 / 生理指標 / 学習環境
個人データ	デモグラフィック / プロフィール / 事前の経験 / パフォーマンス / 心理調査 / 生理/ その他指標
データの型	表 / 数値 / テキスト / Boolean / 複合型

(3) 匿名化ツールの開発

教育研究者および実践者自身が匿名化にたずさわることが可能なツールを開発した。
まず、匿名化ツールの調査をおこない、ARX（GUI および Java で記述）、sdcMicro（R で記述）、Google Data Loss Prevention（DLC/REST API）を有用なツールとして選び、機能を洗い出して、日本語のチュートリアル資料を作成した。

本研究で開発したツール（以下本ツール）の特徴は、①ARX、sdcMicro、DLC を単独または組み合わせて利用することが可能、②匿名化構成情報を設定することによってさまざまな値で識別のリスクと情報損失を計算可能、③Python の標準的データフレーム（pandas）へのアドオンであるために視覚化等既存ツールと組み合わせ可能、の 3 点である。既存ツールの使用に必要な事項は以下の通りである。

本ツールは、表形式のデータの匿名化の手順を以下のように想定している。
1) データの変数について、識別情報と準識別変数を匿名構成情報として指定する。
2) 必要があればその他のメタデータを匿名構成情報に追加する。
3) データのリスクを計算する。
4) 準識別変数の組み合わせについて、k-匿名性、l-多様性などの匿名性の要求を匿名構成情報に追加する。
5) 匿名化の方法（の組み合わせ）を指定する。
6) 匿名化を実行する。
7) 結果のリスクと情報損失を計算する。

匿名化の方法は以下を指定できる：①識別情報の削除、②識別情報の仮名化法の指定、③トップ・ボトムコーディング、④リコーディング、⑤リサンプリング、⑥スワッピング、⑦ノイズの追加、⑧並び替え、⑨テキストデータ内の識別情報および任意の文字列の置換。
本ツールは 2019 年中に公開の予定である。

引用文献

① K. El Emam、Anonymizing Health Data、 O’reilly Media、2014
② ARX https://arx.deidentifier.org/
③ sdcMicro https://cran.r-project.org/web/packages/sdcMicro/index.html
④ Google Cloud Data Loss Prevention (DLP) https://cloud.google.com/dlp/docs/?hl=ja
⑤ H. Drachsler、W. Greller、Privacy and Analytics – it’s a DELICATE Issue A Checklist for Trusted Learning Analytics. 6th Conference on Learning Analytics and Knowledge、89-98、2016、https://doi.org/10.1145/2883851.2883893
⑥ J. Hattie、Visible Learning: A Synthesis of Over 800 Meta-Analyses Relating to Achievement. NY: Routledge、2008
⑦ 武田俊之、教育データ分析のアナリシスパターンの検討、情報処理学会研究報告教育学習支援情報システム(CLE)、2019、pp. 1-4
⑧ 武田俊之、ラーニング・アナリティクスと教育データ、Cybermedia Forum、17 巻、2018、pp. 5-9
https://www.cmc.osaka-u.ac.jp/wp-content/uploads/2018/05/forum2017.pdf

５．主な発表論文等
〔雑誌論文〕（計 1 件）
① 武田俊之、ラーニング・アナリティクスと教育データ、Cybermedia Forum、査読無、17 巻、2018、pp. 5-9、
https://www.cmc.osaka-u.ac.jp/wp-content/uploads/2018/05/forum2017.pdf
〔学会発表〕（計 3 件）
① 武田俊之、教育データ分析のアナリシスパターンの検討、情報処理学会研究報告教育学習支援情報システム(CLE)、2019
② 武田俊之、教育システムと「倫理的に配慮されたデザイン」、教育システム情報学会研究会報告、2019
③ 武田俊之、重田勝介、森秀樹、MOOC データ二次利用のための加工プロセス．情報処理学会研究報告教育学習支援情報システム（CLE)、2016
〔図書〕（計　　件）

〔産業財産権〕
○出願状況（計　　件）

名称：
発明者：
権利者：
種類：
番号：
出願年：
国内外の別：

○取得状況（計　　件）

名称：
発明者：
権利者：
種類：
番号：
取得年：
国内外の別：

〔その他〕
- 本研究の研究成果および知見を以下で公開する。
https://blog.takedato.com/category/個人情報とプライバシー/
- 本研究で開発した匿名化ツールを 2019 年度中に公開、上記ホームページで告知する。

６．研究組織

(1)研究分担者
研究分担者氏名：重田 勝介
ローマ字氏名：SHIGETA, Katsusuke
所属研究機関名：北海道大学
部局名：情報基盤センター
職名：准教授
研究者番号（8 桁）：40451900

研究分担者氏名：森 秀樹
ローマ字氏名：MORI, Hideki
所属研究機関名：東京工業大学
部局名：大学マネジメントセンター
職名：准教授
研究者番号（8 桁）：30527776

(2)研究協力者
研究協力者氏名：
ローマ字氏名：