筑波大学
University of Tsukuba

2023 年 6 月 8 日

報道関係者各位

国立大学法人筑波大学

複数の機関が保有するデータをより安全に統合解析する AI 技術を開発

複数の企業・自治体・病院等がそれぞれ保有する個人情報データに対して、個人を特定可能なデータと容易照合できない抽象化データのみを共有して統合的に解析する、より安全な人工知能技術「容易照合不可データコラボレーション技術」を開発しました。

AI による解析の精度を上げるためには分布の偏りのない十分な数のデータを集めることが重要です。このとき、複数の機関に分散したデータを集めるためには、個人情報やノウハウなどの秘匿したい情報に配慮して安全に統合解析する AI 技術が必要となります。特に個人情報を含む場合、共有するデータと元データに「容易照合性」があるとその利用において制約があることが課題でした。

本研究では、複数の企業・自治体・病院等がそれぞれ保有する個人情報データに対して、個人を特定可能なデータと容易照合できない抽象化データのみを共有し統合的に解析する、より安全な人工知能技術「容易照合不可データコラボレーション技術」を開発しました。この研究では、データ間の容易照合性に対する数学的定義を導入し、検討する枠組みを導入しました。その上で、元データと容易に照合できない抽象化データのみ共有する統合解析アルゴリズムを提案しました。これにより、個人情報を含むデータ解析においてより多くのデータを活用することが可能となり、AI の解析精度の大幅な向上が実現されると考えられます。

具体的な応用例として、複数の医療機関が有する検査・投薬データの統合解析による疾患予測や疾病リスク因子推定や、複数の教育機関の学生データの統合解析による教育効果増進などが挙げられます。また、将来的にはさまざまな機関にある質の高い個人情報データを、元データを保護したまま収集して AI による分析を行う新たなプラットフォームを支える技術としても期待できます。

研究代表者
筑波大学　人工知能科学センター
櫻井 鉄也　センター長

研究の背景

AI による解析の精度を上げるためには分布の偏りのない十分な数のデータを集めることが重要であり、個人情報に配慮して安全に統合解析する多機関分散データ統合解析 AI 技術が求められています。個人を直接特定できないデータであっても、他の情報と容易に照合することにより特定の個人を識別することができる場合、当該情報とあわせて全体として個人情報に該当することがあります（個人情報保護委員会「個人情報の保護に関する法律についてのガイドライン」）。このため、個人情報を含むデータの解析を実施する場合、多機関に分散する元データを変換（または暗号化）して共有し、解析する AI 技術において、共有するデータと元データが「容易照合性」[注1)]を持つ場合には、その利用において制約がありました。

研究内容と成果

今回、複数の企業・自治体・病院等がそれぞれ保有する個人情報データに対して、個人を特定可能なデータと容易照合できない抽象化データ[注2)]のみを共有し統合的に解析する、より安全な人工知能技術「容易照合不可データコラボレーション技術」を開発しました（参考図）。

この研究では、データ間の容易照合性に対する数学的定義を導入し、検討する枠組みを導入しました。その上で、元データと容易に照合できない抽象化データのみ共有する統合解析アルゴリズムを提案しました。データコラボレーション技術では、抽象化データのみを共有することでデータの安全性を担保しつつ、アンカーデータ[注3)]を用いて抽象化データの統合・解析を行います。提案したアルゴリズムでは、元データから抽象化データを生成する処理において、用いられる変換式を破棄することで容易照合を不可能にします。棄却された変換式を必要とすることなく、統合された結果から各機関において高性能な解析モデルを生成する手法を構築しました。これらの処理により、元データに含まれる秘匿性の高い情報の安全性を担保しつつ、元データを共有した場合と同等の解析を実現します。

この技術を、医療分野でのオープンデータを用いた病気リスク診断の AI モデル生成に適用し、既存のデータコラボレーション技術による解析や生データを直接共有した解析と同程度の認識性能を示す AI モデルを構築できることが分かりました。

なお本研究は、国立研究開発法人 新エネルギー・産業技術総合開発機構（NEDO）「人工知能技術適用によるスマート社会の実現」事業において、筑波大学と株式会社 NTT データとの産学共同研究により実施されました。この成果は、Computer Science 分野のトップ 1%以内にランクされる学術誌 Information Fusion に掲載されました。

今後の展開

本研究により、個人情報を含むデータ解析において、個人を特定可能な元データとは容易に照合できないデータのみを共有した統合解析が可能となり、より多くのデータを活用することで AI の解析精度の大幅な向上が実現されると考えられます。具体的な応用例として、複数の医療機関が持つ検査・投薬データの統合解析による疾患予測や疾病リスク因子推定や、複数の教育機関の学生データの統合解析による教育効果増進などが挙げられます。この成果を用いて、さまざまな機関にある質の高いデータを個人情報を保護したまま収集し、AI による分析を行う新たなプラットフォームの開発を進めていきます。

今後は、容易照合不可データコラボレーション技術の社会実装化を進めるとともに、この技術の適用に関して法律面からも検討を行う予定です。本技術は DFFT（Data Free Flow with Trust：信頼性のある自由なデータ流通）[注4)]を支える基盤技術としても期待されます。

参考図

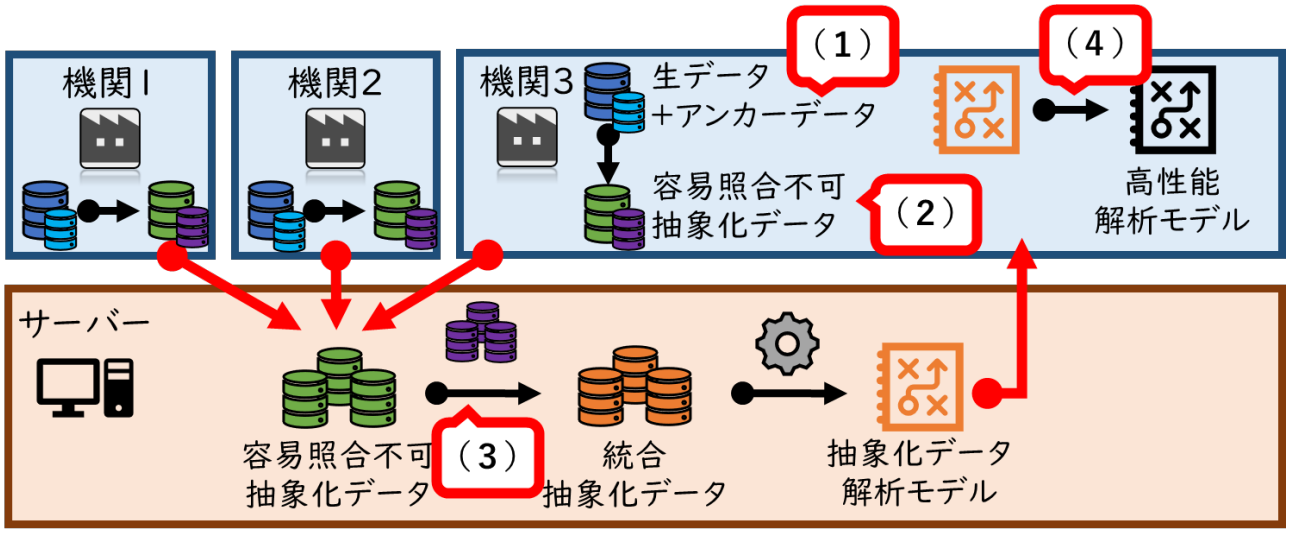

図　本研究で開発した容易照合不可データコラボレーション技術の概要

（1）各機関は、共有可能な疑似データであるアンカーデータを構築・共有する。（2）個人を特定可能な情報を含む生データから、独自の抽象化処理により抽象化データを生成する。抽象化処理の変換式を破棄し、抽象化データをサーバーに共有する。（3）サーバーは抽象化データの統合処理を行い、統合抽象化データを生成し、統合解析を行う。（4）抽象化データ解析モデルをもとに、各機関で生データ統合解析モデルを生成する。

用語解説

注1）　容易照合性

他の情報と容易に照合することができる性質。個人を直接特定できないデータであっても、他の情報と容易に照合することにより特定の個人を識別することができる場合、当該情報とあわせて全体として個人情報に該当することがある（個人情報保護委員会「個人情報の保護に関する法律についてのガイドライン」）。

注2）　抽象化データ

高次元データ（一つの対象について多数の項目を含むデータ）を、元データの情報をできるだけ失わずに低次元データに変換したデータ。

注3）　アンカーデータ

データコラボレーション技術において、各機関の抽象化データを統合するために利用される共有可能な疑似データ。乱数行列等で生成される。

注4）　DFFT（Data Free Flow with Trust：信頼性のある自由なデータ流通）

プライバシーやセキュリティ、知的財産権に関する信頼を確保しながら、ビジネスや社会課題の解決に有益なデータが国境を意識することなく自由に行き来する、国際的に自由なデータ流通の促進を目指すコンセプト。

研究資金

本研究は、国立研究開発法人 新エネルギー・産業技術総合開発機構（NEDO）「人工知能技術適用によるスマート社会の実現」（JPNP18010）の一環として実施されました。

掲載論文

【題　名】 Non-readily identifiable data collaboration analysis for multiple datasets including personal information
（個人情報を含む複数データのための容易照合不可データコラボレーション解析）

【著者名】 Akira Imakura[(1)], Tetsuya Sakurai[(1)], Yukihiko Okada[(1)], Tomoya Fujii[(2)], Teppei Sakamoto[(2)], Hiroyuki Abe [(2)]
(1) University of Tsukuba, (2) NTT DATA Corporation

【掲載誌】 *Information Fusion*
【掲載日】 2023 年 5 月 4 日
【DOI】 https://doi.org/10.1016/j.inffus.2023.101826

問合わせ先

【研究に関すること】
櫻井　鉄也（さくらい　てつや）
筑波大学 人工知能科学センター　センター長／システム情報系　教授
URL: https://www.cs.tsukuba.ac.jp/~sakurai/

【取材・報道に関すること】
筑波大学広報局
TEL: 029-853-2040
E-mail: kohositu@un.tsukuba.ac.jp