公募企画
公募企画シンポジウム6
安心・安全なビッグデータの流通プラットホームとセキュリティ基盤技術
2017年11月22日(水) 08:45 ～ 10:45 C会場 (10F 会議室1001)

[3-C-1-PS6-2] プライバシーを保護した多機関データ突合システムについて

宮地 充子, 中正 和久 （大阪大学大学院工学研究科 電気電子情報工学専攻）

各機関で独立に保管される医療情報を新薬開発や治療法の改善，健康診断結果の追跡による保健指導などに利活用することは重要である．その際，異なる機関が保管する同一患者のデータを突合できることは必須である．しかしながら，医療機関において独立に管理された医療情報を患者の名前，住所などの機微情報を漏らすことなく，突合することは容易ではない．本稿ではプライバシ情報の漏洩を懸念することなく，複数の機関で独立に管理されたデータの突合を可能とするデータ統合システム Privacy-preserving Distributed Data Integration(PDDI)について解説する．

1.高機密性
PDDIシステムでは各機関が独立してデータを保管可能であり，新たに第三の機関によるデータ統合は不要である．各機関に独立に保管されたデータの突合には，各機関で突合データを非可逆圧縮後に暗号化した照合子を用いるため，データ突合に関する一切の情報はどの機関にも流れない．また，各機関は突合に利用するデータや統合したいデータを自由に設定することが可能で，各機関が許可したデータのみが許可された機関でのみ閲覧可能となる．

2.高汎用性
データ数・機関数に依らず高速に動作するスケーラビリティに富んだ設計となっている．またデータ突合を行う機関の新規登録，削除もフレキシブルに運用も可能な汎用性をもつ．データ突合においては，突合対象とする機関や突合する項目，出力項目をフレキシブルに設定可能である．

3.導入容易性
PDDIシステムはデータを保管する機関のみで突合処理が実施でき，新たに全データを管理可能なデータ預託機関を必要としない．このため，現在の医療機関に容易に導入可能である．

4.高速性
主要計算機サーバにより処理速度向上が可能である．計算機サーバはクラウド利用あるいは突合を行うグループでの導入の2方法があるが，どちらの場合においても計算機サーバでは突合データ及び統合データは秘匿される．

©一般社団法人 日本医療情報学会

プライバシを保護した多機関データ突合システムについて

宮地充子[*1*2*3]、中正和久[*1]、河内亮周[*1]
*1 大阪大学, *2 北陸先端科学技術大学院大学,
*3 独立行政法人科学技術振興機構 CREST

Privacy-preserving Multi-party Data Matching System

Atsuko Miyaji[*1*2*3], Kazuhisa Nakasho[*1], Akinori Kawachi[*1]
*1 Osaka University, *2 Japan Advanced Institute of Science and Technology,
*3 Japan Science and Technology Agency CREST

It is an important issue to match and analyze medical records managed at different medical institutions independently without leakage of privacy information such as patient's name, address, etc. In this paper, we introduce two technologies to realize them. The first technology, Privacy-preserving Distributed Data Integration (PDDI), is a data integration system that enables to match and extract records managed by multiple organizations while preserving privacy. The second is a privacy-preserving classification protocol that classifies client information based on a classification model held by a server without knowing partner's information each other.

Keywords: big data, PSI, PDDI, privacy-preserving classification protocol

1. はじめに

ビッグデータの解析結果は新製品開発など様々な活用が期待され，そのデータ収集・解析・利用の促進・定着は重要である．医療分野においては，患者のプライバシを確保しつつ，カルテ情報を医療分野の発展に利活用できることが望まれる．さらに，データ所有者である患者がデータの利活用に合意できる枠組みの構築が必須である．

ビッグデータに関する既存研究では，ビッグデータの効率的な解析手法を改良する研究が多い．一方，本研究課題ではデータ所有者（医療の場合は患者に相当する）に着目し，データのプライバシ保護を実現しつつ，解析結果の適切なデータ所有者への還元・フィードバックを実現し，データ所有者，解析，その利用という 3 つの機能を信頼の環で連結することを目標とする．

具体的に行う研究分野について説明する．

第一は，プライバシを保護しつつ共通データを抽出する方法でプライバシ保護付き共有データ抽出手法 (Privacy Set Intersection)と呼ばれる．PSI では異なる機関が保管するデータのうち各機関が保管する共通データをそれ以外の情報は漏らさずに求める手法である．既存研究では 2 機関のデータの共有データの抽出が頻繁になされていたが，一般に医療機関は複数あることから，本研究課題では複数機関における共有データの抽出方法について議論する．

第二は，サーバとクライアントがお互いの情報を秘匿したまま，サーバの持つ分類モデルに基づいてクライアントが持つ個人情報を分類する秘匿分類プロトコル(Privacy-Preserving Classification Protocol)である．本研究課題については既存研究における分類モデルの一般性とその効率について考察を行い，今後の研究の展望・方針について議論を行う．

さらに医療の分野で上記研究成果を利用する状況を説明し，その具体的な成果について検討する．

本稿の構成は次の通りである．第 2 章で各研究における成果および今後の課題について述べる．第 3 章で医療分野における適用事例について述べる．第 4 章で結論をまとめる．

2. 各研究課題

2.1 PDDI 技術

我々を取り巻く情報社会では，多種多様なデータが多機関で収集される．例えば，小学校で児童がブランコでけがをした事例を考える．このとき，事故が起こった遊具に関するデータは学校，病院への救急搬送データは消防署，傷害・後遺症に関するデータは病院に管理される．つまり，学校での生徒の事故に関する情報では，学校，消防署，病院がそれぞれ同じ事故で異なるデータを管理する．

学校における事故の予防安全の実現には，事故の統計的因果モデルの作成が重要である．これにはこのように異なる機関に分散した関連データの統合が必須である．つまり，異なる機関が独立に収集したデータから生徒の名前などの機微情報は洩れることなく，同じ生徒の事故の情報を突合（分散多機関データ突合）できると，事故の詳細なデータの収集が可能になる．

ここで，異なる機関がもつ医療データの突合方法とプライバシの関係について考える．単純な方法は，1つの医療機関が全データを別の医療機関に渡せば，同じ患者を検索することで，データを突合することができる．しかし，この場合，本来もつはずでなかった患者の情報である名前，住所などの機微情報を別の医療機関が入手することになる．別の方法として，第 3 の機関(データ預託機関)にそれぞれの病院が医療情報を渡し，その第 3 の機関で突合することもできる．しかしこの場合には，第 3 の機関に患者の機微情報が移動することになる．つまり，単純な突合方法は突合に用いる情報が必要となるため，突合を実施する機関に機微情報が移動し，プライバシ保護を実現することが困難になる．

そこで相互の持つ重要な情報を外部に漏らすことなく，必要な情報のみのやり取りを行うための有効な技術が必要である．近年 Private Set Intersection Protocol(PSI) と呼ばれる各機関が持つデータの積集合などの集合演算を，プライバシを保護しつつ実現するプロトコルが注目されている．PSI は様々な種類が考案されているが，一般的な PSI は次のような特性を持つ．$S=\{s_1,...,s_w\}$ と $C=\{c_1,...,c_v\}$ をデータの集合として持つサーバとクライアントの存在を考える．サーバとクラ

イアントはそれぞれ S と C を入力としてプロトコル通りに通信をおこなうと，最終的にサーバは|C| のみを，クライアントは|S|，C ∩ S を得ることができる．上記の病院と学校の例では，病院側はウイルス感染者の人数以外は学校に知られることなく，ウイルスに感染した学生の情報を得ることができる(図 1 参照)．また学校側はウイルスに感染していない学生の情報は学生数以外知られることはない．それぞれのデータ内容がプライバシ，ウイルスに感染した学生情報が病院の求める情報と考えると，互いのプライバシを守りつつ，クライアントは目的の情報を得ることができる．

上述したとおり PSI は様々な種類が考案されており，通信者の入力データ集合の要素数も秘匿する方式 [ACT11] や，プロトコル参加者に任意の行動(プロトコルに従わない，通信を途中で切断するなど) をする者がいた場合でも安全に通信をおこなえる方式[CKT10] ，2 者以上の PSI を可能とした Multi-Party プロトコル[KS05, EFGJST15] などがある．

既存のプロトコル[KS05] では全参加者の入力データ数を一致させなければならないという制限がある．また，既存プロトコル[EFGJST15] での出力は全参加者が入力した集合の和集合の要素数の概数に限定される．さらに，データを所有する参加者以外に第三者機関を3機関導入し，その内1つの機関には出力が漏れる．さらにどの既存方法も，単一の属性を前提としており，複数の属性を統合することができない．

本研究室で提案された方法は，機微情報を他の機関に移動することなくデータ突合を実現する方式である． この方式は，データ統合システム Privacy-preserving Distributed Data Integration(PDDI)においてプログラム実装されている．本システムを利用することによって，ユーザは情報漏洩を懸念することなく，複数機関が所有するデータの統合を実現することができる．次に図 2 に沿って，プライバシを保護しながら同一のユーザのデータ突合を実現する方法について述べる．

ステップ①

各データ集合の突合したいデータ x(図 2 の場合，名前)をハッシュ関数 H 等で圧縮する．例えば機関1のデータの名前「竹田」は

竹田→H(竹田)

と一意に不可逆な情報に変換される．その後，さらに準同型暗号 E で暗号化する．

H(竹田)→E (H(竹田))

同様の処理を機関 2，3 で行い，暗号化データを秘匿計算機サーバに送付する．

ステップ②

準同型暗号は，暗号文の和が，元の文の和の暗号文に一致する性質を持つ．この性質を用いて秘匿計算機サーバでは入手した暗号化データをそれぞれ加えることで，各機関のデータの総和の暗号結果を計算できる．図 2 の場合，下記のように，各機関で計算された E (H(竹田))のような名前の暗号文は秘匿計算機サーバで加えられ，名前の和の暗号文となる．

E (H(竹田)) + E (H(田中)) + E (H(竹田)) + E (H(田中))
+ E (H(竹田)) + E (H(山田))
= E (3H(竹田) + 2H(田中) + H(山田))

この暗号化データの総和を各機関に送付する．

ステップ③

受信した各機関のデータの総和の暗号文を復号して，突合したデータ，図 2 の場合，「竹田」のデータを出力する．

本提案は以下の特徴を持つ．

- 秘匿計算機サーバには暗号化データのみが送付されるので，データは完全に秘匿される．
- 暗号化不可逆データを用いて，突合が実現されるので機微情報はどの機関にも移動しない．
- 各機関の処理時間は機関数に依存しない．
- 各機関のデータ数に制限や条件はない．

既存研究はデータサイズや機関数に依存する処理時間，通信量が大きな課題であるのみならず，図 2 の喫煙，職業，血圧等の複数の属性の統合方法が存在しなかった．当研究室で考案された方式は，機関数に処理時間が依存せず，さらに複数の属性の統合を可能にする．

巻末表 2 に提案方式と既存方式の比較を記載する．提案方式はデータ預託機関が不要で，各機関のデータ数の制限がなく，通信量，計算量を削減した方式となっている．既存方式1と提案方式の処理時間は表 1 に記載されている．機関数の二乗のオーダで処理時間が掛かる既存方式では，データ数が増えると非常に多くの処理時間がかかり，提案方式の優位性がわかると思われる．

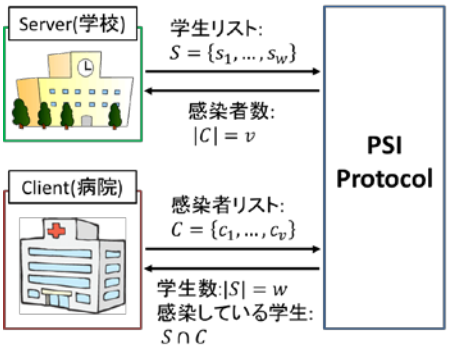
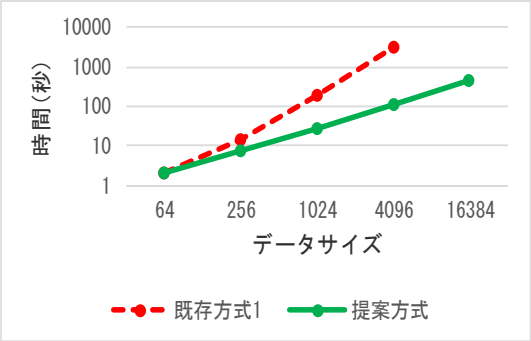


図 1. PSI

表 1. 提案方式と既存方式 1 との比較
(16 機関、 データ数 2^6 – 2^{14})



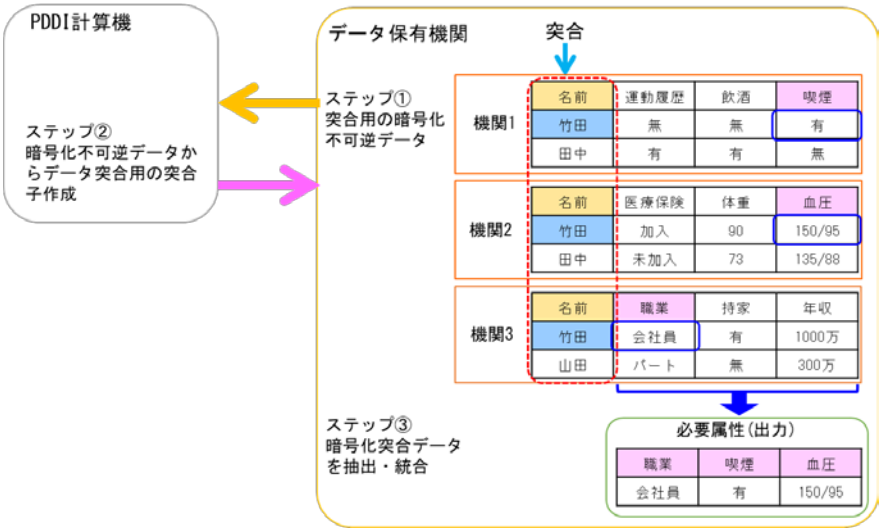


図 2. システムモデル

2.2 秘匿分類プロトコル

近年の機械学習技術の飛躍的発展に伴い，様々な場面で医療情報処理を含め，機械学習の応用事例を見ることができる．これらの例では訓練用データに機械学習アルゴリズムを適用して生成された分類モデルを構築する訓練フェイズとその分類モデルを利用してプライバシ情報から分類結果を得る分類フェイズに分かれる．このような機械学習のような例においてもデータのプライバシ保護が重要であり，暗号基盤技術の応用として様々な技術が開発されている．従来の多くの研究では訓練用データのプライバシ保護に着目しているが，最近では分類フェイズにおいてプライバシ保護を達成できる秘匿分類プロトコル研究も進んでいる．

秘匿分類プロトコルは分類モデルを持つサーバとプライバシ情報を持つクライアントが通信を行い，クライアントのプライバシ情報に分類モデルを適用することによりクライアントが有益な分類結果を得ることができるプロトコルである(図 3)．このとき，サーバは知的財産である分類モデルの構造に関する情報をクライアントに漏らしたくない，その一方クライアントは自身のプライバシ情報をサーバに漏らしたくない，という要求があり，この要求はそのまま単純に互いの持つ分類モデル，プライバシ情報を送信することでは達成できない．そのためモデルの歪曲化技術，紛失通信プロトコル，準同型性公開鍵暗号化方式などの高度な暗号基盤技術を構成要素として利用することが必要となる．簡単に言えば，歪曲化技術はデータを分類する機能を保存したまま分類モデル自身を暗号化する技術，紛失通信プロトコルはクライアントがどの情報が必要かをサーバに隠したままサーバから分類モデル計算のための必要最小限の情報のみを得る技術，準同型性暗号は暗号化前の情報の演算を暗号文の上で実行することが可能となる暗号化技術(例えば数値 a,b の暗号文 E(a), E(b)に対してある演算 Hom(E(a),E(b))を実行すると暗号文を復号することなく a+b の暗号文 E(a+b)を得ることが可能となる)である．

例えば Barni ら[BFL+09][BFL+11]は決定木に類似した分類モデルである線形分岐プログラムに対する歪曲技法および紛失通信プロトコルから線形分岐プログラムの秘匿分類プロトコルを構成し，心電図のプライバシ保護分類器を実装している．また彼らはパーセプトロンの一般化となるニューラルネットワークモデルの秘匿分類プロトコルを研究している．

Bost らの研究[BAT+15]では Barni らの結果を含む形でさらにいくつかの機械学習モデルにおける秘匿分類プロトコルを提案し，その安全性と性能を解析している．この研究では Paillier 公開鍵暗号方式および Goldwasser-Micali 公開鍵暗号方式といった加法準同型性公開鍵暗号パーセプトロン，Fisher 線形判別，サポートベクトルマシンを含む線型判別器およびナイーブベイズ分類器の秘匿分類プロトコルを提案し，また決定木を多項式で表現してその多項式をレベル付き完全準同型性暗号(二つの暗号文の加法だけでなく回数制限付きで乗法も計算できる暗号化技術)で暗号化することにより決定木の秘匿分離プロトコルの構成を与えている．

しかしながらレベル付き完全準同型性暗号は現在のところ実行効率が高くないため，効率化のためには別の暗号基盤技術を利用する必要がある．Wu ら[WFN+16]は紛失通信プロトコルとともに加法準同型性暗号を利用した巧みな方法で決定木の各節点における大小比較演算を実現することによって決定木の秘匿分類プロトコルを実現し，既存結果よりも非常に高速であることを実証している．また彼らは提案プロトコルを機械学習でよく用いられるランダムフォレストモデルにも拡張できることを示している．

以上のようにいくつかの分類モデルにおいて秘匿分類プロトコルが提案されているが，対象とされている分類モデルは決定木やサポートベクトルマシンといった比較的単純なものであり，さらに一般的かつ強力な分類モデルに対する高効率な秘匿分類プロトコルの開発とその安全性証明，効率解析，実装を与えることが重要となる．

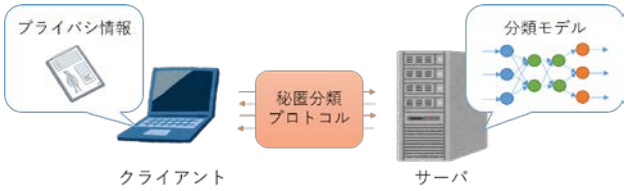


図 3. 秘匿分類プロトコル

3. 医療分野への応用

ビックデータを利用する分野として医療分野がある．

PDDI 技術では，診療所や病院などの複数機関でやりとり

するデータを非開示のまま共有データを抽出することが可能である．例えば，各医療機関が持っている希少な症例データを複数の病院で集めることにより，希少な症例データの統計的な解析が可能となる．

また，医療現場においては，関連するデータが複数の異なる機関で管理されるケースが頻繁に起こる．例えば，同じ患者が異なる疾患にかかった場合，各疾患を専門とする複数の病院に通うことが考えられる．このように独立した 2 つの医療機関で管理された異なる疾患同士には因果関係がある可能性がある．この時，同一の患者のデータを患者のプライバシを保護しつつ，必要な医療データのみ突合できると，異なる病気の因果関係に関する詳細なデータの収集が可能になる．

秘匿分類プロトコル技術は，既存の機械学習の医療分野への応用にそのまま適用してプライバシ保護の機能を付加することが可能である．前述のように Barni らの研究ではプライバシ保護を行いながら心電図の信号を線形分岐プログラムによって，正常洞調律，心房期外収縮，心室期外収縮，心室細動，心室頻拍，上室頻拍の 6 クラスに分類するプログラムを実装している．

図 4 は，PDDI 技術と秘匿分類プロトコルを組み合わせる事例を表す．特定疾患の患者データをプライバシ保護しながら収集し，その後，秘匿分類プロトコルによって疾患を分類する．

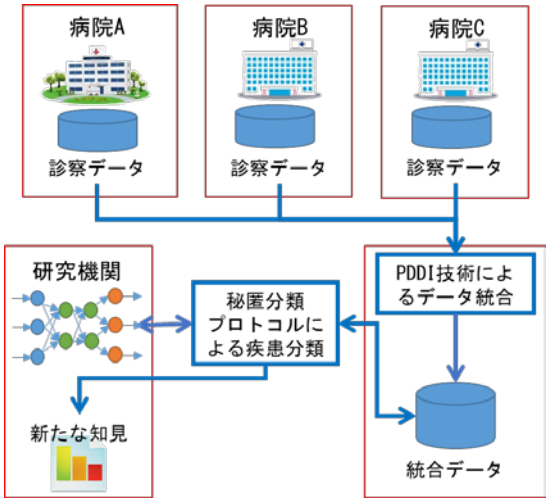


図 4. PDDI 技術と秘匿分類プロトコルを用いた研究

4 まとめ

本研究ではプライバシを保護しつつ医療データの利活用を可能にするセキュリティ技術について紹介し，各研究課題の現状や新しい研究成果について述べた．第一はプライバシ保護付き共有データ抽出手法で，複数機関におけるデータの突合及び抽出方法について説明した．また，当研究室で開発した方法(PDDI)は暗号化不可逆データを用いることで，個人を特定できる情報はどの機関にも移動せずに突合とデータ抽出を可能にし，突合を行う医療機関数が増えても高速に処理できることを紹介した．今後，各種医療データへの利活用により，異なる病気の因果関係の究明や，稀少疾患の治療方法の発展への貢献が望まれる．第二に秘匿分類プロトコルについてその技術を概説した．特に既存研究においてサーバとクライアントのプライバシを保護しつつ利用可能となっている分類モデルの種類と構成要素として利用している暗号基盤技術からその効率について議論を行い，今後の研究の展望を考察した．次にこれらの 2 つのセキュリティ技術の医療分野における適用事例を取り上げた．

謝辞

本研究の一部は JSPS 科研費基盤 C (JP15K00183) と (JP15K00189) 及び科学技術振興機構（JST）の CREST(JPMJCR1404)と国際科学技術協力基盤整備事業及び文部科学省の情報技術人材育成のための実践教育ネットワーク形成事業分野・地域を越えた実践的情報教育協働ネットワークの助成を受けています．

参考文献

[ACT11] G. Ateniese et al. (if) size matters: size-hiding private set intersection. In PKC 2011, volume 6571 of LNCS, pages 156-173, Springer, 2011.

[BAT+15] Raphael Bost, Raluca Ada Popa, Stephen Tu, and Shafi Goldwasser. Machine Learning Classification over Encrypted Data. In Proc. The 2015 Network and Distributed System Security (NDSS) Symposium, 2015.

[BFL+09] Mauro Barni, Pierluigi Failla, Riccardo Lazzeretti, Annika Paus, A-R Sadeghi, Thomas Schneider, and Vladimir Kolesnikov. Efficient privacy-preserving classification of ECG signals. In Proc. 1st IEEE International Workshop on Information Forensics and Security (WIFS 2009), pages 91–95, 2009.

[BFL+11] Mauro Barni, Pierluigi Failla, Riccardo Lazzeretti, Ahmad-Reza Sadeghi, and Thomas Schneider. Privacypreserving ECG classification with branching programs and neural networks. IEEE Transactions on Information Forensics and Security (TIFS), 6(2):452–468, June 2011.

[BL13] K. Bache and M. Lichman. UCI machine learning repository, 2013.

[BLN13] Joppe W. Bos, Kristin Lauter, and Michael Naehrig. Private predictive analysis on encrypted medical data. In Microsoft Tech Report 200652, 2013.

[CKT10] E. De Cristofaro et al. Linear-complexity private set intersection protocols secure in malicious model. In ASIACRYPT 2010, volume 6477 of LNCS, pages 213-231, Springer, 2010.

[EFGJST15] R. Egert et al. Privately computing set-union and set-intersection cardinality via bloom filters. In ACISP 2015, volume 9144 of LNCS, pages 413-430, Springer, 2015.

[KS05] L. Kissner and D. Song. Privacy-preserving set operations. In CRYPTO 2005, volume 3621 of LNCS, pages 241-257, Springer, 2005.

[WFN+16] David J. Wu, Tony Feng, Michael Naehrig, and Kristin Lauter. Privately Evaluating Decision Trees and Random Forests. In Proc. Privacy Enhancing Technologies (4);1-21, 2016.

表 2. 提案方式と既存研究の比較

方式	既存方式 1[1]	既存方式 2[2]	提案方式[3]
データ預託機関	不要	必要	不要
各機関の計算量 （機関数 n）	機関数 n の 2 乗の計算量	機関数に依存しない	機関数に依存しない
通信量	機関数 n の計算量	機関数 n の計算量	機関数 n の計算量
データ数の制限	全機関が同じデータ数	制限なし	制限なし
秘匿される情報	データ集合のみ	データ集合とその個数	データ集合とその個数