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

平成　３０　年　　６　月　１３　日現在

機関番号：　３２６８２
研究種目：　基盤研究(C)（一般）
研究期間：　2015～2017
課題番号：　１５Ｋ００１９４

研究課題名（和文）医療ビッグデータのプライバシー保護ロジスティック回帰の研究

研究課題名（英文）Privacy-Preserving Logistic Regression for Medical Big Data

研究代表者
菊池　浩明（Kikuchi, Hiroaki）

明治大学・総合数理学部・専任教授

研究者番号：２０２６６３６５

交付決定額（研究期間全体）：（直接経費）　　3,600,000円

研究成果の概要（和文）：本研究課題では垂直に分割された２つのデータセットに対するロジスティック回帰を安全に実行する新しいプロトコルを提案する．提案方式は，反復再重み付け最小二乗(IRLS)を適用し，従来よりも少ない繰返し回数で収束する効率の良いプロトコルである．従来の最小勾配法(SGD)の収束回数30,000回と比較して，提案プロトコルは7回で収束する．大規模な国内の患者の診療情報を含む診療情報データベース(DPC)を用いて，その実現可能性を評価している．提案により，対象患者の機微情報を関連組織に漏洩することなく，死亡確率のみを予測することを可能とする．疑似データによる本方式のパフォーマンスと精度の評価を与えている．

研究成果の概要（英文）：In this study, we propose a new secure protocols for privacy-preserving logistic regression of two vertically partitioned datasets. Our protocol is efficient in the sense that coefficients of logistic model are converged in few iterations by using the Iteratively Re-weighted Least Squares (IRLS). In the comparison to one of the existing work using the stochastic gradient descent (SGD), our protocol improved the performance of estimate from 30,000 to 7 iterations. We study the feasibility of the proposed protocol over the the Diagnosis Procedure Combination (DPC) database, a large-scale claim-based database of Japanese hospitals that contains confidential status of patients. Our scheme allows to estimate the probability of death with some patient information without revealing confidential data to the other party. Using the toy dataset and the trial implementation of the proposed scheme, we examine the accuracy of the proposed scheme and study the feasibility.

研究分野：　セキュリティ，プライバシー

キーワード：　医療情報　ビッグデータ　プライバシー保護

１．研究開始当初の背景

人口高齢化, 医療･介護の需要増加を背景に, 疾病リスクの分析, 医療資源の効率的配分を目的とする医療情報の電子化が進んでいる. 医療の大規模データベースの種類は多岐に渡るが, 数十万から数十億人レベルの患者レコードに, 多数の複雑な属性が含まれる. 異なるデータベースには異なる患者属性が含まれるが, 共通する属性も記録されており, それらを統合することで, より多くの属性情報がデータベース間で共有でき, さらに進んだ臨床疫学的な分析や医療サービス品質の評価が可能になる. しかし現状では, 過剰な個人情報保護がそれを拒んでいる.

医療データを用いて, 手術の術式の違いによって, 死亡に至るリスクの違いを評価することとする. まず, データが複数の病院に**分散管理**されている問題がある. これらを統合して解析しなくては十分なデータが得られないが, そのためには十分な漏えい対策をする必要がある. 更に, 喫煙などの生活習慣の違いが解析結果を歪める**交絡因子**となる点も重要である. 正確な術式間の比較を行なうためには, 交絡因子が死亡率に及ぼす影響を排除しなくてはならない.

複数の病院を横断した調査をするためには, 名前を仮名化したり, 属性の一部を削除したりする匿名化の対策を施してデータを集約することが検討されている. **単なる匿名化では精度とプライバシー保護を両立させることが出来ない**と言える. 特に医療情報を用いた臨床疫学においては, 安全性と精度に高い要請があり, 匿名化に代わる安全な技術が望まれている.

２．研究の目的

この問題に対して, 互いのデータセットを秘匿したまま協調して分析を実施して, 共通の知識を安全に抽出するプライバシー保護データマイニング(PPDM)の研究が精力的に進められている. PPDM の主流は, 加法準同型性を満たした公開鍵暗号(Paillier 暗号など)を要素技術として用い, 暗号化されたまま複雑なデータマイニングを実行し, そのマイニングされた結果だけを復号して明らかにする. 暗号化の為の大きな計算時間が必要だが, 匿名化で生じる精度の劣化はなく, 暗号化アルゴリズムの安全性に基づいて再識別されるリスクも低い. これまでに, ベクトル内積, 積集合, ベイズ推論, 決定木学習, 相関ルール抽出など基本的なデータマイニングアルゴリズムの多くが PPDM の枠組みで実行できることが示されている.

しかしながら, 医療ビッグデータの応用においては, 前述した交絡因子の排除が必要であった. この為, 従来の疫学コホート研究では多重ロジスティック回帰が広く利用されている. 死亡に至る確率を, 複数の要因の線形関数を入力とするシグモイド関数(1/1+exp(-Z))で定義し, そのオッズ比を考えることで他の交絡因子を調整した評価を与える. ただし, 通常の回帰分析とは異なり, その尤度推定の為には連立方程式を代数的に解くことが出来ず, 最急降下法などの逐次解法を用いる必要がある. この処理は, 大きな通信コストのかかる PPDM には適しておらず, これまでに効率的な解法は知られていない.

そこで, 本研究では, 複数の病院に分散された医療データを秘匿したままで, ロジスティック回帰を実行する秘匿計算暗号プロトコルを研究する. 病院 A, B, C に分散管理された年齢や術式(P, C)などのデータベースがあり, それらを暗号化して交換し, 暗号化されたままでロジスティック回帰にかけ, その結果であるオッズ比だけを得る. こうして, 例えば開腹出術に対するカテーテル手術のリスクの削減を正確に安全に算出することを可能とする.

３．研究の方法

医療データに対してロジスティック回帰を実行するプライバシー保護データマイニング（PPDM）技術を開発するため，次の

手順で研究を進める．医療情報の実験用疑似データベースを構築するともに，現実の疫学コーポレート研究で要求されるスケーラビリティ等の要求条件を明らかにする．また，論文を中心に PPDM プロトコル技術に必要とされる要素技術を調査すると同時に，多重ロジスティック回帰を計算する数値計算手法を調査，PPDM に適したアルゴリズムを選出する．その上で，水平分割秘匿ロジスティック回帰のための暗号プロトコルの構築を開始する．

４．研究成果

4.1 PPDM プロトコルの調査

従来，加法準同型性暗号には大きな計算コストがかかり，事前に決められたアルゴリズムを実施することに制約があった．そこで，目的変数のブール演算を秘匿したままで行う暗号プロトコルを提案し，プライバシー保護決定木学習を構成した．本研究成果は，雑誌論文１にまとめている．

プライバシーを保護して疫学などの解析を実現する匿名加工技術による研究を進めた．与えられたデータの特性に応じてリスクが変わるため，Zipf モデルを提案し，交通系 IC カードのデータに適用して，そのリスクを解析した．この成果を雑誌論文１と２にて発表している．論文１は，情報処理学会の英文誌 Journal of Information Processing (JIP) の論文賞(JIP Outstanding Paper Award)を受賞した．

4.2 ロジスティック回帰を行うプライバシー保護暗号プロトコル

多大な計算コストのかかる PPDM に適しているアルゴリズムとして，反復再重み付け最小二乗法 IRLS を選出した．IRLS は収束効率が良く，多くのロジスティック回帰で採用されている効果的なアルゴリズムである．加法準同型性を満たした公開鍵暗号で，このアルゴリズムを実施するシステムを試験実装し，そのパフォーマンスを評価した．

4.3 現実データを用いた実証実験

現実の医療データを用いて，線形重回帰を秘匿して行うシステムを開発した．本研究は，Java を用いて実装して，現実の 5000 名の診療情報から成る DPC データセットに適用してその実現可能性を評価した．本研究は，DICOMO の優秀論文賞を受賞している．

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

[雑誌論文] (計 3 件)

1. Hiroaki Kikuchi, Katsumi Takahashi, "Zipf Distribution Model for Quantifying Risk of Re-identification from Trajectory Data", Journal of Information Processing, Vol. 24 (2016) No. 5 pp. 816-823. doi.org/10.2197/ipsjjip.24.816 **(情報処理学会 2017 年度，JIP Outstanding Paper Award 受賞)**
2. 菊池 浩明，匿名加工・再識別コンテスト Ice and Fire：匿名加工方式とその安全性を評価する試み，情報処理学会論文誌,57(9), pp. 1900-1910, IPSJ, 2016. http://id.nii.ac.jp/1001/00174619/
3. Hiroaki KIKUCHI, Kouichi ITOH, Mebae USHIDA, Hiroshi TSUDA, Yuji YAMAOKA, Privacy-Preserving Decision Tree Learning with Boolean Target Class, IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, Vol. E98.A, No. 11, pp. 2291-2300, 2015. doi.org/10.1587/transfun.E98.A.2291

[学会発表] (計　件)

4. H. Kikuchi, H. Yasunaga, H. Matsui and C.I.Fan, "Efficient Privacy-Preserving Logistic Regression with Iteratively Re-weighted Least Squares", 2016 11th Asia Joint Conference on Information Security (AsiaJCIS), pp. 48-54, IEEE, 2016. https://doi.org/10.1109/AsiaJCIS.2016.21
5. 濱永 千佳 , 菊池 浩明 , 康永 秀生 ,

松居 宏樹 , 橋本 英樹, プライバシーを保護した垂直分割線形回帰システムの実装と DPC データセットを用いた評価, マルチメディア, 分散協調とモバイルシンポジウム 2016 論文集, pp. 1471-1478, IPSJ, 2016.
http://id.nii.ac.jp/1001/00177284/
(**Dicomo 2016 優秀論文賞受賞**)
6. 菊池浩明, 康永秀生, 松居宏樹, 橋本秀樹,組織間での分散秘匿ロジスティック回帰 による脳卒中の分析, 第 26 回日本疫学会学術総会講演集, p. 91, P1-026, 日本疫学会, 2016.
7. Hiroaki Kikuchi, Hideki Hashimoto, Hideo Yasunaga, Privacy-Preserving Epidemiological Analysis for a Distributed Database of Hospitals, 2015 10th Asia Joint Conference on Information Security (AsiaJCIS), pp. 85-90, IEEE, 2015.
DOI:10.1109/AsiaJCIS.2015.31

〔図書〕(計 1 件)
1. 菊池浩明, 上原哲太郎, IT Text ネットワークセキュリティ, オーム社, 2017 年.

〔産業財産権〕
該当なし

〔その他〕
ホームページ等
https://www.isc.meiji.ac.jp/~kikn/project.html

６．研究組織
(1)研究代表者
菊池 浩明 (Hiroaki Kikuchi)
明治大学・総合数理学部・教授
研究者番号： 20266365

(2)研究分担者
康永秀生 (Hideo Yasunaga)
東京大学・医学系研究科・教授
研究者番号： 90361485

(3)連携研究者
橋本秀樹 (Hideki Hashimoto)
東京大学・医学系研究科・教授
研究者番号： 50317682

(4)研究協力者
( )