様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２７ 年　 ９ 月 １８ 日現在

機関番号：３２６８９
研究種目：若手研究(B)
研究期間：2013～2014
課題番号：２５８７０８１２

研究課題名（和文）プライバシー保護データマイニングにおける分散型回帰分析の実用化

研究課題名（英文）Distributed Regression Protocols for Privacy Preserving Data Mining

研究代表者
須子　統太（Suko, Tota）

早稲田大学・社会科学総合学術院・講師

研究者番号：４０４０９６６０

交付決定額（研究期間全体）：（直接経費）　　1,600,000円

研究成果の概要（和文）：本研究ではプライバシー保護回帰分析を扱った．複数のユーザがそれぞれ個別にデータを保持している状況において，ユーザが持つそれぞれのデータはユーザ間で共有しないもとで，全てのデータを用いた場合と同等の分析結果を得ることを考える．従来よりいくつかのプロトコルが提案されているが，本研究ではより実用的な状況に適用するために，いくつかの種類の分散計算法を新しく提案した．これにより多重共線性がある場合や，スパース性がある場合にもプライバシー保護回帰分析が適用可能となった．

研究成果の概要（英文）：In this research, we study a privacy-preserving linear regression analysis. We consider the situation that a number of users have different data. They don’t want to show their data each other, but they want to calculate a certain estimator using all users data. Although some protocols conventionally proposed, we proposed some kind of protocols of distributed calculation method for practical use. we became privacy-preserving linear regression analysis available, if there is multicollinearity ,or sparse data.

研究分野：統計的学習理論，データ解析，情報理論

キーワード：データマイニング　プライバシー保護　分散計算　回帰分析

１．研究開始当初の背景

インターネットや情報通信機器の発達に伴い，大量のデータを様々なサーバが保持するようになった現在，これら大量のデータを如何に有効に活用するかが大きな課題となっている．他方，クラウドなどのサーバ上に置かれたデータをサーバ管理者が利用する場合，データ提供者のプライバシーを如何に保護するかという問題も浮上してきており，データの活用方法のみならず，プライバシーの保護方法の重要性が増してきている．

そのような社会的要請を受け，近年，プライバシー保護データマイニングの研究が注目を集めている．この分野では，様々なデータマイニング手法について，プライバシー保護機能を付加する研究が行われている[1]．
本研究課題ではその中でも特に，最も基本的な問題である，プライバシーを保護した回帰分析の分散計算について扱った[1-6]．これは，複数のユーザ（サーバ）が異なるデータセットを保持しているもとで，それぞれのデータセットの中身についてはお互い秘匿したまま，全てのデータセットを用いた場合と同様の回帰分析結果を得るというものである．例えばＡ社とＢ社がそれぞれ保持する顧客データを合わせることで，より精密な回帰分析をしたいと考えた場合，いくら協力関係にあると言っても，そのまま自社の顧客データを他社に見せてしまうことはできない．そこで，この技術を用いることで，それぞれの会社の，顧客のプラバシーを保護したまま，より精度の高い回帰分析を行おうというものである．

従来，この研究分野では，分析したい全体のデータセットをどのように分割して保持されているかについて，水平分割と垂直分割の 2 通りの分割方法が考えられている．水平分割とは，同じ目的変数と説明変数について，異なるデータセットを保持している場合で，垂直分割とは，目的変数については共通のデータを保持しているが，対応する説明変数は異なる変数を保持しいている場合である．
具体的な先行研究としては，水平分割のもとで，最小二乗推定量を求める方法[1]や Lasso 推定量を求める方法[4]，垂直分割のもとで最小二乗推定量を求める方法[2][3][5]などが提案されている．

２．研究の目的

このように，従来この分野では既にいくつかの先行研究が行われていたが，まだまだ実用化には程遠い状況である．その理由は，限られた状況における分析手法しか確立されておらず，実用的に起こりうる様々な状況に対応しきれていない為と考えられた．そこで，本研究課題では様々な一般化を行うことで，実用に適した様々な状況で使用可能な手法を開発することを目的とした．

３．研究の方法

研究方法としては，主に以下の 2 種類の方向性からの一般化を計画していた．

① 分析手法の一般化
プラバシーを保護した回帰分析の分散計算において，従来は主に，最小二乗推定量を求めるプロトコルについての研究がほとんどで，一般の回帰分析で用いられる他の推定量についての研究やがほとんど行われていなかった．また，回帰分析を実用的に用いる際に議論される，変数選択等の問題についてもあまり議論されていなかった．そこで，最小二乗推定量以外の良く用いられる推定量に関するプロトコルの開発や，変数選択法に関する検討を行う．

② 分割モデルの一般化
より現実的な状況を考慮するため，単純な水平分割モデルや垂直分割モデルだけではなく，それらを組み合わせた分割モデルについてのプロトコルの開発を行う．プロトコルの開発には，研究代表者が既に提案している，水平と垂直を合わせた分割モデルに対する，最尤推定量の分散計算プロトコル[6]を参考に，①で提案した最小二乗推定量以外の推定量に対するプロトコルを拡張することで行う．

４．研究成果

本研究におけるおもな研究成果は以下の３点である．

① 最小二乗推定プロトコルの収束速度の改善
推定量の一般化や分割モデルの一般化を行う上で，基とする垂直分割型の最小二乗推定量を求めるプロトコルは，繰り返し型のプロトコルであるが，研究を進める過程で，条件によってこのプロトコルの収束速度が非常に遅くなってしまうという問題を発見した．そこで，当初の研究計画にはなかったが，推定量の一般化を行う前に，まず基となる最小二乗推定量を求めるプロトコルの収束速度の改善を検討した．その結果，分析に用いるデータを標準化することで，収束速度が劇的に改善し，条件に左右されず実用的な速度で推定量を求めることができるようになった．

② 事後確率最大推定量を求めるプロトコルの開発
推定量の一般化として，垂直分割型の事後確率最大推定量を求めるプロトコルを開発した．これはパラメータの事前分布において，変数間の相関が無い分布を

仮定した場合，リッジ回帰と等価な推定量で，変数間に多重共線性がある場合などにも有効な推定量である．プロトコルは①のプロトコルに ADMM（The alternating direction method of multipliers）アルゴリズムを組み合わせることで実現した．

③ L1 正則化最小二乗推定量を求めるプロトコルの開発

得られる説明変数が高次元であるのに対し，実際に目的変数に寄与する説明変数が少ない場合（スパース性がある場合）には，上手く変数を選択することで過学習を抑える必要がある．そのような場合に良く用いられる推定量として，L1 正則化最小二乗推定量（Lasso 推定量）があり，近年注目を集めている．本研究では ADMM アルゴリズムを①に組み込むことで，垂直分割型の L1 正則化最小二乗推定量を求めるプロトコルを開発した．

以上のような成果が得られたが，当初の研究計画にはない①の研究を追加したため，申請時に予定していた，分割モデルの一般化まで研究を進めることができなかったが，これについては今後の課題とし，本研究課題終了後も継続して続けていきたいと考えている．

＜引用文献＞

[1] J. Vaidya, C.W. Clifton and Y.M. Zhu, Privacy Preserving Data Mining, Springer-Verlag, 2005.

[2] W. Du, Y.S. Han and S. Chen, "Privacy-Preserving Multivariate Statistical Analysis: Linear Regression and Classication," In 2004 SIAM International Conference on Data Mining, Lake Buena Vista, Florida, Apr. 22-24 2004.

[3] A.P. Snail, A.F. Karr, X. Kin and J.P. Reiter, "Privacy Preserving Regression Modelling Via Distributed Computation," Proc. the tenth ACM SIGKDD international conference on Knowledge discovery and data mining pp.667-682, New York, NY, USA, 2004.

[4] G. Mateos, J.A. Bazerque and G.B. Giannakis, "Distributed Sparse Linear Regression," IEEE Trans. SIGNAL PROCESSING, VOL. 58, NO. 10, pp.5262-5267, OCTOBER 2010.

[5] 須子統太，堀井俊佑，小林学，後藤正幸，松嶋敏泰，平澤茂一,プライバシー保護を目的とした線形回帰モデルにおける最小二乗推定量 の分散計算法について,電子情報通信学会技術研究報告，vol. 112，no. 279,

[6] 須子統太，堀井俊佑，小林学，松嶋敏泰，平澤茂一,プライバシー保護を目的とした回帰分析の拡張について，第 35 回情報理論とその応用シンポジウム予稿集,pp. 562-567,2012.

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 2 件）

① 須子統太，堀井俊佑，小林学，後藤正幸，松嶋敏泰，平澤茂一，"プライバシー保護機能を持つ線形回帰モデルにおける最小二乗推定量の分散計算法について，" 日本経営工学会論文誌， Vol.65, No.2,pp.78-88，2014.（査読有り）

② Nozomi Miya, Tota Suko, Goki Yasuda, Toshiyasu Matsushima, “Asymptotics of Bayesian Inference for a Class of Probabilistic Models under Misspecification, ” IEICE Trans. FUNDAMENTALS, Vol.E97-A, No.12,pp.2352-2360, 2014.（査読有り）

〔学会発表〕（計 9 件）

① 須子統太，堀井俊佑，小林学，"プライバシー保護機能を持つ分散型正則化最小二乗法について，" 第 37 回情報理論とその応用シンポジウム予稿集 (SITA2014),pp.300-305，2014.

② 中井 祥人，須子統太，松嶋敏泰，"プライバシー保護を目的とした線形回帰モデルにおける事後確率最大推定量の分散計算法について，" 電子情報通信学会技術研究報告.IBISML，112(454)，pp.47-54，2013.

③ 後藤正幸，須子統太，小林 学，平澤茂一：“判別を目的としたプライバシー保護データ解析に関する一考察”，日本経営工学会 平成 25 年春季大会予稿集，pp.54-55，2013.

④ Goki Yasuda; Nozomi Miya; Tota Suko; Toshiyasu Matsushima, "Asymptotics of MLE-based Prediction for Semi-supervised Learning," Proc. of 2012 International Symposium on

Information Theory and its Applications (ISITA2014), Melbourne, p.343, 2014.

⑤ 都築遼馬, 須子統太, 松嶋敏泰, "線形回帰モデルにおけるベイズ決定理論に基づく予測の近似手法," 第 36 回情報理論とその応用シンポジウム予稿集 (SITA2013), pp.438-441, 2013.

⑥ 山本粋士 , 須子統太, 松嶋敏泰, "次数未知の多変数多項式回帰モデルにおけるベイズ予測," 第 36 回情報理論とその応用シンポジウム予稿集 (SITA2013), pp.520-524, 2013.

⑦ 安田豪毅, 宮希望, 須子統太, 松嶋敏泰, "半教師付き学習における一致推定量に基づく予測の漸近評価," 第 36 回情報理論とその応用シンポジウム予稿集 (SITA2013), pp.659-664, 2013.

⑧ 宮希望, 須子統太, 安田豪毅, 松嶋敏泰, "真の分布を含むとは限らない階層モデル族に対するベイズ推定の漸近評価," 第 36 回情報理論とその応用シンポジウム予稿集(SITA2013), pp.665-670, 2013.

⑨ Shunsuke Horii, Tota Suko, Toshiyasu Matsushima, Shigeichi Hirasawa, Iterative Multiuser Joint Decoding based on Augmented Lagrangian Method, 電子情報通信学会技術研究報告 IT2013-34, pp.13-17, 2013.

〔図書〕（計 0 件）

〔産業財産権〕
○出願状況（計 0 件）

○取得状況（計 0 件）

６．研究組織
(1)研究代表者
須子 統太（SUKO, Tota）
早稲田大学・社会科学総合学術院・専任講師
研究者番号：40409660