DPCデータベースからのプライバシーを保護した線形回帰による入院日数モデルの学習

菊池 浩明[1]　橋本 英樹[†1]　康永 秀生[†1]

概要：患者の併存症やそれらに対する治療法とその結果に関する情報を統合することにより疾病に対する医療疫学や病院マネージメントなどに有益である．しかしながら，年齢や性別などの個人情報，喫煙やBMIなどのプライベート情報，そして治療行為に関する機微な情報を含むため，統合により個人が特定されてしまうリスクがある．この課題に対して，加法準同型性を満たした公開鍵暗号を用いて個々のデータセットを暗号化し，復号することなく様々なデータマイニングのみを実行するプライバシー保護データマイニングが注目されている．しかし，暗号化に係わる大きなコストがかかるため，属性の種類やデータレコード数などの観点でスケーラビリティの成約があり，いわゆるビッグデータに適用するのは時期尚早と言われている．そこで，本研究では，最もシンプルな線形回帰分析のアルゴリズムを取り上げ，分散された状態で患者の年齢，性別，肺炎，脳血管障害，認知症関連病態，糖尿病などの併存症の状態から，対象患者の在院日数を予測するモデルをプライバシー保護して実施することを試みる．その計算過程において，適用範囲を支配するボトルネックが患者数と病院数のどこにあるのかを明らかにすること，および，その改善方法について提案することを研究の目的とする．

Privacy-Preserving Linear Regression to Estimate Length of Hospital Stay From Distributed Diagnosis Procedure Combination Database

Hiroaki Kikuchi[1]　Hideki Hashimoto[†1]　Hideo Yasunaga[†1]

1. はじめに

日本版診断群分類 (DPC) データベースは，病名や治療行為の表コードによる患者の大規模データベースである．従来は各病院に分散されていた患者の併存症やそれらに対する治療法とその結果に関するこれらの情報は，統合することにより疾病に対する医療疫学や病院マネージメントなどに有益である．しかしながら，年齢や性別などの個人情報，喫煙やBMIなどのプライベート情報，そして治療行為に関する機微な情報を含むため，統合により個人が特定されてしまうリスクがある．

この課題に対して，加法準同型性を満たした公開鍵暗号を用いて個々のデータセットを暗号化し，復号することなく様々なデータマイニングのみを実行するプライバシー保護データマイニングが注目されている．これまでに，秘匿積集合や決定木学習などにより医療疫学に応用できることが研究されている．ただし，暗号化に係わる大きなコストがかかるため，属性の種類やデータレコード数などの観点でスケーラビリティの成約があり，いわゆるビッグデータに適用するのは時期尚早と言われている．例えば，本研究で対象としているDPCデータセットでは患者数は一施設あたり6000人規模，病院数は1000を超える．

そこで，本研究では，最もシンプルな線形回帰分析のアルゴリズムを取り上げ，分散された状態で患者の年齢，性別，肺炎，脳血管障害，認知症関連病態，糖尿病などの併

[1] 明治大学総合数理学部
School of Interdisciplinary Mathematical Sciences, Meiji University
4-21-1 Nakano, Nakano Ku, Tokyo, 164-8525 Japan
[†1] 現在，東京大学 大学院医学系研究科
Presently with Graduate School of Medicine, The University of Tokyo
Presently with 7-3-1, Hongo, Bunkyo, Tokyo, 113-8555 Japan
[†2] 現在，マルチメディア, 分散, 協調とモバイルシンポジウム運営委員会
Presently with DICOMO2014

存症の状態から，対象患者の在院日数を予測するモデルをプライバシー保護して実施することを試みる．その計算過程において，適用範囲を支配するボトルネックが患者数と病院数のどこにあるのかを明らかにすること，および，その改善方法について提案することを研究の目的とする．

プライバシー保護データマイニングの基本的な方式については多くの研究があり，線形回帰を実現する要素技術には困難性がない．しかし，実データに近い大規模な医療診断データを用いることで，値の分布や種類数などがより現実のものに近づき，そこから方式の実用性に関する知見が得られることが期待できる．パフォーマンスを改善する方向にも，基本アルゴリズムの改良だけではなくて，Apach Hadoop など並列計算技術や Bloom Filter などの近似アルゴリズムの適用などの様々な方向性が有る．

2. 疫学調査と秘匿計算

2.1 DPC データセット

DPC データセットは，病院の経営や診療の質を公平に評価する目的で，疾患 (Diseases), 治療 (Procedure) の組み合わせ (Combination) のデータからなる [1]．1015 施設から収集された，700 万人分の患者データが構築されている．緊急入院患者のほぼ 50%がカバーされている．データセットには，病院コード，データ識別番号，性別，年齢，郵便番号，在院日数，手術名，疾病名，身長，体重，喫煙歴，がんステージ分類，重症度などの診療情報を含む．

2.2 線形回帰

m 個の説明変数（属性） $x_{i,1}, x_{i,2}, \ldots, x_{i,m}$ についての標的変数 y_i が，$i = 1, \ldots, n$ 組与えられている．これらを近似する線形式

$$y = f(X) = \alpha + \beta_1 x_1 + \cdots + \beta_m x_m \tag{1}$$

を次のように求める．

まず，線形式の予測値 $f(X)$ と標的変数 y の差の二乗の総和を求め，それらを最小化する各係数 β を，総和の微分を 0 とおいて次の方程式を立てる．

$$\frac{\partial}{\partial \alpha} \sum_i^n (y_i - f(X_i))^2 =$$
$$\sum_i^n 2(y_i - \alpha - \beta_1 x_{i,1} - \cdots - \beta_m x_{i,m})(-1) = 0$$
$$\frac{\partial}{\partial \beta_1} \sum_i^n (y_i - f(X_i))^2 =$$
$$\sum_i^n 2(y_i - \alpha - \beta_1 x_{i,1} - \cdots - \beta_m x_{i,m})(-x_{i,1}) = 0$$
$$\cdots$$

これらを整理して，$m+1$ 個の連立方程式が得られる．行列で

$$AX = B \tag{2}$$

となる X を求めればよい．ただしここで，

$$A = \begin{pmatrix} \sum x_{i,1}^2 & \sum x_{i,1}x_{i,2} & \cdots & \sum x_{i,1} \\ \sum x_{i,1}x_{i,2} & \sum x_{i,2}^2 & \cdots & \sum x_{i,2} \\ \vdots & & & \vdots \\ \sum x_{i,1} & \sum x_{i,2} & \cdots & \sum 1 \end{pmatrix},$$

$$X = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_m \\ \alpha \end{pmatrix}, \qquad B = \begin{pmatrix} \sum x_{i,1}y_i \\ \vdots \\ \sum x_{i,m}y_i \\ \sum y_i \end{pmatrix}$$

である．A の逆行列を左からかけて，係数を得る．

仮説検定は，統計量 $Z = \hat{\beta}_j / S.E.(\hat{\beta}_i)$ が正規分布 $N(0,1)$ に従うことで確かめられる．

3. プライバシー保護線形回帰

3.1 問題定義

m 個の説明変数 $x_1, \ldots, x_m$ と標的変数 y についての値の組 $(a_1, \ldots, a_m, y)$ が n 個ある集合 S が与えられている．標的変数を m 個の説明変数の線形和

$$y = \alpha + \beta_1 x_1 + \cdots + \beta_m x_m \tag{3}$$

で表す時，誤差の総和を最小化する係数 $\alpha, \beta_1, \ldots, \beta_m$ を求めたい．しかし，S が次の様に分割されて管理されているとする．

(1) (水平分割) S が，$S = S_1 \cup \cdots \cup S_N$ となる N の組織に直和分割されている．各分割された集合の大きさを，$n_1, \ldots, n_N$ とする．

(2) (垂直分割) 説明変数の集合 X が，$X_1 \cup \cdots \cup X_M$ となる M 個の集合に分割されている．ここで，それぞれの変数の集合の大きさを $m_1, \ldots, m_M$ とする．

N 個の組織は，各病院に該当しており，患者の年齢や疾病のデータを全病院が共通した変数で管理している．n_i は病院 i の持つ患者の数を表している．実際の病院数は $N = 1000$, 患者数 $n_i = 1, .., 6000$ 程度であり，病院によって大きな開きがある．

一方，M の組織は，病院と学会の様に，患者の情報と手術症例といった異なるリソースを持つ独立した組織である．何らかの方法で共通の患者を識別できるとする．

3.2 従来研究

Hall らは，加法準同型性を満たした公開鍵アルゴリズムを用いることで，複数者間で分散されたデータを公開することなく，線形回帰と検査統計量を算出するアルゴリズムを提案している [10]．彼らの方法では，線形回帰の為の中間情報を漏らすことなく，線形モデルの係数のみを算出する．逆行列を計算するために，Newton 法により逐次的に

係数を求めている．提案方法のスケーラビリティを示すために，5 万人分の患者からなる 22 変数のデータセットで実験を行っている．

3.3 研究目標

Hall らが秘匿している情報は，各属性値のその病院ごとの総和である．例えば，1000 名の分の年齢の総和は，個人が特定されることが困難な統計情報であり，これらを秘匿して性能を下げるよりも，この秘匿を妥協して実現性を高めることを試みる．その制約のもとで，どの位の規模のデータを秘匿したまま解析できるかそのスケーラビリティを明らかにする．

3.4 水平分割線形回帰プロトコル

(入力) i 番目の患者の説明変数と標的変数の値を $(x_{i,1},\ldots,x_{i,m},y_i)$ とする．患者集合は $U=U_1\cup\cdots\cup U_N$ となる N 個の集合に分割されている．

ある代表組織が秘密鍵を管理し，対応する公開鍵は全員で共有する．

(1) 組織 $h=1,\ldots,N$ は，(それぞれ並列に) 各属性 $j=1,\ldots,m$ についての U_h の総和 $\sum_{i\in U_h} x_{i,j}$ と，二つの属性 $j,k\in\{1,\ldots,m\}$ の U_h の総和 $\sum_{i\in U_h} x_{i,j}x_{i,k}$ を求め，それらを暗号化して同報する．

(2) 代表組織が N 個の組織からの暗号文から，$E(\sum_{i\in U_1} x_{i,j})\cdots E(\sum_{i\in U_N} x_{i,j}) = E(\sum_{i\in U} x_{i,j})$ を（効率よく）求める．同様に，$E\sum_{i\in U_1} x_{i,j}x_{i,k})\cdots E(\sum_{i\in U_N} x_{i,j}x_{i,k}) = E(\sum_{i\in U} x_{i,j}x_{i,k})$ を求め，それらを復号して $\sum_{i\in U} x_{i,j}x_{i,k}$ を公開する．

(3) 任意の組織が総和から，式 (2) によって，係数 $\alpha,\beta_1,\ldots,\beta_m$ を求める．

加法が可換であるので，N 個の暗号文の積を求めるのに，任意の順番で計算をしてよいことに注意せよ．例えば，N 個の組織を木構造に構成すると，$\log N$ 回の暗号文の積の時間で総和が計算できる．これを，Step (2) で「効率よく」と記している．

3.5 垂直分割線形回帰プロトコル

(入力) 属性値の集合 $A=\{a_1,\ldots,a_m\}$ が $A=A_1\cup\cdots\cup A_M$ となる M 個の部分集合に分割されている．M 番目の組織が, 標的変数 b についての値を持っている．

(1) 各組織 $\ell=1,\ldots,M$ は，それぞれ並列に，各属性 $j\in A_\ell$ についての U の総和 $\sum_{i\in U} x_{i,j}$ と，二つの属性 $j,k\in A_\ell$ の U の総和 $\sum_{i\in U} x_{i,j}x_{i,k}$ を求めて公開する．

(2) 二つの組織 $\ell_1,\ell_2\in\{1,\ldots,M\}$ の間で，それぞれの持つ属性集合の全ての組 $j\in A_{\ell_1},k\in A_{\ell_2}$ について，秘匿内積プロトコルを用いて，n 次元ベクトル $\boldsymbol{x}_j=(x_{1,j},\ldots,x_{n,j})$ と $\boldsymbol{x}_k=(x_{1,k},\ldots,x_{n,k})$ の内積 $\boldsymbol{x}_j\cdot\boldsymbol{x}_k$ を計算して，公開する．

(3) 任意の組織が総和から，式 (2) によって，係数 $\alpha,\beta_1,\ldots,\beta_m$ を求める．

Step 2 において，内積 $\boldsymbol{x}_j\cdot\boldsymbol{x}_k=\sum_{i\in U} x_{i,j}x_{i,k}$ であることに注意せよ．

4. 評価

4.1 安全性

公開鍵暗号の安全性に依存して，交換した暗号文からの漏えいの確率は無視できるほど小さい．提案方式では，各属性値の総和が公開されてしまう．従って，非常に頻度の少ない属性がある場合は，総和から入力に関する情報が特定される恐れが生じる．入力値に対する十分な考慮を必要とする．

4.2 スケーラビリティ

4.2.1 水平分割の場合

プロトコルで大きな負荷がかかるのは，ステップ (1) における暗号化である．m 個の属性があるとき，必要な暗号文の数は

$$\frac{m^2}{2}+\frac{3}{2}m+1=O(m^2)$$

であり，データ（患者）の数 n には依存しない．分散の数 N を増やすことで多少負荷が増えるが，ステップ (1) は並列に行うことが出来るので $N=1$ の時と変わらない．ステップ (2) では，N 個の暗号文の積と復号化が必要であるが，暗号化の処理時間と比較して無視できるほど小さい．

4.2.2 垂直分割の場合

プロトコルで大きな負荷がかかるのは，ステップ (2) の n 次元ベクトルの秘匿内積プロトコルである．データの規模 n に依存してしまうし，m_1 と m_2 に分割しているときは，$m_1\times m_2$ の全ての組について計算しなくてはならない．更に，M 組織に分散しているときは，M 個の中の任意の二つの組織間でこの内積プロトコルを繰り返す必要がる．

従って，トータルで必要とする計算量 (暗号化回数) は，

$$n(\frac{m}{M}^2+\frac{m}{M})\frac{M(M-1)}{2}=O(nm^2M^2)$$

である．後述するようにこれは大変大きな制約である．

4.3 DPC 疑似データへの適用

$n=250{,}000$, $m=13$ の疑似 DPC データセットを用意して，提案方式を適用した結果を報告する．

表 1 は，このデータを線形回帰を行い，m 個の属性の中でも標的変数に対する寄与率の大きなものを表している．この例では，年齢や疾病による在院日数を予測しており，透析の有無と心疾患が有意であることを表している．

図 1 は，R 言語を用いて行った線形回帰モデルの予測値

表 1 線形回帰モデルの係数と提案方式の比較

属性（説明変数）	Estimate	Std. Error	t value	$Pr(>\|t\|)$	提案方式
α	37.1454	19.7892	1.877	0.0625	36.7032
性別	−7.0761	5.127	−1.38	0.1696	−7.0661
ステント	4.9625	4.744	1.046	0.2973	4.9642
バイパス手術	−6.933	4.5578	−1.521	Y0.1304	−6.9535
年齢	0.2915	0.2437	1.196	0.2335	0.2967
実施年度	−5.0528	5.7254	−0.883	0.3789	−5.0282
肺炎	0.4686	12.4077	0.038	0.9699	0.4222
脳卒中	−2.1286	9.9463	−0.214	0.8308	−2.1514
不整脈	−0.5005	7.1754	−0.07	0.9445	−0.4541
認知症	6.8778	20.1738	0.341	0.7336	6.7364
糖尿病	−5.024	6.25	−0.804	0.4228	−5.0428
虚血性心疾患	−15.4635	7.3467	−2.105	0.037∗	−15.3963
透析	42.9962	7.5196	5.718	5.88^{-08}∗∗∗	43.0425

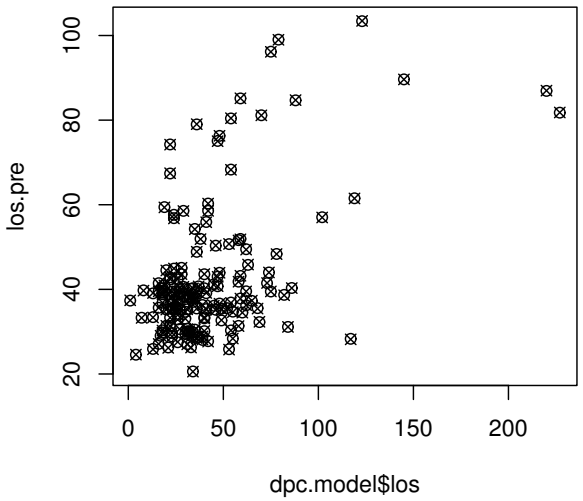

図 1 線形回帰モデルの予測値と真値の分布

と提案方式の予測値の散布図を表している．X 軸が真の値，○が R 言語，×が提案方式の値である．ほぼ二つの予測が一致していることが分かる．

図 2 と 3 に，水平分割と垂直分割のそれぞれにおける提案方式の処理時間を属性数 m について表している．水平分割の処理時間に比べて，垂直分割が著しい増加をしていることが示されている．水平分割では，$n = 25$ 万人分のデータに対して，26 個の説明変数の線形回帰を約 11 秒で終えている．分割数を実際の規模の数千にあげても処理時間は変わらない．それに対して，垂直分割は n に比例する計算時間がかかり，$n = 1$ 万人で約 2 分かかっているので，疑似 DPC データの 25 万人は約 1 時間という見積もりが出来る．なお，分割数 M を増やすと単独で計算できる属性も増えるので，トータルの処理時間をある程度抑える効果がある．

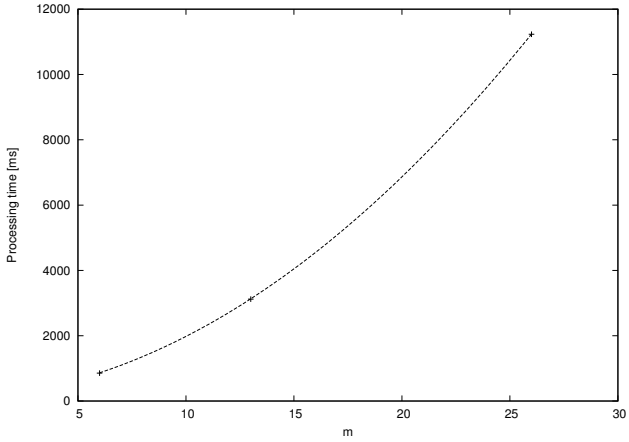

図 2 水平分割線形回帰の処理時間（$N = 1$, $n = 257,997$）

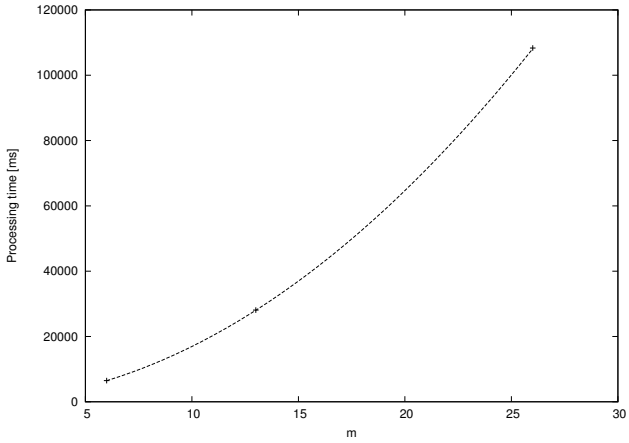

図 3 垂直分割線形回帰の処理時間（$N = 1$, $n = 10,000$, $M = 1$）

5. おわりに

病院や検査組織などに分散した機微な情報を，それらを平文で外部に取り出すことなく，暗号化したまま線形回帰を求めるプロトコルを示し，その性能を評価した．水平分割線形回帰ならば，その負荷は属性値の種類にのみ依存し，データの量には依存しないために大規模なデータに適用可

能であることを示した．$n = 25$ 万人の 26 説明変数による複数の病院に分散された線形回帰を 11 秒で計算できる．

参考文献

[1] 松田，伏見，診療情報による医療評価，DPC データから見る医療の質，東京大学出版会.

[2] 丹後，山岡，高木，「ロジステック回帰分析，SAS を利用した統計解析の実際」朝倉書店.

[3] 椿，岩崎，「R による健康科学データの統計分析」，朝倉書店.

[4] 高橋 信, 井上 いろは, トレンドプロ, マンガでわかる統計学 回帰分析編, オーム社, 2005.

[5] H. Yasunaga, H. Horiguchi, K. Kuwabara, S. Matsuda, K. Fushimi, H. Hashimoto, “Outcomes After Laparoscopic or Open Distal Gastrectomy for Early-Stage Gastric Cancer: A Propensity-Matched Analysis”, Annals of Surgery, Volume 257, Issue 4, pp. 640 646, 2012.

[6] L. Sweeney, “k-anonymity: A model for protecting privacy”, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10.05, pp. 557-570, 2002.

[7] 菊池，佐久間，三上， “プライバシーを保護したピロリ菌疫学調査”, 第 26 回人工知能学会，3I2-OS-20-9, pp. 1-4, 2012.

[8] Vaidya, J. and C. Clifton, “Privacy preserving association rule mining in vertically partitioned data”, The Eighth ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining, SIGKDD, ACM Press, Edmonton, Canada, pp. 639-644, 2002.

[9] S. Wu, T. Teruya, et. al, “Privacy-preservation for Stochastic Gradient Descent”, The 27th Annual Conference of the Japanese Society for Artificial Intelligence, 3L1-OS-06a-3, 2013. (`https://kaigi.org/jsai/webprogram/2013/paper-596.html`)

[10] Rob Hall, Stephen E. Fienberg and Yuval Nardi, “Secure Multiple Linear Regression Based on Homomorphic Encryption”, Journal of Official Statistics, Vol.27, No.4, pp. 669-691, 2011.