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評価

濱永 千佳 †

明治大学総合数理学部 先端メディアサイエンス学科 菊池研究室 †

1 はじめに

2015 年に個人情報保護法が改正され，マイナンバー制度も導入された．ベネッセ社で大規模な個人情報流出事件 [1] も発生しており，個人情報保護に関して，社会の関心が高まっている．本研究では，医療データベースを対象とする．医療データベースには，年齢，性別などの基本情報に加え，既往歴や入退院時の状況，診療情報など，多くの情報が登録されており，ほとんどの項目が個人情報と成り得るからである．プライバシー保護をしつつ情報を活用する方法には匿名加工などのプライバシー保護データパブリッシクングや秘密分散などの様々な手法が提案されている．しかし，特異な症例のデータは匿名加工しても，登録情報から特定可能という問題があった．

そこで，本研究では，準同型性公開鍵暗号を用いることで，データの価値を失うことなく活用するプライバシー保護データマイニング (Privacy Preserving Data Mining, PPDM) を試みる．

5000 例規模の脳卒中患者の実在する患者のデータを用いて，同一の患者群についての情報を持つ 2 つのデータセットを生成する．両データセットは分散管理されており，平文のままで情報を持ち出すこと，活用することが認められていないという状況を想定し，2 者間における垂直分割方式での PPDM を行うことを目的とする．

本稿では，最もシンプルな統計計算である線形回帰を取り上げる．個人情報を秘匿したままで線形回帰を行う際の正確性，パフォーマンスの 2 点を明らかにすることを試み，漏洩の可能性のあるデータについても検討する．

2 DPC

2.1 DPC データセット

DPC データセットは，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り，疾患，治療の組合せのデータから成る [2]．年齢，性別，疫病名，重症度，退院時転帰などの診療情報を含んでいる．

表 1　脳卒中データ

変数		最小・最大値	平均値	管理者
Death	y	0 - 1	0.12	A
Age	x_1	40 - 106	72.03	A
Sex	x_2	1 - 2	1.431	B
Japan Coma Scale	x_3	0 - 3	0.957	B
modified Rankin Scale	x_4	0 - 5	3.556	B
Stroke Type	x_5	1 - 3	1.432	B
Liver Disease	x_6	0 - 1	0.022	B

2.2 実験使用データ

本稿の実験で用いるデータとその統計量を表 1 に示す．表 1 は実際の患者のデータである．

4 節で後述する実験において，表 1 に示す脳卒中の患者のデータを使用する．患者の個人情報 3 項目，脳卒中の分類 1 項目，既往歴 6 項目，入院時の状態 2 項目の中から 7 項目を抽出する．年齢，性別，患者の退院時の生死を表す説明変数 Death，脳卒中の分類を表す Stroke Type に加えて，病歴から肝疾患 Liver Disease の有無を表す項目と入院時の病状から 2 項目を選んでいる．

Japan Coma Scale は，入院時の意識レベルを 4 段階に分類したものであり，数字が小さいほど意識が明瞭である．modified Rankin Scale は，脳卒中により生じる身体障害についての項目であり，数値が大きいほど寝たきりなど重度の障害である．Stroke Type は脳卒中の病型を示しており，脳梗塞，脳内出血，くも膜下出血の 3 分類である [3].

2.3 データの分割について

ユーザ A を官公庁などの公的機関，ユーザ B を病院などの医療機関を想定している．表 1 のデータは，y，x_1 を持つユーザ A と，それ以外の全てのデータを持つユーザ B に所有されている．

†Chika Hamanaga, Department of Frontier Media Science, School of Interdisciplinary Mathematical Science, Meiji University, Kikuchi Laboratory.

3　提案方式

各々異なるデータセットを有するユーザ A とユーザ B が，互いにデータセットを参照させることなく，安全に正しく線形回帰を実行するプロトコルを提案する．

加法準同型性を満たした公開鍵暗号（本研究では Paillier 暗号 [4]）を用いて，互いのデータを暗号化したままで，線形回帰 $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_m x_m$ を実行し，線形式の係数 $\alpha, \beta_1, \beta_2, ..., \beta_m$ を得る．

単回帰

$$y = \alpha + \beta x \tag{1}$$

2 変数の重回帰

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 \tag{2}$$

3 変数以上の多変数に対応した重回帰

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_m x_m \tag{3}$$

の，3 つに分けて方式を提案する．ここで，レコード数を n，説明変数 x の数を m とする．

3.1　重回帰

垂直分割方式での重回帰を提案する．計算方法を以下に示す．(単回帰と 2 変数の重回帰は，[7] に示す．)

$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m$ について，偏微分を用いる．行列式での計算を用いて，線形式の係数 $\alpha, \beta_1, \beta_2, \cdots, \beta_m$ を算出する（Algorithm 1）．

$S = \sum (y_i - \alpha - \beta_1 x_{1i} - \cdots - \beta_m x_{mi})^2$ 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．

$$\begin{cases} \dfrac{\partial S}{\partial \alpha} = -2 \displaystyle\sum_i^n (y_i - \alpha - \beta_1 x_{i,1} - \cdots - \beta_m x_{i,m}) = 0 \\ \dfrac{\partial S}{\partial \beta_1} = -2 \displaystyle\sum_i^n x_{i,1} (y_i - \alpha - \beta_1 x_{i,1} - \cdots - \beta_m x_{i,m}) = 0 \\ \cdots \end{cases}$$

これらを整理して，$m + 1$ 個の連立方程式を

$$FX = G \tag{4}$$

と行列で表し，これを満たす X を求めればよい．ここで，

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

Algorithm 1 : scLinear(多変数の重回帰)		
	A	暗号鍵を生成
	$A \rightarrow B$	公開鍵を共有
1.	A	データ $y_1, y_2, ..., y_n$ を暗号化．
	$A \rightarrow B$	$Enc(y_1), ..., Enc(y_n)$　を送る．
2.	B	行列 F，G を求める．
		(ここで，A のみで求められるものは計算しない．)
3.	$B \rightarrow A$	$Enc(F)$，$Enc(G)$ を送る．
4.	A	復号し，F，G を求め，
		$FX = G$ となる X を求める．

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

とする．F の逆行列を左からかけて，係数を得る．

4　実験

提案方法を表 2 の環境上に scLinear システムとして実装した．

4.1　実験目的

システム scLinear を用いて，以下の 2 項目を評価する．

1. 本システムの計算結果の正確性．
2. 本システムのパフォーマンス．

4.2　実験方法

DPC データセットについて $m = 3$，4，5，6 の重回帰を実施する．$n = 1000$ 行，2000 行，5000 行のデータセットについて測定し，提案方式について評価する．

ユーザ間でのデータのやりとりはネットワーク通信で行われることが多いが，本システムでは通信の過程を省略し，通信内容を一次ファイルに出力する手法をとっている．

4.3　実験の結果

4.3.1　正確性

表 3 に 6 変数の重回帰での計算結果を示す．秘匿計算を行った結果を scLinear に，R での実行結果を coefficient に示している．

scLinear の実行結果は，R の結果と差が見られなかった．また，$n = 1000$，2000 においても差がなかったため，scLinear は正確に計算できていると言える．

表 2　実験環境

	実験	改良の検討
OS	Windows 7	OS X El Capitan
メモリ	11.7 GB	8 GB
CPU	Intel Xeon X5460	Intel(R) Core(TM) i5
クロック	3.16 GHz	2.9 GHz
使用言語	Java(1.8.0_45-b15) R(3.1.2)	C/C++
鍵長	2048[bit]	

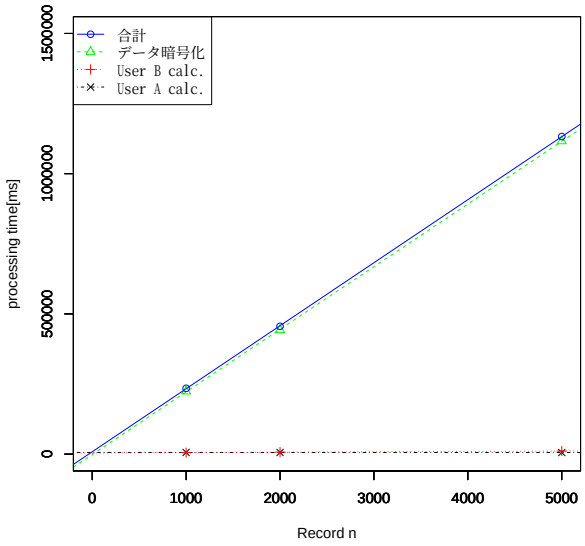

図 1　レコード数 n におけるシステムの実行時間

4.3.2　パフォーマンス

図 1 に scLinear の処理時間を示す．図 1 は，サイズの異なる 3 つのデータセットにおけるシステムの実行時間についてである．

図 1 より，レコード数 n に線形に，システムの処理時間が増加している．システム全体の実行時間 Total は $y = 224.576n + 9023.692$ で，暗号化処理にかかる時間は $y = 223.374n - 1649.538$ である．n はレコード数である．

4.4　実験考察

多変数に対応させた場合，ユーザ各々でデータを処理する時間が増加しているが，2 変数の重回帰では，ユーザ A の処理時間は最大でも 2000[ms] と，n に依存しない．

表 3 より，死亡という目的変数に対して，Japan Coma Scale，modified Rankin Scale，Stroke Type だけでなく，年齢，性別が生死に影響を与えていると分析できた．Japan Coma Scale の係数は 0.1283596 であり，入院時の意識レベルが 1 大きく（より悪く）なると，死に近づきやすくなることを示している．

scLinear を用いて，脳卒中の患者の DPC データについて考察し，入院時の意識の状態 Japan Coma Scale が，係数の絶対値が最大であることを示した．すなわち，死亡 Death という結果に対して最も支配的である．

5　改良の検討

提案方式では，計算途中の数値やデータのうち統計量を相手ユーザに提示していた．しかし，医療データであっても統計量を相手に公開することができない場合にもデータを活用できるよう，全ての過程を暗号化したままで実装したい．

そこで，完全準同型暗号ライブラリに着目した．提案手法で完全準同型暗号を用いた場合のパフォーマンスを検討する．

5.1　HElib

HElib は IBM 社が公開している，完全準同型暗号を C++ 上に実装したライブラリ [6] である．提案システムで用いていた Paillier 暗号は，暗号文同士の足し算は行えるが，掛け算は平文と暗号文での計算であった．完全準同型暗号ライブラリの HElib は，暗号文同士での足し算，掛け算の両方を行うことができる．

5.2　推定パフォーマンス

HElib を用いて，暗号化，足し算，掛け算，復号のそれぞれにおける時間を計測した．計測を行った環境は，表 2 である．100 回行い，それぞれのフェーズでの平均時間を求めた．結果を表 4 に示す．

表 4 の結果を用いて，HElib 上でシステムを実装できた場合の実行時間を推定し，提案方式と比較する．1000 行のデータにおける，表 4 での平均時間より推定した実行時間を表 5 に示す．$m = 6$，$n = 1000$ の場合において，暗号化 7008 回，加算 6000 回，乗算 5994 回，復号 34 回とし，HElib での平均処理時間を用いて導出した．

HElib での推定時間のほうが，暗号化，線形回帰計算を行う部分の処理時間が長くかかる結果になった．

表 3　線形回帰モデルの係数と提案方式の比較 ($n = 5000$)

	提案方式	R				
variables	scLinear	coefficient	Std. Error	t value	$Pr(>\|t\|)$	
α	−0.1731982	−0.1731982	0.0290099	−5.970	$2.53e-09$	***
Age	0.0015410	0.0015410	0.0003576	4.310	$1.67e-05$	***
Sex	−0.0217865	−0.0217865	0.0083993	−2.594	0.009519	**
JapanComaScale	0.1283596	0.1283596	0.0049296	26.039	$< 2e-16$	***
modifiedRankinScale	0.0121227	0.0121227	0.0034845	3.479	0.000507	***
StrokeType	0.0292522	0.0292522	0.0073582	3.975	$7.12e-05$	***
LiverDisease	0.0095770	0.0095770	0.0324591	0.295	0.767970	

表 4　HElib での実行時間 [sec]

	暗号化	復号	加算	乗算
最長	0.119506	0.139702	0.001142	0.144726
最短	0.100093	0.118716	0.000424	0.121745
平均	0.106915	0.128065	0.000541	0.130216
標準偏差	0.003698	0.004574	0.000089	0.003918

表 5　処理時間の比較 [sec]($m = 6$，$n = 1000$)

	scLinear	HElib での推定
暗号化	223.742	749.262
計算	6.010	783.765
復号	5.736	4.354
合計	235.488	1537.382

6　おわりに

医療機関と公的機関を例とした 2 組織間で，各々が情報を暗号化して秘匿したまま線形回帰を求めるプロトコルを実装し，その性能を評価した．実際の DPC データを用いて実験を行い，どの方法でも正確性のある結果を算出することができることを示し，パフォーマンスを評価した．

統計値のみの公開であっても，相手ユーザに提示する情報があり，暗号化したままではあるが公開している．そこで，相手に漏れてしまう情報をなくすべく，提案システムの改良についても検討した．HElib を用いた場合，すべての情報を暗号化して計算を行うため，提案システムより処理時間が長くなることが想定される．今後，HElib を用いたシステムを実装する場合においては，どう計算処理の高速化を図るかが課題である．

参考文献

[1] ベネッセホールディングス，“事故の概要”（http://www.benesse.co.jp/customer/bcinfo/01.html, 2015 年 6 月参照）

[2] 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．2012.

[3] 日本脳卒中学会，“脳卒中ガイドライン 2009”（http://www.jsts.gr.jp/jss08.html，2016 年 5 月参照）

[4] P. Paillier, Public-Key Cryptosystems Based on Composite Degree Residuosity Classes, EUROCRYPT 1999, pp.223-238，1999.

[5] 菊池，橋本，康永，“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”，DICOMO2014 シンポジウム，pp．219-223，2014.

[6] shaih/HElib（https://github.com/shaih/HElib, 2016 年 5 月参照）

[7] 濱永，菊池，康永，松居，橋本，“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”，DICOMO2016 シンポジウム，情報処理学会，pp.1471-1478，2016.