DPC データセットからのプライバシーを保護した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†

1　はじめに

2015 年に個人情報保護法が改正され，マイナンバー制度の導入が開始された．ベネッセ個人情報流出[1]も発生しており，情報の個人情報保護に関して，社会の関心が高まっている．

プライバシー保護をしつつ情報を活用する方法には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が，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を取り上げる．2 者間における垂直分割方式での秘匿計算システムの実装と，本システムのパフォーマンスの評価について報告する．

2　提案システム“scLinear”

本研究の目的である線形回帰を，暗号を用いた秘匿計算で行うシステムを実装した．線形単回帰，3 変数までの線形重回帰を実現している．本実装では，様々な暗号方式から，加法準同型性を持つ Paillier 暗号[2]を使用した．scLinear は，単回帰では$\mathrm{y} = \mathrm{a}x + \mathrm{b}$，重回帰では$\mathrm{y} = \mathrm{a}x_1 + \mathrm{b}x_2 + \mathrm{c}$の式を近似する．

2.1　使用したデータ

ＤＰＣデータセット[3]に基づく擬似データを使用した．本実験のデータは，表 1 に示す 655 行 3 列のデータである．

2.2　システムの動作

本システムは，表 2 に示す暗号化・復号化を行うユーザ A と計算を行うユーザ B との 2 者間で動作する．

2.3　システムのプログラム

プログラム開発は，表 3 の環境で行った．鍵生成の Key，暗号処理を行う paillier，線形回帰の計算を行う Linear の 3 クラスから構成されている．

2.4　計算システム

単回帰分析は$y_i = \mathrm{a}x_i + \mathrm{b}$，データ数はn と表す．

表 1：実験データ

		データ数	最小・最大値
年齢(歳)	x_1	655	0 - 95
入院日数(日)	x_2	655	0 - 101
総入院費用(円)	y	655	597 - 1291937

表 2：ユーザ情報

	ユーザ A	ユーザ B
担当	暗号・復号	回帰計算
所持データ	$y_1, y_2, \ldots, y_n$	$x_{1,1}, x_{2,1}, \ldots, x_{n,1}$
所持鍵	秘密鍵(p, q), n	公開鍵(n, g)

ここで，傾き a は，

$$\mathrm{a} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum {x_i}^2 - \sum x_i \sum x_i} = \frac{D}{C} \qquad (1)$$

切片 b は，

$$\mathrm{b} = \frac{\sum {x_i}^2 \sum y_i - \sum x_i \sum x_i y_i}{n\sum {x_i}^2 - \sum x_i \sum x_i} = \frac{E}{C} \qquad (2)$$

により算出される[4][5]．

プライバシー保護という観点から，B は秘匿計算を用いて， B には xを見せずに，傾き a，切片 b の 2 値の暗号文だけを計算し，ユーザ A は復号することで傾き a，切片 b だけを得る．

垂直分割での実装にあたり，加算と乗算は秘匿して行ったが，商は平文で求めた．傾き a や切片 b は，割り切れない数をとることが多く，乗法逆元を求めて掛け算としても，元の数値に戻らない．従って，プログラムとして実装できた計算は，和，差，積である．

本システムでは，この問題を解決するため，傾き a，切片 b についての計算は，ユーザ B が(1)，(2)式の C，D，E の暗号文をユーザ A に渡し，ユーザ A が復号した後で，(1)，(2)より傾き a，切片 b を算出する（Algorithm1）．

3　実験

3.1　実験目的

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.

3.2　実験方法

実験環境を表 3 に示す.

Estimation and implementation of Privacy-Preserving Linear Regression

†Chika HAMANAGA

†School of Interdisciplinary Mathematical Science, Meiji University

Algorithm 1:scLinear

	A	暗号鍵を生成
	A→B	公開鍵を共有
1	A	データ $y_1, y_2, \ldots, y_n$ を暗号化.
	A→B	$\mathrm{Enc}(y_1), \ldots, \mathrm{Enc}(y_n)$を送る.
2	B	データ$x_{1,1}, x_{2,1}, \ldots, x_{n,1}$から, $\mathrm{Enc}(C) = \mathrm{Enc}(n\sum {x_i}^2 - (\sum x_i)^2)$ Enc(y), データ x から, $\mathrm{Enc}(D) = \left(\left(\Pi \mathrm{Enc}(yi)^{\Sigma xi}\right)^n\right) *$ $((\Pi \mathrm{Enc}(yi))\hat{}(-\Sigma xi))$ $Enc(E) = \left(\Pi \mathrm{Enc}(yi)^{\Sigma xi^2}\right) *$ $((\Pi \mathrm{Enc}(y)^{\sum xi})\hat{}(-\Sigma xi))$ を出力.
3	B→A	Enc(C), Enc(D), Enc(E)を送る.
4	A	復号し, C, D, E と傾き a, 切片 b を求める.

表 3：実験環境

OS	Windows 7
メモリ	128GB
使用環境	Eclipse 4.4
使用言語	Java(1.8.0_40)
鍵長	2048[bit]

用意した DPC データについて, 線形単回帰, 線形重回帰をそれぞれ 10 回ずつ実施する. データについて, 所要時間を 100 行, 300 行, 500 行, 655 行のデータセットについて測定する.

3.3 実験 1 結果

表 4 に計算結果を示す. scLinear の実行結果は, R の計算結果と差がなかった. scLinear は, 高い精度で計算を正確に実行できると分かった.

表 4:実行結果（単回帰分析, 655 行）

行数	傾き a		切片 b	
	scLinear	R	scLinear	R
655	5002.521	5002.521	5099.358	5099.358
500	1021.751	1021.751	67751.810	67751.810
300	322.378	322.378	72369.082	72369.082
100	786.790	786.790	89508.076	89508.076

3.4 実験 2 結果

図 1 と図 2 に scLinear の処理時間を示す. ここで, 点線は重回帰分析を, 実線は単回帰分析の結果を示している.

Algorithm1 の 2 を行なう Linear クラスにおいて, 単回帰分析は 1 データあたりの平均処理時間が 1.9 ミリ秒であったが, 煩雑な計算を要求される重回帰分析は平均 16.7 ミリ秒を要した. 暗号化の所要時間には平均 320 ミリ秒と単回帰, 重回帰に差が見られなかった. 復号化については, 単回帰は平均 1.0 ミリ秒, 重回帰は平均 0.6 ミリ秒であり, 大きな差はない. 従って, Linear クラスでの処理時間が, システム実行時間に大きな影響を与えていると考えられる.

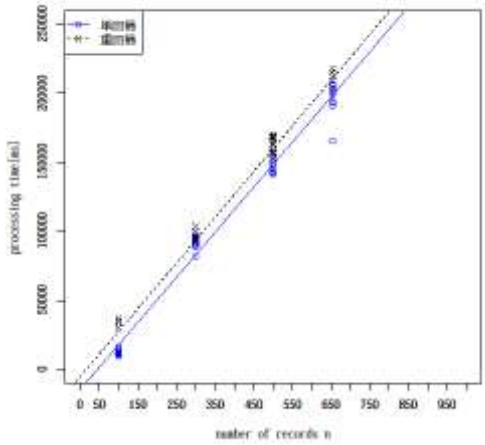

図 1：システム実行時間

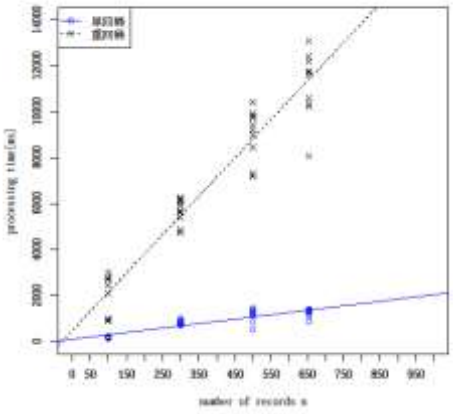

図 2：Linear クラス実行時間

4 おわりに

本研究では, プライバシー保護を目指した線形回帰システムの開発と, その評価を行った. scLinear は, 傾き, 切片の 2 情報を求めるために, 3 情報を相手ユーザに提示する. ここで, もれてしまう秘匿情報の安全性評価を今後の課題とする.

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) P. Paillier, Public-Key Cryptosystems Based on Composite Degree Residuosity Classes, EUROCRYPT 1999, pp.223-238, 1999.
3) 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.
4) 蓑谷, “線形回帰分析（統計ライブラリー）”, 朝倉書店, 2015.
5) 前野, “直線と曲線でデータの傾向をつかむ 回帰分析超入門”, 技術評論社, 2012.
6) 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, 312-OS-20-9, pp. 1-4, 2012.
7) 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.