2B5-OS-15b-3

プライバシを保護した尤度比検定

Privacy-preserving likelihood-ratio test

川崎 将平[*1] 呉 双[*1] 佐久間 淳[*1*2]
Shohei Kawasaki Shuang Wu Jun Sakuma

[*1]筑波大学 システム情報工学研究科 コンピュータサイエンス専攻
Dept. of Computer Science, Graduate Scool of SIE, University of Tsukuba

[*2]科学技術振興機構 CREST
Japan Science and Technology Agency CREST

Feature selection is an important task for risk model building. In this manuscript, we consider the problem of feature selection from private samples by means of privacy-preserving likelihood ratio test of the logistic regression model. We propose a cryptographically secure protocol which evaluates log likelihood of the logistic regression model in the vertical partitioned model. Privacy-preserving feature selection for the logistic regression model is realized by performing likelihood-ratio tests with using the privacy-preserving log likelihood evaluation protocol. We experimentally demonstrate our protocol with several benchmark data sets.

1. はじめに

近年では遺伝子分野の技術の発展により，個人ゲノムを比較的少ない金銭コストで簡単に検査することが可能となっている (e.g. 23 and me [*1])．個人ゲノムの情報に対して分析を行うことにより，医療において有益な情報を得ることができると期待されている．その一例として疾患リスク予測が挙げられる [1]．これはゲノム情報を利用して個人がある病気を発症するリスクを予測する手法であり，予防医療などに役立てることができる．このような予測を行うには，まず膨大なゲノム情報の中から注目する疾患に関係する特徴を探し出し，予測モデルを生成する必要がある．このように，ある目的に関係する特徴を選択する手法は特徴選択と呼ばれ，機械学習や統計分析の分野においてモデル生成の際の処理として重要である．

特徴選択を行うためには，大きく 2 つの方法がある [2]．1 つはデータセットに含まれる特徴の部分集合を使って実際に分析を行った後，交差検定などによって汎化性能が最も良い部分集合を選択する方法である．このような方法はラッパー法と呼ばれる．2 つ目は特徴の良さを表す規準を使って選択する方法であり，フィルター法と呼ばれる．本研究では，フィルター法に焦点を当てる．フィルター法による特徴選択では，特徴の出力への寄与を測る必要がある．疫学などの分野では 2 つのロジスティック回帰モデルの間で尤度比検定を行い，ある特徴が出力にどれだけ寄与しているかを測る手法が古くから用いられており，その結果を用いて特徴選択を行うことができる．

しかしながら個人ゲノムを用いた疾患リスク予測には，データの取り扱いについてプライバシ上の問題がある [3]．個人ゲノムはそれ自体が個人の識別子であり，同時に個人のセンシティブな情報を内包している．そのため個人ゲノムが漏洩すると多くの問題を引き起こす可能性があり，簡単にデータを公開することはできない．複数のパーティ間で個人ゲノムの情報を取り扱う場合，情報の安全性を保証できる手法が必要となる．

本研究では，準同型性を持つ公開鍵暗号を用いることでデータのプライバシを保護しながら，2 つのパーティ間で尤度比検定を実現するプロトコルを提案する．提案するプロトコルでは，ロジスティック回帰モデルの対数尤度をプライバシを保護しながら計算し，その結果を用いて尤度比検定を行う．尤度比検定の結果から，ある特徴の出力への寄与を測ることができ，特徴選択等に応用することができる．

連絡先: 川崎 将平，筑波大学 システム情報工学研究科 コンピュータサイエンス専攻，茨城県つくば市天王台 1-1-1，029-853-3826，kawasaki@mdl.cs.tsukuba.ac.jp

*1 https://www.23andme.com/

2. 尤度比検定による特徴の出力への寄与の評価

本節では，ロジスティック回帰モデルの対数尤度を用いた尤度比検定によって分類問題における特徴量の予測への寄与を評価する方法を示す．そのため，まずはじめにロジスティック回帰について導入する．

2.1 ロジスティック回帰

ロジスティック回帰とは，分類問題を解くための一般化線形モデルの 1 つである．ロジスティック回帰では，二値変数 $y \in \{0, 1\}$ を予測する分類問題を解くために，確率 $Pr(y = 1)$, $Pr(y = 0)$ をシグモイド関数を用いた予測モデルを用いて予測する．シグモイド関数は式 (1) で表される実関数であり，出力は $[0, 1]$ となる．

$$\sigma(a) = \frac{1}{1 + \exp(-a)} \tag{1}$$

予測モデルの入力を $\boldsymbol{x} \in \mathbb{R}^D$, モデルパラメータ $\boldsymbol{w} \in \mathbb{R}^D$ とすると，予測モデルは式 (2) で表される．

$$f(y|\boldsymbol{x}, \boldsymbol{w}) = \sigma(\boldsymbol{w}^T\boldsymbol{x})^y(1 - \sigma(\boldsymbol{w}^T\boldsymbol{x})^{(1-y)}) \tag{2}$$

モデルの最適化は，与えられたデータ $\{(\boldsymbol{x}_k, y_k)_{k=1}^N\}$ に対する最尤推定の枠組みで行われる．与えられたデータに対する式 (2) の予測モデルの対数尤度は式 (3) となる．

$$\ln \mathcal{L}(\boldsymbol{w}) = \sum_{k=1}^{N} \ln \mathcal{L}_k(\boldsymbol{w}) = \sum_{k=1}^{N} \ln f(y|\boldsymbol{x}_k, \boldsymbol{w}) = \sum_{k=1}^{N} \{y_k \ln \sigma(\boldsymbol{w}^T \boldsymbol{x}_k) + (1 - y_k) \ln \sigma(-\boldsymbol{w}^T \boldsymbol{x}_k)\} \tag{3}$$

最尤推定量は式 (3) の最大化によって $\boldsymbol{w}^* = \mathrm{argmax}_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{w})$ として得られる．

2.2 尤度比検定による特徴の出力への寄与の評価

ある特徴の出力への寄与の推定は特徴選択を行うために必要であり，また疫学等の医学分野においても重要視されている．モデルの最適化に最尤推定を用いる場合，特徴量の寄与を推定するには尤度比検定を用いる事ができる．尤度比検定とは尤度比に基づく統計的検定の 1 つであり，2 つのモデルの尤度に有意な差があるかどうかを検定することができる．この 2 つのモデルにおいて，一方ではある特徴 (あるいは特徴集合) を用い，もう一方では用いずに最尤推定を行い，それら 2 つのモデルの尤度に有意な差があるかどうかを検定することで，ある特徴 (あるいは特徴集合) の出力への寄与を推定する．

尤度比検定を用いて，分類問題における特徴量の寄与の大きさを評価する手法を述べる．2 つのロジスティック回帰モデルの最尤推定パラメータ $\boldsymbol{w}_1 \in \mathbb{R}^D$，$\boldsymbol{w}_2 \in \mathbb{R}^{D+\nu}$ とする．$\boldsymbol{w}_2$ は $\boldsymbol{w}_1$ の最尤推定に用いた特徴に加え，ν 個の特徴を用いて最尤推定を行ったパラメータである．このとき，この尤度比検定の帰無仮説は，2 つのモデルの尤度には差がないという仮説であり，検定統計量 G は式 (4) で定義される．

$$G = -2 \ln \left(\frac{\mathcal{L}(\boldsymbol{w}_1)}{\mathcal{L}(\boldsymbol{w}_2)} \right) = -2(\ln \mathcal{L}(\boldsymbol{w}_1) - \ln \mathcal{L}(\boldsymbol{w}_2)) \tag{4}$$

この検定統計量は漸近的に自由度 ν の χ^2 分布に従うことが知られている．そのため，この尤度比検定の p 値は式 (5) で求めることができる．

$$p = Pr[\chi^2(\nu) > G] \tag{5}$$

この p 値は，2 つのモデルの尤度には差がないという仮説のもとで式 (4) の検定統計量を観測する確率である．ゆえに，得られた p 値が小さいほど 2 つのモデルの尤度には差がないという仮説のもとでは稀にしか観測されないことを示し，実際には 2 つの尤度に統計的有意差があったということを示す．2 つの尤度に統計的な有意差が認められれば，$\boldsymbol{w}_2$ の最尤推定を行う際に追加した特徴 (特徴集合) は出力に対して寄与が大きいと判断できる．

3. 要素技術

本節では提案するプロトコルで用いる技術について述べる．

3.1 Paillier 暗号

Paillier 暗号 [4] は，加法準同型性を持つ公開鍵暗号である．ここで，加法準同型性とは平文同士の加算を暗号文における演算によって計算することができる性質である．Paillier 暗号の公開鍵 pk と乱数 $r \in \mathbb{Z}_n$ を用いて平文 $m \in \mathbb{Z}_n$ を暗号化する処理を $c = \mathrm{Enc}_{\mathrm{pk}}(m; r)$ と表し，秘密鍵 sk を用いて復号する処理を $m = \mathrm{Dec}_{\mathrm{sk}}(c)$ と表す．Paillier 暗号は以下の性質を持つ．

$$\mathrm{Enc}_{\mathrm{pk}}(m_1; r_1)\mathrm{Enc}_{\mathrm{pk}}(m_2; r_2) = \mathrm{Enc}_{\mathrm{pk}}(m_1 + m_2; r_1 + r_2)$$

$$\mathrm{Enc}_{\mathrm{pk}}(m_1; r)^{m_2} = \mathrm{E}_{\mathrm{pk}}(m_1 m_2; m_2 r)$$

本研究では，暗号プロトコル中で Paillier 暗号を用いた秘密計算を行う．以降では暗号化と復号の表記を省略し $\mathrm{E}_{\mathrm{pk}}(\cdot), \mathrm{D}_{\mathrm{sk}}(\cdot)$ と表記する．

3.2 多項式フィッティング

Paillier 暗号上では非線形な演算を行うことができない．この問題を解決するため，本研究では多項式フィッティングを利用して近似を行う．多項式フィッティングとは，与えられたデータ点の近傍を通る d 次元多項式の係数 α_i $(i = 0, \cdots, d)$ を回帰分析によって求める手法である．非線形関数 $f(z)$ の d 次元多項式による近似は，多項式フィッティングによって得られた係数を用いて $f(z) \simeq \tilde{f}(z) = \sum_{i=0}^{d} \alpha_i z^i$ と表すことができる．提案プロトコルでは，非線形関数 $\ln \sigma(z)$ を暗号上で計算するために多項式による近似を用いる．

非線形関数 $\ln \sigma(z)$ の入力 z 及び，d 次元多項式の係数 α_i $(i = 0, \cdots, d)$ は共に実数値を取り得るが，Paillier 暗号上での演算では全ての平文は整数であることを前提とする．そのため多項式の入力を $\lfloor Mz \rfloor$ とし，多項式の評価は式 (6) のように行う．

$$M' \widetilde{f_M}(\lfloor Mz \rfloor) = \sum_{i=0}^{d} M'(\alpha_i M^{-i})(\lfloor Mz \rfloor)^i = \sum_{i=0}^{d} M' \beta_i (\lfloor Mz \rfloor)^i \tag{6}$$

ここで M は入力 z を整数にするために十分大きな整数，M' は $\alpha_i M^{-i}$ を整数にするために十分大きな整数である．式 (6) において，$M' \widetilde{f_M}(\lfloor Mz \rfloor) = M' \tilde{f}(z)$ であり，M' の要素は後に除去することができる．

3.3 プライバシ保護ロジスティック回帰

本研究ではロジスティック回帰モデルの最尤推定パラメータをプライバシを保護したまま求める必要がある．そのため，プライバシ保護ロジスティック回帰 (以下 PPLR) [5] の手法を用いる．この手法では，2 つのパーティの間でプライバシを保護したままロジスティック回帰の最尤推定および予測を行うことができる．

2 パーティ A，B はプライベートなデータを垂直分割した形で所有し，それぞれ $\mathfrak{D}^A = \{\boldsymbol{x}_k^A | \boldsymbol{x}_k^A \in \mathbb{R}^{D_1}\}_{k=1}^N$，$\mathfrak{D}^B = \{\boldsymbol{x}_k^B | \boldsymbol{x}_k^B \in \mathbb{R}^{D_2}\}_{k=1}^N$ とする．ここで $D_1 + D_2 = D$ である．また B は分類ラベル $\{y_k | y_k \in \{0, 1\}\}_{k=1}^N$ を持ち，それぞれのデータと分類ラベルは一意に結合可能であるとする．

$$\boldsymbol{X} : \begin{pmatrix} \overbrace{x_{1,1}^A \cdots x_{1,d_1}^A}^{\mathfrak{D}_A} & \overbrace{x_{1,D_1+1}^B \cdots x_{1,D}^B}^{\mathfrak{D}_B} \\ \vdots \quad\quad \vdots & \vdots \quad\quad \vdots \\ x_{i,1}^A \cdots x_{i,d_1}^A & x_{i,D_1+1}^B \cdots x_{i,D}^B \\ \vdots \quad\quad \vdots & \vdots \quad\quad \vdots \\ x_{N,1}^A \cdots x_{N,d_1}^A & x_{N,D_1+1}^B \cdots x_{N,D}^B \end{pmatrix}$$

$$\boldsymbol{y} : (y_1, \cdots, y_k, \cdots, y_N)^T$$

PPLR ではこのような分割モデルのもとで $\mathfrak{D}^A, \mathfrak{D}^B$ および $\boldsymbol{y}$ をプライベートな入力として最尤推定を行い，推定パラメータ $\boldsymbol{w}^T = (\boldsymbol{w}_1^T | \boldsymbol{w}_2^T)$ となる $\boldsymbol{w}^A \in \mathbb{R}^{D_1}, \boldsymbol{w}^B \in \mathbb{R}^{D_2}$ を出力することが可能である．2 パーティの入出力を，$(A's\ input, B's\ input) \rightarrow (A's\ output, B's\ output)$ の形式で記述すると以下のようになる．

$$\mathrm{PPLR} : (\mathfrak{D}^A, (\mathfrak{D}^B, \boldsymbol{y})) \rightarrow (\boldsymbol{w}^A, \boldsymbol{w}^B)$$

4. プライバシを保護した尤度比検定による特徴量評価

本節では，分類問題における特徴量評価を 2 パーティ間でプライバシを保護したまま実行するプロトコルの提案を行う．また，プライバシを保護した特徴量評価プロトコルのためのサブプロトコルとなるロジスティック回帰の対数尤度計算プロトコルを提案する．提案するプロトコルでは，Paillier 暗号による秘密計算を行う．

4.1 問題定義

本研究の目標は，尤度比検定を用いて 2 つのパーティ間で特徴量の出力への寄与を評価することである．プライバシを保護した尤度比検定による特徴量評価を行うプロトコル (PPLT) の入出力を以下のように定義する．

$$\mathrm{PPLT} : ((\boldsymbol{w}_1^A, \boldsymbol{w}_2^A, \mathfrak{D}^A), (\boldsymbol{w}_1^B, \boldsymbol{w}_2^B, \mathfrak{D}^B, \boldsymbol{y})) \rightarrow (\emptyset, p)$$

また，PPLT で用いるサブプロトコルとしてロジスティック回帰モデルの対数尤度をプライバシを保護しながら計算するプロトコル (PPLC) を導入する．PPLC の入出力を以下に示す．

$$\mathrm{PPLC} : ((\boldsymbol{w}^A, \mathfrak{D}^A), (\boldsymbol{w}^B, \mathfrak{D}^B, \boldsymbol{y})) \rightarrow (\emptyset, \mathrm{E}_{\mathrm{pk}}(M' \ln \mathcal{L}(\boldsymbol{w})))$$

ここで，$\boldsymbol{w}^A$ および $\boldsymbol{w}^B$ は PPLR によって求めた結果を用いる．

4.2 プライバシを保護したロジスティック回帰モデルの対数尤度計算プロトコル

プライバシを保護しながら，ロジスティック回帰モデルの対数尤度を計算するプロトコルを Algorithm1 に示す．このプロトコルでは，式 (3) のロジスティック回帰モデルの対数尤度において $\ln \sigma(\cdot) \simeq \widetilde{\ln \sigma}_M(\cdot)$ と多項式近似し，Paillier 暗号上で評価する．

Algorithm1 では，まず分類ラベル y_k を持つ B がその値に応じて z_k を設定する．z_k は $\widetilde{\ln \sigma}_M(\cdot)$ の入力となる．このとき Paillier 暗号上では z_k^j 計算をすることができないため，z_k に乱数 r_k を加えた値を A が復号し，$\mathrm{E}_{\mathrm{pk}}((z_k + r_k)^j)$ を B に送信する．B は準同型暗号を用いて $\mathrm{E}_{\mathrm{pk}}((z_k + r_k)^j)$ から余剰な項を減算し，$\mathrm{E}_{\mathrm{pk}}(M' \ln \mathcal{L}_k(\boldsymbol{w})) = \widetilde{\ln \sigma}_M(z_k)$ を評価する．最後に，N サンプル分の加算をすることで $\mathrm{E}_{\mathrm{pk}}(M' \ln \mathcal{L}(\boldsymbol{w}))$ を得る．このプロトコルの時間計算量は，近似多項式の入力 z のべき乗を計算するための式 (7) の評価のために $\ln \mathcal{L}_k(\boldsymbol{w})$ の計算に対して $O(d^2)$ である．また $\ln \mathcal{L}(\boldsymbol{w})$ の計算には $\ln \mathcal{L}_k(\boldsymbol{w})$ を N 回計算する必要があるため $O(Nd^2)$ となる．

4.3 プライバシを保護した尤度比検定による特徴量評価プロトコル

プライバシを保護した尤度比検定によって特徴量の出力への寄与を測るプロトコルを Algorithm2 に示す．Algorithm2 では入力である 2 つのロジスティック回帰のパラメータ $\boldsymbol{w}_1, \boldsymbol{w}_2$ から Algorithm1 を用いて $\mathrm{E}_{\mathrm{pk}}(M' \ln \mathcal{L}(\boldsymbol{w}_1))$，$\mathrm{E}_{\mathrm{pk}}(M' \ln \mathcal{L}(\boldsymbol{w}_2))$ を計算する．B はこの 2 つの値から検定統計量を M' 倍した値の暗号文 $\mathrm{E}_{\mathrm{pk}}(M'G)$ を計算し，乱数を加えて A に送信する．A は $\mathrm{E}_{\mathrm{pk}}(M'G + r)$ を復号し B に送信する．その後，B は $M'G + r$ から r, M' の要素を取り除き $p \leftarrow Pr[\chi^2(\nu) > G]$ を計算することで，特徴量の出力への寄与を推定することができる．

Algorithm 1 プライバシを保護した対数尤度計算プロトコル

- public input: the polynomial coefficient $\beta_i (i = 0, \cdots, d)$
- the input of A: {pk, sk}, $\boldsymbol{x}_k^{\mathrm{A}}$, $\boldsymbol{w}^{\mathrm{A}}$
- the input of B: {pk}, $\boldsymbol{x}_k^{\mathrm{B}}$, $\boldsymbol{w}^{\mathrm{B}}$, $\boldsymbol{y}$
- the output of A: $\emptyset$
- the output of B: $\mathrm{E}_{\mathrm{pk}}(M' \ln \mathcal{L}(\boldsymbol{w}))$

1: **for** $k = 1$ to N **do**
2: A computes $\mathrm{E}_{\mathrm{pk}}(M\boldsymbol{w}^{\mathrm{A}} \cdot \boldsymbol{x}_k^{\mathrm{A}})$ and sends it to A
3: B evaluates:

$$\mathrm{E}_{\mathrm{pk}}(z_k) = \begin{cases} \mathrm{E}_{\mathrm{pk}}(M\boldsymbol{w}^{\mathrm{B}} \cdot \boldsymbol{x}_k^{\mathrm{B}}) \cdot \mathrm{E}_{\mathrm{pk}}(M\boldsymbol{w}^{\mathrm{A}} \cdot \boldsymbol{x}_k^{\mathrm{A}}) \text{ if } y_k = 1 \\ \mathrm{E}_{\mathrm{pk}}(-M\boldsymbol{w}^{\mathrm{B}} \cdot \boldsymbol{x}_k^{\mathrm{B}}) \cdot \mathrm{E}_{\mathrm{pk}}(M\boldsymbol{w}^{\mathrm{A}} \cdot \boldsymbol{x}_k^{\mathrm{A}})^{-1} \text{ o.w.} \end{cases}$$

4: B generates $r_k \in \mathbb{Z}_n$, and sends $\mathrm{E}_{\mathrm{pk}}(z_k + r_k) = \mathrm{E}_{\mathrm{pk}}(z_k)\mathrm{E}_{\mathrm{pk}}(r_k)$ to A
5: A decrypts $\mathrm{Enc}_{\mathrm{pk}}(z_k + r_k)$ to get $z_k + r_k$, and compute $(z_k + r_k)^j$ $(j = 1, ..., d)$
6: A encrypts $(z_k + r_k)^j$ $(j = 1, ..., d)$, and send $\mathrm{E}_{\mathrm{pk}}((z_k + r_k)^j)$ $(j = 1, ..., d)$ to B
7: B computes (7), for $j = 1$ to d

$$\mathrm{E}_{\mathrm{pk}}(z_k^j) = \mathrm{E}_{\mathrm{pk}}((z_k + r_k)^j) \cdot \prod_{i=0}^{j-1} \mathrm{E}_{\mathrm{pk}}(z_k^i)^{-\binom{j}{i} r^{j-i}} \quad (7)$$

8: B computes (8)

$$\mathrm{E}_{\mathrm{pk}}(M' \ln \mathcal{L}_k(\boldsymbol{w})) \leftarrow \prod_{j=0}^{d} \mathrm{E}_{\mathrm{pk}}(z_k^j)^{M' M^{-i} \beta_i} \quad (8)$$

9: **end for**
10: Finally, B computes (9) to get $\mathrm{E}_{\mathrm{pk}}(M' \ln \mathcal{L}(\boldsymbol{w}))$

$$\mathrm{E}_{\mathrm{pk}}(M' \ln \mathcal{L}(\boldsymbol{w})) \leftarrow \prod_{k=1}^{N} \mathrm{E}_{\mathrm{pk}}(M' \ln \mathcal{L}_k(\boldsymbol{w})) \quad (9)$$

5. 実験

Algorithm2 では多項式による近似を用いる．そのため，本節では Algorithm2 を用いて計算した p 値の精度を測るための実験とその結果について述べる．

5.1 実験設定

データセット: 2008.05.07 から 2010.04.20 の東京の気象に関するデータセット (以降 weather データセットとする) を用いた．このデータの特徴は，気温や湿度などの気象データの観測値である．このデータセットは特徴数 13，サンプル数 699 である．

Algorithm 2 プライバシを保護した尤度比検定による特徴量評価プロトコル

- the input of A: {pk, sk}, $\boldsymbol{x}_k^{\mathrm{A}}$, $\boldsymbol{w}_1^{\mathrm{A}}$, $\boldsymbol{w}_2^{\mathrm{A}}$
- the input of B: {pk}, $\boldsymbol{x}_k^{\mathrm{B}}$, $\boldsymbol{w}_1^{\mathrm{B}}$, $\boldsymbol{w}_2^{\mathrm{B}}$, $\boldsymbol{y}$
- the output of A: $\emptyset$
- the output of B: p-value, p

1: B gets $\mathrm{E}_{\mathrm{pk}}(M' \ln \mathcal{L}(\boldsymbol{w}_1))$, $\mathrm{E}_{\mathrm{pk}}(M' \ln \mathcal{L}(\boldsymbol{w}_2))$ by using PPLC

2: B computes:

$$\mathrm{E}_{\mathrm{pk}}(M'G) \leftarrow \mathrm{Enc}_{\mathrm{pk}}(M'\mathcal{L}(\boldsymbol{w}_1))^{-2} \cdot \mathrm{Enc}_{\mathrm{pk}}(M'\mathcal{L}(\boldsymbol{w})_2))^2$$

3: B generate $r \in \mathbb{Z}_n$, and sends $\mathrm{E}_{\mathrm{pk}}(M'G + r) = \mathrm{Enc}_{\mathrm{pk}}(M'G) \cdot \mathrm{E}_{\mathrm{pk}}(r)$ to A

4: A decrypts $\mathrm{E}_{\mathrm{pk}}(M'G + r)$ to get $m = M'G + r$ and sends it to B

5: B computes $G = (m - r)/M'$ and $p \leftarrow Pr[\chi^2(\nu) > G]$ (where ν is degree of freedom, equals to the difference between the dimension of $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$)

評価: 定数項のみのロジスティック回帰モデルの対数尤度 $\ln \mathcal{L}(\boldsymbol{w}_0)$ と，データセット中の 1 つの特徴を用いるロジスティック回帰モデルの対数尤度 $\ln \mathcal{L}(\boldsymbol{w}_i)(i = 1, \cdots, D)$ を用いて尤度比検定を行い，Algorithm2 による p 値と真の p 値の比較を行う．ここで，D はデータセットに含まれる特徴数である．

5.2 結果

真の $\ln \sigma(\cdot)$ を用いた場合の p 値 (赤) と 10 次元近似多項式を用いたときの p 値 (緑) の比較結果を図 1 に示す．図 1 から，Algorithm2 によって計算した p 値は真の値とおよそ等しい値となっていることがわかる．Algorithm2 による値と真の値との誤差は多項式近似の影響であるため，特徴毎に誤差の大きさや真の値との大小関係が異なっている．特徴の出力への寄与を測り，その結果を用いて特徴選択を行う場合にはこの p 値を比較する必要がある．そのため，誤差の大きさや真の値との大小関係を保証する必要があると考えられる．

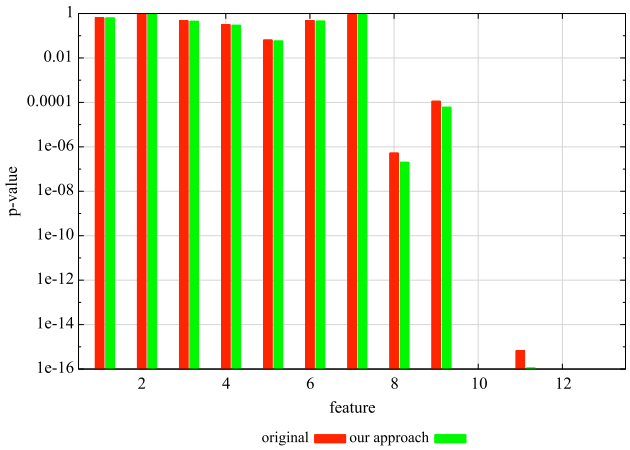

図 1: 10 次元近似多項式を用いた p 値の比較

6. まとめ

プライバシを保護した尤度比検定プロトコルによって，分類問題における特徴の出力への寄与の推定を実現した．提案プロトコルでは，多項式による近似を用いてロジスティック回帰の対数尤度を秘密計算により計算し，その結果を用いて尤度比検定を行う．また，実験の結果，我々のプロトコルは真の p 値と十分に近い値を出力することができた．我々のプロトコルを用いて特徴の出力への寄与を測ることによって，プライバシを保護したまま特徴選択を行うことが可能である．

しかしながら，プロトコルの中で多項式による近似を行っているため，真の p 値との誤差や大小関係が定まらないという問題があった．今後，用いる近似式等を工夫し，誤差の大きさや大小関係を保証することが課題となると考えられる．

謝辞

本研究は，JST CREST「ビッグデータ統合利活用のための次世代基盤技術の創出・体系化」領域におけるプロジェクト「自己情報コントロール機構を持つプライバシ保護データ収集・解析基盤の構築と個別化医療・ゲノム疫学への展開」の助成を受けました．

参考文献

[1] Naomi R Wray, Michael E Goddard, and Peter M Visscher. Prediction of individual genetic risk to disease from genome-wide association studies. *Genome research*, 17(10):1520–1528, 2007.

[2] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *J. Mach. Learn. Res.*, 3:1157–1182, March 2003.

[3] Erman Ayday, Emiliano De Cristofaro, Jean-Pierre Hubaux, and Gene Tsudik. The chills and thrills of whole genome sequencing. *CoRR*, abs/1306.1264, 2013.

[4] Pascal Paillier. Public-key cryptosystems based on composite degree residuosity classes. In *Proceedings of the 17th international conference on Theory and application of cryptographic techniques*, EUROCRYPT'99, pages 223–238, Berlin, Heidelberg, 1999. Springer-Verlag.

[5] Shuang Wu, Junpei Kawamoto, Hiroaki Kikuchi, and Jun Sakuma. Privacy-preserving online logistic regression based on homomorphic encryption. In *IEICE Tech. Rep.*, volume 113 of *IBISML2013-10*, pages 67–74, Tokyo, July 2013. Thu, Jul 18, 2013 : Nishiwaseda Campus (Waseda univ.) (IBISML).