Vol.2022-DPS-190 No.41
Vol.2022-CSEC-96 No.41
2022/3/11

プライバシ保護とデータ利活用の可能性検証

長尾 佳高[1,a)]　高橋 朋伽[1,b)]　大久保 佑弥[1,c)]　山月 達太[1,d)]　三本 知明[1,2,e)]　宮地 充子[1,3,f)]

概要：IoT 機器の普及に伴い，心拍数，運動量，歩数，脈拍，酸素摂取量，消費カロリーなど，様々な生活データが収集されている．これらのデータを解析することで病気の予兆などを知ることが可能になるといわれている一方で，本人の活動状況が本人以外のサーバで管理されることはプライバシの観点から危険である．そこで，プライバシを保護する方法として，データにランダムにローカルでノイズを加える技術であるローカル差分プライバシ（LDP）が提案されている．しかし，LDP によるプライバシの確保は重要だが，ノイズを付加したデータの解析の有用性の劣化という諸刃の剣ともいえる．本論文では，乳がん検診データのユースケースを用いて，LDP を適用したデータのプライバシプライバシ保護と乳がん判定における有用性について検証する．

キーワード：プライバシ，データ有用性

Feasibility Study between Privacy Protection and Data Utilization

Abstract: With the spread of IoT devices, various data about our lives are being collected, such as heart rate, physical activity, number of steps, pulse, oxygen intake, calorie consumption, etc. If these data can be analyzed, it will be possible to learn the signs of disease. However, it is dangerous for a person's activity status to be managed on an external server with the view of privacy. To solve this problem, local differential privacy (LDP), which is a technique for randomly adding local noise to data, has been proposed. While ensuring privacy by LDP is certainly important, it degrades the usefulness of the analysis of data with added noise. In this paper, we examine the privacy protection of data with added noise by LDP and its usefulness in determining breast cancer, using a use case of breast cancer screening data.

Keywords: privacy, data availability

1. はじめに

1.1 背景

情報技術の発達に伴い様々なデータが活用されている近年，データのプライバシの担保が課題となっている．特にビッグデータの解析において，ユーザの個人情報の適切な秘匿と利便性のバランスは重要な課題である．例えばデータを暗号化している場合，個人情報は秘匿されるが，それゆえに収集したデータを解析することは非常に困難である．

そこで，データを秘匿しつつ解析に用いるため，データにランダムにローカルでノイズを加える既存研究として LDP [1] が提案されている．これは入力データを部分的に秘匿することで，データのプライバシを担保しながらデータの出現頻度分析に用いることができる手法である．一方問題点として，ノイズを付加することによる利便性の劣化や，データの活用法が限定的になるなどがあげられる．さらに，入力するデータが離散地でなく連続値をとる場合，性能や利便性の担保がより困難になることも知られている．

本論文では乳がん検診データのユースケースを用いて，LDP とデータ汎化などによる複数のアプローチにより，プライバシを担保したデータ解析の手法とその有用性につい

1 大阪大学
Osaka University
2 国際電気通信基礎技術研究所
Advanced Telecommunications Research Institute International
3 北陸先端科学技術大学院大学
Japan Advanced Institute of Science and Technology
a) nagao@cy2sec.comm.eng.osaka-u.ac.jp
b) takahashi@cy2sec.comm.eng.osaka-u.ac.jp
c) okubo@cy2sec.comm.eng.osaka-u.ac.jp
d) yamatsuki@cy2sec.comm.eng.osaka-u.ac.jp
e) to-mimoto@atr.jp
f) miyaji@comm.eng.osaka-u.ac.jp

て検証する．LDP のパラメータの設定や満たしている差分プライバシについても示し，実環境においてどの程度実用的であるのかについても実験的に検証する．

ここで，本論文の構成について説明する．2 章では本論文で使用する学習手法である SVM とランダムフォレストについて解説し，3 章では既存研究である LDP や連続値に対する LDP，およびデータ汎化手法について説明する．4 章では実験に使用するデータと提案方式の出力を詳細に解説し，5 章でその性能についての評価を行う．6 章ではその結果を解析し，最後に 7 章にて結論を述べる．

2. 準備

2.1 線形 SVM

線形 SVM とは分類や回帰問題に適応できる機械学習モデルの 1 つである．データを 2 つのクラスに分離する超平面のうち各データから最も離れているものを学習し，少ないデータ量でも高い制度のモデルを得ることが可能といわれている．

$\boldsymbol{x} = (x_1, x_2, \cdots x_p)^T$ を p 個の説明変数からなるベクトル，$y \in \{-1, +1\}$ を目的変数とする．切片 b と係数ベクトル $\boldsymbol{w} = (w_1, w_2, \cdots, w_p)^T$ を導入し，超平面を表す関数 $f(\boldsymbol{x})$ を以下のように定義する．

$$f(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + b = \sum_{j=1}^{p} w_j x_j + b$$

この関数の出力値の符号に基づき

$$\begin{cases} f(\boldsymbol{x}) < 0 & \Rightarrow y = -1 \\ f(\boldsymbol{x}) \geq 0 & \Rightarrow y = 1 \end{cases}$$

と目的変数の予測を行う．

超平面の学習に利用する n 個の学習データの集合を $\{(\boldsymbol{x_i}, y_i) \mid i \in [n]\}$ とする．ここで $x_i = (x_{i1}, x_{i2}, \cdots x_{ip})^T$ は事例 i の説明変数のデータ，$y_i \in \{-1, +1\}$ は事例 i の目的変数のデータとする．一般には全てのデータを超平面で完全に分離できず，線形分離可能とは限らない．そこで分類の違反度を表す非負決定定数 $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \ldots \epsilon_n)^T$ を導入し，一部データが分類できないことを許容して超平面の学習をおこなう．分類の違反度最小化の優先度を表す非負パラメータ C を導入すると，線形 SVM のモデル学習は $y_i(\boldsymbol{w}^T\boldsymbol{x_i} + b) \geq 1 - \epsilon_i$ の制約条件のもと以下の最適化問題として記述される．

$$\text{minimize} \left(\frac{1}{2}\|x\|_2^2 + C\sum_{i=1}^{n} \epsilon_i \right)$$

本研究では $C = 1$ とする．

2.2 決定木

決定木は与えられたデータ集合の各データをある特徴量に対する条件式によって分割することを複数回繰り返すことでデータの分類や回帰分析のためのモデルを生成する手法である．アルゴリズムがシンプルであり，モデルの解釈がしやすいこと，計算量が他のアルゴリズムと比べて小さいことは利点であるが，先に述べた SVM などの他のアルゴリズムと比較すると精度が下がる傾向にある点が欠点である．

今回の実験では決定木の中でも分類を行う分類木を用いるため，分類木の生成方法について述べる．分類木を生成する際は，与えられた訓練データの集合を不純度が最小となるように分割していくという考え方で生成される．不純度とは，ある分割先に含まれるデータの集合の中にどの程度異なるクラスのデータが混ざり合っているかを示す値であり，誤分類率やジニ指数などを用いて表現される．

今回の実験ではジニ指数を用いて不純度が計算されるため，ジニ指数を用いた不純度の導出について説明する．ジニ指数は，分類されるクラスが $c = 1, 2, \cdots, n$ で表され，ある分割先の領域 D 内のデータの個数が m 個，領域 D 内のクラスが c であるデータの個数が D_c 個であった場合，以下のように求められる．

$$g_D = 1 - \sum_{c=1}^{n} \left(\frac{D_c}{m} \right)^2$$

このようにして計算されるジニ係数を，ある親ノード N の子ノード $C_1, \cdots, C_l$ に対して，ノード C 内のデータの個数を $|C|$ と表すとき，以下の数式

$$\sum_{i=1}^{l} \frac{|C_i|}{|N|} g_{C_i}$$

が最小となるように分割することを決定木の生成では繰り返し行われる．

2.3 ランダムフォレスト

ランダムフォレストとは分類や回帰問題に適応できる機械学習モデルの 1 つである．訓練データから複数の決定木を生成し，これらの決定木の出力の平均で分類先のクラスの選択や推定値の出力を行うことを可能とするアンサンブル学習の 1 つであり，その構造は木の深さ $depth$ と木の個数 $tree$ の 2 つの変数によって決定される．特徴として，調整する必要のあるハイパーパラメータの数が少ないことと事前に変数選択をする必要がない点が利点として挙げられるが，訓練データの個数が少ないと過学習が発生しやすい欠点がある．

ランダムフォレストを構成するうえで複数の決定木を生成するが，各決定木を生成するための訓練データの集合は，ランダムフォレストの生成のために与えられた訓練データ集合からブートストラップサンプリングと呼ばれる手法で選択される．この手法では，与えられた n 個のデータから

n 個のデータをランダムに重複を許して選択し，これらを用いて決定木を構築することを $tree$ 回繰り返すことで $tree$ 個の異なる決定木を生成している．また，木の深さ $depth$ と木の個数 $tree$ については $depth = 5$，$tree = 20$ として実験を行った．

3. 既存研究

3.1 RAPPOR Algorithm

RAPPOR (Randomized Aggregatable Privacy-Preserving Ordinal Response) アルゴリズム [1] は，匿名かつ強力なプライバシの保証のもとでエンドユーザ側のソフトウェアが統計情報を取得する方式である．主にデータのランダマイズにより匿名化を実現しており，以下のように構成される．

1. **Signal.** クライアントの入力 v について，Bloom filter B を構成する．ただし，Bloom filter のサイズを k，ハッシュ関数の個数を h とする．
2. **Permanent randomized response.** 各入力 v とインデックス i $(0 \leq i < k)$ について，以下の $B_i^{'}$ を計算する．
$$B_i^{'} = \begin{cases} 1 & (P(B_i^{'} = 1) = \frac{1}{2}f) \\ 0 & (P(B_i^{'} = 0) = \frac{1}{2}f) \\ B_i & (P(B_i^{'} = B_i) = 1 - f) \end{cases}$$
ただし、f はユーザが調整可能な縦断的なプライバシー保証のレベルを制御するパラメータである．
3. **Instantaneous randomized response.** 全要素が 0 に初期化された長さ k の配列 S について，各要素を以下の確率に従って設定する．
$$P(S_i = 1) = \begin{cases} q & (B_i^{'} = 1 \text{ の場合}) \\ p & (B_i^{'} = 0 \text{ の場合}) \end{cases}$$
4. **Report.** 生成したレポート S をサーバに送信する．

データの匿名化は，主に上記アルゴリズムの 2 つの Step にて行われる．まず，Step2 では元の値 B_i をランダムなノイズ $B_i^{'}$ に置き換える．$B_i^{'}$ は確率 f で元の値をランダムな 0,1 に置き換えており，平均化攻撃を避け，プライバシーを確保している．Step3 はリクエストに対する処理であり，クライアントは $B_i^{'}$ ではなく，さらにランダム化された S を報告する．この変更により，報告データを一意の識別子として扱うことができなくなる．最終的に，これら 2 つの Step と各パラメータ f, p, q によって短期的なリスクと長期的なリスクのバランスをとることができる。これらについて，RAPPOR アルゴリズムが満たすプライバシ要件は以下となる．

Theorem1 $\epsilon_\infty = 2h \ln \frac{1-\frac{1}{2}f}{\frac{1}{2}f}$ について，恒久的ランダマイズされたレスポンス (RAPPOR アルゴリズムの Step1,2) は ϵ_∞-差分プライバシを満たす．

Theorem2 $\epsilon' = h \ln \frac{q^*(1-p^*)}{p^*(1-q^*)}$ について，瞬間的ランダマイズされたレスポンス (RAPPOR アルゴリズムの Step3) は ϵ'-差分プライバシを満たす．ただし，$q^* = P(S_i = 1|b_i = 1) = \frac{1}{2}f(p+q) + (1-f)q$, $p^* = P(S_i = 1|b_i = 0) = \frac{1}{2}f(p+q) + (1-f)p$ とする．

また，RAPPOR アルゴリズムはデータ収集のシナリオの特性に応じて様々な方式で使うことができる．例えばサンプルサイズが小さい場合，アルゴリズムの一部を省略することでより効率的な学習につながることが知られており，以下はその代表的な例である．

- **One-time RAPPOR.** データ収集が一度しか行われない場合は縦断的なプライバシ保護の必要性がなくなり，Step3 を省略することができる．
- **Basic RAPPOR.** 収集する文字列の集合が小さく，well-defined かつ各文字列をビット配列の 1 要素に決定的にマッピングできる場合，ブルームフィルタを使用する必要はない．例えば性別データを収集する場合，2 ビットの配列を使用し各ビットを「男性」「女性」に対応付ければ十分である．この場合，ハッシュ関数が 1 で配列長が収集データの候補数である Bloom filter を使用しているとみなすことができる．
- **Basic One-time RAPPOR.** これは上記 2 つの方式を組み合わせた単純な構成であり，収集文字列を独自のビットに決定論的にマッピングしてランダム化し，Step3 を省略する．

3.2 連続値の LDP

RAPPOR アルゴリズムは主に離散値の入力を前提としているが，これに対して連続値を取り扱う LDP も複数提案されている．Wang らが提案した Piecewise Mechanism [2] は連続値 $t_i \in [-1, 1]$ を入力とし，$t_i^* \in [-C, C]$ を出力する方式である．t_i^* の計算にはラプラスメカニズムと Duchi の手法 [3] を用いており，適度に大きな確率で t_i と近づくことを可能にしている．また，一般の $[-r, r]$ から生成されるデータについても，入力を $t_i' = t_i/r$ とすることでこのアルゴリズムを適用することができる．

本研究では連続値データを離散値データに汎化することにより，RAPPOR アルゴリズムを連続値の入力に対して拡張する．

3.3 データ汎化

RAPPOR アルゴリズム [1] は連続値の入力データに対する使用を想定していないため，2 つの類似した入力に対して全く異なる出力になってしまう問題がある．この問題を解決するため，データ汎化によりデータをラベリングすることで，類似するデータを同一の入力データとみなし，RAPPOR アルゴリズムを適用する．ここではその汎化手法について，データの分類が異なる 2 つの手法を紹介する．また，各パラメータを以下のように設定する．

- $V = \{v_i\}_{i=1}^m$: 入力データ

- L_j: j 番目のラベル
- l: 分割数

まず，入力データを区間長が等しくなるように分割する手法について説明する．具体的には以下の手順でデータをラベリングする．

1. 入力 $V = \{v_i\}_{i=1}^m$ について，$d = \max(V) - \min(V)$ を計算する．
2. データ区間を $[\min(V), \min(v) + d/l) \cup ... \cup [\min(V) + (l-1)d/l, \min(V) + ld/l = \max(V)] = S_1 \cup ... \cup S_l$ に分割する．
3. データ区間 S_j に含まれるデータをラベル L_j にラベリングする．

また，分割した区間にデータが同数だけ存在するように分割する手法については，以下の手順で行われる．

1. 入力 $V = \{v_i\}_{i=1}^m$ について昇順にソートし，それを $V' = v'^m_{i i=1}$ とする ($i \leq j$ なら $v'_i \leq v'_j$ となる)．
2. データ区間を $[v'_1, v'_{\lfloor m/l \rfloor}] \cup ... \cup [v'_{\lceil (l-1)m/l \rceil}, v'_{lm/l} = v'_m] = S_1 \cup ... \cup S_l$ に分割する．
3. データ区間 S_j に含まれるデータをラベル L_j にラベリングする．

このようにデータを汎化することにより，連続値についても RAPPOR アルゴリズムを適用できることが期待される．一方で，学習データとテストデータのように事前に計算した分割データ区間を他のデータに適用したい場合，$x \notin S_1 \cup ... \cup S_l$ となるデータ x をラベリングする必要が出てくる．この場合，x は最も近い区間に分類することでこの問題を解決する．つまり，$x < \min V$ の場合はラベル L_1 に，$x > \max V$ の場合はラベル L_l に分類する．

4. 比較基準データ

4.1 Reference

本研究では，乳がん腫瘤の画像データから抽出された検診データを用いて有用性を検証する．良性・悪性を表す 2 値の診断結果（目的変数）と 30 種類の説明変数をもつ総数 569 のデータセットを用い，そのうち 357 が良性，212 が悪性である．説明変数の内容は

a) radius (mean of distances from center to points on the perimeter)
b) texture (standard deviation of gray-scale values)
c) perimeter
d) area
e) smoothness (local variation in radius lengths)
f) compactness (perimeter^2 / area -1.0)
g) concavity (severity of concave portions of the contour)
h) concave points (number of concave portions of the contour)
i) symmetry
j) fractal dimension ("coastline approximation" -1)

の 10 の項目に対し，それぞれの mean standard error, worst or largest の値となっている．今後，各説明変数を 1 から 30 までの整数で表す．具体的には 1 は平均半径，11 は半径の標準誤差，21 は最大半径とする．各説明変数の平均と分散は以下のとおりである．ただし，有効数字 3 桁とする．

表 1 説明変数の平均・分散

	生データ		汎化データ	
	平均	分散	平均	分散
1	14.1	12.4	14.5	20.4
2	19.3	18.5	20.1	36.7
3	92.0	5.90×10^2	94.8	9.54×10^2
4	6.55×10^2	1.24×10^5	7.47×10^2	2.55×10^3
5	9.64×10^{-2}	1.98×10^{-4}	9.80×10^{-2}	4.76×10^{-4}
6	1.04×10^{-1}	2.80×10^{-3}	1.14×10^{-1}	4.98×10^{-3}
7	8.88×10^{-2}	6.36×10^{-3}	1.04×10^{-1}	9.75×10^{-3}
8	4.89×10^{-2}	1.51×10^{-3}	5.57×10^{-2}	2.25×10^{-3}
9	1.81×10^{-1}	7.52×10^{-4}	0.184	1.56×10^{-3}
10	6.28×10^{-2}	4.98×10^{-5}	6.43×10^{-2}	9.83×10^{-5}
11	4.05×10^{-1}	7.69×10^{-2}	5.80×10^{-1}	0.326
12	1.22	3.04×10^{-1}	1.44	8.59×10^{-1}
13	2.87	4.09	4.24	18.9
14	40.3	2.07×10^3	78.7	1.20×10^4
15	7.04×10^{-3}	9.02×10^{-6}	8.56×10^{-3}	3.46×10^{-5}
16	2.55×10^{-2}	3.21×10^{-4}	3.16×10^{-2}	7.89×10^{-4}
17	3.19×10^{-2}	9.11×10^{-4}	6.14×10^{-2}	6.44×10^{-3}
18	1.18×10^{-2}	3.81×10^{-5}	1.42×10^{-2}	1.13×10^{-4}
19	2.05×10^{-2}	6.83×10^{-5}	2.41×10^{-2}	2.12×10^{-4}
20	3.79×10^{-3}	7.00×10^{-6}	5.70×10^{-3}	3.44×10^{-5}
21	16.3	23.4	16.9	38.0
22	25.7	37.8	26.4	64.9
23	1.07×10^2	1.13×10^3	1.12×10^2	1.89×10^3
24	8.81×10^2	3.24×10^5	1.07×10^3	7.64×10^5
25	1.32×10^{-1}	5.21×10^{-4}	1.35×10^{-1}	9.65×10^{-4}
26	2.54×10^{-1}	2.48×10^{-2}	2.94×10^{-1}	4.94×10^{-2}
27	2.72×10^{-1}	4.35×10^{-2}	3.20×10^{-1}	7.67×10^{-2}
28	1.15×10^{-1}	4.32×10^{-3}	1.17×10^{-1}	4.96×10^{-3}
29	2.90×10^{-1}	3.83×10^{-3}	3.08×10^{-1}	9.87×10^{-3}
30	8.39×10^{-2}	3.26×10^{-4}	9.13×10^{-2}	9.86×10^{-4}

5. Feasibility Study

5.1 実証方針

本実験では，機械学習の学習及び評価に用いるデータとして，次の 3 つのデータを扱う．

- 生データ: 乳がん検診データそのもの
- 汎化データ: 生データをラベリングし，各汎化区間の中央値に対応付けたもの
- LDP: 汎化データを RAPPOR アルゴリズムでノイズ付与し，元のデータ空間に戻したもの

汎化データ，LDP データの詳細な取り扱いについては次節にて説明する．また，使用する RAPPOR アルゴリズムのパラメータは 6 パターン用意した．これらはプライバシ指標 ϵ が同じパラメータとなる 3 つのグループに分類され，グループ内の 2 パターンは同じプライバシ指標について Basic RAPPROR および Basic One-time RAPPOR の 2 種類にそれぞれ対応している．各パラメータに対するプ

表 2 各パラメータに対する LDP のプライバシ指標 (ただし，有効数字 4 桁)

パラメータ	プライバシ指標 ϵ
LDP($f=0.1, q=0.9, q=0.1$) LDP($f=0.28$)	3.631
LDP($f=0.1, q=0.75, q=0.25$) LDP($f=0.55$)	2.531
LDP($f=0.3, q=0.75, q=0.25$) LDP($f=0.65$)	1.462

ライバシ指標は以下のとおりである．

以降，簡単のため $\epsilon_1 = 3.631, \epsilon_2 = 2.531, \epsilon_3 = 1.462$ と表記する．

注意点として，一般に複数の属性にそれぞれ LDP を通す場合，すべての属性に対するプライバシ指標はその属性数とそれぞれのプライバシ指標の積になることが知られている．これは，属性数とハッシュ関数の個数が一致しており，プライバシ指標がハッシュ関数の個数に比例するためである．つまり，例えば $\epsilon = 0.1$ で各属性にそれぞれノイズ付与した場合，今回のケースでの全体のプライバシ指標 ϵ_m は $\epsilon_m = 30\epsilon$ で表されることになる．

5.2 Implementation

モデル生成に利用する学習データの形式は様々なケースが想定される．評価データの形式も利用シーンによって様々だ．本節では生データ，汎化データ，LDP によるノイズ付与後の各データで生成した学習モデルに対して，各種データの性能評価を行う．LDP によるノイズ付与後のデータはラベル値となっている．そのため，各種データを双方向に解析を行うには，LDP によるノイズ付与後のデータを生データ汎化データ空間に戻す必要がある．今回はデータ汎化区間の中央値に戻すこととする．

また，機械学習を行う際，データの前処理として正規化を行う．データの正規化により，各特徴量のもつ値の重みを平等にし，学習コストを削減することが可能となる．生成したモデルの評価は交差検定を用いて行い，データの分割数は 10 とする．

本研究では Python3.7 をプログラミング言語として使用し，機械学習ライブラリの scikit-learn を使用して SVM，ランダムフォレストの実装を行った．

5.3 汎化と LDP について

ここではデータ汎化と LDP の関係性，およびその出力について説明する．3.3 節で解説したとおり，データ汎化は連続データを離散値にラベリングすることで LDP [1] の入力として使用するために用いられる．本論文ではラベルの個数を $l = 5$ とし，ラベルの候補数が十分に小さく well-defined であるため，LDP として Basic RAPPOR を用いる．これにより，ノイズ付加前にビット配列から元のラベル値に戻すことが可能になり，入力データ数が十分に大きい場合，最終的な出力から元のラベル値をある程度推測可能であることが期待できる．図 1 はその大まかな流れを示している．この例では入力ラベルに対応する要素が最終的な出力に残ったままとなっているが，必ずしもそうなるとは限らないことに注意する．

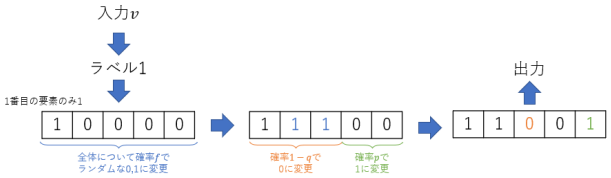

図 1 汎化と LDP によるデータの変化

また，LDP によりノイズが付与されたデータからラベルに戻す際は，1 が立っているビットからランダムに 1 つ選択し，これに対応するラベルを出力とする．すべてのビットが 0 になっている場合は，ランダムなラベルを出力とする．

5.4 属性数の削減

本項では，属性数の削減について説明する．5.1 節で述べたとおり，一般に属性数が増えると全体でのプライバシ指標 ϵ_m は増加してしまう．そのため，学習の精度を保ちつつ属性数を削減する手法についての評価も重要である．本研究では主成分分析と共分散による属性数の削減を行い，それぞれの学習精度についても比較する．属性数は 5 まで削減し，共分散による削減では絶対値の大きいものから 5 つ選択する．共分散により選択した属性は以下の 5 つである．

- radius (worst)
- perimeter (mean)
- perimeter (worst)
- concave points(mean)
- concave points (worst)

5.5 feasibility:SVM

本節では，生データ，汎化データ，LDP によるノイズ付与後のデータについて線形 SVM での性能評価を行う．ただし，学習データと評価データに用いる LDP のパラメータは同一とし，正答率は有効数字 3 桁とする．また，学習データと評価データに生データを用いた場合の精度を基準値とし，各学習結果に対して基準値との差も記載する．

5.5.1 SVM: パラメータによる影響

本項では，RAPPOR アルゴリズムの各パラメータやその方式，およびプライバシ指標 ϵ による性能の変化についての実験結果を記載する．

表 3 学習データを生データ，評価データを生データ，汎化データ，LDP($\epsilon = \epsilon_1,\epsilon_2,\epsilon_3$) によるノイズ付与後のデータとしたときの SVM の正答率 [%]．(有効数字 3 桁，() 内は基準値との差 [%])

評価データ	平均	最高値	最小値
生データ	96.7(±0)	-	-
汎化データ	95.4(−1.3)	-	-
Basic RAPPOR			
LDP $(f,p,q) = (0.1, 0.9, 0.1)$	76.4(−20.3)	81.4	70.8
LDP $(f,p,q) = (0.1, 0.75, 0.25)$	62.5(−34.2)	68.4	56.8
LDP $(f,p,q) = (0.3, 0.75, 0.25)$	58.2(−38.5)	66.1	51.8
Basic One-time RAPPOR			
LDP($f = 0.28$)	76.4(−20.3)	81.7	69.8
LDP($f = 0.55$)	62.5(−34.2)	68.5	56.2
LDP($f = 0.65$)	58.2(−38.5)	64.3	51.3

表 4 学習データを LDP($\epsilon = \epsilon_1$) によるノイズ付与後のデータ，評価データを生データ，汎化データ，LDP($\epsilon = \epsilon_1$) によるノイズ付与後のデータとしたときの SVM の正答率 [%]．(有効数字 3 桁，() 内は基準値との差 [%])

評価データ	平均	最高値	最小値
Basic RAPPOR			
生データ	93.4(−3.3)	95.2	91.3
汎化データ	95.6(−1.1)	97.4	94.0
LDP $(f,p,q) = (0.1, 0.9, 0.1)$	89.1(−7.6)	93.0	84.9
Basic One-time RAPPOR			
生データ	93.4(−3.3)	94.9	91.0
汎化データ	95.6(−1.1)	97.4	93.8
LDP($f = 0.28$)	89.1(−7.6)	93.0	84.7

表 5 学習データを LDP($\epsilon = \epsilon_2$) によるノイズ付与後のデータ，評価データを生データ，汎化データ，LDP($\epsilon = \epsilon_2$) によるノイズ付与後のデータとしたときの SVM の正答率 [%]．(有効数字 3 桁，() 内は基準値との差 [%])

評価データ	平均	最高値	最小値
Basic RAPPOR			
生データ	90.9(−5.8)	93.7	87.5
汎化データ	93.4(−3.3)	95.3	91.3
LDP $(f,p,q) = (0.1, 0.75, 0.25)$	78.3(−18.4)	84.0	73.5
Basic One-time RAPPOR			
生データ	90.9(−5.8)	94.1	87.9
汎化データ	93.5(−3.2)	95.3	91.3
LDP($f = 0.55$)	78.4(−18.3)	83.7	72.9

表 6 学習データを LDP($\epsilon = \epsilon_3$) によるノイズ付与後のデータ，評価データを生データ，汎化データ，LDP($\epsilon = \epsilon_3$) によるノイズ付与後のデータとしたときの SVM の正答率 [%]．(有効数字 3 桁，() 内は基準値との差 [%])

評価データ	平均	最高値	最小値
Basic RAPPOR			
生データ	89.9(−6.8)	93.5	84.7
汎化データ	92.9(−3.8)	95.3	90.2
LDP $(f,p,q) = (0.3, 0.75, 0.25)$	72.8(−23.9)	78.5	67.3
Basic One-time RAPPOR			
生データ	89.9(−6.8)	92.7	86.1
汎化データ	92.9(−3.8)	95.2	90.5
LDP($f = 0.65$)	72.8(−23.9)	78.9	66.6

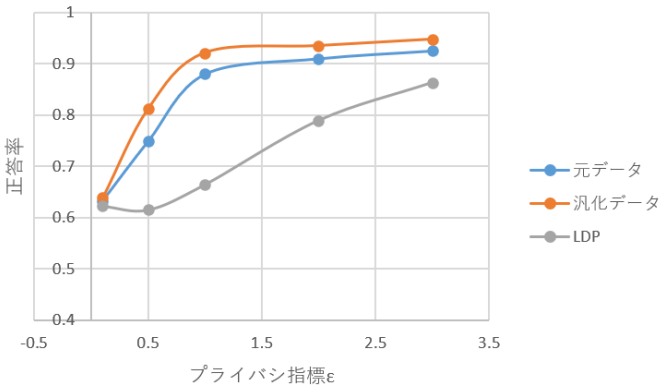

図 2 プライバシ指標 ϵ による SVM の学習性能の変化．(学習データは LDP によるノイズ付与後のデータとし，LDP として Basic One-time RAPPOR を使用)

5.5.2 SVM: 属性数の削減

本項では，全体のプライバシ指標 ϵ_m を揃えた場合について，生データ，属性削減したデータのそれぞれについて学習した結果を示す．つまり，各属性に対するプライバシ指標は ϵ_m を属性数で割った値となる．ここで，学習，評価に用いる生データは属性削減後のデータ (もしくは乳がん検診データそのもの) を表しており，汎化データ，LDP データはそのデータを基準に生成されていることに注意する．例えば共分散についての学習では，生データとしてもとの乳がん検診データを共分散により属性削減したものを用いている．

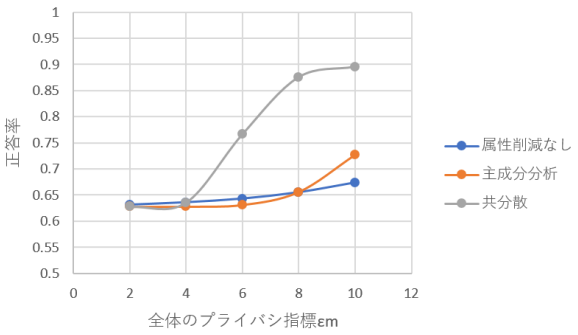

図 3 各属性削減データに対するプライバシ指標 ϵ_m による SVM の学習性能の変化．(学習データは LDP によるノイズ付与後のデータであり，評価データとして生データを使用)

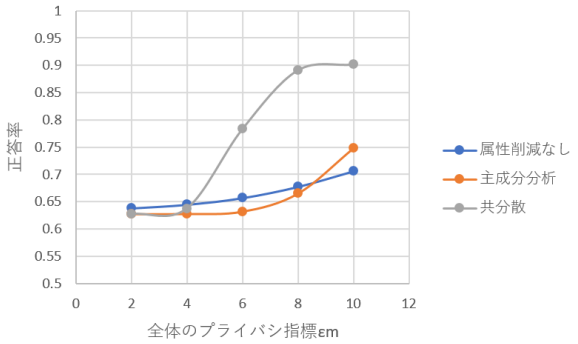

図 4 各属性削減データに対するプライバシ指標 ϵ_m による SVM の学習性能の変化．(学習データは LDP によるノイズ付与後のデータであり，評価データとして汎化データを使用)

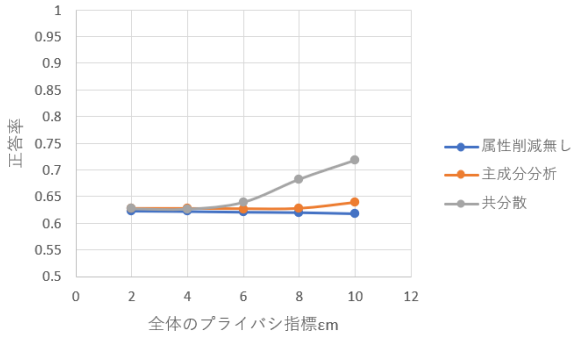

図 5 各属性削減データに対するプライバシ指標 ϵ_m による SVM の学習性能の変化．(学習データは LDP によるノイズ付与後のデータであり，評価データとして LDP によるノイズ付与後のデータを使用)

5.6 feasibility:forest

本節では，生データ，汎化データ，LDP によるノイズ付与後のデータについてランダムフォレストでの性能評価を行う．ただし，学習データと評価データに用いる LDP のパラメータは同一とし，正答率は有効数字 3 桁とする．

表 7 学習データを生データ，評価データを生データ，汎化データ，LDP($\epsilon = \epsilon_1,\epsilon_2,\epsilon_3$) によるノイズ付与後のデータとしたときの forest の正答率 [%]．(有効数字 3 桁，() 内は基準値との差 [%])

評価データ	平均	最高値	最小値
生データ	94.9(±0)	-	-
汎化データ	92.1(−2.8)	-	-
Basic RAPPOR			
LDP $(f,p,q) = (0.1, 0.9, 0.1)$	75.9(−19.0)	91.2	57.9
LDP $(f,p,q) = (0.1, 0.75, 0.25)$	64.4(−30.5)	80.7	50.9
LDP $(f,p,q) = (0.3, 0.75, 0.25)$	60.7(−34.2)	74.2	50.0
Basic One-time RAPPOR			
LDP($f = 0.28$)	75.8(−19.1)	91.2	59.6
LDP($f = 0.55$)	64.7(−30.2)	82.5	50.9
LDP($f = 0.65$)	60.9(−34.0)	78.9	50.4

表 8 学習データを LDP($\epsilon = \epsilon_1$) によるノイズ付与後のデータ，評価データを生データ，汎化データ，LDP($\epsilon = \epsilon_1$) によるノイズ付与後のデータとしたときの forest の正答率 [%]．(有効数字 3 桁，() 内は基準値との差 [%])

評価データ	平均	最高値	最小値
Basic RAPPOR			
生データ	79.5(−15.4)	93.0	68.4
汎化データ	78.7(−16.2)	91.2	66.7
LDP $(f,p,q) = (0.1, 0.9, 0.1)$	89.5(−5.4)	98.2	77.2
Basic One-time RAPPOR			
生データ	79.4(−15.5)	89.5	68.4
汎化データ	78.6(−16.3)	91.2	64.9
LDP($f = 0.28$)	89.5(−5.4)	98.2	75.4

表 9 学習データを LDP($\epsilon = \epsilon_2$) によるノイズ付与後のデータ，評価データを生データ，汎化データ，LDP($\epsilon = \epsilon_2$) によるノイズ付与後のデータとしたときの forest の正答率 [%]．(有効数字 3 桁，() 内は基準値との差 [%])

評価データ	平均	最高値	最小値
Basic RAPPOR			
生データ	77.3(−17.6)	89.6	61.4
汎化データ	76.7(−18.2)	89.5	64.9
LDP $(f,p,q) = (0.1, 0.75, 0.25)$	75.1(−19.8)	93.0	57.1
Basic One-time RAPPOR			
生データ	77.5(−17.4)	91.2	63.2
汎化データ	76.8(−18.1)	91.2	64.9
LDP($f = 0.55$)	75.8(−19.1)	93.0	56.1

表 10 学習データを LDP($\epsilon = \epsilon_3$) によるノイズ付与後のデータ，評価データを生データ，汎化データ，LDP($\epsilon = \epsilon_3$) によるノイズ付与後のデータとしたときの forest の正答率 [%]．(有効数字 3 桁，() 内は基準値との差 [%])

評価データ	平均	最高値	最小値
Basic RAPPOR			
生データ	74.6(−20.3)	87.7	56.1
汎化データ	74.3(−20.6)	89.5	57.9
LDP $(f, p, q) = (0.3, 0.75, 0.25)$	69.7(−25.2)	89.5	50.9
Basic One-time RAPPOR			
生データ	74.2(−20.7)	89.5	57.9
汎化データ	74.2(−20.7)	89.5	59.6
LDP($f = 0.65$)	69.4(−25.5)	87.7	51.8

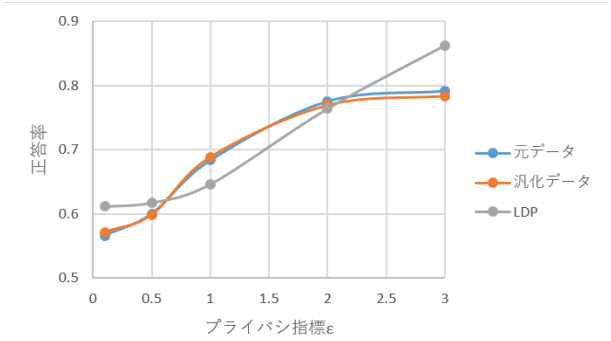

図 6 プライバシ指標 ϵ による forest の学習性能の変化．(学習データは LDP によるノイズ付与後のデータとし，LDP として Basic One-time RAPPOR を使用)

6. 結果・考察

6.1 Analysis

まず，表 3-6 の Basic RAPPOR と Basic One-time RAPPOR の項目をそれぞれ比較すると，プライバシ指標が同じ場合は正答率がほとんど変わらないことが確認できる．また，評価データとして LDP によるノイズ付与後のデータを用いる場合，学習データに生データや汎化データを用いるよりも，評価データと同様にノイズ付与したデータを用いたほうが性能が高いことも特徴的である．さらに，このように学習データにノイズ付与後のデータを用いた場合，生データや学習データに対する正答率も 9 割前後に保たれている．一方で図 2,6 からも分かる通り，プライバシ指標 ϵ を小さな値，つまり大きくノイズを付与した場合，ノイズ付与後のデータに対する学習精度がまず急激に下落し，続いて生データや汎化データに対する性能が下落する．

また，図 3-5 は各プライバシ要件を緩和するため属性数を削減した場合の精度を表している．特に学習データとして生データ，汎化データを用いた場合に，同じプライバシ指標 ϵ_m であっても共分散による属性削減で高い精度となることが確認できる．

次に，表 7-10 の Basic RAPPOR と Basic One-time RAPPOR の項目をそれぞれ比較すると，SVM での学習と同様にプライバシ指標が同じ場合は正答率が大きく変わらないことが確認できる．また，評価データとして LDP によるノイズ付与後のデータを用いる場合，学習データに生データや汎化データを用いるよりも，ノイズ付与したデータを用いたほうが性能が高いことも SVM での学習結果と同様である．

7. 結論

本論文では乳がん検診データのユースケースを用いて，データ汎化による RAPPOR の拡張アルゴリズムの提案と性能評価を示した．離散値データのみでなく連続値データを扱えるようにデータ汎化を行い，ノイズを付与した．さらにラベルと生データの空間の対応付けを行い，ノイズを付与したデータを SVM などの機械学習に用いることを可能にした．また，生データ，汎化データ，LDP によるノイズを付与した後のデータについて SVM とランダムフォレストでの学習精度の評価を行い，ノイズを付与したデータについて学習した場合に高い精度でデータの分類が成功することを示した．

また，共分散による属性の削減により，同じプライバシ指標でもより高い精度で学習ができることも示した．一方で，このように属性を削減するには生データを参照する必要があり，プライバシを侵害する要因となる．そのため実用上はランダムサンプリングによる共分散の安全な計算などの工夫が必要となることが今後の課題として挙げられる．

謝辞 本研究の一部は文部科学省「Society5.0 に対応した高度技術人材育成事業成長分野を支える情報技術人材の育成拠点の形成 (enPiT)」，文部科学省の平成 30 年度「Society 5.0 実現化研究拠点支援事業」，さらに JSPS 科研費 JP21H034438 の助成を受けています．

参考文献

[1] Úlfar Erlingsson, Pihur, V. and Korolova, A.: RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response, *Proceedings of the 21st ACM Conference on Computer and Communications Security*, Scottsdale, Arizona, (online), available from ⟨https://arxiv.org/abs/1407.6981⟩ (2014).

[2] Wang, N., Xiao, X., Yang, Y., Zhao, J., Hui, S. C., Shin, H., Shin, J. and Yu, G.: Collecting and Analyzing Multidimensional Data with Local Differential Privacy, *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, pp. 638–649 (online), DOI: 10.1109/ICDE.2019.00063 (2019).

[3] Duchi, J., Wainwright, M. and Jordan, M.: Minimax Optimal Procedures for Locally Private Estimation (2017).