LDPを用いた機械学習フレームワーク

宮地 充子[1,2,a)]　高橋 朋伽[1,b)]　Wang Ping-Lui[1,c)]　山月 達太[1,d)]　三本 知明[3,e)]

Abstract: 心拍数，運動量，歩数，脈拍，酸素摂取量，消費カロリーなど，私たちの生活に関するさまざまなデータが収集されています．これらのデータをプライバシーを保護しながら解析できれば，様々な問題を解決が可能になる．この問題を解決するために、データに局所的なノイズをランダムに加える手法である局所差分プライバシー（LDP）が提案されているが，ノイズが付加されたデータの解析の有用性を低下させる．本論文では，LDP に基づき，学習とテストの両フェーズでデータのプライバシーを保護する新しい仕組みを提案する．また，本機構の実現可能性を実験的に検証する．

Keywords: Privacy, Data Analysis

1. 背景

IoT 機器の普及に伴い，様々なデータが各地で分散的に収集されるようになった．我々の生活の日々のデータでは心拍数，運動量，歩数，脈拍，消費カロリーなどのデータも収集されるし，一方，医療機関には診察データ，さらには保健所には定期的な検査の結果が収拾されている．これらの分散されたデータを収集し，解析することで，病気の予兆の検知など各種課題解決が可能になる考えられている．この際，重要になることは各データに紐づくプライバシの取り扱いである．

分散したデータをプライバシ保護をしながらデータ解析に用いる方法にはいくつかの手法がある．大きく 2 つのアプローチに分けられる．第一のアプローチはセキュリティ技術を用いるアプローチである．機械学習をターゲットとした場合，第一のアプローチでは例えば学習モデルの構築時に最適化問題を暗号化，あるいは分散した状態で繰り返し解く必要がある．このボトルネックに対して，活性化関数の近似処理などによって高速化を図るアプローチなども存在する [1] が，依然として計算量と通信量が大きな問題となる．また，このアプローチは最終出力結果を得るユーザとデータ保有者が一致していない場合，本質的に保有データのプライバシを保護しているとはいえない．一方，第二のアプローチとしてデータの非識別化が考えられる．これはデータ自体，あるいはパラメータに加工処理を施すことで個々データのプライバシを保護するものである．特に最近の機械学習モデルへの攻撃の複雑さを鑑みて [2]，本論文では情報理論に基づく定量的なプライバシ強度を保証する差分プライバシに着目する．差分プライバシは中央型と局所型の大きく二通りの構成が存在する．中央型はデータサーバへのクエリに対するレスポンスに，局所型はデータ保有者がデータに直接確率的メカニズムを適用することでプライバシ保護を実現する．差分プライバシを満たす最も基本的なメカニズムとしてラプラスメカニズム [3] が知られているが，直接的にこのメカニズムを利用すると，データや分析結果に悪影響を及ぼす強いノイズが付与されうる．したがってユースケースに応じてメカニズムのチューニングが必要であり，例えばヘビーヒッター検知を始めとした頻度分析 [4,5] や仮説検定 [6,7]，さらには機械学習を対象としたメカニズム [8] などが提案されている．特に局所差分プライバシメカニズムを利用したアプリケーションは参考文献 [9] にてまとめられている．

本論文では信頼すべき機関が不要な局所差分プライバシに注目し，局所差分プライバシメカニズムを通したデータの機械学習への適用を検討する．これまで離散値，および連続値に対して高い汎用性がある局所差分プライバシメカニズム [10,11] が提案されている．しかしこれらは主に統計分析への利用が想定されているが，個々のデータにおける属性間の相関等の維持は考慮しないため，機械学習を前提とした利用は適さない可能性がある．我々は局所差分プライバシを満たすデータの機械学習への利用を想定したフレームワークを提案する．特にデータの次元数 K および各次元のクラス数 L をパラメータとしたデータ構造の簡略化，すなわち弱匿名化，により，データのプライバシと学習モデルの精度のバランスを取ることが可能となる．我々の提案は [11] と同様離散値および連続値を持つ任意のデータ形式の入力が可能であるが，提案する弱匿名化の目的は頻度分析のみではなく，必要十分な情報量の削減にある．また提案するアルゴリズムは学習モデルではなくデータに対して適用されるため，学習フェーズだけではなく評価フェーズにおいても機能する．先行研究の多くは学習モデル生成時の教師データのプライバシ保護を目的としているが，我々は評価フェーズにおけるテストデータのプライバシも考慮する．本研究ではフィージビリティスタディとして乳がんデータおよび電解データに対して提案フレームワークを適用し，その有効性を確認した．実験結果は弱匿名化のパラ

[1] 大阪大学　大学院工学研究科
Graduate School of Engineering, Osaka University
[2] 北陸先端科学技術大学院大学
JAIST
[3] 国際電気通信基礎技術研究所
Advanced Telecommunications Research Institute International
[a)] miyaji@.comm.eng.osaka-u.ac.jp
[b)] takahashi@cy2sec.comm.eng.osaka-u.ac.jp
[c)] ming@cy2sec.comm.eng.osaka-u.ac.jp
[d)] yamatsuki@cy2sec.comm.eng.osaka-u.ac.jp
[e)] to-mimoto@atr.jp

メータをコントロールすることで，適当なプライバシ強度で十分な精度のモデルを生成することが可能であることを示唆している．また評価フェーズでは，学習モデルへの入力として，生データではなく学習モデルの生成に利用したメカニズムを適用したデータを与える方が高い精度となることを確認した．

本稿は以下の構成からなる．2 章では SVM の概要および提案フレームワークを構成する LDP メカニズムを紹介する．3 章では本稿の提案メカニズムを述べる．続く 4 章では提案フレームワークを利用した実験結果をまとめる．最後に 5 章で実験結果にもとづく考察を述べる．

2. 準備

この章ではまず使用する機械学習モデルである SVM について紹介し次に LDP について紹介する．

2.1 サポートベクターマシン (SVM)

SVM は分類問題や回帰問題を解く機械学習モデルの一種である．学習データセットには n 個のレコード $\mathsf{D_i}$ $(\mathsf{i} = 1, \cdots, \mathsf{n})$ があり，各レコード $\mathsf{D_i} = [\mathsf{D_{i,1}}, \mathsf{D_{i,2}}, \cdots, \mathsf{D_{i,m-1}}, \mathsf{TA_i}]$ には $m-1$ 個の属性と目的変数 $\mathsf{TA_i} \in \{-1, 1\}$ があるとする．線形 SVM の学習段階では切片 b とベクトル $\boldsymbol{w} = (w_1, w_2, \cdots, w_{m-1})$ で定義される超平面を表す関数 $f(D_i)$ を以下のように計算する．

$$f(\mathsf{D_i}) = \mathsf{w} \cdot \mathsf{D_i^T} + \mathsf{b} = \sum_{\mathsf{j}=1}^{\mathsf{m}-1} \mathsf{w_j} \mathsf{D_{i,j}} + \mathsf{b}$$

この超平面を用いて未知データ D_i' を $f(\mathsf{D_i'})$ の出力に従って分類する．

$$\begin{cases} f(\mathsf{D_i'}) < 0 & \Rightarrow \mathsf{TA_i'} = -1 \\ f(\mathsf{D_i'}) \geq 0 & \Rightarrow \mathsf{TA_i'} = 1 \end{cases}$$

線形 SVM モデルの限界は線形分離不可能なデータセットを正しく分類できないことである．そこで非線形 RBF カーネルを用い，ドット積演算を新しいカーネル関数 $\exp(-\gamma||\mathsf{D_i} - \mathsf{D_i'}||^2)$ (ただし γ は非負のパラメータ) に置き換えることにより，この制限を克服している [12].

また超平面によってデータを完全に分離できない場合，超平面の計算時にマージンを用いることができる．これは一部の学習例が誤って分類されることを許容するものであり，超平面の滑らかさを制御する非負のパラメータ C が存在する．

実験ではパラメータ γ を $\frac{1}{m-1 \times Var(\mathsf{D})}$ に設定し，パラメータ C を生データで良好に動作するように選択した．

2.2 局所差分プライバシー

局所差分プライバシー [13] では n 個のデータレコードはそれぞれデータ $\mathsf{D_i}(1 \leq \mathsf{i} \leq \mathsf{n})$ を持つ．各データは m 個の属性 $A_1, \cdots, A_m$ を含む．各属性は離散値でも連続値でも良く，離散的ならば属性は k 個のカテゴリ $1, 2, \cdots, k$ を持ち，連続値ならば正規化され $[-1, 1]$ の領域を持つ．このとき，各データ提供者はランダムノイズ関数 f を用い，データ収集者に $f(D_i)$ を送る．

定義 1 関数 f が全てのあり得る入力の組合せ x, x' に対して以下を満たすとき関数 f は ϵ-局所差分プライバシーを満たすという．

$$Pr[f(x) = y] \leq exp(\epsilon) \cdot Pr[f(x') = y]$$

Piecewise mechanism [11] は連続値に適用する局所差分プライバシーを満たすランダム化関数である．そのアルゴリズムを 1 で示す．このアルゴリズムの出力の確率密度関数は以下のようになる

$$\mathrm{pdf}(y|x) = \begin{cases} p, & \text{if} \quad x \in [l, r], \\ \frac{p}{exp(\epsilon)}, & \text{if} \quad x \in [-H, l) \cup (r, H], \end{cases}$$

Algorithm 1 Piecewise mechanism (PM) [11]

Input: continuous value x_j, range $[-t, t]$, privacy budget ϵ
Output: perturbed data y_j
1: Adapt range $[-t, t]$ to $[-1, 1]$
2: Sample t uniformly at random from $[0, 1]$
3: Compute $H = \frac{exp(\epsilon/2)+1}{exp(\epsilon/2)-1}$, $\ell = \frac{H+1}{2} \cdot x_j - \frac{H-1}{2}$, $r = \ell + H - 1$
4: **if** $t \leq \frac{e^{\frac{\epsilon}{2}}}{e^{\frac{\epsilon}{2}}+1}$ **then**
5: Sample y_j uniformly at random from $[\ell, r]$
6: **else**
7: Sample y_j uniformly at random from $[-H, \ell) \cup (r, H]$
8: **end if**
9: **return** y_j

Randomised Response メカニズム [14] は離散値に対する局所差分プライバシーを満たすためのランダムノイズ関数である．入力 x と出力 y は同様に L 種類の値を取る．RR メカニズムは以下のようにノイズを加える．

$$p(y|x) = \begin{cases} \frac{exp(\epsilon)}{L-1+exp(\epsilon)}, & \text{if} \quad y = x, \\ \frac{1}{L-1+exp(\epsilon)}, & \text{if} \quad y \neq x, \end{cases}$$

RR メカニズムは $\frac{exp(\epsilon)}{n-1+exp(\epsilon)}$ の確率で元の値と等しい値を出力し，$\frac{n}{n-1+exp(\epsilon)}$ の確率で元の値と異なる値を出力する．RR メカニズムは Algorithm 2 によって与えられる．

Algorithm 2 Randomised Response mechanism (RR) [14]

Input: discrete value x_j of A_j, L values $\{A_j[1], \cdots, A_j[L]\}$, privacy budget ϵ
Output: perturbed data y_j
1: Sample x uniformly at random from $[0, 1]$
2: **if** $x \leq \frac{exp(\epsilon)}{L-1+exp(\epsilon)}$ **then**
3: $y_j = x_j$
4: **else**
5: Sample y_j uniformly at random from $\{0, \cdots, l\}$ except x_j
6: **end if**
7: **return** y_j

3. 提案方式

機械学習は一般に，モデルを構築する学習フェーズとモデルを検証するテストフェーズの 2 つのフェーズで構成される．一般には，モデル構築サーバや運用サーバは信頼する，つまり，Trusted Third Party（TTP）を仮定することが多い．しかし，サイバー攻撃によるデータ漏洩のリスクをゼロにすることは難しく，絶対安全な TTP の構築は現実的ではない．つまり，TTP に基づくデータ利活用では，実質的なプライバシー保護の実現は困難といえる．

本論文では，構築されたモデルからのプライバシー漏洩だけでなく，モデル構築サーバや，データをテストするために

モデルを運用するサーバからのプライバシ保護を目的とした，プライバシを保護した機械学習フレームワークを提案する．提案フレームワークは，学習段階，テスト段階のいずれにおいても，TTP を仮定せず，各データ所有者は自らデータを制御可能となる．つまり，包括的なプライバシー保護付き機械学習フレームワークを提案する．

3.1 Main Concepts of Our Privacy Mechanism

フレームワークを示す前に，本論文で使用する表記法を紹介する．

- LDP: 局所差分プライバシ
- PM: the piecewise mechanism [11]
- RR: the randomized response mechanism
- Agg: 集約者
- ϵ: プライバシバジェット
- ϵ_K: プライバシバジェット ϵ/K
- n: レコード総数
- m: 1 レコードに含まれる属性数（次元数）
- K: m 属性から利用する属性数（利用次元数）
- TA: 目的属性
- A_j:j-番目の属性 (連続・離散値の両方) ($j \in [1, m-1]$) (目的属性 TA を含めない)
- D, $\mathsf{D_i}$: i 番目のレコード, $i = 1, \cdots, n$, m 属性が含まれる．$\mathsf{D_i} = [D_{i,1}, \cdots, D_{i,m-1}, \mathsf{TA}_i]$
- $D_{i,j}$: i 番目のレコード $\mathsf{D_i}$ の j 番目の属性データ ($j \in [1, m-1]$).
- ta_i レコード $\mathsf{D_i}$ の目的属性.
- $\max(A_j), \min(A_j)$: 属性データ A_j の最大値または最小値.（連続値・離散値いずれの属性でも使用可能）離散値データを準連続データに変換することで，連続値属性と離散値属性の両方で利用することができる (後述).
- $\mathsf{Range}_j = \max(A_j) - \min(A_j)$: 属性 A_j の範囲
- L: 弱匿名化時の分類数
- $A_{j_1}, \cdots, A_{j_K}$: 選択属性
- $\mathsf{WA}_j[1], \cdots, \mathsf{WA}_j[L]$: 属性 A_j の弱匿名化変換された属性

本提案フレームワークは，生データを収集する TTP を前提にせずに，各データ所有者は自分自身のデータを管理する．一方，既存のプロトコルの多くは，生データを収集または取り扱う信頼できる機関が必要である [8, 15]. 一般に，プライバシー保護と機械学習の性能は相反する性質を持つ．LDP-mechanism は，データプライバシー保護の観点から強力であるが，機械学習の性能を低下させる．そのため，既存のプロトコルは LDP-mechanism を用いて，(信頼できる) ローカルクライアント [8] や TTP サーバ [15] で学習パラメータにノイズを付加することで，学習モデルからの学習データのプライバシ保護を実現する．ここで，LDP で変換されたデータを LDP データと呼ぶ. 生データと LDP データだけでは，プライバシーと機械学習性能のバランスを取ることが難しい．TTP を前提としない機械学習に適した新たなプライバシーメカニズムを構築するために, データの特性, すなわち,1 レコードが複数の属性から構成されていることに着目する．属性の特性は多様であり，[16, 17] に見られるように，連続的な値もあれば，離散的な値もある．一方，連続データ用の LDP-mechanism [11] と離散データ用 [10, 14] は独立に構築されている．一つのフレームワークで両方のデータを統一的かつバイデザインで扱えると，多様なデータに対してスケーラブルといえる．そこで，我々は生データから LDP データのの中間的な位置づけである WA で示される弱匿名化データの概念を提案する．中間的な概念を経ることで，任意の連続・離散データは WA に変換され，統一的に WALDP と呼ばれる LDP データに変換する新しいプライバシーメカニズムを提案する．

次に, 機械学習のプライバシーと精度を制御する方法について検討する．特に本論文では，学習とテストの両フェーズで利用するデータを制御する．1 つのレコード D_i は複数の属性 $\{A_j\}$ から構成される．プライバシーの観点から，各属性のプライバシーバジェットを ϵ，属性の総数を m とすると，1 レコードのプライバシーバジェットは $m\epsilon$ となり，属性数が多いほど，プライバシーが浪費される．PM [11] では，使用する属性数 K はプライバシーバジェットに従って決定される．データ所有者は全属性 m から K 属性をランダムに選択し，LDP でノイズを付加し，残りの $m-K$ の属性を 0 に設定し，すべての m 個の属性を集約者 Agg に送信する．この方法は，平均のような目的には有効であるが，機械学習では，0 に意味があるため，精度をうまく制御することが難しい．また，各データ所有者がランダムにデータを扱うため，機械学習で必要な学習と評価の両フェーズを制御することも難しい．

以上のことから，我々はプライバシーを保護した機械学習フレームワークを提案する．提案フレームワークは，次元削減，学習，テストフェーズの 3 フェーズからなる機械学習フレームワークになる．

3.2 Our Proposed Privacy Mechanism

本節では，任意のデータに対して統一的なプライバシーメカニズム WALDP を提案する．WALDP は，3 つの変換関数から構成される．第一ステップは，離散データから順序付き離散データへの変換 DTO (discrete-to-ordered-discrete data) である．第二ステップは，弱匿名化データ WA に変換する, 最後に, 弱匿名化データ WA に対して，LDP ノイズを付加する．3 ステップの変換手法により，連続データ，離散データのいずれも一律にプライバシーメカニズムを実行できる．特に，データ所有者自身が全ての変換を実行できことが特徴である．

では各ステップを詳細に説明する．まず，離散データから順序付き離散データへの変換 DTO を説明する．ここで，A_j を離散属性とする．離散的属性は連続的属性と異なり大小比較が困難な場合がある．例えば，北，東，南，西のデータからなる**方向**を考えると，それぞれの方向の順序は単純に比較できない．そこで，離散データを比較するために，離散データにラベル $i = 1, 2, \cdots$ を形式的に付与する．このラベルにより，離散データも**順序付き離散データ**と見なせる．この結果，離散データも最小値または最大値を定義でき，最小，最大離散データを，それぞれ $\min(A_j), \max(A_j)$ で表す．この結果，離散データも連続値データと同様，$\mathsf{Range}_j = \max(A_j) - \min(A_j)$ と範囲を定義できる．なお，離散属性 A_j のクラス数は，$\min(A_j) > 0$ のとき $\mathsf{Range}_j + 1$ に等しくなる．すなわち，連続値属性における $\min(A_j), \max(A_j), \mathsf{Range_j}$ は離散値属性でも使用可能となる．方向の例では，$A_j[1] =$ 北, $A_j[2] =$ 東，$A_j[3] =$ 南 , $A_j[4] =$ 西 とし，$\min(A_j) = 1$, $\max(A_j) = 4$, $A_j[4] =$, $\mathsf{Range_j} = 3$ となる．つまり，Algorithm 3(DTO) に属性の { 西 } と全属性 { 北，東，南，西 } を入力すると，$(2, \{1, 2, 3, 4\})$ を出力する．ここで，DTO は最初に決定し，データ所有者に通知する．

順序付き離散データの概念とその変換関数 DTO を用いることで，その後のデータから弱匿名化データ WA への変換は以

下のように定義される．

Algorithm 4 弱匿名化データ変換 (WAT)

Input: data x_j of attribute A_j, $\min(A_j)$, Range_j, number of classes L
Output: weak anonymized data y_j and L weak-anonymized classes of A_j $\{\mathsf{WA}_j[i]\}$

if A_j is continuous data **then**
  $\mathsf{WA}_j[1] \leftarrow \min(A_j) + \mathsf{Range}_j/2L$
  for $i = 2$ **to** L **do**
    $\mathsf{WA}_j[i] \leftarrow \mathsf{WA}_j[i-1] + \mathsf{Range}_j/L$
  end for
  $i \leftarrow \lceil \frac{(x_j - \min(A_j))L}{\mathsf{Range}_j} \rceil$
  $y_j \leftarrow \min(A_j) + (2i-1)\mathsf{Range}_j/2L$
else
  $(ox_j, \{\min(A_j), \cdots, \mathsf{Range}_j + 1\}) \leftarrow \mathsf{DTO}(\mathsf{x_j}, \mathsf{A_j})$.
  if $\mathsf{Range}_j \le L$ **then**
    $\ell \leftarrow \mathsf{Range}_j + 1$
  else
    $\ell \leftarrow L$
  end if
  for $i = 1$ **to** ℓ **do**
    $\mathsf{WA}_j[i] \leftarrow j_i$
  end for
  $i' \leftarrow ox_j \pmod L$,
  $y_j \leftarrow \mathsf{WA_j}[\mathsf{i'}]$,
end if
return y_j and $\{\mathsf{WA}_j[i]\}$

Algorithm 3 離散データから順序付き離散データへの変換 (DTO)

Input: data x_j of discrete attribute A_j and all available discrete data A_j
Output: order index ox_j of x_j and indexes of ordered-discrete A_j $\{\min(A_j), \cdots, \mathsf{Range}_j{+}1\}$

我々の提案フレームワークでは，連続データも離散データも，順序付き離散値の弱匿名化データ WA に変換する．つまり，本プライバシメカニズムは連続データと離散データの両方を WA を通じて統一的に扱うことができる．本プライバシメカニズムは，WALDP と呼び，あらゆるデータ型に対してスケーラブルな統一的プライバシーメカニズムである．既存の連続データに対するフレームワークでは，ϵ のみで有用性とプライバシを制御する [11]．注意したいのは，機械学習の有用性とプライバシはトレードオフの関係にあるため，制御が難しい．一方，本メカニズムでは，プライバシーバジェット ϵ と提案したプライバシメカニズムで用いる 弱匿名化データ変換 (WAT)（Algorithm 4) で変換される WA のパラメータ L の 2 つのパラメータを用いることで，機械学習の有用性とプライバシーをよりスムーズに制御することができる．Algorithm 5 は提案の LDP-メカニズム（Unified LDP-メカニズム) を表す．Algorithm 5 は生データ x を離散データを順序付き離散値に変換する DTO(Algorithm 3) と離散データと連続データを弱匿名化データに変換する WAT(Algorithm 4) と RR(Algorithm 2) を利用して，WALDP(x) に変換する．

Algorithm 5 Unified LDP-Mechanism (UPM)

Input: data x_j of attribute A_j, $\min(A_j)$, Range_j, number of classes L, privacy budget ϵ
Output: data z_j

if A_j is discrete data **then**
  $(x_j, \{\min(A_j), \cdots, \mathsf{Range}_j + 1\}) \leftarrow \mathsf{DTO}(x_j, A_j)$.
end if
$(y_j, \mathsf{WA}_j[1], \cdots, \mathsf{WA}_j[L]) \leftarrow \mathsf{WAT}(x_j, \min(A_j), \mathsf{Range}_j, L)$
$z_j \leftarrow \mathsf{RR}(y_j, \mathsf{WA}_j[1], \cdots, \mathsf{WA}_j[L], \epsilon)$
return z_j

3.3 プライバシを保護した機械学習フレームワーク

ここでは，3.2 章で提案した LDP-mechanism (UPM) を用いることにより，スケーラブルなプライバシを保護した機械学習フレームワーク (Scalable Unified Privacy-preserving Machine Learning Framework) SUP.ML を提案する．SUP.ML は，大きく分けて 3 つのフェーズ, 次元削減フェーズ，学習フェーズ，テストフェーズで構成される．学習段階とテスト段階の 3 段階からなる．次元削減のフェーズは，属性数が多いとプライバシーバジェットを浪費することを回避することが目的である．一方，機械学習に有用な各属性の優先順位が分かっているユースケースもあるだろう．この場合，次元削減フェーズを行わずに，選択した属性に対して学習とテストを行う．本研究では，決定した属性数までランダムに属性を削減した学習モデルの構築についても実験を行う．ランダムに属性を削減する手法を DR.Rand と呼ぶ．ここでは，データのプライバシーを保護しながら属性を選択する 2 つの方法を提案する．1 つの方法は，ユーザが PM でデータにノイズを付加し，そのデータを用いて利用する属性を決定する方法で，DR.PM と呼ばれる．もう一つは，提案した WA を適用する方法であり,DR.WA と呼ばれる．まず，DR.PM について説明する．

Algorithm 6 PM による次元削減 (DR.PM)

Input: m-dimension raw data D= $[D_{i,1}, \cdots, D_{i,m-1}, \mathsf{TA_i}]$, privacy budget ϵ, the number of used attribute K

Output: chosen attributes $A_{j_1}, \cdots, A_{j_K}$

1: $\epsilon_{K+1} \leftarrow \epsilon/(K+1)$
2: Sample K values of $(D_{i,j_1}, \cdots, D_{i,j_K})$ and target attribute TA_i from m-dimension data $\{\mathsf{D_i}\}$. uniformly, execute $\mathsf{PM}(\mathsf{x_{i,j_1}}, \mathsf{Range}, \epsilon_{\mathsf{K+1}}), \cdots, \mathsf{PM}(\mathsf{x_{i,j_K}}, \epsilon_{\mathsf{K+1}})$, $\mathsf{PM}(\mathsf{TA}_i, \epsilon_{K+1})$, and send them to Agg.
3: Agg collects $\{(\mathsf{PM}\ (D_{i,j_1}, \epsilon_{K+1}), \cdots, \mathsf{PM}\ (D_{i,j_K}, \epsilon_{K+1}), \mathsf{PM}\ (\mathsf{TA}_i, \epsilon_{K+1})\}$, determines K-attribute $A_{j_1}, \cdots, A_{j_K}$ by evaluating these correlation coefficients without seeing any raw data.
4: **return** K-attribute $A_{j_1}, \cdots, A_{j_K}$

次に DR.WA について，説明する．なお，DR.WA ではデータにノイズを付加せずに，データを秘匿する．2 と 4 章では，2 値分類の SVM に適用するため，DR.WA では二値分類を前提として記載する．よって，TA = $\{-1, 1\}$ とする．

Algorithm 7 WA による次元削減 (DR.WA)

Input: m-dimension raw data D= $[D_{i,1}, \cdots, D_{i,m-1}, \mathsf{TA_i}]$, $\min(A_j)$, Range_j, the number of setting classes of attribute L, the number of used attribute K

Output: chosen attributes $A_{j_1}, \cdots, A_{j_K}$

1: Sample K values of $(D_{i,j_1}, \cdots, D_{i,j_K})$ and target attribute $\mathsf{TA}_i \in \{-1, 1\}$ uniformly, get $\{y_{j_s}\}$ by executing $\{\mathsf{WAT}(D_{i,j_s}, \min(A_{j_s}), \mathsf{Range}_{j_s}, L\}$ compute $y_{j_s} \cdot \mathsf{TA}_i$ for $s = 1, \cdots, K$, send K-tuple data to Agg.
2: Agg collects perturbed parts of correlation coefficients and determines K-attribute $A_{j_1}, \cdots, A_{j_K}$ without seeing any raw data.
3: **return** K-attribute $A_{j_1}, \cdots, A_{j_K}$

次に，学習フェーズとテストフェーズについて説明する．次元削減フェーズでは，学習とテストに用いる属性を決定した．学習フェーズとテストフェーズでは，決定した属性を用いることで，プライバシーを保護した学習 PPTraining，テスト PPTesting を行う．また，PPTraining と PPTesting では，データ所有者は決定した属性の生データ x を WA または WALDP に変換し，変換したデータを集約者 Agg に送る．Agg は，データ所有者から送られたデータを用いて，学習とモデルの構築を行う．つまり，PPTraining と PPTesting で利用するデータタイプとしては，4 つの組み合わせがある．(PPTraining, PPTesting) = (WA, WA), (WA, WALDP). (WALDP, WALDP), (WALDP, WA) また，これらの組み合わせにさらに，3 種類の次元縮小アルゴリズム DR.Rand，DR.WA，DR.PM を組み合わせる．

また，我々の SUP.ML と PM を比較するために，生データを PM で変換した場合，すなわち，(training, testing)= (PM, PM) についても実験を実施した．PPTraining と PPTesting を以下で与える．

Algorithm 8 プライバシーを保護した学習モデル構築 PPTraining

Input: K data and target data, $[D_{i,j_1}, \cdots, D_{i,j_K}, \mathsf{TA_i}]$, the number of setting classes of attribute L, the privacy budget ϵ

Output: trained model

1: $\epsilon_{K+1} \leftarrow \epsilon/(K+1)$
2: Sample K data of $(x_{i,j_1}, \cdots, x_{i,j_K})$ and target data TA_i.
3: $y_{j_s} \leftarrow \mathsf{WALDP}(x_{i,j}, \min(A_j), \mathsf{Range}_j, \epsilon_{K+1})$ for $j = 1, \cdots, K$.
4: $y_{j_{K+1}} \leftarrow \mathsf{WALDP}(\mathsf{TA}_i, -1, 2, \epsilon_{K+1})$.
5: Send $K+1$-tuple perturbed data to Agg.
6: Agg collects perturbed $K+1$ data and constructs training model.
7: **return** Training model.

Algorithm 9 プライバシーを保護した学習モデルテスト PPTesting

Input: K data and target data $[D_{i,j_1}, \cdots, D_{i,j_K}, \mathsf{TA_i}]$, the number of setting classes of attribute L, the privacy budget ϵ

Output: Result.

1: $\epsilon_{K+1} \leftarrow \epsilon/(K+1)$
2: Sample K data of $(x_{i,j_1}, \cdots, x_{i,j_K})$ and target data TA_i.
3: $y_s \leftarrow \mathsf{WALDP}(x_{i,s}, \min(A_s), \mathsf{Range_s}, \epsilon_{K+1})$ for $s = j_1, \cdots, j_K$.
4: $y_{j_{K+1}} \leftarrow \mathsf{WALDP}(\mathsf{TA}_i, -1, 2, \epsilon_{K+1})$,
5: Send $K+1$-tuple perturbed data to a training model.
6: A training model executes perturbed $K+1$ attributes and gets the result.
7: **return** Result.

4. 実験結果

4.1 実験

本研究では UCI Machine Learning Repository より Breast Cancer Wisconsin (Diagnostic) dataset (WDBC) [16] と Ionosphere dataset [17] の二つを取得しメカニズムを評価した．これらのデータセットは特に SVM に適した二値分類問題のため用いた．しかしながら，我々が提案したメカニズムはどのような機械学習データセットに対しても適用可能であり二値分類問題に限定されるものではない．ここでは評価のため二値分類問題と SVM を選択した．本メカニズムの性能を評価するために、以下の 3 種類のデータを用いて SVM モデルを学習させ、その精度を比較した。

(1) SUP.ML **data**: 提案メカニズムを適用し，属性数の削減とデータの匿名化を行う SUP.ML データに対し，$K = [2, 10]$、$L = [2, 5]$ の全ての可能な組み合わせについて実

験を行った．本セクションでは良好な結果が得られた 2, 3 の組合せについて報告する．

(2) **Raw data**: データに対し何のノイズも加えていないデータである．一般にプライバシー保護を適用すると学習の精度は低下するため，このデータは達成できる最大の精度を示している．

(3) PM **data**:データに加えるノイズとして PW メカニズムを用いる．この結果は我々のメカニズムと比較するためのベーうラインとして用いる．

(4) **Mixed data**: PPTraining と PPTesting では、WA と WALDP の両データを使用した場合、4 つの組み合わせがある。そこで，これらの組合せをテストし，性能を確認する．

正規化されたデータについて 10 分割の交差検定を行いモデルの精度を評価した．なお交差検定前にデータセットをランダムにシャッフルしている．シャッフル，次元削減ノイズ付加のランダムシードを固定することで 3 種類のデータに対し同じ学習，テストの処理を行えるようにした．実験は、Intel Xeon Gold 5120 CPU と 48GB RAM を搭載した Ubuntu 20.04 マシンで行った.Python 3.8 と機械学習ライブラリ scikit-learn [12] を使用して SVM モデルを構築した．

4.2 WDBC Data Set

WDBC data set は乳がんの検診データセットである．乳がん腫瘍の画像データから抽出された 30 の説明変数により，良性・悪性を判断する．なお，要素数は 569 である．WDBC data set では SVM の正規化パラメータは $C = 2.1$ を用いる．

4.2.1 SUP.ML data

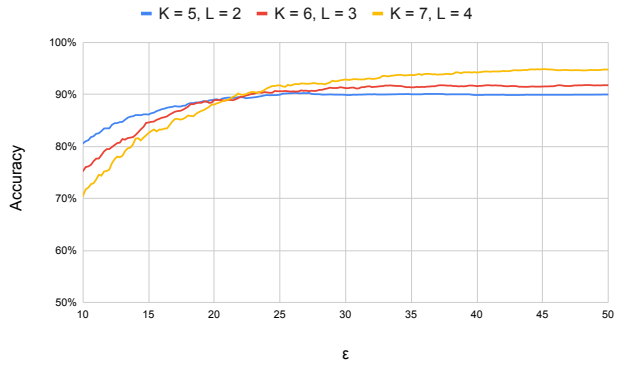

Fig. 1 WDBC data set における K, L の変化による accuracy の推移．x 軸は学習・評価フェーズを通じて用いたプライバシ指標，y 軸は正答率を表す．DR.Rand により属性を削減し，WALDP データにより学習・評価を行うものとする．

表 1 では $(K, L) = \{(5, 2), (6, 3), (7, 4)\}$ における正答率の変化を示す．なお，説明変数は $DRrand$ により決定したものとする．多くの属性・クラスを用いると，ノイズによる強い影響を受ける．そのためプライバシ指標の変化により，$(K, L) = (6, 3)$ はで $(K, L) = (5, 2)$ よりも，$(K, L) = (7, 4)$ では $(K, L) = (6, 3)$ よりも急激な正答率の変化が観測された．$(K, L) = (5, 2)$ では，大きな正答率の変化は観測されなかった．すなはち，厳しいプライバシ指標下においても，少ない属性数・クラスであれば高い正答率となる．一方大きなプライバシ指標下では多くの属性・クラスを用いた方が高い正答率となる傾向が見られ，実験では $\epsilon > 21.4$ において $(K, L) = (7, 4)$ が最も高い正答率となった． と同様の傾向が DR.PM, DR.WA でも観測された．

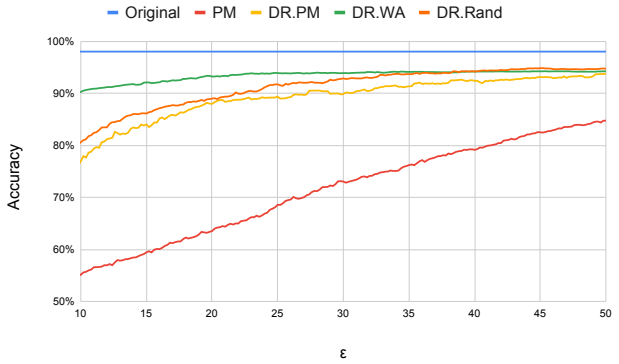

Fig. 2 WDBC data set における生データ，PM，SUP.ML data の正答率比較．x 軸は学習・評価フェーズを通じて用いたプライバシ指標，y 軸は正答率を表す．

表 2 の DR.WA, DR.PM では $(K, L) = \{(2, 2), (4, 4)\}$, ${}_D RRand$ では $(K, L) = \{(5, 2), (7, 4)\}$ における最大の正答率を表す．なお，学習・評価データはいずれも WALDP とする．DR.WA では DR.Rand と DR.PM と比較し高い正答率となった．今回の実験で最も厳しいプライバシ指標 $\epsilon = 10$ においても DR.WA では正答率 90.29% が確認された．

4.2.2 Raw data

WDBC data set にノイズ付与や属性削減を行わず学習・評価を行った場合，正答率は 98.04% であった．観測された 98.04% を WDBC data set における最大の正答率とし，他の実験結果と比較を行う．

4.2.3 PM data

プライバシ指標 ϵ が 50 よりも小さい場合，PM の最大の正答率は 84.77% であった．すべてのプライバシ指標において，SUP.ML よりも大幅に低い結果となった．

4.2.4 Mixed data

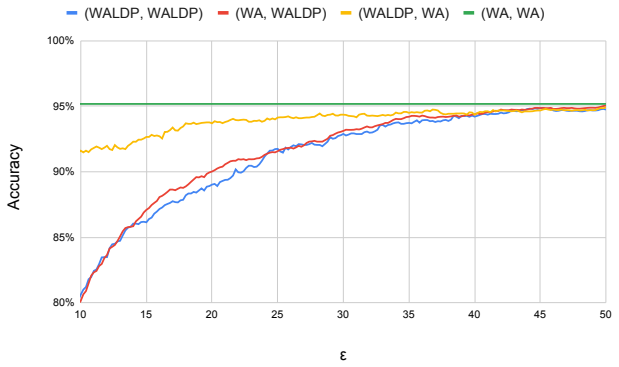

Fig. 3 WDBC data set における (PPTraining, PPTesting) の正答率の比較．なお，DR.Rand により属性を削減するものとする．x 軸は学習・評価フェーズを通じて用いたプライバシ指標，y 軸は正答率を表す．

表 3 では，PPTraining, PPTesting における比較を行い．(WALDP, WALDP)，(WALDP, WA)，(WA, WALDP)，(WA, WA) ではそれぞれ $(K, L) = \{(5, 2), (7, 4), (6, 3)\}$, $\{(6, 2), (8, 4)\}$, $\{(6, 2), (7, 4)\}$, $(7, 2)$ における最大の正答率を表す．ただし，$DRRand$ により属性は決定するものとする． 学習データが WALDP の場合，評価データに関係せず正答率に大きな変化は

見られませんでした．(WA, WA) と (WA, WALDP) を比較した場合も，正答率の差は最大で 3.57%と大きな性能劣化は確認されなかった．すなはち学習データに対してノイズを加えても正答率に大きな影響をもたらさないといえる．同様の傾向が DR.PM, DR.WA でも確認された．

4.3 Ionosphere Dataset

Ionosphere Dataset は Goose Bay システムで収集されたレーダーのデータで構成されている．このデータセットは 17 個のパルスと複素数の値からなり，パルスが電離層に何らかの構造がある証拠を示しているかどうかを示すラベルが付けられている．17 個のパルスは複素数値を持ち，2 つの実数値に分割できるため，このデータセットは 34 個の連続した属性を持つことになる．しかしこのデータセットを検査した結果，属性の 1 つが常に 0 という値を持っていることがわかった．属性は 0 か 1 のどちらかの値を持っている．したがって，このデータセットには 32 の連続属性と 1 つの離散属性があると考える，電離層データセットの実験では，SVM のパラメータとして $C = 3.9$ を用いた．

4.3.1 SUP.ML Data

SUP.ML の適用方法として，学習とテストに使用する属性数を削減する．その際，削減時の変数 (K, L) の値を決める必要があり，最適な値を決定するために実験を行う．(K, L) を設定した後，セクション**??**の次元削減の 3 つの方法，DR.WA, DR.PM，DR.Rand を比較する．

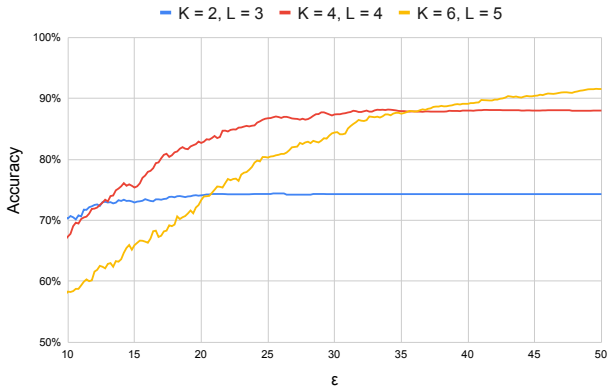

Fig. 4 Ionosphere データセットで K と L の異なる値を設定した場合の精度．x 軸はトレーニングとテストに使用するプライバシーバジェット，y 軸は精度である．学習とテストには WALDP データを用い、次元削減には DR.Rand データを用いる。

Optimal values of (K, L).

(K, L) の値をそれぞれ $[2, 10]$, $[2, 5]$ の範囲で探索しその最適値を求めた．その結果プライバシーバジェットが大きい場合 K と L の値を大きくすると精度が向上することがわかった．一方プライバシーバジェットが厳しい場合は K と L の値を小さくすることで精度が向上することがわかった．表 4 はこのことを表している．ここでは DR.Rand を用い $(K, L) = (2, 3), (4, 4), (6, 5)$ の 3 つの構成でモデルを学習させる．その結果プライバシーバジェットが 36 より大きい場合 $(K, L) = (6, 5)$ が最も良い結果を与えることがわかった．また ϵ が 13 よりも小さい場合 $(K, L) = (2, 3)$ がより高い精度を維持することができることが確認できた．ϵ の値の中間の範囲では，代わりに $(K, L) = (4, 4)$ を使用する必要がある．さらに，次元削減の方法を DR.PM や DR.WA に変更しても，同じ状況が現れることを発見した．提案メカニズムで高い精度を実現するためにプライバシバジェットに応じて (K, L) を決定した．そして表 5 が示すように DR.PM では $(K, L) \in \{(2, 2), (3, 2), (4, 2)\}$ が，DR.WA では $(K, L) \in \{(2, 2), (4, 2)\}$ が，そして DR.Rand では $(K, L) \in \{(2, 3), (4, 4), (6, 5)\}$ が高い精度であることが確認できた．また DR.Rand と DR.WA の性能を比較すると DR.WA は $\epsilon < 25$ のときかなり高い精度が得られるが DR.Rand は $\epsilon > 47$ のとき若干高い精度が得られる．したがって最も高い精度を得るためには，次元削減を行う DR.WA を使用することを推奨する．

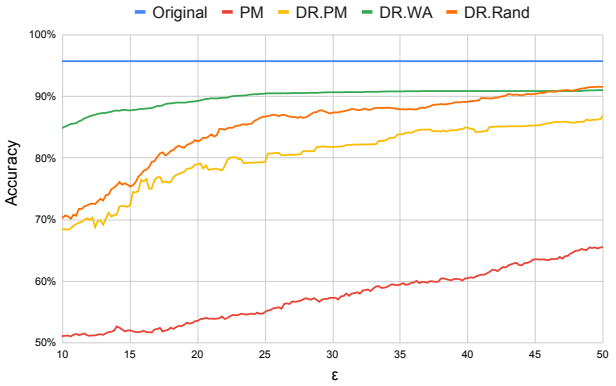

Fig. 5 Ionosphere データセットにおける生データ，PM データ，SUP.ML データの比較．x 軸は学習とテストフェーズで使用するプライバシーバジェット，y 軸は精度である．

4.3.2 Raw Data

このデータセットを直接使用して SVM モデルを学習させた場合 95.71%の精度を達成することができる．したがってこのデータセットで達成できる最大精度は 95.71%となる．この結果から我々の機構は $\epsilon = 50$ のときわずか 4.17%しか精度を劣化させないことがわかる．

4.3.3 PM Data

表 5 に示すように PM データはプライバシーバジェットとして $\epsilon = 50$ を用いても 65% の精度しか得られず生データで得られる精度に比べてはるかに低い．この結果は PM ノイズを直接データセットに適用する、学習済みモデルの精度が著しく低下することを示唆している．そこで高い精度を維持しつつデータにプライバシーを与えることができる新たな機構が必要であり本機構はこの要求を満たすものである。

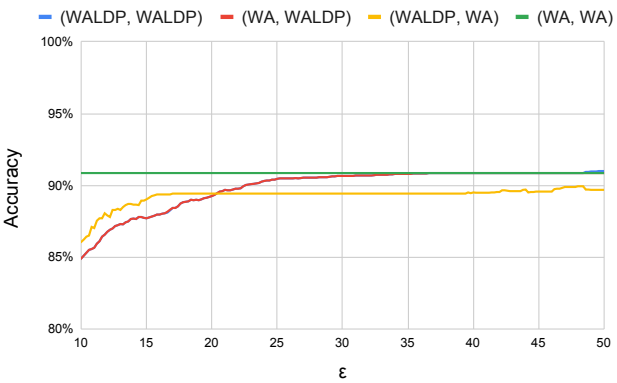

Fig. 6 Ionosphere データセットにおいて (PPTraining, PPTesting) の 4 つの組み合わせで DR.WA を次元削減に使用した場合の比較である．x 軸は WALDP のプライバシーバジェット，y 軸は精度を表す．

4.3.4 Mixed data

表 6 では 4 種類の (PPTraining, PPTesting) の組み合わせを比較した．(WA, WA) が常に最高の精度を示した．また、(WA, WA) は、(WA, WALDP) と (WA, WALDP) より精度が高いことがわかった．$\epsilon < 20$ の場合 (WALDP, WA) は (WALDP, WALDP) や (WA, WALDP) よりも良い精度を示す．また WA は弱い匿名化であるため学習モデルをダウンロードし我々の環境でテストできる場合はテストデータに WA を使用することができる．

なお今回は DR.WA を用いた次元削減の結果のみを示したが，他の次元削減手法でもテストしており同様の結果であった．

5. 解析

プライバシ強化型の機械学習モデルを構築する場合，LDP メカニズムのようなプライバシ保護技術をデータに直接適用する方法と，学習モデルの構築途中でプライバシ保護技術を適用する方法などが考えられる．前者は非効率であることが実験的に判明した．PM などの LDP メカニズムは主に統計解析を目的としており，データ固有の特性が失われることが原因として考えられる．後者については多くの研究が報告されているが，それぞれ特定のユースケースに特化したものであり，汎用性に欠く．ここでは過剰な情報を削減するために弱匿名化を行うことで，高精度な機械学習モデルを構築するフレームワークを提案した．本稿では実験を通じて以下の知見を得た．

5.1 K と L の選択

SUP.ML データの次元削減を行う際，プライバシバジェットが十分に大きい場合，各データに適用するノイズを削減するよりも，より多くの属性を使用する方が効果的であることが分かった．これはノイズが一定以下である場合，それ以上のノイズの削減はモデルの精度には寄与せず，一方で属性を増加することでモデルの精度向上が見込めるからである．またプライバシバジェットが小さい場合，属性数が多いと各属性値に対するノイズが大きくなり，モデルの精度が著しく劣化する．そのため，利用する属性数を削減することで，各属性値に対して低スケールのノイズに抑えることができ，より高い精度を達成することができる．なお，クラス数もモデルの精度とノイズのスケールに関係性があるため同様のことが言える．今後は最適な K および L の決定方法を確立し，提案フレームワークを利用することで，要求されるプライバシレベルに応じた最適なモデルの構築が可能となる．

5.2 次元削減手法の比較

次元削減手法を比較した結果より，DR.Rand は DR.PM と比較して同程度のプライバシレベルにおいてより高い精度を提供することが分かった．これは相関係数を計算する際，DR.PM はプライバシバジェットを消費することに起因する．この特徴はデータセットの属性数やデータ数によって逆転する場合があり，これは DR.PM の次元削減において，よりよい属性の選択ができているためであると考えられる．DR.Rand と DR.WA を比較すると，常に DR.WA が高い精度を持つことがわかる．しかし DR.WA は機密データに対して弱いプライバシ保護しか提供しないため，DR.Rand を使用することでデータセットに対して LDP を提供することができ，プライバシの観点からより良い手法として機能することが分かった．

6. 結論

本稿ではデータ型によらず統一的にデータを扱うためのプライバシメカニズム WALDP を提案した．WALDP はプライバシバジェット以外にデータの属性数とクラス数を扱うことでプライバシと有用性を制御することが可能である．また本稿では次元削減，学習，テストの各フェーズで使用するデータ全体を制御可能なプライバシ強化型機械学習フレームワーク SUP.ML を提案した．これは TTP などの信頼された機関の存在が不要な，次元削減，学習，テストを可能とする初めてのプライバシ強化型機械学習フレームワークである．

Acknowledgments 本研究の一部は文部科学省の平成 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．

References

[1] P. Xie, M. Bilenko, and e. Finley, "Crypto-nets: Neural networks over encrypted data," *arXiv preprint arXiv:1412.6181*, 2014.

[2] H. Hu, Z. Salcic, L. Sun, and e. Dobbie, "Membership inference attacks on machine learning: A survey," *ACM Computing Surveys (CSUR)*, 2021.

[3] C. Dwork, "Differential privacy," in *Proc. of ICALP 2006, LNCS*, vol. 4052, 2006, pp. 1–12.

[4] P. Kairouz, K. Bonawitz, and D. Ramage, "Discrete distribution estimation under local privacy," in *International Conference on Machine Learning*. PMLR, 2016, pp. 2436–2444.

[5] T. Wang, J. Blocki, and e. Li, "Locally differentially private protocols for frequency estimation," in *USENIX Security 17*, 2017, pp. 729–745.

[6] M. Gaboardi and R. Rogers, "Local private hypothesis testing: Chi-square tests," in *International Conference on Machine Learning*, 2018, pp. 1626–1635.

[7] B. Ding, H. Nori, and e. Li, "Comparing population means under local differential privacy: with significance and power," in *Proceedings of the AAAI*, vol. 32, no. 1, 2018.

[8] K. Wei, J. Li, M. Ding, C. Ma, H. H. Yang, F. Farokhi, S. Jin, T. Q. Quek, and H. V. Poor, "Federated learning with differential privacy: Algorithms and performance analysis," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 3454–3469, 2020.

[9] M. Yang, L. Lyu, and e. Zhao, "Local differential privacy and its applications: A comprehensive survey," *arXiv preprint arXiv:2008.03686*, 2020.

[10] N. Holohan, D. J. Leith, and O. Mason, "Optimal differentially private mechanisms for randomised response," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 11, pp. 2726–2735, 2017.

[11] N. Wang, X. Xiao, and e. Yang, "Collecting and analyzing multidimensional data with local differential privacy," in *IEEE ICDE*, 2019, pp. 638–649.

[12] F. Pedregosa, G. Varoquaux, and e. Gramfort, A., "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[13] J. C. Duchi and e. Jordan, "Local privacy and statistical minimax rates," in *54th Annual Symposium on Foundations of Computer Science*. IEEE, 2013, pp. 429–438.

[14] P. Kairouz, S. Oh, and P. Viswanath, "Extremal mechanisms for local differential privacy," in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, Eds. Curran Associates, Inc.

[15] B. I. Rubinstein, P. L. Bartlett, and e. Huang, "Learning in a large function space: Privacy-preserving mechanisms for svm learning," *arXiv preprint arXiv:0911.5708*, 2009.

[16] "Breast cancer wisconsin (diagnostic) data set," UCI Machine Learning Repository https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(diagnostic).

[17] "Ionosphere data set," UCI Machine Learning Repository https://archive.ics.uci.edu/ml/datasets/ionosphere.