プライバシ保護ディープラーニングのためのニューラルネットワークモデル構築手法の提案

代表研究者　清雄一　電気通信大学大学院情報理工学研究科　助教
共同研究者　大須賀昭彦　電気通信大学大学院情報理工学研究科　教授

1 はじめに

多数の個人に関するデータを保有していれば，年齢，性別，職業や身体の症状等の情報から，年収や罹患している病名等のセンシティブな情報を推測する機械学習モデルを構築することができる．さらにこのモデルを第三者に提供することで，受領者はある人物の年齢，性別，職業や身体の症状等の情報を持っていれば，その人が罹患している病名等を高精度に推測することが可能となる．これは病気の診断等に非常に有益である．しかしモデル構築には個人のデータが利用されており，個人の同意なく第三者にモデルを提供することには注意が必要である．

近年，プライバシを保護したままデータベースを共有するための匿名化手法が盛んに研究されている．匿名化の指標として様々なものが提案されているが「差分プライバシ」と呼ばれる指標が最も有望視されている．しかし，差分プライバシに限らず，プライバシを保護したまま機械学習を適用するための手法はこれまで開発されていない．特に，ニューラルネットワークの一形態である「ディープラーニング」は最も注目を浴びている機械学習手法の一つであり，これに対応することが望まれる．

本研究では，差分プライバシを厳密に保証したままニューラルネットワークモデルを構築する手法を開発する．これにより，ニューラルネットワーク（ディープラーニング含む）の機械学習モデルを第三者に提供するシナリオにおいて，モデル作成に用いられた情報のプライバシを保護できるようになる．

2 背景

2-1 システムモデル

多数の個人に関するデータを保有している事業者がディープラーニング等のニューラルネットワークモデルを構築してそれを第三者に提供するシナリオを想定し，差分プライバシを厳密に満たすような機械学習モデル構築アルゴリズムを開発することを研究の目的とする．システムモデルを図 1 に示す．

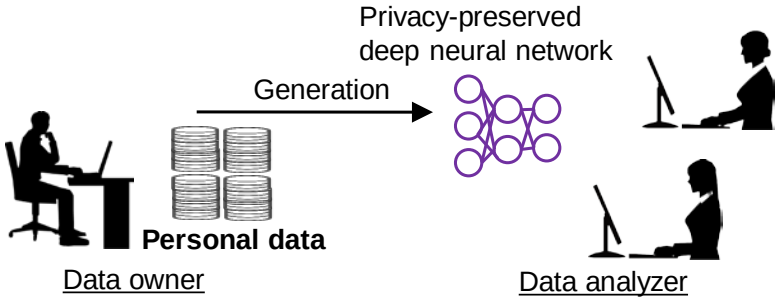


図 1 システムモデル

2-2 ディープニューラルネットワーク

図 2 にディープニューラルネットワーク（DNN: Deep Neural Network）の構造を示す．
$L^{(l)}$は DNN のl番目の層を表す．全体で$\mathcal{L}+1$個の層があり，入力層は$L^{(0)}$，出力層は$L^{(\mathcal{L})}$である．

$N_i^{(l)}$は層$L^{(l)}$のi番目のノードを表し，$n^{(l)}$は層$L^{(l)}$におけるノードの個数を表す．層$L^{(l)}$にはノード$N_1^{(l)}$，$N_2^{(l)}$，…，$N_{n^{(l)}}^{(l)}$がある．

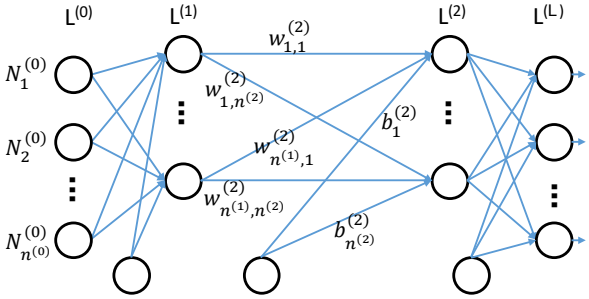


図 2 ディープニューラルネットワーク（$\mathcal{L}=3$）

$w_{ij}^{(l)}$はノード$N_i^{(l-1)}$とノード$N_j^{(l)}$の間の重みパラメータを表し，$b_j^{(l)}$はノード$N_j^{(l)}$へのバイアスパラメータを表す．

$F^{(l)}$は層$L^{(l)}$の活性化関数を表す．$x_i^{(l)}$はノード$N_i^{(l)}$への入力を表し，$y_i^{(l)}$はノード$N_i^{(l)}$からの出力を表す．これら入出力の値は以下の式で計算される：

$$\begin{aligned} x_i^{(l)} &= \sum_{j=1}^{n^{(l-1)}} y_j^{(l-1)} w_{ji}^{(l)} + b_i^{(l)} \\ y_i^{(l)} &= F^{(l)}(x_i^{(l)}). \end{aligned} \tag{1}$$

t_iはノード$N_i^{(\mathcal{L})}$の目標出力値を表し，Mは誤差関数を表す．誤差関数Mは入力として$y_i^{(\mathcal{L})}$及びt_iを取り，その誤差の値を返す．

学習データは，いくつかのバッチと呼ばれるまとまりに分割される．以下のプロセスは各バッチに対して行われる．

バッチ内の各レコードに対して，DNN は$y_i^{(\mathcal{L})}$を計算する（$i=1,\dots,n^{(\mathcal{L})}$）．次に，DNN は各ノード$N_i^{(l)}$における誤差信号（$\delta_i^{(l)}$とおく）を計算する．$l=\mathcal{L}$のとき，$\delta_i^{(\mathcal{L})}$は以下のように計算される：

$$\delta_i^{(\mathcal{L})} = \sum_{k=1}^{n^{(\mathcal{L})}} \frac{\partial M}{\partial y_k^{(\mathcal{L})}} \frac{\partial y_k^{(\mathcal{L})}}{\partial x_i^{(\mathcal{L})}} \tag{2}$$

$l=1,\dots,\mathcal{L}-1$に対しては，$\delta_i^{(l)}$は以下のように計算される：

$$\delta_i^{(l)} = \frac{\partial F^{(l)}}{\partial x_i^{(l)}} \sum_{j=1}^{n^{(l+1)}} w_{jk}\, \delta_j^{(l+1)}. \tag{3}$$

DNN は$\delta_i^{(l)}$をバッチ内の各レコードに対して計算し，その総和を新たに$\delta_i^{(l)}$とおく．

次に，変動量$\Delta w_{ij}^{(l)}$を以下のように定義する：

$$\Delta w_{ij}^{(l)} = y_i^{(l-1)} \delta_j^{(l)}. \tag{4}$$

最後に，DNN は各重みパラメータ$w_{ij}^{(l)}$ for $l=1,\dots,\mathcal{L}$，$i=1,\dots,n^{(l-1)}$，and $j=1,\dots,n^{(l)}$を以下のように更

新する：

$$w_{ij}^{(l)} \leftarrow w_{ij}^{(l)} - \alpha(\Delta w_{ij}^{(l)} + \lambda w_{ij}^{(l)})/B, \tag{5}$$

ここで，αは学習率，λは正則項を表し，これは事前に決定しておく．

バイアスパラメータに関しては，以下のように更新する：

$$b_j^{(l)} \leftarrow b_j^{(l)} - \alpha \Delta b_j^{(l)}/B, \tag{6}$$

ここで，$\Delta b_j^{(l)} = \delta_j^{(l)}$である．

このプロセスを全てのバッチに対して行う．

また，上記プロセスを複数回繰り返す．この繰り返し回数をエポック数と呼ぶ．このエポック数は，事前に，または，学習を進めながら決定する必要がある．

2-3 差分プライバシ

プライバシ保護データマイニングの研究分野では，k-anonymity[12]やl-diversity[13]と呼ばれるプライバシ保護指標が提案されている．これらの指標を拡張した指標も数多く提案されている[14,15]．

しかし近年では，ϵ-差分プライバシ[11]というプライバシ指標が最も注目を浴びている．

パラメータϵに基づき以下のように定義される：

Definition 1. *ϵ-differential privacy* DとD′は最大1レコードだけ異なるデータベースであるとする．ランダム機構$\mathcal{A}$は，出力の全ての集合Yについて以下が成り立つとき，またそのときのみ，ϵ-差分プライバシを実現する：

$$P(\mathcal{A}(D) \in Y) \leq e^{\epsilon} P(\mathcal{A}(D') \in Y) \quad \text{for all} \quad D, D' \tag{7}$$

Dwork ら[16]は， Laplace mechanism と呼ばれる， Laplace distribution に基づくノイズを与えることでϵ-差分プライバシを実現する手法を提案している． Laplace mechanism を説明するために，まず global sensitivity という概念を説明する．

Definition 2. *Global sensitivity* DとD′を，1レコードだけ異なるデータベースであると考える．$\mathcal{D}$を，入力のデータベースとして理論上可能性のある全てのデータベースの集合であるとする．fを，$f: \mathcal{D} \rightarrow \mathbb{R}$である関数とする．全てのD及びD′に対して以下が成立するとき，

$$\Delta f = \max_{D,D'} ||f(D) - f(D')||_1, \tag{8}$$

Δfをfの global sensitivity であると定義する．

ここで，ラプラスメカニズムと呼ばれる，ϵ-差分プライバシを満たすメカニズムを紹介する．

Theorem 1. *Laplace Mechanism* Lap(v)を，平均0，スケールがvであるラプラス分布に基づいてランダムな誤差を出力する関数であるとする．ある関数fに対して，ランダムメカニズム$\mathcal{A}$が，f(D) + Lap(Δf/ϵ)を出力するとき，$\mathcal{A}$はϵ-差分プライバシを満たす．

ϵ-差分プライバシは様々な領域において適用されている [7,8,17]．

3 関連研究

3-1 プライバシ保護 DNN

分散したデータベース内の情報を保護しつつ，中央サーバで DNN を生成する手法が提案されている．[18,19,20] ．生成した DNN を第三者と共有する場合には，1 章で述べたような問題が生じる恐れがある．

Abadi ら[21]は本研究と類似の目標を持っている．しかし，ϵ-差分プライバシを対象とはしておらず，これを緩和した (ϵ, δ)-差分プライバシ [22]を対象としている．

3-2 差分プライバシを満たすデータベース公開

差分プライバシを満たすようにデータベースを匿名化し，匿名化した結果を公開する手法が数多く提案さ

れている[23, 8, 25, 26, 27, 7, 28]．これらの研究成果は，後述する AnonymizingFirst の第一ステップで活用することが可能である．

4 提案手法

4-1 パラメータベース差分プライバシ

深層学習の重みパラメータやバイアスパラメータは複数存在する．これらパラメータの集合に対してϵ-差分プライバシを満たすこともできるが，個々のパラメータに対してϵ-差分プライバシを満たすようにすることもできる．本論文では後者をパラメータベースϵ-差分プライバシと呼び，パラメータベースϵ-差分プライバシを対象とする．

Theorem 2. パラメータベースϵ-差分プライバシ DとD′は最大1レコードだけ異なるデータベースであるとする．ランダム機構$\mathcal{A}$は，各パラメータにおける出力の全ての集合Yについて以下が成り立つとき，またそのときのみ，パラメータベースϵ-差分プライバシを実現する：

$$P(\mathcal{A}(D) \in Y) \leq e^{\epsilon} P(\mathcal{A}(D') \in Y) \quad \text{for all} \quad D, D'. \tag{9}$$

4-2 3 つのアプローチ

図 3 に示すように 3 つのアプローチを提案する．

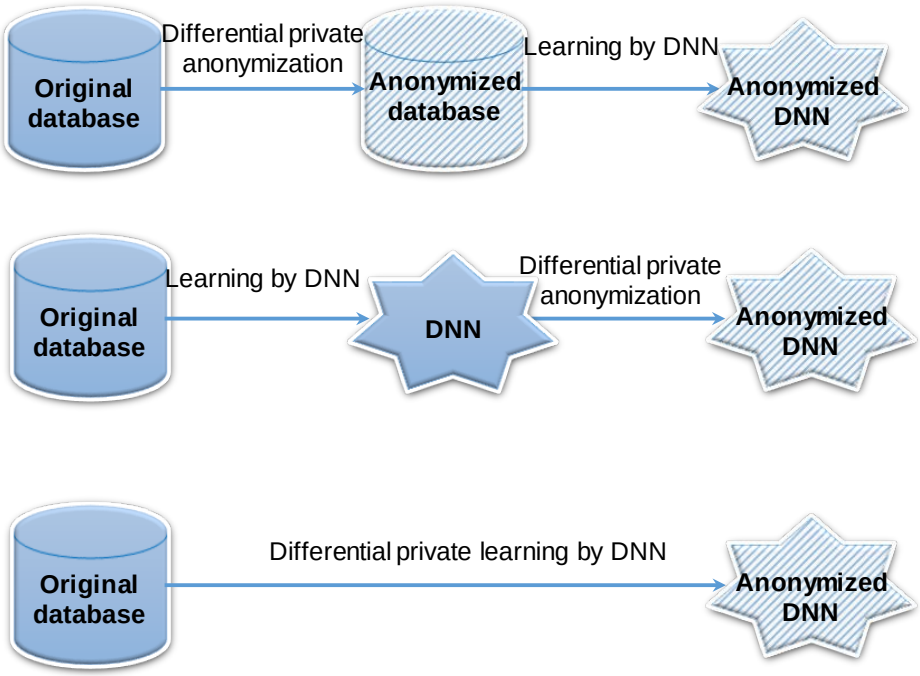


図 3 3 つのアプローチ

i) AnonymizingFirst: まず生データを差分プライバシに基づいて匿名化する．匿名化した結果のデータ（匿名データ）に対して通常の機械学習を行う．

ii) LearningFirst: まず生データに対して通常の機械学習を行い，生モデルを生成する．生モデルを差分プライバシに基づいて匿名化する．

iii) AnonymizedLearning: 生データに対して，差分プライバシに基づく匿名化を行いながら機械学習を行う．

(1) AnonymizingFirst

ヒストグラム匿名化手法[7, 28]を応用して匿名化を行い，その結果に対してニューラルネットワークによる学習を行うアルゴリズムを開発する．本アプローチのアルゴリズムは，どのような機械学習に対しても共通に利用できる．

(2) LearningFirst

まず，重みパラメータ$w_{ij}^{(l)}$とバイアスパラメータ$b_j^{(l)}$に対して値の閾値を設定する．これは，global sensitivity（1 レコードだけ異なるときに変わりうる値の，理論上の最大値）を減少させることで，パラメータに与える誤差を減少させ，これにより，深層学習モデルの精度低下を軽減するためである．w_{max}とw_{min}は，$w_{ij}^{(l)}$の最大値，及び，最小値を表すものとする．また，b_{max}とb_{min}は，$b_j^{(l)}$の最大値と最小値を表すものとする．

また，同様に，DNN への入力値（学習データ）にも閾値を設定する．本論文では[0,1]とする．

深層学習後，学習済重みパラメータに対して誤差を与える．つまり，全てのi，j，and lに対して，$w_{ij}^{(l)} + Lap((w_{max} - w_{min})/\epsilon)$を計算する．この計算結果を$r_{ij}^{(l)}$ とおく．もし$r_{ij}^{(l)}$の値がw_{max}を超えた場合，$w_{ij}^{(l)}$の値をw_{max}に設定する．同様にもし$r_{ij}^{(l)}$の値がw_{min}を下回った場合，$w_{ij}^{(l)}$の値をw_{min}に修正する．いずれにも当てはまらない場合，$w_{ij}^{(l)}$の値を$r_{ij}^{(l)}$に設定する．同様のことをバイアスパラメータに対しても行う．つまり，$b_j^{(l)}$の値を$\min(b_{max}, \max(b_{min}, b_j^{(l)} + Lap((b_{max} - b_{min})/\epsilon))$に設定する．

以下の定理が成り立つ：

Theorem 3. LearningFirstで得られたパラメータは，パラメータベースϵ-差分プライバシを満たす．

Proof. 全てのi，j，lについて，$w_{ij}^{(l)}$の global sensitivity は$(w_{max} - w_{min})$であり，また，全てのi，jについて，$b_j^{(l)}$の global sensitivity は$(b_{max} - b_{min})$である．従って，Lemma 1 より，深層学習で得られた学習済みの重みパラメータ$w_{ij}^{(l)}$の値を$\min(w_{max}, \max(w_{min}, w_{ij}^{(l)} + Lap((w_{max} - w_{min})/\epsilon))$に設定し，また，学習済みのバイアスパラメータ$b_j^{(l)}$の値を$\min(b_{max}, \max(b_{min}, b_j^{(l)} + Lap((b_{max} - b_{min})/\epsilon))$に設定することで，パラメータベース$\epsilon$-差分プライバシが満たされる． □

Lemma 1. ランダムメカニズム$\mathcal{A}$が$\min(f_{min}, (\max(f_{max}, f(D) + Lap(\Delta f/\epsilon)))$を出力するとき，$\mathcal{A}$は$\epsilon$-差分プライバシを実現する．ここで，$f_{max}$及び$f_{min}$は$f(D)$が取り得る理論上の最大値と最小値である．

Proof. 1 レコードだけ異なるデータベースをDとD'とおく．

また，$F(D) = f(D) + Lap(\Delta f/\epsilon)$とおく．$F(D)$の値が$[f_{min}, f_{max}]$の範囲に入るとき，定理 1 より，式 7 が成立する．

次に，$F(D)$の値がf_{min}を下回る場合を考える．このとき，$\mathcal{A}(D)$の出力値はf_{min}となる．$\mathcal{A}(D)$の出力がf_{min}となる確率は以下の式で表される：

$$\int_{t=-\infty}^{f_{min}-f(D)} \mathfrak{Lap}(\Delta f/\epsilon, t) = \frac{1}{2}\exp\frac{\epsilon}{\Delta f}(f_{min} - f(D)), \tag{10}$$

ここで，$\mathfrak{Lap}(v, u)$は，スケールパラメータがvであり，平均との差がuである，ラプラス分布の確率密度関数の値を表す．

同様に，$\mathcal{A}(D')$の出力値がf_{min}となる確率は以下の式で表される：

$$\int_{t=-\infty}^{f_{min}-f(D')} \mathfrak{Lap}(\Delta f/\epsilon, t) = \frac{1}{2}\exp\frac{\epsilon}{\Delta f}(f_{min} - f(D')). \tag{11}$$

式 10 と式 11 の比は最大で，

$$\exp \frac{\epsilon|f(D)-f(D')|}{\Delta f}. \tag{12}$$

となる．

$|f(D)-f(D')| \leq \Delta f$であるから，式 12 の値は$\exp\ \epsilon$以下である．

次に，$F(D)$の値がt_{max}以上となる場合を考える．このとき，$\mathcal{A}(D)$の出力値はf_{max}となる．$\mathcal{A}(D)$の出力がf_{max}となる確率は以下の式で表される：

$$\int_{t=f(D)-f_{max}}^{\infty} \mathfrak{Q}\,\mathfrak{ap}(\Delta f/\epsilon, t) = \frac{1}{2}\exp\left(-\frac{\epsilon}{\Delta f}(f(D)-t_{max})\right). \tag{13}$$

同様に，$\mathcal{A}(D')$の出力値がf_{max}となる確率は以下の式で表される：

$$\int_{t=f(D')-f_{max}}^{\infty} \mathfrak{Q}\,\mathfrak{ap}(\Delta f/\epsilon, t) = \frac{1}{2}\exp\left(-\frac{\epsilon}{\Delta f}(f(D')-t_{max})\right) \tag{14}$$

式 13 と式 14 の比は最大で，

$$\exp \frac{\epsilon|f(D)-f(D')|}{\Delta f}. \tag{15}$$

となる．

$|f(D)-f(D')| \leq \Delta f$であるから，式 15 の値は$\exp\ \epsilon$以下である．

この議論は全てのパラメータに対して成立する．したがって，パラメータベースϵ-差分プライバシが成り立つ． □

(3) AnonymizedLearning

本論文では，活性化関数と誤差関数を事前に決めて，AnonymizedLearning を行う．

$f(x) = \max(0, x)$で定義される ReLU が，深層学習の，最終層を除く活性化関数として広く利用されている[29]．

深層学習の利用目的として，カテゴリ分類（例：将来破産しそうか否か，将来の借金額が 100 万円以下・100－200 万・200-300 万・それ以上）の場合，最終層の活性化関数（$F^{(\mathcal{L})}$）としてソフトマックス関数が，また，誤差関数としてクロスエントロピー誤差関数が広く利用されている（[30,31]）．

ソフトマックス関数は以下のように定義される：

$$F^{(\mathcal{L})}((x_1^{(\mathcal{L})}, \ldots, x_{n^{(\mathcal{L})}}^{(\mathcal{L})}), j) = \frac{e^{x_j^{(\mathcal{L})}}}{\sum_{k=1}^{n^{(\mathcal{L})}} e^{x_k^{(\mathcal{L})}}}, \tag{16}$$

また，クロスエントロピー誤差関数は以下のように定義される：

$$M(t_1, \ldots, t_{n^{(\mathcal{L})}}, y_1, \ldots, y_{n^{(\mathcal{L})}}) = -\sum_{i=1}^{n^{(\mathcal{L})}} t_i \ln\ y_i^{(\mathcal{L})}. \tag{17}$$

本論文では，AnonyizedLearning を行う場合，最終層を除く層は活性化関数として ReLU を，最終層の活性化関数としてソフトマックス関数を，誤差関数としてクロスエントロピー誤差関数を利用することを想定する．

最終層の活性化関数がソフトマックス関数であり，誤差関数がクロスエントロピー誤差関数の場合，誤差信号$\delta_j^{(\mathcal{L})}$ for $j = 1, \ldots, n^{(\mathcal{L})}$の値は次のように計算される：

$$\delta_j^{(\mathcal{L})} = y_j^{(\mathcal{L})} - t_j^{(\mathcal{L})}, \tag{18}$$

ここで$y_j^{(\mathcal{L})}$はノード$N_j^{(\mathcal{L})}$の出力値を表し，$t_j^{(\mathcal{L})}$はノード$N_j^{(\mathcal{L})}$の目標出力値を表す．

最終層以外の層において活性化関数として ReLU を使っている場合，最終層以外の各ノードの誤差信号$\delta_j^{(l)}$

for $l = 1, \dots, \mathfrak{L} - 1$は次のように計算される：

$$\delta_j^{(l)} = \begin{cases} \sum_{k=1}^{n^{(l+1)}} w_{jk}\, \delta_k^{(l+1)} & (x_j^{(l)} > 0) \\ 0 & (otherwise.) \end{cases} \tag{19}$$

$x_j^{(1)}$の値として取り得る範囲は，$[b_j^{(1)} + \sum_i \min(w_{i,j}^{(1)}, 0), b_j^{(1)} + \sum_i \max(w_{i,j}^{(1)}, 0)]$である．また，$x_j^{(2)}$の値として取り得る範囲は，$[b_j^{(2)} + \sum_i (b_i^{(1)} + \sum_k \max(w_{k,i}^{(1)}, 0)) \min(w_{i,j}^{(2)}, 0), b_j^{(2)} + \sum_i (b_i^{(1)} + \sum_k \max(w_{k,i}^{(1)}, 0)) \max(w_{i,j}^{(2)}, 0)]$.

深層学習では，$x_j^{(l)}$ for $l = 1, \dots, \mathcal{L}$は次のように計算される：

$$\begin{cases} \min(x_j^{(l)}) = b_j^{(l)} + \sum_{i=1}^{n^{(l-1)}} \max(y_i^{(l-1)}) \min(w_{i,j}^{(l)}, 0) \\ \max(x_j^{(l)}) = b_j^{(l)} + \sum_{i=1}^{n^{(l-1)}} \max(y_i^{(l-1)}) \max(w_{i,j}^{(l)}, 0). \end{cases} \tag{20}$$

ここで，$\min(y_i^{(0)}) = 0$であり，また，$\max(y_i^{(0)}) = 1$である．何故なら，深層学習の第 1 層目への入力値を 0 以上 1 以下の範囲に正規化しているためである．また，最終層以外の層の活性化関数として ReLU を使っているので，$l = 1, \dots, \mathcal{L} - 1$において，$y_j^{(l)}$は次のように計算される：

$$\begin{cases} \min(y_j^{(l)}) = \max(\min(x_j^{(l)}), 0) \\ \max(y_j^{(l)}) = \max(x_j^{(l)}). \end{cases} \tag{21}$$

$\max(y_j^{(l)})$の値は常に 0 以上であることがわかる．

次に誤差信号$\delta_j^{(l)}$の取り得る値の範囲を計算する．DNN の出力値の範囲は-1 から 1 までであるから，

$$\begin{cases} \min(\delta_j^{(\mathcal{L})}) = -1 \\ \max(\delta_j^{(\mathcal{L})}) = 1. \end{cases} \tag{22}$$

である．
$l = 1, \dots, \mathcal{L} - 1$について，

$$\begin{cases} \min(\delta_j^{(l)}) = \sum_{i=1}^{n^{(l+1)}} \begin{cases} w_{ji}^{(l+1)} \max(\delta_i^{(l+1)}) & w_{ji} < 0 \\ w_{ji}^{(l+1)} \min(\delta_i^{(l+1)}) & otherwise \end{cases} \\ \max(\delta_j^{(l)}) = \sum_{i=1}^{n^{(l+1)}} \begin{cases} w_{ji}^{(l+1)} \max(\delta_i^{(l+1)}) & w_{ji} > 0 \\ w_{ji}^{(l+1)} \min(\delta_i^{(l+1)}) & otherwise. \end{cases} \end{cases} \tag{23}$$

である．ここで，for all j and lについて，$\min(\delta_j^{(l)}) \leq 0$であり，また，$\max(\delta_j^{(l)}) \geq 0$である．

最終的に以下が得られる：

$$\begin{cases} \min(\Delta w_{ij}^{(l)}) = \max(y_i^{(l-1)}) \min(\delta_j^{(l)}) \\ \max(\Delta w_{ij}^{(l)}) = \max(y_i^{(l-1)}) \max(\delta_j^{(l)}). \end{cases} \tag{24}$$

$b_j^{(l)}$については，

$$\begin{cases}\min\left(\Delta b_j^{(\mathcal{L})}\right) = -1 \\ \max\left(\Delta b_j^{(\mathcal{L})}\right) = 1,\end{cases} \tag{25}$$

であり，また，$l = 1, \ldots, \mathcal{L}-1$について

$$\begin{cases}\min\left(\Delta b_j^{(l)}\right) = \min\left(\delta_j^{(l)}\right) \\ \max\left(\Delta b_j^{(l)}\right) = \max\left(\delta_j^{(l)}\right).\end{cases} \tag{26}$$

である．

前述のように，重みパラメータの変動量$\Delta w_{ij}^{(l)}$と，バイアスパラメータの変動量$\Delta b_j^{(l)}$に基づいて，重みパラメータとバイアスパラメータを式 5 と式 6 に基づいて更新する．AnonymizedLearning では，この変動量にラプラス分布に基づく誤差を与える．

重みパラメータの変動量$\Delta w_{ij}^{(l)}$と，バイアスパラメータの変動量$\Delta b_j^{(l)}$についても，global sensitivity を減少させるために値の閾値を設定する．Δw_{max}とΔw_{min}を，重みパラメータの変動量$\Delta w_{ij}^{(l)}$の最大値と最小値とする．また，Δb_{max}とΔb_{min}をバイアスパラメータ$\Delta b_j^{(l)}$の最大値と最小値とする．

DNN のエポック数をEとおく．各バッチに対して学習を行う際に，for each $w_{ij}^{(l)}$ and $b_j^{(l)}$に対して，重みパラメータの変動量$\Delta w_{ij}^{(l)}$を$\min\left(\Delta w_{max}, \left(\min\left(\Delta w_{max}, w_{ij}^{(l)} + Lap((\Delta w_{max} - \Delta w_{min}) \cdot E/\epsilon)\right)\right)\right.$に設定し，バイアスパラメータの変動量$\Delta b_j^{(l)}$を$\min\left(\Delta b_{max}, \left(\max\left(\Delta b_{min}, b_j^{(l)} + Lap((\Delta b_{max} - \Delta b_{min}) \cdot E/\epsilon)\right)\right)\right.$に設定する．

Theorem 4. AnonymizedLearningにより生成されたモデルはパラメータベースϵ-差分プライバシを満たす．

Proof. 各重みパラメータとバイアスパラメータは，式 5 と式 6 に基づいて更新される．式 5 および式 6 において，重みパラメータの変動量$\Delta w_{ij}^{(l)}$とバイアスパラメータの変動量$\Delta b_j^{(l)}$は学習の入力値に依存して変わるが，それ以外の値は入力値に依存しない．したがって，Lemma 1 より，$\Delta w_{ij}^{(l)}$を$\min\left(\Delta w_{max}, \left(\max\left(\Delta w_{min}, w_{ij}^{(l)} + Lap((\Delta w_{max} - \Delta w_{min}) \cdot E/\epsilon)\right)\right)\right.$に設定し，また，$\Delta b_j^{(l)}$を$min(\Delta b_{max}, \left(\max\left(\Delta b_{min}, b_j^{(l)} + Lap((\Delta b_{max} - \Delta b_{min}) \cdot E/\epsilon)\right)\right)\right.$に設定することで，各エポックのイテレーションは，パラメータベース(ϵ/E)-差分プライバシを満たす．

全体でEエポックあるので，Lemma 2 より，最終的にϵ-差分プライバシを満たす． □

Lemma 2. ランダムメカニズム$\mathcal{A}$が，d個のランダムメカニズム$\mathcal{A}_1, \ldots, \mathcal{A}_d$から成り立っており，これを1回ずつ続けて実施するものとする（$i \geq 2$において$\mathcal{A}_i$は入力として$\mathcal{A}_{i-1}$の出力値を取る．$\mathcal{A}_d$の出力値が，$\mathcal{A}$の出力値となる）．ここで，各$\mathcal{A}_i$はパラメータベースϵ_i-差分プライバシを満たすものとする．このとき，$\mathcal{A}$はパラメータベース$(\sum_{i=1}^{d} \epsilon_i)$-差分プライバシを実現する．

Proof. [32]より，ランダムメカニズム$\mathcal{A}$が，d個のランダムメカニズム$\mathcal{A}_1, \ldots, \mathcal{A}_d$から成り立っておりこれを 1 回ずつ続けて実施するものとする（$i \geq 2$において$\mathcal{A}_i$は入力として$\mathcal{A}_{i-1}$の出力値を取る．$\mathcal{A}_d$の出力値が，$\mathcal{A}$の出力値となる）．ここで，各$\mathcal{A}_i$はϵ_i-差分プライバシを満たすものとする．このとき，$\mathcal{A}$は$(\sum_{i=1}^{d} \epsilon_i)$-差分プライバシを実現する．

$\mathcal{A}$ in Lemma 2 は各パラメータに対して実行されるので，Lemma 2 における$\mathcal{A}$はパラメータベース$(\sum_{i=1}^{d} \epsilon_i)$-差分プライバシを実現する． □

5 評価

5-1 データセットとネットワーク構造

評価用のデータセットとして，プライバシ保護データマイニングの分野で広く利用されている Adult data set [33]を利用する．Adult data set は 15 属性(e.g.，Age，Sex，Race，Salary)から構成されており，欠損値を含むレコードを除外して，45,222 レコードから成る．属性 Salary は，各レコードの人物の年収が 50K ドルを超えているか（50K 以下 or 50K より多い）どうかの 2 値を取る．

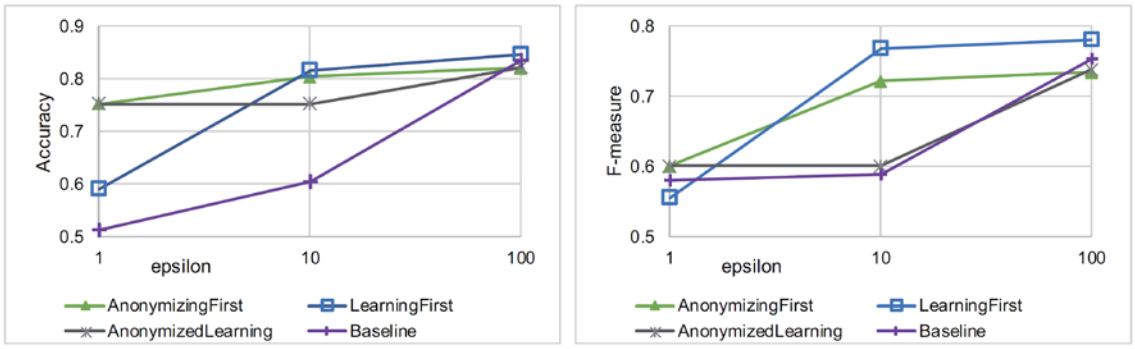
評価として，Salary を除く 14 属性から，Salary が 50K ドルを超えるかどうかを予測する，DNN を構築する．

差分プライバシを満たすような匿名化を行わない，生データに対して事前実験を行い，DNN の精度が高くなるような DNN の構造を決定した．学習率は 0.01，バッチサイズは 50，エポック数は 500，正則項は 0.001，中間層の数は 4（入力層，出力層を含めると，全部で 5 層）が良い結果を出した．

10 分割交差検定を行って精度を計測した．精度を計測する指標として，accuracy と f-measure を用いた(どちらも 0 から 1 までの値を取り，1 に近いほど精度が高い)．この，生データに対して実施した結果では，accuracy は 0.85，f-measure は 0.79 となった．

5-2 結果

3 アプローチの比較結果を図 5 に示す．



(a) Accuracy　　(b) F-measure

図 5 3 アプローチの比較

3 アプローチとも概ねベースラインを上回る結果となった．

6 おわりに

近年，プライバシを保護したままデータベースを共有するための匿名化手法が盛んに研究されており，「差分プライバシ」と呼ばれるプライバシ保護指標が最も有望視されている．しかし，差分プライバシに限らず，プライバシを保護したまま機械学習を適用するための手法はこれまでほとんど開発されていない．特に，深層学習は最も注目を浴びている機械学習手法の一つであり，これに対応することが望まれる．本研究では，差分プライバシを厳密に保証したまま深層学習モデルを構築する手法を提案した．これにより，深層学習モデルを第三者に提供するシナリオにおいて，モデル作成に用いられた情報のプライバシを保護できるようになるベースライン手法と比較して，同レベルのプライバシ保護レベルにおいて，精度を向上させることを，実データを用いたシミュレーション評価において確認した．

【参考文献】

[1] H.-I. Suk and D. Shen, “Deep Learning-Based Feature Representation for AD/MCI Classification,”

in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2013, pp. 583–590.

[2] J. Fombellida, S. Torres-Alegre, J. Pinuela-Izquierdo, and D. Andina, "Artificial Metaplasticity for Deep Learning: Application to WBCD Breast Cancer Database Classification," in *Bioinspired Computation in Artificial Systems*.Springer International Publishing, 2015, pp. 399–408.

[3] R. S. P. Huang, E. Nedelcu, Y. Bai, A. Wahed, K. Klein, H. Tint, I. Gregoric, M. Patel, B. Kar, P. Loyalka, S. Nathan, R. Radovancevic, and A. N. Nguyen, "Post-Operative Bleeding Risk Stratification in Cardiac Pulmonary Bypass Patients Using Artificial Neural Network," *Annals of Clinical & Laboratory Science*, vol. 45, no. 2, pp. 181–186, 2015.

[4] D. D. Wu, D. L. Olson, and C. Cuicui Luo, "A Decision Support Approach for Accounts Receivable Risk Management," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 44, no. 12, pp. 1624–1632, 2014.

[5] I. S. Duma, B. Twala, and T. Marwala, "Predictive modeling for default risk using a multilayered feedforward neural network with Bayesian regularization," in *Proc. International Joint Conference on Neural Networks (IJCNN)*, 2013, pp. 1–10.

[6] J. Yi, J. Wang, and R. Jin, "Privacy and Regression Model Preserved Learning," in *Proc. AAAI*, 2015, pp. 1341–1347.

[7] J. Soria-Comas, J. Domingo-Ferrer, D. Sánchez, and S. Martínez, "Enhancing data utility in differential privacy via microaggregation-based k-anonymity," *The VLDB Journal*, vol. 23, no. 5, pp. 771–794, 2014.

[8] J. Xu, Z. Zhang, X. Xiao, Y. Yang, G. Yu, and M. Winslett, "Differentially private histogram publication," *VLDB Endowment*, vol. 22, no. 6, pp. 797–822, 2013.

[9] M. Terrovitis, N. Mamoulis, and P. Kalnis, "Local and global recoding methods for anonymizing set-valued data," *The VLDB Journal*, vol. 20, no. 1, pp. 83–106, 2011.

[10] M. Xue, P. Karras, C. Raïssi, J. Vaidya, and K.-L. Tan, "Anonymizing set-valued data by nonreciprocal recoding," in *Proc. ACM KDD*, 2012, pp. 1050–1058.

[11] C. Dwork, "Differential Privacy," in *Automata, Languages and Programming*, 2006, vol. 4052, pp. 1–12.

[12] P. Samarati, "Protecting respondents' identities in microdata release," *IEEE Trans. Knowl. Data Eng.*, vol. 13, no. 6, pp. 1010–1027, 2001.

[13] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "l-diversity: Privacy Beyond k-Anonymity," in *Proc. IEEE ICDE*, 2006, pp. 24:1–24:12.

[14] F. Zhang, V. E. Lee, and R. Jin, "k-CoRating: filling up data to obtain privacy and utility," in *Proc. AAAI*, 2014, pp. 320–327.

[15] Y. Sei, T. Takenouchi, and A. Ohsuga, "(l1, ..., lq)-diversity for Anonymizing Sensitive Quasi-Identifiers," in *Proc. IEEE TrustCom*, 2015, pp. 596–603.

[16] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating Noise to Sensitivity in Private Data Analysis," in *Proc. Theory of Cryptography (TCC)*, 2006, pp. 265–284.

[17] C. Task and C. Clifton, "A Guide to Differential Privacy Theory in Social Network Analysis," in *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2012, pp. 411–417.

[18] J. Jiawei Yuan and S. Shucheng Yu, "Privacy Preserving Back-Propagation Neural Network Learning Made Practical with Cloud Computing," *IEEE Transactions on Parallel and Distributed Systems*, vol. 25, no. 1, pp. 212–221, 2014.

[19] T. Tingting Chen and S. Sheng Zhong, "Privacy-Preserving Backpropagation Neural Network Learning," *IEEE Transactions on Neural Networks*, vol. 20, no. 10, pp. 1554–1564, 2009.

[20] Y. Kokkinos and K. G. Margaritis, "Confidence ratio affinity propagation in ensemble selection of neural network classifiers for distributed privacy-preserving data mining," *Neurocomputing*, vol. 150, pp. 513–528, 2015.

[21] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep Learning with Differential Privacy," *arXiv*, vol. 1607.00133, no. 1, 2016.

[22] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor, "Our data, ourselves: privacy via distributed noise generation," in *Proc. Eurocrypt*, vol. 4004, 2006, pp. 486–503.

[23] G. Acs, C. Castelluccia, and R. Chen, "Differentially Private Histogram Publishing through Lossy Compression," in *Proc. IEEE ICDM*, dec 2012, pp. 1–10.

[24] B. C. M. F. Rui Chen, Bipin C. Desai, Noman Mohammed, Li Xiong, "Publishing set-valued data via differential privacy," *Proc. VLDB*, vol. 4, no. 11, pp. 1087–1098, 2011.

[25] J. Zhang, G. Cormode, C. M. Procopiuc, D. Srivastava, and X. Xiao, "PrivBayes: private data release via bayesian networks," in *Proc. ACM SIGMOD*, 2014, pp. 1423–1434.

[26] W. Qardaji, W. Yang, and N. Li, "PriView: practical differentially private release of marginal contingency tables," in *Proc. ACM SIGMOD*, 2014, pp. 1435–1446.

[27] R. Chen, Q. Xiao, Y. Zhang, and J. Xu, "Differentially Private High-Dimensional Data Publication via Sampling-Based Inference," in *Proc. ACM KDD*, 2015, pp. 129–138.

[28] D. Sánchez, J. Domingo-Ferrer, S. Martínez, and J. Soria-Comas, "Utility-preserving differentially private data releases via individual ranking microaggregation," *Information Fusion*, vol. 30, pp. 1–14, 2016.

[29] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[30] D. Yu, L. Deng, and F. Seide, "The Deep Tensor Neural Network With Applications to Large Vocabulary Speech Recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 2, pp. 388–396, 2013.

[31] S. Shaofei Xue, O. Abdel-Hamid, H. Hui Jiang, L. Lirong Dai, and Q. Qingfeng Liu, "Fast Adaptation of Deep Neural Network Based on Discriminant Codes for Speech Recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, 2014.

[32] S. P. Kasiviswanathan, H. K. Lee, K. Nissim, S. Raskhodnikova, and A. Smith, "What Can We Learn Privately ?" *SIAM Journal on Computing*, vol. 40, no. 3, pp. 793–826, 2013.

[33] UCI Machine Learning Repository, "Adult Data Set, http://archive.ics.uci.edu/ml/datasets/Adult."

〈発　表　資　料〉

題　　名	掲載誌・学会名等	発表年月
Privacy-Preserving Publication of Deep Neural Networks	Proc. IEEE International Conference on Data Science and Systems	2016 年 12 月
Differential Private Data Collection and Analysis Based on Randomized Multiple Dummies for Untrusted Mobile Crowdsensing	IEEE Transactions on Information Forensics and Security	2017 年 4 月