差分プライバシを満たすニューラルネットワークモデル構築手法の提案

清 雄一　大須賀 昭彦

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属性値から，年収や罹患している病名等のセンシティブ属性値を推測するニューラルネットワークモデルを構築することができる．さらにこの構築されたモデルを第三者に提供することで，受領者は研究やマーケティング等にモデルを活用することが可能となる．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，データベースを匿名化して第三者に提供するシナリオにおいて，ϵ-差分プライバシと呼ばれる匿名化指標が盛んに研究されている．本研究では，ϵ-差分プライバシを満たすニューラルネットワークモデルを構築する手法を提案する．

1　はじめに

近年，ニューラルネットワーク等の機械学習が注目を集めている．大規模データベースに対して機械学習を行うことにより，データに潜在的に存在する構造を抽出したり，未知の入力に対して出力を推測したりすることができる．

機械学習を行った結果**モデル**が得られ，そのモデルを用いて新しい入力に対応する出力を推測することができる．学習した結果得られるモデルは第三者と共有することもできる．たとえば，複数の基本的検査の結果から甲状腺機能異常を発見するモデル構築 [12] や，下顎前突患者の側貌パターンを入力として述語の軟組織側貌を予測するモデル構築 [11] 等，ニューラルネットワークを用いたモデル構築の研究も盛んに行われている．しかしモデル構築には個人のデータが利用されているため，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 [10]．

データベースを匿名化して第三者に提供し，第三者がその匿名化されたデータベースを用いてデータマイニングを行うことを，プライバシ保護データマイニングと呼び，近年盛んに研究が行われている．しかし，これまで対象とされているデータマイニングとはクロス集計分析 [4] [8] やある属性値を持つユーザ数の分析 [9] といった基本的な分析が対象であり，ニューラルネットワーク等の機械学習を行ってプライバシ保護データマイニングを行う研究はほとんど存在しない．

本研究では，プライバシ保護データマイニングの分野で最も研究されている **ϵ-差分プライバシ**をプライバシ指標として用いる．この指標はプライバシパラメータ ϵ を持ち，この値を調整することでプライバシ保護レベルを調整する．本研究では ϵ-差分プライバシを満たした上でニュラルネットワークモデルを構築する手法を検討する．

本論文の構成を示す．2 章で関連研究について述べる．3 章で提案手法について述べ，4 章で評価を行う．5 章で本論文をまとめる．

2　関連研究

2.1　差分プライバシ

近年，匿名化の分野で広く利用されている指標に ϵ-差分プライバシ（ϵ-Differential Privacy）[1] がある．データ解析者がデータ保有者にクエリを投げ，そのク

Construction of Anonymized Neural Network Model with Differential Privacy.

Yuichi Sei and Akihiko Ohsuga, 電気通信大学大学院情報システム学研究科, Graduate School of Information Systems, The University of Electro-Communications.

エリの回答に誤差を与えることで，データベースに格納されている個人のプライバシを保護する．直感的には，ある一人のデータがあってもなくても匿名化の結果がほとんど変わらないことを要請する．より詳細には，プライバシパラメータ ϵ を使って以下のとおり定義される．

定義 2.1 (ϵ-差分プライバシ) D と D' を高々*1* レコードだけ異なるデータベースとする．確率的メカニズム $\mathcal{A}$ が，出力のすべての集合 Y について以下を満たすとき，ϵ-差分プライバシが満たされる．

$$P(\mathcal{A}(D) \in Y) \leq e^{\epsilon} P(\mathcal{A}(D') \in Y) \quad \text{for all} \quad D, D' \quad (1)$$

[2] ではラプラス分布に従って生成されたノイズを加算することによって ϵ-差分プライバシを実現する手法が提案されており，多くの研究がこの Laplace mechanism を採用している．Laplace mechanism を説明するために，まず global sensitivity という概念を導入する．

定義 2.2 (global sensitivity) 高々*1* レコードのみ異なるデータベース D_1 及び D_2 を考える．d を非負整数とし，関数 $f : \mathcal{D} \rightarrow \mathbb{R}^d$ について，全ての D_1 及び D_2 に対して，

$$\Delta f = \max_{D_1, D_2} ||f(D_1) - f(D_2)||_1 \quad (2)$$

が成り立つとき，この Δf を f の *global sensitivity* と定義する．

たとえば 1 レコードだけ異なるデータベース D_1 と D_2 に対して，ある属性値を持つレコード数を取得するクエリを投げた場合，その回答は最大 1 だけ異なる．したがって，Δf の値は 1 である．別の例として，ある属性値の最大値を取得するクエリを投げた場合を考える．その属性値が取り得る理論上の最小値を v_{min}，理論上の最大値を v_{max} とすると Δf=$v_{max} - v_{min}$ となる．何故なら，あるデータベース D_1 における当該属性の最大値が v_{min} であり，1 レコードだけ異なるデータベース D_2 における当該属性値の最大値が v_{max} である可能性があるためである．

定義 2.3 (Laplace Mechanism) データベース D に対して $f(D) + RandomVariate(Lap(\Delta f/\epsilon))$ を返すランダム機構 $\mathcal{R}$ は，ϵ-差分プライバシを実現する．

ここで $Lap(\Delta f/\epsilon)$ は，平均 0，尺度指数 σ が $\Delta f/\epsilon$ であるラプラス分布を表し，$RandomVariate$ は引数の確率分布から生成した乱数を返す関数である．

表 1 データ保有者が保有しているデータベースの例

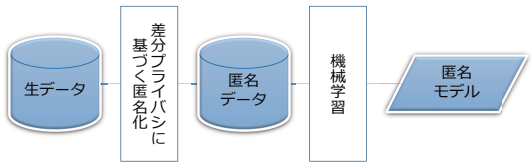
Name	Age	Job	Disease
Alex	35	Lawyer	Fever
Bob	38	Lawyer	Asthma
Carl	40	Lawyer	Hepatitis
Daniel	45	Doctor	Hepatitis
Edward	42	Doctor	Hepatitis

表 1 に示すデータベース D を使った例を示す．このデータベースに対して，[年齢が 30 代かつ職業が法律家である人数] というクエリを発行したとする．真の値は 2 である．この値を ϵ-差分プライバシで保護すると匿名化された値として $2 + RandomValue(Lap(1/\epsilon))$ が得られる．2 に限りなく近い値になる場合もあるし，マイナス値や真の値から数倍離れた値を取る場合もある．

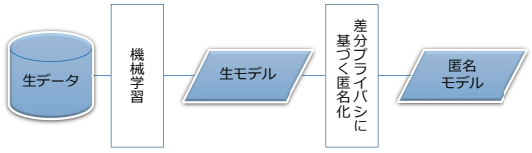
この ϵ-差分プライバシ（プライバシパラメータ ϵ の値に特に注目していない場合は，以下では単に「差分プライバシ」と呼ぶ）は，元々は上述のようにデータベースに対するクエリへの回答に対して提案されているものである．近年では差分プライバシがプライバシ保護データマイニングにおけるプライバシ指標として標準的になっており，ソーシャルネットワーク解析 [6] やヒストグラム共有 [4] [8] 等のように，あらゆるデータ共有のために利用され始めている．しかし，ニューラルネットワーク等の機械学習モデルに差分プライバシを適用させた例はほとんど無い．

2.2 プライバシを考慮した統計モデル構築

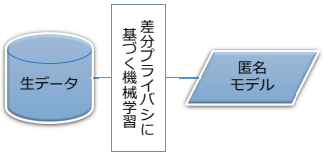
Yi らは本論文と同じ問題意識を持ち，プライバシを保護したまま回帰モデルを構築する手法を提案している [10]．彼らの手法は回帰モデルには有効であるが，ニューラルネットワークを対象としたものではない．また，プライバシ保護データマイニングの分野で標準的な指標となっている差分プライバシにも適用していない．



(a) AnonymizationFirst



(b) LearningFirst



(c) AnonymizedLearning

図 1　3 つの方針

3　提案手法

データ保有者が保有するデータベースを**生データ**，生データに基づいて機械学習を行い得られたモデルを**生モデル**，生データを差分プライバシに基づいて匿名化したデータベースを**匿名データ**，差分プライバシを満たすモデルを**匿名モデル**と呼ぶ．匿名モデルを得るための方針として，図 1 に示す 3 つの方針が考えられる．

- 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．
- 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．
- 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．

AnonymizingFirst 及び LearningFirst については，差分プライバシに基づく誤差の付与と機械学習は独立して行われる．前者は入力データに対して差分プライバシに基づく誤差を与え，後者は機械学習した結果のモデルに対して誤差を与える．一方で AnonymizedLearning は誤差を与えながらモデルを構築することになる．次節以降で，各手法の詳細を述べる．

3.1　AnonymizingFirst

Soria-Comas らの手法である eMDAV [4] を応用する．eMDAV ではまずデータベースに対して k-匿名化 [3] を行う．ここでの k-匿名化は，全ての属性値が同一となるレコードが少なくとも k レコード存在するように，データベース内の属性値を一般化することを言う．一般化とは，たとえば年齢の属性である場合，“33” 歳や “34 歳” を “30 代” とするように，複数の属性値を区別できないように一つにまとめる操作を言う．

各レコードに対して ϵ-差分プライバシに基づく誤差を与える場合は大きな誤差を与える必要があるが，k 個のまとまったレコードに対して誤差を与える場合は比較的小さい誤差を与えれば良いという，差分プライバシの特徴を活かした手法である．

この eMDAV を入力データに対して適用させることで，ϵ-差分プライバシを満たした匿名データが得られる．この匿名データに対して通常の機械学習を行うことで，ϵ-差分プライバシを満たす機械学習のモデルが得られる．

3.2　LearningFirst

ニューラルネットワークモデルにおいて，i 層における j 番目のユニットを u^i_j，i 層におけるユニット数を n_i と表す．

u^{i-1}_j から u^i_k への結合荷重を $W^i_{j,k}$ とおき，ユニット u^i_k におけるバイアスを b^i_k とおくと，これらの結合荷重及びバイアスの値が，学習すべきパラメータである．

結合荷重及びバイアスの各値に対して，入力データが一つ違った場合に変わり得る値の最大値をまず導出する必要がある．このそれぞれの最大値が，式 2 における Δf に相当する．

各 Δf の値を用い，定義 2.3 に従ってラプラス分布に基づく誤差を結合荷重及びバイアスの各値に対して与えることで，ϵ-差分プライバシを満たす匿名化を実現する．

3.3 AnonymizedLearning

まず第 1 層目において各活性化関数が取り得る値の最大値と最小値を求める，その差が各々の活性化関数における Δf となる．

各 Δf の値を用い，第 1 層目における各活性化関数の出力値に対して定義 2.3 に従ってラプラス分布に基づく誤差を与えながら学習することで，ϵ-差分プライバシを満たす匿名化を実現する．

4 評価

AnonymizingFirst のみを実装して評価を行った．

4.1 データセット

プライバシ保護データマイニングの研究分野で頻繁に利用されているデータセットである，UCI Machine Learning Repository の Adult Data Set [7] を利用した．欠損値を含むレコードを除外して 45,222 レコードある．年齢，性別，人種等の 14 属性から，年収が 500 万ドルを上回るか 500 万ドル以下かの 2 択を推定するモデルを構築した．

4.2 ニューラルネットワークモデル

本論文では差分プライバシを満たすニューラルネットワークモデル構築に焦点を当て，ニューラルネットワークの最適化は行わなかった．確率的勾配降下法におけるバッチサイズを 100，エポック数を 20，中間層の数を 3，各層におけるユニット数を 100 に設定した．また，目的関数として交差エントロピー関数を，活性化関数として Rectified linear unit (ReLU) $f(x) = \max(x, 0)$ を利用し，ドロップアウト [5] の率を 0.5 に設定した．

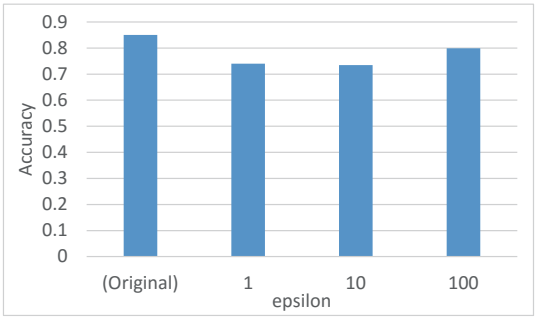


図 2 AnonymizationFirst の評価結果

4.3 評価結果

評価は 10 交差検定で実施した．まず匿名化しない場合の精度を計測したところ，精度は 0.848 であった．

AnonymizationFirst に用いる eMDAV ではまず k-匿名化を行い，その結果に対して ϵ-差分プライバシに基づく誤差を与える．パラメータ k の値の決定方法は示されていないが，本評価では 2，10，100，500 に設定してそれぞれ評価を行い，最も良い結果となった設定を採用した．結果を図 2 に示す．

ϵ の値が増加するほど，つまりプライバシ保護レベルが低下するほど精度が向上していることがわかる．しかし，ϵ の値が 100 の場合でも全く匿名化しない場合と比べて精度が約 0.05%低下しており，向上の余地がまだあると考えられる．

5 おわりに

差分プライバシに基づく匿名化を行いながらニューラルネットワークモデルを生成することを目的とし，AnonymizingFirst，LearningFirst，AnonymizedLearning の 3 つの方針を提案した．

本論文では AnonymizaingFirst のみ具体的なアルゴリズムを示したが，今後残り 2 手法についても詳細なアルゴリズムを提案する予定である．また，3 つの方針を組み合わせて差分プライバシを満たすことも可能だと考えられる．さらに，各方針に従うことで差分プライバシに基づく匿名化が実現できていることを証明する必要がある．また，他のデータセットに提案手法を適用させるとともに，より最適化されたニューラルネットワークに対して評価を行う．

謝辞 本研究は JSPS 科研費 24300005, 26330081, 26870201 の助成を受けたものです.

参 考 文 献

[1] Dwork, C.: Differential Privacy, *Automata, Languages and Programming*, Lecture Notes in Computer Science, Vol. 4052, Springer, 2006, pp. 1–12.

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

[3] Samarati, P.: Protecting respondents' identities in microdata release, *IEEE Trans. Knowl. Data Eng.*, Vol. 13, No. 6(2001), pp. 1010–1027.

[4] Soria-Comas, J., Domingo-Ferrer, J., Sánchez, D., and Martínez, S.: Enhancing data utility in differential privacy via microaggregation-based k-anonymity, *The VLDB Journal*, Vol. 23, No. 5(2014), pp. 771–794.

[5] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting, *The Journal of Machine Learning Research*, Vol. 15, No. 1(2014), pp. 1929–1958.

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

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

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

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

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

[11] 坂井真実子, 玉置幸雄, 石川博之: 骨格性下顎前突患者の側貌パターンの違いが術後の軟組織側貌の予測に与える影響, 日本顎変形症学会雑誌, Vol. 24, No. 4(2014), pp. 305–317.

[12] 青木空眞, 佐藤研, 星憲司, 川上準子, 佐藤憲一, 齋藤芳彦, 森弘毅, 吉田克己: 複数の基本的検査を組み合わせて甲状腺機能異常を発見する診断支援ツール－人間ドックにおけるスクリーニングの実際－, 人間ドック, Vol. 1, No. 26(2011), pp. 9–16.