L-003

表形式データ向け拡散モデルへの差分プライバシーの導入とその評価

Differentially Private Diffusion Models for Tabular Data and Their Evaluation

小林 龍生 [1]　　亀谷 由隆 [1]

Ryusei Kobayashi　　Yoshitaka Kameya

1 はじめに

近年，敵対的生成ネットワーク（Generative Adversarial Networks，以下 GAN）などの生成モデルが表形式データの生成にも利用されている．CTGAN (Conditional Tabular GAN) [6] などの生成モデルを利用した表形式データ生成の取り組みは，医療分野に代表されるプライバシーを含むデータを扱う分野の進歩への貢献が期待される．一方で，これらの生成モデルは学習データに含まれるプライバシー情報を漏洩する可能性が指摘されている．

そこで，学習手続きに差分プライバシー（Differential Privacy，以下 DP） [3] を導入することで，プライバシーを保証した CTGAN である DP-CTGAN [4] が提案された．しかし，DP を導入した生成モデルによる合成データの品質とプライバシー保護の強さの間にはトレードオフの関係があり，DP-CTGAN の提案論文では合成データの品質が CTGAN と比較して低いことが示されている．また，CTGAN の提案論文では生成モデルが出力する合成データの品質に差があることが示されている．

そこで本研究では，より品質の高いデータを出力するモデルに DP を導入することで，合成データの品質を向上させられると考えた．すなわち，本研究では，拡散モデルをベースとした表形式データ生成モデル TabDDPM [5] に DP を導入し，合成データの品質を評価する．

本論文は以下の構成をとる．はじめに 2 節で本研究で使用するモデルやフレームワークなどについて紹介する．そして，3 節で本論文で提案するモデルについて述べる．4 節では合成データの有用性の評価実験について述べ，5 節で実験結果を示す．最後に 6 節でまとめを行い，今後の課題を述べる．

2 準備

2.1 差分プライバシー

差分プライバシー [3] とは，数学的に厳密なプライバシー保護の強さを示すプライバシー基準である．差分プライバシーでは，あるメカニズムにおいて入力がわずかに異なる場合に出力の区別を難しくすることを考える．2 つの同じサイズのデータベース $D, D' \in \mathcal{D}$ において，同一でないレコードの数 $d(D, D') = 1$ という関係が成り立つとき，この 2 つを隣接データベースと呼び，それぞれを入力としたメカニズム m の出力が区別困難であれば，同一でないレコードの情報のプライバシーは保証されていると考える．差分プライバシーは，パラメータ ϵ を用いて以下のように定義される．

定義 1 (ϵ-差分プライバシー)．クエリ $q \in Q$ において，$d(D, D') = 1$ なる任意のデータベースの組 $D, D' \in \mathcal{D}$，およびメカニズム m の出力空間の任意の部分空間 S について，

$$\frac{\Pr(m(q, D) \in S)}{\Pr(m(q, D') \in S)} \leq \exp(\epsilon)$$

ならば，m は ϵ-差分プライバシーを満たす．ここで $\epsilon > 0$ である．■

[1]名城大学大学院理工学研究科情報工学専攻, Division of Information Engineering, Graduate School of Science and Technology, Meijo University

ϵ-差分プライバシーの拡張として，(ϵ, δ)-差分プライバシーがあり，以下のように定義される．

定義 2 ((ϵ, δ)-差分プライバシー)．クエリ $q \in Q$ において，$d(D, D') = 1$ なる任意のデータベースの組 $D, D' \in \mathcal{D}$，およびメカニズム m の出力空間の任意の部分空間 S について，

$$\Pr(m(q, D) \in S) \leq \exp(\epsilon)\Pr(m(q, D') \in S) + \delta$$

ならば，m は (ϵ, δ)-差分プライバシーを満たす．ここで $\epsilon > 0, \delta \geq 0$ である．■

(ϵ, δ)-差分プライバシーは，$\delta = 0$ の場合に ϵ-差分プライバシーと等価であり，$\delta > 0$ の場合に ϵ-差分プライバシーの緩和になる．一般的には，$\delta < \frac{1}{|D|}$ であることが望ましいとされている [8].

2.2 DP-SGD

DP-SGD (Differentially Private Stochastic Gradient Descent) [1] は DP の下で機械学習モデルを訓練するために確率的勾配降下法（Stochastic Gradient Descent，以下 SGD）に修正を加えたものである．DP-SGD では，クリッピングとノイズ追加の操作により，DP を実装する．具体的には，SGD の各ステップで，学習データの各サンプルの勾配を計算し，指定した最大ノルムに勾配をクリップした後，各勾配を集約し，ノイズを追加する操作を行う．

クリッピングは，各タイプステップ t における入力データのサンプル x_i に対する勾配 $\mathbf{g}_t(x_i)$ について，クリッピング定数 C を用いて以下の式で表される．

$$\mathrm{clip}(\mathbf{g}_t(x_i)) = \min\left(1, \frac{C}{\|\mathbf{g}_t(x_i)\|_2}\right)\mathbf{g}_t(x_i)$$

また，ノイズの追加はデータ全体からランダムサンプリングされたバッチ B を用いて行われる．ノイズが追加された勾配は以下の式で表される [2].

$$\bar{\mathbf{g}}_t = \frac{1}{\beta}\left(\sum_{i \in B}\mathrm{clip}(\mathbf{g}_t(x_i)) + \mathcal{N}(0, \sigma^2 C^2 \boldsymbol{I})\right)$$

ここで $\beta = |B|$ はバッチサイズ，σ は追加するノイズの大きさの指標である．以降では σ をノイズスケールと呼ぶ．

2.3 Opacus

Opacus [7] は，DP を使用して深層学習モデルをトレーニングするためのオープンソース PyTorch ライブラリである．Opacus の機能を利用することで，DP-SGD 実装のために既存モデルに加える変更を少なく抑えられる．Opacus を利用して DP-SGD を実装するためのメソッドには，make_private() と make_private_with_epsilon() がある．make_private() では，先述の DP-SGD の導入を行うが，ノイズスケール σ の値を引数に指定する [3]．一方で，make_private_with_epsilon() では，ϵ_0-差分プライバシーを実現するためのプライバシー予

[2]この式中の i に関する和記号の位置は論文 [1] ではなく arXiv の第 2 版 (`https://arxiv.org/abs/1607.00133v2`) の記述に従っている．

[3]Opacus ではノイズスケール σ は noise_multiplier と呼ばれる．

算 ϵ_0 と学習エポック数を引数に指定することで，指定エポック数でプライバシー予算を消費するように Opacus 内部でノイズスケール σ の値が計算される．議論を簡潔にするために，本研究では make_private() を利用して，ノイズスケール σ やその他条件の複数の組み合わせに対する振る舞いを比較する．

2.4 TabDDPM

本研究では，DP を組み込む対象の表形式データ生成モデルとして TabDDPM を用いる．TabDDPM は，拡散モデルの代表的なモデルである Denoising Diffusion Probabilistic Models (DDPM) を表形式データのタスクに対応できるよう拡張されたモデルであり，TabDDPM の提案論文での実験により，数値特徴やカテゴリ特徴などの混合データタイプに対応でき，GAN に基づくものや変分オートエンコーダに基づくものなどの代表的な表形式データ生成モデルと比較して優れていることが示されている．

拡散モデルには，入力データにノイズを追加する拡散過程 (forward process) とノイズデータからノイズを除去して出力データを得る逆拡散過程 (reverse process) があり，TabDDPM では，数値特徴についてはガウス拡散モデル (Gaussian diffusion model)，カテゴリ特徴については多項拡散モデル (multinomial diffusion model) で拡散過程をそれぞれモデル化し，逆拡散過程を多層パーセプトロンでモデル化する．損失関数には，ガウス拡散モデルで加えたノイズと逆拡散過程で除去したノイズの全タイムステップとの平均二乗誤差の合計と，多項拡散モデルの各タイムステップにおける拡散過程の確率分布と逆拡散過程の確率分布間の Kullback-Leibler ダイバージェンスの合計をカテゴリ特徴数で割った値との合計損失が設定される．これを最小化することによって学習する．

2.5 XGBoost

本研究では，評価実験にて使用する予測モデルとして XGBoost (eXtreme Gradient Boosting) [2] を用いる．XGBoost は，アンサンブル学習の 1 つである勾配ブースティングを使用したモデルである．勾配ブースティングとは，弱学習器を複数学習させ，前の学習器の誤りを修正するように次の学習器を学習させることを繰り返すことで性能を向上させる手法である．

3 提案手法

拡散モデルをベースとした表形式データ生成モデルである TabDDPM に対して DP を導入する．モデルへの DP の導入には，TabDDPM の学習部分に DP-SGD を導入することで実現する．具体的には，拡散モデルの逆拡散過程の学習において先述の勾配のクリッピングおよびノイズの追加を行う．また，学習手続きの中でプライバシー予算の消費量を観察し，消費量が予め指定したプライバシー予算 ϵ_0 に達したら学習を停止する．

実装においては，先述したように Opacus ライブラリの make_private() メソッドを用いて DP-SGD を導入する．本研究では最適化器として TabDDPM で用いられている AdamW を利用する．

4 実験方法

4.1 基本的な設定

TabDDPM の提案論文における評価実験で用いられた Cardio データセットを用いて実験を行う．このデータセットは 5 つの数値特徴と 6 つのカテゴリ特徴で構成された 70000 レコードのデータセットであり，学習用，検証用，テスト用に 7 : 1 : 2 で分割する．生成モデルの学習に学習用データを使用し，出力する合成データのサイズは学習用データと同サイズとする．

最終的には，以下の 3 通りのデータについて品質を評価するために，それぞれを学習データとし，XGBoost によってテスト用データに対する予測精度を計測して，プライバシー保護度合いに対する合成データの有用性を評価する．

条件 1: 元データの 7 割の学習用データ
条件 2: DP なし TabDDPM による合成データ
条件 3: DP あり TabDDPM による合成データ

4.2 パラメータチューニング

実験で XGBoost を用いる際のパラメータとして，**条件 1** のデータを学習データ，検証用データをテストデータとして Optuna によるパラメータチューニングを行った結果得られたものを共通のチューニング済みパラメータとする．チューニングに当たり、Optuna の疑似乱数の seed 値を 9 通り，XGBoost の疑似乱数の seed 値を 9 通り用意する．XGBoost の 9 通りの予測結果の F1 値の中央値を Optuna の目的関数として 9 通りのチューニング済みパラメータを求め，更に，これらのパラメータにおける予測結果の F1 値の中央値をとる疑似乱数の seed 値のパラメータを最終的な XGBoost のパラメータとする．

また，TabDDPM を用いる際のパラメータとして，XGBoost のチューニングと同様に**条件 1** で用いたデータと同じものを学習データ，検証用データをテストデータとして Optuna によるパラメータチューニングを行った結果得られたものを DP あり/なし共通のパラメータとする．チューニングに当たり、Optuna の seed 値を 9 通り，TabDDPM の学習用 seed 値を 9 通り，TabDDPM の生成用 seed 値を 9 通り，XGBoost の seed 値を 9 通り用意する．TabDDPM 学習 1 試行ごとにデータ生成を 9 試行行い，生成されたデータのそれぞれに対して XGBoost で 1 試行ずつ評価し，この 9 通りの予測結果の F1 値の中央値を Optuna の目的関数として 9 通りのチューニング済みパラメータを求め，更に，これらのパラメータにおける予測結果の F1 値の中央値をとる seed 値のパラメータを最終的な TabDDPM のパラメータとする．

4.3 有用性の評価

条件 1 の元データに対する有用性の評価として，チューニング済みパラメータを用いてこのデータを学習した XGBoost により，テスト用データに対する予測精度を計測する．このとき，9 通りの seed において予測結果を求め，F1 値が中央値となる seed における結果を最終的な予測精度とする．

条件 2 の合成データに対する有用性の評価として，まず，チューニング済みパラメータを用いて**条件 1** のデータを学習した TabDDPM により合成データを生成する．更に，チューニング済みパラメータを用いてこの合成データを学習した XGBoost により，テスト用データに対する予測精度を計測する．このとき，9 通りの seed において TabDDPM を学習させ，それぞれのモデルで 9 通りの合成データを出力し，それぞれの合成データを 1 通りの XGBoost で学習させて予測精度を計測する．81 通りの結果の F1 値が中央値をとる seed での結果を最終的な予測精度とする．

条件 3 の合成データに対する有用性の評価として，まず，チューニング済みパラメータを用いて**条件 1** のデータを学習した DP あり TabDDPM により合成データを生成する．チューニング済みパラメータを用いてこの合成データを学習した XGBoost により，テスト用データに対する予測精度を計測する．このとき，DP あり TabDDPM の学習では，DP におけるプライバシー予算 ϵ_0，DP を緩和する大きさ δ，学習データのバッチサイズ β，クリップする勾配の最大ノルムの指標 C，追加するノイズの大きさの指標 σ の値を，複数の

表 1: 元データを用いた XGBoost の予測精度（**条件 1**）

Accuracy	AUROC	AUPRC	Precision	Recall	F1
0.735	0.805	0.792	0.750	0.711	0.730

表 2: DP なし TabDDPM の合成データを用いた XGBoost の予測精度（**条件 2**）

Accuracy	AUROC	AUPRC	Precision	Recall	F1
0.732	0.804	0.788	0.754	0.695	0.723

表 3: DP あり TabDDPM の合成データを用いた XGBoost の予測精度（**条件 3**，$\epsilon_0 = 2$）

β	C	σ	Accuracy	AUROC	AUPRC	Precision	Recall	F1
1024	1.0	3.2	0.577	0.669	0.671	0.552	0.847	0.669
4096	1.0	9.2	0.505	0.648	0.650	0.505	0.965	0.663
4096	0.1	9.2	0.514	0.598	0.600	0.510	0.925	0.657

表 4: DP あり TabDDPM の合成データを用いた XGBoost の予測精度（**条件 3**，$\epsilon_0 = 5$）

β	C	σ	Accuracy	AUROC	AUPRC	Precision	Recall	F1
2048	0.1	2.0	0.542	0.730	0.721	0.525	0.974	0.682
1024	1.0	1.5	0.541	0.742	0.732	0.524	0.975	0.682
4096	0.1	3.2	0.688	0.744	0.726	0.704	0.658	0.680

表 5: DP あり TabDDPM の合成データを用いた XGBoost の予測精度（**条件 3**，$\epsilon_0 = 10$）

β	C	σ	Accuracy	AUROC	AUPRC	Precision	Recall	F1
2048	1.0	1.3	0.722	0.778	0.762	0.726	0.720	0.723
4096	0.1	1.6	0.644	0.779	0.764	0.596	0.909	0.720
2048	1.0	1.2	0.706	0.764	0.747	0.705	0.718	0.711

表 6: DP あり TabDDPM の合成データを用いた XGBoost の予測精度（**条件 3**，$\epsilon_0 = 20$）

β	C	σ	Accuracy	AUROC	AUPRC	Precision	Recall	F1
4096	1.0	1.4	0.718	0.784	0.767	0.710	0.745	0.727
4096	1.0	1.7	0.718	0.785	0.773	0.716	0.729	0.722
2048	1.0	1.0	0.646	0.775	0.752	0.598	0.909	0.721

組み合わせで学習させて予測精度を観察する．また，seed 値の扱いについては**条件 2** のデータの評価と同様とする．

本実験では，Cardio データを分割して得られた学習用データのサイズ 49000 より，$\delta = 1/49000$ と固定する．また ϵ_0, β, C についてはそれぞれ $\epsilon_0 \in \{2, 5, 10, 20\}$, $\beta \in \{1024, 2048, 4096\}$, $C \in \{0.1, 1.0\}$ の範囲を試した．そして σ は XGBoost の予測精度や各特徴量の分布を見ながら調整した．

予測精度の指標には，Accuracy，AUROC，AUPRC，Precision，Recall，F1 を用いる．AUROC と AUPRC は決定閾値を動かしたときの ROC (Receiver Operating Characteristic) 曲線下，Precision-Recall 曲線下の面積である．また，出力した**条件 2** のデータと元データ，**条件 3** のデータと元データの各特徴量の分布を相対度数密度を用いて重ねて表して比較する．

5 実験結果

条件 1，**条件 2** での XGBoost の予測精度をそれぞれ表 1 および表 2 に，各 ϵ_0 において F1 値の降順上位 3 つまでの，その DP 条件の組み合わせにおける**条件 3** での予測精度を表 3〜 6 に示す．また，合成データと元データを比較するヒストグラムを図 1 に示す．**条件 2** のデータと元データとの比較が左から 1 列目，**条件 3** のデータと元データとの比較が左から 2 列目〜5 列目である．ここでは，各 ϵ_0 において F1 値が最大となった条件における数値特徴量 age とカテゴリ特徴量 cholesterol のみを示す．

条件 2 のデータについて，表 2 により，Cardio データセットにおいては**条件 1** のデータの予測精度と比較して，6 種の評価指標でほぼ変わらない精度が確認できた．また，図 1 の左から 1 列目により，各特徴量の出現分布も再現できていることが確認でき，合成データの品質が高いことが示された．

条件 3 のデータについて，$\epsilon_0 = 10, 20$ では F1 値が 0.7 以上を取り，元データの予測精度に近い結果を確認できた．図 1 の左から 4，5 列目からも，1 列目ほどではないが元データの分布を再現できていることが確認できた．$\epsilon_0 = 5$ では F1 値が 0.68 程度となり，全体的な予測精度の低下と Recall 値の上昇が確認できた．図 1 の左から 3 列目からも，$\epsilon_0 = 10, 20$ の場合と比較して再現できていないことが確認できた．$\epsilon_0 = 2$ では予測精度が更に低下し，図 1 の左から 2 列目からも元データの分布が再現できていないことが確認できた．

上記の結果から，TabDDPM に DP を導入したモデルにおいてもプライバシー保護の強さと合成データの品質との間にトレードオフの関係を確認でき，$\epsilon_0 = 10, 20$ といった緩いプライバシーの下では高い品質を保っていたが，$\epsilon_0 = 2, 5$ の厳しいプライバシーの下では品質の低下が確認され，実用は難しいと考えられる．

6 おわりに

本研究では，拡散モデルをベースとした表形式データ生成モデル TabDDPM に差分プライバシーを導入して合成データ

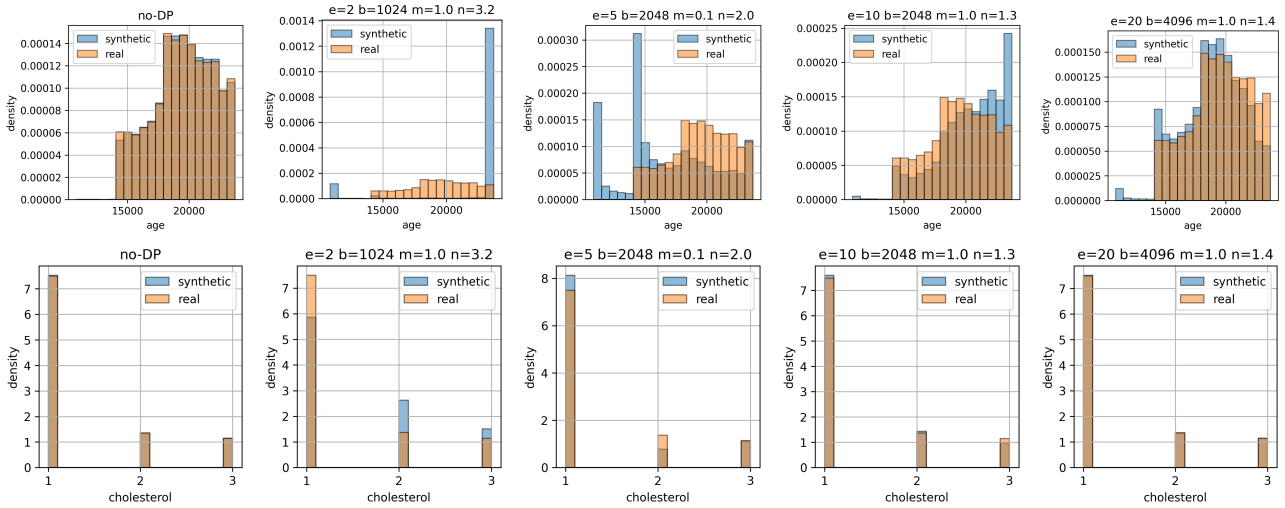

図 1: 元データと合成データにおける特徴量分布の比較．左から順に，DP なし，$\epsilon_0 = 2$，$\epsilon_0 = 5$，$\epsilon_0 = 10$，$\epsilon_0 = 20$ について，数値特徴量 age（上段）とカテゴリ特徴量 cholesterol（下段）の比較ヒストグラム．

の品質を評価した．その結果合成データの品質とプライバシー保護の強さとの間にトレードオフを確認でき，更に $\epsilon_0 \geq 10$ において，予測精度が保たれていることも確認できた．

今後の課題としては，複数データセットにおける評価，他の DP を導入した表形式データ生成モデルとの比較評価，本モデルに対するプライバシー攻撃によるプライバシー評価などが挙げられる．

参考文献

[1] Abadi, M., Chu, A., Goodfellow, I., McMahan, H.B., Mironov, I., Talwar, K., Zhang, L.: Deep Learning with Differential Privacy. In: Proc. of the 2016 ACM SIGSAC Conf. on Computer and Communications Security (CCS 2016). pp. 308–318 (2016)

[2] Chen, T., Guestrin, C.: XGBoost: A Scalable Tree Boosting System. In: Proc. of the 22nd ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining (KDD 2016). pp. 785–794 (2016)

[3] Dwork, C.: Differential Privacy. In: Proc. of Int'l Colloquium on Automata, Languages, and Programming (ICALP 2006). pp. 1–12 (2006)

[4] Fang, M.L., Dhami, D.S., Kersting, K.: DP-CTGAN: Differentially Private Medical Data Generation Using CTGANs. In: Proc. of Int'l Conf. on Artificial Intelligence in Medicine (AIME 2022). pp. 178–188 (2022)

[5] Kotelnikov, A., Baranchuk, D., Rubachev, I., Babenko, A.: TabDDPM: Modelling Tabular Data with Diffusion Models. In: Proc. of the 40th Int'l Conf. on Machine Learning (ICML 2023). pp. 17564–17579 (2023)

[6] Xu, L., Skoularidou, M., Cuesta-Infante, A., Veeramachaneni, K.: Modeling Tabular Data Using Conditional GAN. Proc. of the 33rd Int'l Conf. on Neural Information Processing Systems (NeurIPS 2019) (2019)

[7] Yousefpour, A., Shilov, I., Sablayrolles, A., Testuggine, D., Prasad, K., Malek, M., Nguyen, J., Ghosh, S., Bharadwaj, A., Zhao, J., et al.: Opacus: User-Friendly Differential Privacy Library in PyTorch. arXiv:2109.12298 (2021)

[8] 佐久間淳: データ解析におけるプライバシー保護. 講談社 (2016)