Adversarial Autoencoder を用いた Vertical Federated Learning

黄　凱
Huang Kai
指導教員 劉　慶豊
法政大学大学院理工学研究科システム理工学専攻修士課程

1. はじめに

機械学習技術が各分野で広く応用されるにつれ、データのプライバシー保護の問題がますます重要になっている。機械学習モデルの訓練には通常、大量の高品質なデータが必要であるが、データの集中化はプライバシー漏えいやデータセキュリティの問題を引き起こす可能性がある。特に医療や金融といった敏感な分野では、プライバシーを損なうことなくデータ共有を実現する方法が緊急の課題となっている。

連合学習（Federated Learning, FL）は、新興の機械学習フレームワークであり、複数のデータ保有者間で協調的なモデリングを可能にする。データ保有者は生のデータを共有せずにモデル訓練に参加できるため、データのプライバシーを効果的に保護できる。しかし、現行の連合学習手法では、プライバシー保護とモデル性能の間に依然としてトレードオフが存在し、そのバランスをどのように実現するかが重要な研究課題となっている。

Cha [1] は、オートエンコーダを用いて垂直連合学習におけるデータの特徴表現と共有を実現した。この方法はプライバシー保護に一定の効果があるものの、再構成攻撃に対して潜在的なリスクが残っており、特に高次元で複雑なデータを扱う場合、モデルの性能に影響を及ぼす可能性がある。

これらの問題を解決するため、本研究では敵対的オートエンコーダ（Adversarial Autoencoder, AAE）に基づく方法を提案する。AAE はオートエンコーダと敵対的生成ネットワーク（GAN）の利点を組み合わせており、敵対的訓練を通じてエンコードされた潜在表現が事前の分布に従うようにすることで、プライバシー保護効果を向上させる。本研究は、敵対的オートエンコーダーを垂直連合学習に導入することで、既存の方法よりも優れた連合学習の運用手法を提供することを目的としている。これにより、データプライバシー保護を強化し、モデルの有効性と実用性を確保し、センシティブなデータを扱うシナリオでの連合学習の適用に対するより優れたソリューションを提供する。

2. 連合学習の紹介

連合学習は、中央サーバーで複数のクライアントが協力して機械学習の問題を解決する枠組みである。この設定により、トレーニングデータが分散化され、各デバイスのデータプライバシーが確保される。連合学習は、ローカル計算とモデルの伝達という二つの主要なアイデアに基づいており、従来の集中型機械学習手法がもたらすシステムのプライバシーリスクとコストを削減する。クライアントの元データはローカルに保存され、交換や移行はできない。連合学習を適用することで、各デバイスはローカルデータを用いてローカルトレーニングを行い、その後モデルをサーバーにアップロードして集約する。最終的にサーバーはモデルの更新を参加者に送信し、学習目標を達成する。

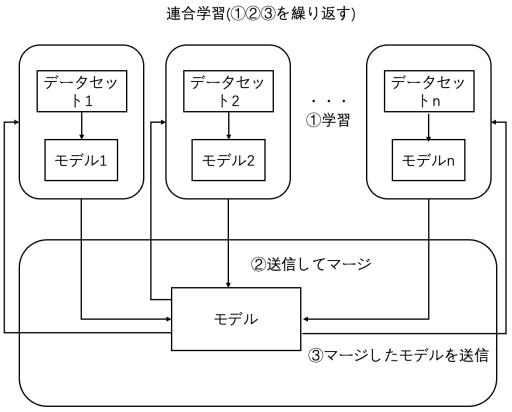

図 1 連合学習のプロセス

.

Yang et al. [2] による分類では、連合学習は参加者が所有するデータセットがどのように分散しているかで 3 種類に分けられる。具体的には各データセットの標本空間、特徴空間、正解ラベルに注目する。

- 標本空間（sample space）は、データのサンプル全体の集合である。例えば与えられた画像が犬の画像であるか判定する分類問題の場合、問題の具体例（インスタンス）である 1 個の画像が 1 個のサンプルである。また、例えば各行がユーザに関する情報を表す表形式データを使用したタスクの場合、各行（つまり各ユーザ）がそれぞれサンプルである。
- 特徴空間（feature space）は、サンプルが取りうる特徴量全体の集合である。画像の場合、サンプル画像の各ピクセルの数値が特徴量である。表形式データの場合、各列（属性）がそれぞれ 1 個の特徴量を表す。

- 正解ラベル（labels / ground truth）は、インスタンスの正解データである。学習では与えられたサンプルに対して推論を行い、その予測値と正解ラベルの誤差を小さくするようにモデルの改善を行う。

これらの要素に基づいて、連合学習は次の 3 つに分類される。

- 水平連合学習（HFL: Horizontal Federated Learning）：各データセットの標本空間が異なり、特徴空間に共通部分がある。例えば、Google は HFL を使用して、携帯電話のユーザーが彼らのデータセットを使用して次の単語予測モデルを共同で訓練することを可能にしている [3]。
- 垂直連合学習（VFL: Vertical Federated Learning）：各データセットの標本空間に共通部分があり、特徴空間が異なる。例えば、銀行は VFL を使用して、請求書機関と協力して企業顧客の金融リスクモデルを構築している [4]。
- 連合転移学習（FTL: Federated Transfer Learning）：各データセットにわずかに共通した標本空間と特徴量空間がある。例えば、異種分布を持つ複数の被験者からの EEG（electroencephalogram）データは、FTL を使用して BCI（brain–computer interface）モデルを共同で構築する [5]。

3 つの連合学習のデータセットの状態を簡単にまとめた図 2 を次に示す。

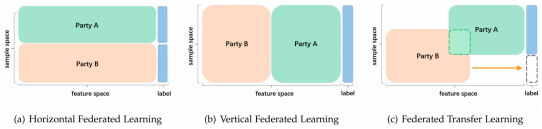

図 2 連合学習の分類 [2]

.

3. 垂直連合学習の詳細

垂直連合学習は、異なるデータ保有者が同じサンプルを持ちながら、特徴が異なるデータセットで主に使用される。この場合、参加者 A と B がいて、それぞれのデータセットは

$$X_A \in \mathbb{R}^{n \times d_A}, \quad X_B \in \mathbb{R}^{n \times d_B}$$

となる。ここで、n はサンプル数、d_A と d_B はそれぞれ A と B の特徴数である。両者は同じサンプル ID を共有し、サンプルごとにデータを対応付ける。

垂直連合学習では、原始データを共有せずに勾配を交換することで、全体のモデル $f(X_A, X_B)$ を共同で訓練することを目指す。一般的な最適化問題は以下のように表される：

$$\min_{\theta_A, \theta_B} \sum_{i=1}^{n} L\left(f\left(X_A^{(i)}; \theta_A\right), f\left(X_B^{(i)}; \theta_B\right), y^{(i)}\right)$$

ここで、θ_A と θ_B はそれぞれ A と B のモデルパラメータ、L は損失関数である。

4. 敵対的オートエンコーダの紹介

敵対的オートエンコーダは、敵対的生成ネットワーク（GAN）の思想を組み合わせたオートエンコーダであり、対敵的訓練を通じて潜在表現が特定の事前分布に従うようにし、プライバシー保護を強化する。

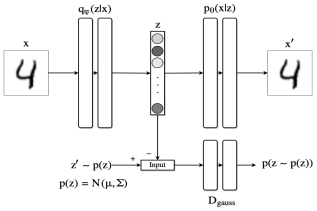

図 3 AAE の基本アーキテクチャ[6]

AAE の構造は、エンコーダ、デコーダ、判別器で構成される。入力データを X とすると、エンコーダはそれを潜在表現 Z にマッピングする：

$$Z = E(X; \theta_E)$$

ここで、θ_E はエンコーダのパラメータである。デコーダは潜在表現 Z を入力データ空間に再マッピングする：

$$\hat{X} = D(Z; \theta_D)$$

θ_D はデコーダのパラメータである。生成器の損失関数は次のようになる：

$$L_G = -\frac{1}{m} \sum_{k=1}^{m} \log(D(z))$$

ここで、m はミニバッチサイズ、z はエンコーダによって生成される。

AAE は判別器 D_{adv} を導入し、潜在表現 Z が事前分布 $p(Z)$ に従っているかどうかを判別する。事前分布は標準正規分布が利用され、このようにエンコーダーが生成したデータの分布が標準正規分布に近づく。そのメリットは、標準正規分布に従うデータは統計的特性が均一であり、そのため機械学習モデルは学習過程においてデータの変動に適応しやすく、偏りや異常なデータがモデル性能に与える負の影響を回避することができる。さらに、正規化されたデータはモデルの汎化能力を向上させ、未知のデータに対してもより安定したパフォーマンスを発揮できるようになる。判別器の損失は次のようになる：

$$L_D = -\frac{1}{m} \sum_{k=1}^{m} \left(\log(D(z')) + \log(1 - D(z))\right)$$

ここで、m はミニバッチサイズ、z はエンコーダによって生成され、z' は真の事前分布からのサンプルである。

この敵対的訓練により、AAE は潜在表現の特徴空間やデータ分布の両方において元のデータと異なり、再構成攻撃に対する防御能力を向上させ、プライバシー保護を強化することができる。

5. 実装

垂直連合学習において、データの整合プロセスは、異なるデータ保有者間でデータサンプルが正しく一致することを保証する。まず、ID の照合によって、各データ保有者が共有するサンプルの集合を確認できるようになる。例えば、クライアント A の第 3 行目は、クライアント B および C の第 3 行目に対応する。同手法の汎用性を検証するため、データセットを様々な数に分割した（表 2）。

データの垂直分割後、各クライアントに対して敵対的オートエンコーダに基づくモデル訓練を行った（図 3a, b, c）。訓練後、各クライアントの潜在データ（図 3a', b', c' の隠れ層での表現）を集約して中央サーバーに送信してモデル訓練を行った。本タスクには PyTorch [7] を使用した。各サイトのデータは垂直分割され、異なるサイト間での垂直に分割されたデータをシミュレートした。分類タスクの評価指標として、正解率および ROC 曲線の下の面積（AUROC）を使用した (図 4)。

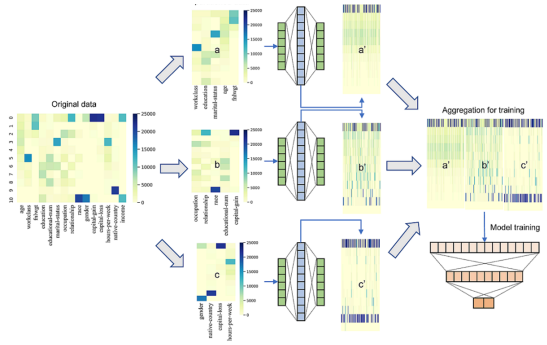

図 4 トレーニングの流れ [1]

.

訓練には、カテゴリカル埋め込みを使用した表形式のニューラルネットワークモデルが使用された。集中型モデルがベースラインモデルとして使用された。各垂直分割データに対して、2 つのモデルを訓練した。1 つは垂直に分割されたデータセットに基づく機械学習モデル（図 4a, b, c）、もう 1 つは分割されたデータセットの潜在表現に基づく機械学習モデル（図 4a', b', c'）である。最後に、集中型モデルと潜在データを集約したモデルを比較し、垂直連合学習ニューラルネットワークモデルのベンチマークを行った。

6. 実証研究

(1) Adult Income Dataset

表 1 に示されているように、Adult Income Dataset [8] には、ラベルとして年間所得が 5 万ドルを超えるかどうかを表すダミー変数、および特徴量として 8 つのカテゴリカル変数と 6 つの連続変数が含まれている。このデータセットには、年収が 5 万ドル以下の個人が 37,155 名、5 万ドル以上の個人が 11,687 名含まれている。

(2) データの垂直分割

データセットを均衡させるため、年収 5 万ドル以下の 37,155 名のサンプルから 11,687 名を無作為にサンプリングし、年収 5 万ドル以上の 11,687 名の個人データと合わせて、最終的に合計 23,374 名のサンプルで構成し、予測の確率レベルを 50%に設定した。このデータセットは、個人ごとに異なる部分データを所有する 3 つの異なる組織を仮定し、垂直に 3 つの部分に分割した（表 2）。

ID	age	workclass	⋯	native-country	income
1	39	State-gov		United-States	$\leq 50K$
2	50	Self-emp-not-inc		United-States	$\leq 50K$
3	38	Private		United-States	$\leq 50K$
⋯					
48842	52	Self-emp-inc		United-States	>50K

表 1 データセットの特徴量とラベル

Dataset	Division	Dataset size	Feature dimension	Autoencoder layers
Adult income	3 sites	23,374	5, 5, 4	64-128-64

表 2 垂直に分割されたデータをシミュレートするためのデータセット構成およびトレーニングパラメータ

7. 実験結果

Adversarial Autoencoder を用いた垂直連合学習における有効性を検証するために、Adult Income Dataset を用いて比較実験を行った。実験では、集中型モデル（Central Model）、Cha [1] によって提案された Overcomplete Autoencoder、Adversarial Autoencoder、そして Adversarial Autoencoder(Gaussian Mixture Model,GMM) の手法を比較し、4 つの手法のモデル性能における違いを評価した。モデルの性能を評価するために、正解率（Accuracy）および ROC 曲線の下の面積（AUROC）といった指標を使用した。

ここで、集中型モデルは、すべてのデータを集中して処理する従来の方法を指す。Overcomplete Autoencoder は、元のデータを高次元の潜在表現に変換するオートエンコーダであり、隠れ層の次元を増やすことでモデルの表現力を強化し、連合学習におけるデータの安全性を向上させるが、わずかな性能低下を引き起こす可能性がある。Adversarial Autoencode(GMM) は、事前分布の標準正規分布をガウス混合モデル（GMM）に置き換えた手法である。

Adult income dataset	Accuracy	AUROC
Central	0.83	0.91
Overcomplete Autoencoder	0.82	0.90
Adversarial Autoencoder	0.83	0.83
GMM	0.85	0.85

表 3 分類結果

表 3 の結果から見ると、AAE を導入する方法は、モデルの精度において集中型モデルと同等の性能を維持しているが、AUROC においては低下している。この原因としては、敵対的訓練がプライバシーを保護する一方で、潜在表現の特徴抽出に一定の影響を与えた可能性がある。しかし、AAE はプライバシー保護の面で顕著な利点を持ち、特に異なるクライアントにおけるセンシティブなデータの学習時に、敵対的

訓練を通じて潜在表現をより安全にし、元のデータへの復元を困難にしている。ガウス混合モデル（GMM）を組み合わせた AAE は、元々の AAE の事前分布を標準正規分布から GMM に変更することで実現されており、これにより潜在空間の表現がより柔軟かつ複雑になり、モデルの性能が向上している。これらの実験結果は、AUROC における性能損失があるものの、AAE はプライバシー保護とモデル性能の間で良好なバランスを達成している。AAE（GMM）はより柔軟な潜在空間の分布を通じて、プライバシー保護とモデル性能の間でより良いバランスを達成している。

8. 今後の課題

本研究において、AAE（GMM）は精度および AUROC において他のモデルよりも優れた性能を示したが、その原因はまだ完全には理解されていない。GMM を AAE の事前分布として使用することでなぜこれほどの改善が見られるのかは未解明である。今後の研究では、さらなる実験と理論的な分析を通じて、GMM が潜在空間に与える影響を検証し、それがモデルの性能をどのように向上させるのかを分析する予定である。これにより、AAE（GMM）をより深く理解し、モデル設計を最適化し、より優れたプライバシー保護と性能を実現することに貢献する。

参考文献

[1] Dongchul Cha, MinDong Sung, Yu-Rang Park, et al. Implementing vertical federated learning using autoencoders: Practical application, generalizability, and utility study. *JMIR medical informatics*, 9(6):e26598, 2021.

[2] Yang Liu, Yan Kang, Tianyuan Zou, Yanhong Pu, Yuanqin He, Xiaozhou Ye, Ye Ouyang, Ya-Qin Zhang, and Qiang Yang. Vertical federated learning: Concepts, advances, and challenges. *IEEE Transactions on Knowledge and Data Engineering*, 2024.

[3] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

[4] Yong Cheng, Yang Liu, Tianjian Chen, and Qiang Yang. Federated learning for privacy-preserving ai. *Communications of the ACM*, 63(12):33–36, 2020.

[5] Ce Ju, Dashan Gao, Ravikiran Mane, Ben Tan, Yang Liu, and Cuntai Guan. Federated transfer learning for eeg signal classification. In *2020 42nd annual international conference of the IEEE engineering in medicine & biology society (EMBC)*, pages 3040–3045. IEEE, 2020.

[6] Paperspace Blog. Adversarial autoencoders with pytorch, 2023. Accessed: 2024-10-25.

[7] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[8] Becker B. and Kohavi R. Adult [dataset], 1996.