差分プライバシーを保証したモデル説明DPGD-Explainに対するレコード再構築リスクの実験評価

當麻 僚太郎[1,a)]　菊池 浩明[1,b)]

概要：機械学習モデルの公平性や学習の透明性を保証し，ユーザに納得感を与えるために機械学習モデルの出力を説明する説明可能性技術が注目されている．機械学習モデルを用いたサービスの多くは Machine Learning as a Service（MLaaS）と呼ばれるプラットフォーム上で提供されており，これらの MLaaS プラットフォームでは，モデルの出力に加えて，モデルを説明するいくつかの指標を提供している． 2022 年に Patel らは，差分プライバシーを保証したブラックボックスモデルの説明可能性指標 DPGD-Explain を提案している．しかし，モデル説明は入力データの重要な因子を提供するため，参照データを学習した攻撃モデルを用いて，説明可能性指標から元のレコードを再構築されるリスクが疑われる．そこで，本研究では，2 つのオープンデータを用いて DPGD-Explain に対するレコード再構築リスクを調べ，プライバシー予算や説明の質と安全性との関係を明らかにする．

キーワード：説明可能性，差分プライバシー，レコード再構築攻撃，DPGD-Explain

Empirical Evaluation of Record Reconstruction Risk from DPGD-Explain Model Explanations with Differential Privacy

Ryotaro Toma[1,a)]　Hiroaki Kikuchi[1,b)]

Abstract: Explainability has gained attention to ensure fairness and transparency in machine learning models, providing users with a sense of understanding. Most services for machine learning models are offered in a style of Machine Learning as a Service (MLaaS) platforms, which provide several methods to explain model outputs. Patel et al. (2022) proposed DPGD-Explain, model explanations with differential privacy. Nevertheless, it remains unclear how model explanations with guarantee of differential privacy is vulnerable against the record reconstruction attack that trains the behavior between input data and the model explanations. In this study, we investigate the record reconstruction risk of DPGD-Explain in terms of the privacy budget and the quality.

Keywords: XAI, Differential Privacy, Record Reconstruction Attack, DPGD-Explain

1. はじめに

近年，機械学習モデルは金融や雇用などの重要な意思決定の場面で活用されることが増えている．それらの多くのモデルはニューラルネットワークやアンサンブルモデルなどの複雑な構造を持ち，入力に対する内部構造が不明な，いわゆるブラックボックスであった．そのため，モデルの公平性や透明性を保証し，モデルの出力に対して説明を与えるための説明可能性技術 eXplainable AI（XAI）が注目されている [1], [2], [3].

2022 年に Patel ら [24] は差分プライバシーを保証した XAI 手法である DPGD-Explain を提案した．しかし，XAI には説明に用いられたプライベートな入力を漏洩するリスクがあることが知られている．例えば，2022 年に Luo

[1] 明治大学大学院先端数理科学研究科
Graduate School of Advanced Mathematical Sciences, Meiji University
[a)] cs242022@meji.ac.jp
[b)] kikn@meiji.ac.jp

ら [6] は Shapley 値に基づく説明から本来秘匿されているモデルへの入力レコードを推論出来ることを示した．また，XAI のプライバシーリスクが差分プライバシーによってどれだけ低減されるのかは明らかでない．

そこで，本研究では，DPGD-Explain に対して Luo ら [6] の手法を基にしたレコード再構築攻撃を行い，説明の有用性や安全性がどう変化するかを明らかにする．この提案方式を，Adult データセット [8] と IMDB データセット [9] について適用した結果を報告する．

2. 基本定義

2.1 DPGD-Explain

DPGD-Explain[24] は，2022 年に Patel らによって提案された，差分プライバシーを保証した XAI 手法の一つである．説明対象のブラックボックスモデルを f，入力ベクトルを $\boldsymbol{z} = (z_1, ..., z_n)$，$\boldsymbol{z}$ に対するモデル説明を $\boldsymbol{\phi} = (\phi_1, ..., \phi_n)$ とするとき，DPGD-Explain は $\boldsymbol{z}$ の周辺での f のふるまいを近似する線形モデル $g(\boldsymbol{x}, \boldsymbol{z}) = \boldsymbol{\phi}^T(\boldsymbol{x} - \boldsymbol{z})$ を学習する．ここで，損失関数 $\mathcal{L}$ を次のように定義する．

$$\mathcal{L}(\boldsymbol{\phi}, \boldsymbol{z}, \mathcal{X}, f) := \frac{1}{|\mathcal{X}|} \sum_{\boldsymbol{x} \in \mathcal{X}} \alpha(\|\boldsymbol{x} - \boldsymbol{z}\|)(\boldsymbol{\phi}^T(\boldsymbol{x} - \boldsymbol{z}) - f(\boldsymbol{x}))^2 \tag{1}$$

$\mathcal{X}$ は m 行 n 列の説明データセットである．$\alpha : \mathbb{R}^+ \to \mathbb{R}^+$ は重み関数であり，2 つのベクトル間の距離 $\|\boldsymbol{x} - \boldsymbol{z}\|$ が大きくなればなるほど小さな値を取る関数である．例えば，$\alpha(d) = \exp(-d)$ が挙げられる．この損失関数に対して，DPGD-Explain の最適解 $\boldsymbol{\phi}^*$ は次式により求められる．

$$\boldsymbol{\phi}^* = \arg\min_{\boldsymbol{\phi} \in \mathcal{C}} \mathcal{L}(\boldsymbol{\phi}, \boldsymbol{z}, \mathcal{X}, f) \tag{2}$$

ここで，$\mathcal{C} = \{\boldsymbol{\phi} : \|\boldsymbol{\phi}\| \leq 1\}$ である．

DPGD-Explain では，損失を最小化する過程でノイズを付加することで差分プライバシーを保証する．ノイズの大きさを決定し DPGD-Explain を実行する過程を次のアルゴリズム 1 に示す．DPGD-Explain は，説明データセット $\mathcal{X}$ を用いて生成した説明ベクトル $\boldsymbol{\phi}$ と 1 行のみ異なる隣接データセット $\mathcal{X}'$ を用いて生成した説明ベクトル $\boldsymbol{\phi}'$ に対し，ある説明ベクトル集合 S について次の式で示される (ϵ, δ)-差分プライバシーを保証する．

$$\Pr[\boldsymbol{\phi} \in S] \leq e^{\epsilon} \cdot \Pr[\boldsymbol{\phi}' \in S] + \delta \tag{3}$$

2.2 Feature Inference Attack on Shapley Values[6]

Luo ら [6] は Shapley 値に対する Feature Inference Attack を示した．Luo らの提案手法の概要を図 1 に示す．サービスプロバイダは訓練データセット $\mathcal{X}_{train}$ に基づいてブラックボックスモデル f を訓練し，MLaaS 上にモデルをデプロイして提供する．攻撃者はサービスに対して，標

Algorithm 1 差分プライバシーを保証したモデル説明 [24]

Input: 説明対象のデータ点 $\boldsymbol{z} \in \mathbb{R}^n$，説明データセット $\mathcal{X}$，損失関数 $\mathcal{L}$，プライバシー予算 (ϵ, δ)，ステップ数 t，学習率関数 $\eta(t) \in \mathbb{R}^+$

Output: t ステップ目の説明ベクトル $\boldsymbol{\phi}^t$

Set the parameter for the Gaussian mechanism $\sigma \leftarrow \frac{1}{m\epsilon}\sqrt{16t \log\left(e + \frac{\sqrt{t}\epsilon}{\delta}\right) \log\frac{t}{\delta}}$
Initialize $\boldsymbol{\phi}^0$ with an arbitrary vector in C
return DPGD-Explain$(\boldsymbol{\phi}^0, \sigma, t)$

procedure DPGD-Explain$(\boldsymbol{\phi}^0, \sigma, t)$
  for $\tau \in 0, \ldots, t-1$ **do**
    $\xi_t \leftarrow \boldsymbol{\phi}^t - \eta(t)\left(\nabla\mathcal{L}(\boldsymbol{\phi}^\tau) + \mathcal{N}(0, \sigma^2 I)\right)$
    $\boldsymbol{\phi}^{\tau+1} \leftarrow \arg\min_{\boldsymbol{\phi} \in \mathcal{C}} \|\boldsymbol{\phi} - \xi_\tau\|$
  end for
  return $\boldsymbol{\phi}^t$
end procedure

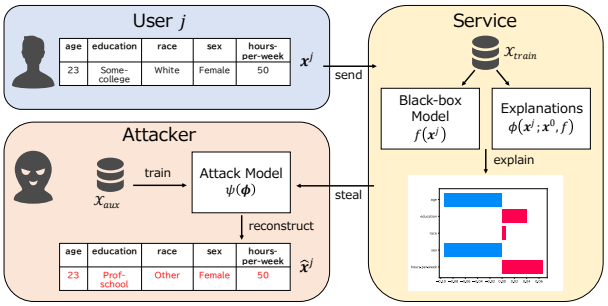

図 1 システムモデル概要図

的ユーザ j の入力に対する説明（Shapley 値）にアクセスできる．与えられた説明ベクトル $\phi(\boldsymbol{x}^j; \boldsymbol{x}^0, f)$ に対して，攻撃者は参照ベクトル $\boldsymbol{x}^0$ と攻撃モデル $\psi(\boldsymbol{\phi})$ を用いて元の入力ベクトル $\boldsymbol{x}^j$ を再構築する．

2.3 関連研究

2.3.1 XAI

説明可能性手法は大域的な手法と局所的な手法に分けられる．大域的な手法では全体的なモデルのふるまいを説明し，各特徴量の重要度を算出する [13], [14]．一方で局所的な手法ではそれぞれの入力に対して各特徴量の重要度を説明する [7], [11], [12], [15]．Amazon SageMaker[4] や Microsoft Azure[5]，Google Cloud Platform[16] などの主要な MLaaS プラットフォームでは，局所的な説明手法として LIME[7] や Shapley 値 [10], [11], [12] が提供されている．

本研究で注目する DPGD-Explain[24] は局所的手法に分類され，LIME と同様にモデルのふるまいを近似する線形モデルを学習する．

一方で，メンバーシップ推論攻撃 [17], [18], [19] やモデル窃取攻撃 [17], [20], [21]，再構築攻撃 [6], [22]，敵対的攻撃 [17], [23] など，様々な攻撃手法に対する説明可能性のプ

ライバシーリスクが調査されている．

2.3.2 プライバシー保護技術

データプライバシーを保護する技術として，差分プライバシー [27], [28], [29] や合成データ [27], [30], [31] がある．加えて，プライバシーを保護する XAI 手法の研究 [24], [25], [26] も進められている..

3. 問題設定

3.1 レコード再構築攻撃

Luo らの手法 [6] に従い，レコード再構築攻撃（または特徴推論攻撃）を定義する．

f をブラックボックスモデル, ψ を攻撃モデルとし, $\mathcal{X}_{train}$ を f の訓練データセット，$\mathcal{X}_{aux}$ を ψ を訓練するための補助データセット，$\mathcal{X}_{test}$ をテストデータセットとする．ユーザ $j = 1, ..., m$ の入力ベクトルを $\boldsymbol{x}^j = (x_1^j, ..., x_n^j)$，説明の生成に用いる参照サンプルを $\boldsymbol{x}^0$，入力 $\boldsymbol{x}^j$ に対する説明を $\boldsymbol{s}^j = \phi(\boldsymbol{x}^j; \boldsymbol{x}^0, f)$ とする．すべての $\boldsymbol{x}_{aux} \in \mathcal{X}_{aux}$ とブラックボックスモデル f について，与えられた説明データセット $\mathcal{S}_{aux}$ に対して，攻撃者は攻撃モデル $\psi : \mathcal{S}_{aux} \to \mathcal{X}_{aux}$ を訓練する．

3.2 評価指標

説明の安全性を評価するためにレコード再構築攻撃の攻撃成功率 SR を定める．説明の有用性を，最適解の説明ベクトルと比較した MSE で評価する．

3.2.1 レコード再構築の攻撃成功率

再構築された特徴量のうち，再構築に成功した特徴量の割合を攻撃成功率とする．対象が質的変数の場合は結果が一致しているかどうか，量的変数の場合は一定の閾値内に誤差が収まっているかどうかで攻撃の成功判定を行う．m 行 n 列の入力データセット X に対して各レコードごとに再構築されたデータセット $\hat{X}$ の攻撃成功率 SR は

$$SR(X, \hat{X}) = \frac{success(X, \hat{X})}{mn} \tag{4}$$

とする．ここで，$success(X, \hat{X})$ は正しく再構築された特徴量の数である．

3.2.2 説明ベクトルの有用性

DPGD-Explain の $\epsilon = \infty$（ノイズなし）における説明ベクトルを $\boldsymbol{\phi}^*$ と ϵ で学習された説明ベクトル $\boldsymbol{\phi}$ の MSE を有用性損失として評価する．

$$MSE(\boldsymbol{\phi}, \boldsymbol{\phi}^*) = \frac{1}{n}\left((\phi_1 - \phi_1^*)^2 + \cdots + (\phi_n - \phi_n^*)^2\right) \tag{5}$$

4. 実験

4.1 データセット

実験に用いるデータセットを表 1 に示す．

Adult は 14 個の特徴量と年収からなるデータセットである．IMDB データセットは映画レビューの自由記述のテキストデータと positive/negative のラベルからなるデータセットである．IMDB データセットは出現頻度上位 500 語に限定して，その語を含むか否かを示すバイナリベクトルに変換する．

表 1 実験に用いたオープンデータセット

Dataset	No. of Records	Classes
Adult [8]	48,842	2
IMDB Moview Reviews [9]	50,000	2

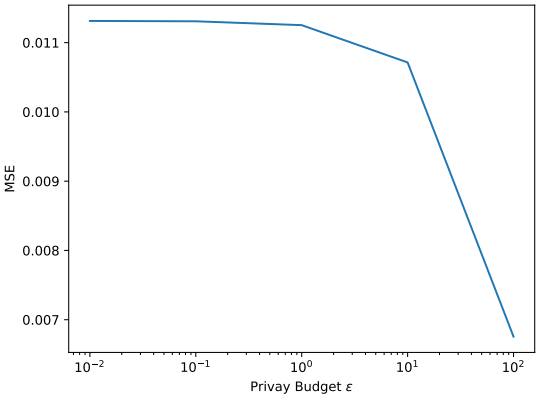

図 2 Adult データセットに対する ϵ に応じた説明ベクトルの有用性損失の変化

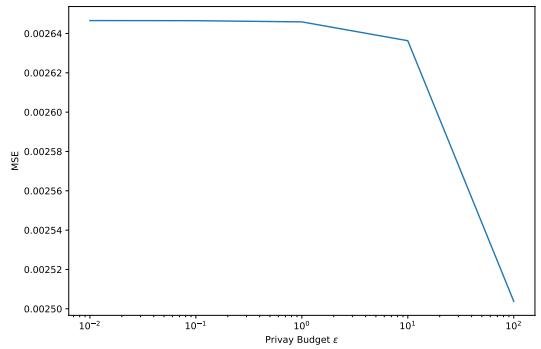

図 3 IMDB データセットに対する ϵ に応じた説明ベクトルの有用性損失の変化

4.2 結果 1：説明の有用性

プライバシー予算 ϵ についての DPGD-Explain の有用性を図 2，3 に示す．

Adult データセットと IMDB データセットの双方で，ϵ が大きくなるにつれて有用性損失は小さくなっている．一方で，$\epsilon = 10$ という大きな設定でも有用性損失は $\epsilon = 0.01, 0.1, 1$ の場合とあまり変わらない．

4.3 結果 2：説明の安全性

ϵ についてのレコード再構築の攻撃成功率を図 4，5 に

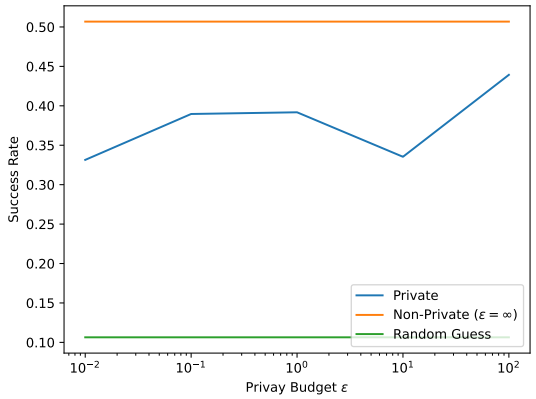

図 4 Adult データセットに対する ϵ に応じた攻撃成功率の変化

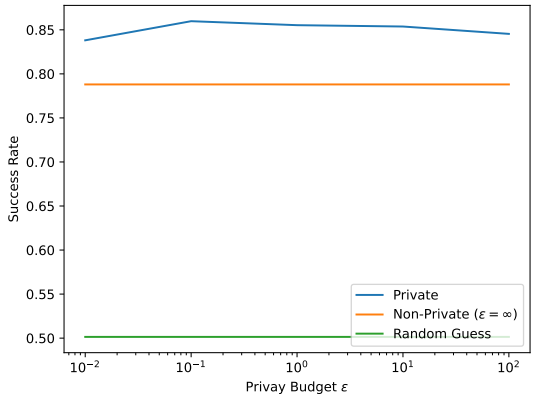

図 5 IMDB データセットに対する ϵ に応じた攻撃成功率の変化

示す．

ϵ に応じて攻撃成功率が単調増加することを期待したが，図 4 では不規則な増減があり，図 5 では変化が見られなかった．Adult データセットに関しては，どの ϵ についても，$\epsilon = \infty$ の説明より攻撃成功率が低く，ランダムな予測より攻撃成功率が高く，妥当な結果になった．しかし，IMDB データセットに関しては，どの ϵ についても，ノイズなしの $\epsilon = \infty$ の説明ベクトルよりも攻撃成功率が高いという結果になった．

4.4 考察

図 2，3 では，ϵ が大きくなるにつれて真の説明ベクトルと比較した MSE は小さくなった．しかし，$\epsilon = 10$ という大きな設定でも有用性損失は $\epsilon = 0.01, 0.1, 1$ の場合とあまり変わらず，$\epsilon \leq 10$ のとき説明の質が十分に高いとは言えない．この原因としては，DPGD-Explain を計算する際に用いられる勾配に対し，付加するノイズが大きいことが考えられる．例えば，$\epsilon = 10, \infty$ それぞれについて，同じ入力に対して説明ベクトルの損失をステップ数に応じてプロットした例を図 6，7 に示す． この例では，$\epsilon = 1$ のとき説明の学習が上手くいっていない．

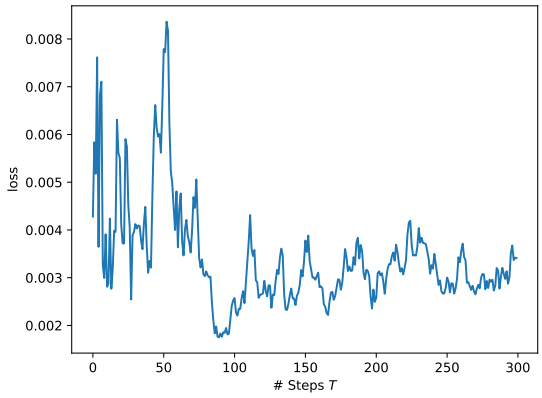

図 6 勾配降下法のステップ数に対する $\epsilon = 10$ のときの説明ベクトルの損失

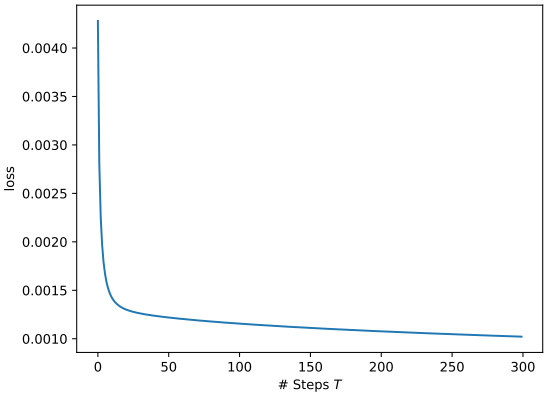

図 7 勾配降下法のステップ数に対する $\epsilon = \infty$ のときの説明ベクトルの損失

また，図 5 では，どの ϵ についても，$\epsilon = \infty$ の説明より攻撃成功率が高いという結果になった．これは，IMDB データセットの入力ベクトルが 500 次元のバイナリベクトルであり，要素のほとんどが 0 である不均衡なデータセットであることが要因の一つであると考えられる．攻撃モデル ψ が学習するのは単語が入力に含まれていたかどうかの二値分類であるため，ほとんどの教師データが 0 となる状況下における学習が難しく，ノイズを付加された説明を入力とした際の学習に失敗して 0 を多く出力することで見かけ上の攻撃成功率が上昇していると言える．

この仮説を検証するため，$\epsilon = 10, \infty$ の場合における攻撃モデルの ROC 曲線を図 8，9 に示す． 図 9 において断続的に変化している部分は，ある値に攻撃モデル ψ の予測値が集中しているため，そこをまたぐように閾値を変化させた際に大きく変化していると考えられる．

5. 評価

5.1 LIME との安全性比較

LIME に対する攻撃成功率と DPGD-Explain に対する

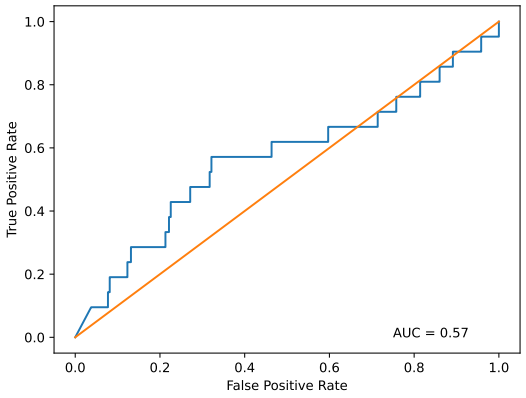

図 8 $\epsilon = 10$ のときの攻撃モデル ψ の予測に対する ROC 曲線の例

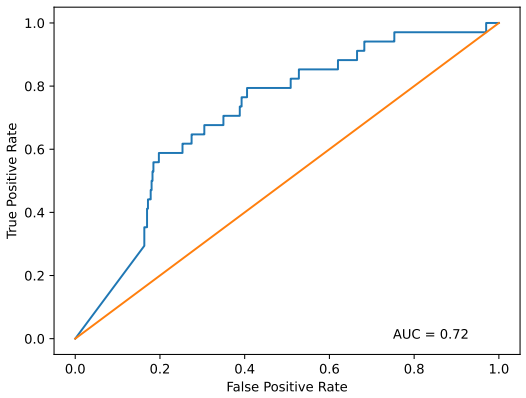

図 9 $\epsilon = \infty$ のときの攻撃モデル ψ の予測に対する ROC 曲線の例

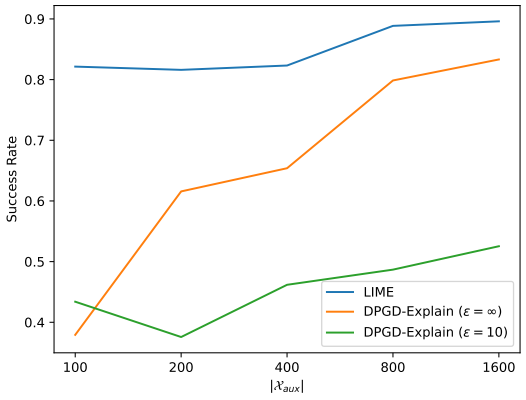

図 10 攻撃者の持つ補助データセットの大きさ $\mathcal{X}_{aux}$ に対する LIME と DPGD-Explain の攻撃成功率

攻撃成功率を比較したものを図 10 に示す．$|\mathcal{X}_{aux}| = 100$ のときは LIME の方が脆弱であるが，$|\mathcal{X}_{aux}| \geq 200$ のとき，ノイズなしの DPGD-Explain が最も脆弱であり，ノイズを付加することで LIME よりも安全になると言える．

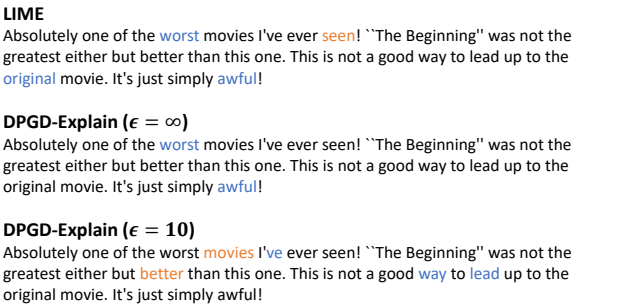

図 11 それぞれの説明において，正の影響を及ぼすと説明される単語を橙色，負の影響を及ぼすと説明される単語を青色で示した例

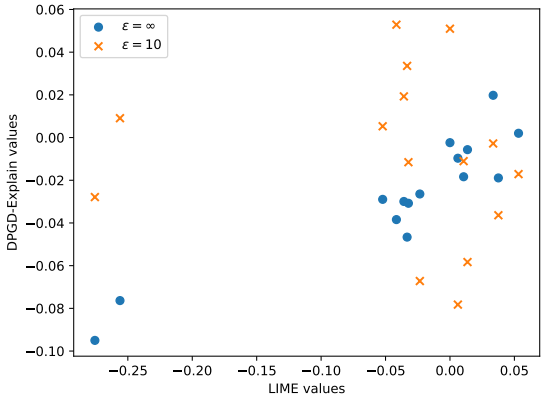

図 12 LIME と DPGD-Explain ($\epsilon = 10, \infty$) における入力に含まれる単語の説明値

5.2 DPGD-Explain の説明可能性

LIME と比較した DPGD-Explain の説明の違いを表 11 に示す．入力クエリは負に分類される文章であるため，特に影響を及ぼしているであろう “awful” と “worst” の 2 単語に注目すると，LIME とノイズなしの DPGD-Explain ではどちらも負の説明値が得られているが，ノイズありの DPGD-Explain では上手く説明できていない．また, LIME と DPGD-Explain における入力クエリ中に含まれる単語ごとの説明値の散布図を図 12 に示す．LIME と $\epsilon = \infty$ における DPGD-Explain には強い相関が見られるが，$\epsilon = 10$ のときは相関が弱いことが分かる．

6. おわりに

本研究では，差分プライバシーを保証したモデル説明 DPGD-Explain について安全性と有用性の 2 つの観点から実験評価を行った．DPGD-Explain による説明ベクトルは，プライバシー予算 ϵ が大きくなるほど $\epsilon = \infty$ における真の説明ベクトルに近づいた．また，入力クエリに対する説明ベクトルの例では，LIME とノイズなしの DPGD-Explain では似た説明が得られたが，ノイズありの DPGD-Explain では上手く説明できていなかった．一方

で，DPGD-Explain のレコード再構築攻撃に対する安全性は，$\epsilon = \infty$ のとき LIME より脆弱であるが，ノイズを付加することで LIME より安全になった．

今後の課題として，説明の有用性を保ちながら安全性を高めることが挙げられる．特に，DPGD-Explain のような勾配降下法を用いる手法はステップ数に応じてプライバシー損失が累積するため，勾配降下法を用いず差分プライバシーを保証した XAI 手法を提案することで，より説明の有用性と安全性を改善できると考えられる．

参考文献

[1] Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence **1**(5), 206–215 (2019)

[2] Sakai, A., Komatsu, M., Komatsu, R., Matsuoka, R., Yasutomi, S., Dozen, A., Shozu, K., Arakaki, T., Machino, H., Asada, K., Kaneko, S., Sekizawa, A., Hamamoto, R.: Medical Professional Enhancement Using Explainable Artificial Intelligence in Fetal Cardiac Ultrasound Screening. Biomedicines 2022 **10**(3), 551 (2022)

[3] Chen, J., Song, L., Wainwright, M., Jordan, M.: Learning to Explain: An Information-Theoretic Perspective on Model Interpretation. In: 35th International Conference on Machine Learning, pp. 882–891. PMLR 80, Stockholm, Sweden (2018)

[4] Amazon SageMaker Documentation. `https://docs.aws.amazon.com/sagemaker/latest/dg/clarify-model-explainability.html`, last accessed 2024/04/19

[5] Azure Machine Learning Documentation. `https://learn.microsoft.com/en-us/azure/machine-learning/how-to-machine-learning-interpretability`, last accessed 2024/04/19

[6] Luo, X., Jiang, Y., Xiao, X.: Feature Inference Attack on Shapley Values. In: 2022 ACM SIGSAC Conference on Computer and Communications Security (CCS '22), pp. 2233–2247. Association for Computing Machinery, Los Angeles, CA, USA (2022)

[7] Ribeiro, M., Singh, S., Guestrin, C.: "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In: 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16), pp. 1135–1144. Association for Computing Machinery, San Francisco, California, USA (2016)

[8] Becker, B., Kohavi, R.: Adult. UCI Machine Learning Repository. (1996). DOI: 10.24432/C5XW20

[9] Maas, A., Daly, R., Pham, P., Huang, D., Ng, A., Potts, C.: Learning Word Vectors for Sentiment Analysis. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pp. 142–150. Association for Computational Linguistics, Portland, Oregon, USA (2011)

[10] Shapley, L.: 17. A Value for n-Person Games. Contributions to the Theory of Games (AM-28) **II**, 307–318 (1953)

[11] Lundberg, S. Lee, S.: A Unified Approach to Interpreting Model Predictions. In: 31st International Conference on Neural Information Processing Systems (NIPS'17), pp. 4768–4777. Curran Associates Inc., Long Beach, California, USA (2017)

[12] Štrumbelj, E., Kononenko, I.: Explaining prediction models and individual predictions with feature contributions. Knowl. Inf. Syst. **41**, 647–665 (2016). DOI: 10.1007/s10115-013-0679-x

[13] Covert, I., Lundberg, S., Lee, S.: Understanding Global Feature Contributions With Additive Importance Measures. In: 34th International Conference on Neural Information Processing Systems (NIPS '20), pp. 17212–17223. Curran Associates Inc., Vancouver, BC, Canada (2020)

[14] Fisher, A., Rudin, C., Dominici, F.: All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. Journal of Machine Learning Research **20**(177), 1–81 (2019)

[15] Ribeiro, M., Singh, S., Guestrin, C.: Anchors: High-Precision Model-Agnostic Explanations. In: Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence (AAAI'18/IAAI'18/EAAI'18), pp. 1527–1535. AAAI Press, New Orleans, Louisiana, USA (2018)

[16] Introduction to Vertex Explainable AI, https://cloud.google.com/vertex-ai/docs/explainable-ai/overview, last accessed 2024/04/19

[17] Kuppa, A., Le-Khac, N.: Adversarial XAI Methods in Cybersecurity. IEEE Transactions on Information Forensics and Security **16**, 4924–4938 (2021)

[18] Shokri, R., Strobel, M., Zick, Y.: On the Privacy Risks of Model Explanations. In: 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21), pp. 231–241. Association for Computing Machinery, Virtual Event, USA (2021)

[19] Liu, H., Wu, Y., Yu, Z., Zhang, N.: Please Tell Me More: Privacy Impact of Explainability through the Lens of Membership Inference Attack. In: 2024 IEEE Symposium on Security and Privacy (SP), pp. 119–138. IEEE Computer Society, San Francisco, CA, USA (2024)

[20] Yan, A., Hou, R., Liu, X., Yan, H., Huang, T., Wang, X.: Towards explainable model extraction attacks. International Journal of Intelligent Systems **37**(11), 9936–9956 (2022)

[21] Yan, A., Huang, T., Ke, L., Liu, X., Chen, Q., Dong, C.: Explanation leaks: Explanation-guided model extraction attacks. Information Sciences: an International Journal **632**(C), 269–284 (2023)

[22] Fredrikson, M., Jha, S., Ristenpart, T.: Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures. In: 22nd ACM SIGSAC Conference on Computer and Communications Security (CCS '15), 1322–1333. Association for Computing Machinery, Denver, Colorado, USA (2015)

[23] Baniecki, H., Biecek, P.: Adversarial attacks and defenses in explainable artificial intelligence: A survey. Information Fusion **107**, 102303 (2024)

[24] Patel, N., Shokri, R., Zick, Y.: Model Explanations with Differential Privacy. In: 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), pp. 1895–1904. Association for Computing Machinery, Seoul, Republic of Korea (2022)

[25] Nguyen, T., Lai, P., Phan, H., Thai, M.: XRand: Differentially Private Defense against Explanation-Guided Attacks. Proceedings of the AAAI Conference on Artificial Intelligence **37**(10), 11873–11881 (2023)

[26] Bozorgpanah, A., Torra, V., Aliahmadipour, L.: Privacy and Explainability: The Effects of Data Protection on Shapley Values. Technologies **10**(6), 125 (2022)

[27] Shlomo, N.: Integrating Differential Privacy in the Statistical Disclosure Control Tool-Kit for Synthetic Data Production. In: Domingo-Ferrer, J., Muralidhar, K. (eds) Privacy in Statistical Databases (PSD 2020), LNCS, vol. 12276, pp. 271–280. Springer, Cham. (2020). DOI: 10.1007/978-3-030-57521-2_19

[28] Wang, G., Gehrke, J. Xiao, X.: Differential Privacy via Wavelet Transforms. IEEE Transactions on Knowledge & Data Engineering **23**(8), 1200–1214 (2011)

[29] Ito, S., Miura, T., Akatsuka, H., Terada, M.: Differential Privacy and Its Applicability for Official Statistics in Japan - A Comparative Study Using Small Area Data from the Japanese Population Census. In: Domingo-Ferrer, J., Muralidhar, K. (eds) Privacy in Statistical Databases (PSD 2020), LNCS, vol. 12276, pp. 337–352. Springer, Cham. (2020). DOI: 10.1007/978-3-030-57521-2_24

[30] Slokom, M., Wolf, P., Larson, M.: When Machine Learning Models Leak: An Exploration of Synthetic Training Data. In: Domingo-Ferrer, J., Laurent, M. (eds.) Privacy in Statistical Databases (PSD 2022), LNCS, vol. 13463, pp. 283–296. Springer, Cham (2022). DOI: 10.1007/978-3-031-13945-1_20

[31] Tritscher, J., Ring, M., Schlr, D., Hettinger, L., Hotho, A.: Evaluation of Post-hoc XAI Approaches Through Synthetic Tabular Data. In: Helic, D., Leitner, G., Stettinger, M., Felfernig, A., Raś, Z.W. (eds) Foundations of Intelligent Systems (ISMIS 2020), LNCS, vol 12117, pp. 422–430. Springer, Cham. (2020). DOI: 10.1007/978-3-030-59491-6_40