F-045

異常検知 AI と XAI を用いた表形式合成データに対する品質評価手法の検討
A Study on Operational-Grade Assessment Methods for Synthetic Tabular Data Using Anomaly Detection AI and XAI

宮野 咲紀† 関 弘翔† 辻 泰弘† 細野 裕行†
Saki Miyano Hiroto Seki Yasuhiro Tsuji Hiroyuki Hosono

1. はじめに

近年，生成モデルを用いた表形式データ生成が多く研究されている一方で，金融や医療分野などで用いられる表形式データの合成における評価指標が課題となっている．表形式合成データの評価には，主に類似性・有用性・プライバシー保護の 3 点が用いられている．これらの指標は，生成モデルが学習データにどれだけ類似したデータを生成できるかに重点を置いている．しかし，生成された個々のデータが実用に耐えうる品質であるかについては，明確な評価がなされていない．特に，医療データを評価する際には，属性の組み合わせが医学的に妥当であるかが重要だが，その点に明確な尺度を設けることは困難である．

本稿では，合成表形式データ自体の評価をするため，特徴抽出器と異常検知 AI，Explainable AI（XAI：説明可能な AI）を組み合わせた新たな評価手法を構築し，従来手法との比較検証を行った．

2. 方法

2.1 提案手法

図 1 に本報告で提案する評価方法の概要図を示す．

本手法では，本物のデータを用いて異常検知 AI を学習し，合成データ全体に対する異常度の平均値を評価指標とする．ここで言う異常度は，各データに対して異常である確率を算出している．よって，評価値が小さいほど，正常，すなわち本物に類似していると判定され，高品質とみなされる．この時，表形式データをそのまま入力とした場合，属性間の関係性や組み合わせに着目した評価にならない可能性がある．そこで，特徴抽出器により抽出した合成データの特徴を異常検知手法の入力にすることで問題点の解決になると考えた．また，異常度が高いデータに対して何を根拠に異常と判断されたかを判別できるように，XAI による可視化を導入した．

本稿では，特徴抽出器として，深層学習に決定木の考えを取り入れ，表形式データの特徴を解釈可能な状態で抽出できる TabNet[1]を採用した．また，異常検知 AI の一種で決定木の概念を用いて異常度の算出が可能な Isolation Forest（IF）[2]を使用した．さらに，XAI には協力ゲーム理論の Shapley 値を機械学習に応用し，モデルが出力する予測値に対して各属性の貢献度を確認可能な SHapley Additive exPlanations（SHAP）[3]を採用した．

2.2 比較手法

提案する評価手法と比較するための評価指標には SynMeter[4]に搭載されている忠実性評価を採用した．

†日本大学 Nihon University

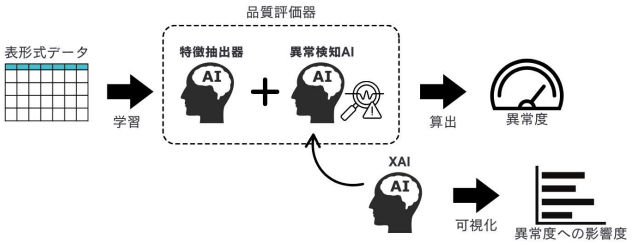

図 1 提案する評価方法の概要

表 1 Adult データ属性（一部）

属性名	データ型	内容
fnlwgt	連続（int）	サンプリング重み（その人が何人分を代表しているか）
education	カテゴリ	学歴（Bachelors, HS-grad, Some-college など）
occupation	カテゴリ	職業（Tech-support, Craft-repair, Sales など）
relationship	カテゴリ	家庭状況（Husband, Wife, Not-in-family など）
race	カテゴリ	人種（White, Black, Asian-Pac-Islander など）
capital-gain	連続（int）	資本利益
capital-loss	連続（int）	資本損失
native-country	カテゴリ	出身国（United-States, Mexico, Philippines など）

SynMeter は，表形式データ生成モデルを体系的に評価するためのフレームワークであり，10 種類の生成モデル，12 種類のデータセット，3 種類の評価指標が実装されている．

本稿で使用する忠実性評価は実データと合成データ間の統計的類似性を評価するものである．この評価は連続値に対して Wasserstein 距離を算出し，カテゴリ値に対しては実データと合成データそれぞれの条件付確率を算出し，それらの平均値が最終的な評価として出力される．つまり，この評価は値が小さいほど高評価であると言える．

2.3 Dataset

検証に用いる Dataset として，Adult Income Dataset を採用する．Adult は UCI Machine Learning Repository にて公開されており，労働時間や出身国，年練，職業などから年収を予測する分類タスク用のデータセットである．このデータセットは連続値 6 属性，カテゴリ値 9 属性の合計 15 属性で構成されている．表 1 にデータを構成する属性の一部を示す．

2.4 生成モデル

表形式合成データを生成する生成モデルとして，Tabular Variational Autoencoder （TVAE） [5]， Conditional Tabular Generative Adversarial Networks （CTGAN） [6]， Tabular Denoising Diffusion Probabilistic Model（TabDDPM）[7]の 3 種を使用した．

TVAE は Variational Autoencoder（VAE）[8]に基づく生成モデルで，Encoder と Decoder，2 つのニューラルネットワークを用いて潜在変数を介して生成を行う手法である．

CTGAN は Generative Adversarial Networks（GAN）[9]に基づき，条件付き生成を行うことで不均衡なカテゴリ属性に対しても対応が可能な表形式生成モデルである．

TabDDPM は Diffusion Model（拡散モデル）[10]を応用し，ノイズと実データ間の関係を学習することで高品質な表形式生成を可能としたモデルである．

2.5 検証方法

本稿では，Adult データセットを学習し合成データを生成した後に，SynMeter の忠実性評価と，提案した評価方法それぞれ評価を行い，結果の比較を行う．この比較において，重要な点は各生成モデルの良し悪しが 2 つの評価方法で変動しないかである．評価結果に変動がない場合は，提案手法も忠実性評価として有効であると判断できる．一方で，異なる傾向が見られる場合は，評価指標としての妥当性に疑問が生じる可能性がある．なお，今回正確に評価を行うために，1 つの生成モデルにつき 10 回合成データ生成および評価を行い，10 回平均と標準偏差を算出し，生成モデルの評価とした．

また，提案手法が属性間の関係性や組み合わせも考慮した評価が可能になっているのかの検証が必要である．そこで，提案手法で最も評価が良い生成モデルの合成データの中で異常度が高いデータに対して SHAP を適用し，実際に異常度が高いデータは属性の組み合わせも異常であるのかを確認する．

3. 結果考察

表 2 に各生成モデルに対する 2 つの評価方法による評価結果を示す．表 2 の忠実性評価より，最も実データと類似した合成データを生成したのは TabDDPM であり，次点で TVAE であった．ここで提案手法の評価を確認すると，最良の生成モデルは TabDDPM，次点で TVAE となっており，忠実性評価と一致した．以上のことから，提案手法は忠実性評価と同様に評価指標として有効であることが確認できた．

続いて，最も高評価であった TabDDPM が生成した合成データの中で異常度が高かったデータに対して，SHAP を用いて異常判定の根拠を可視化した．図 2 に異常度根拠を可視化した結果を示す．この結果は，異常度への影響が大きい属性ほど上位に表示され，正の寄与が大きいほど異常度を高くし，負の寄与が大きいほど異常度を低くすることを示している．

図 2 より，この合成データは家庭状況，出身国，学歴などに影響を受けていることがわかる．このデータの内容を確認すると，アメリカ人の地方公務員で修士課程を修了した既婚白人女性のデータであった．このデータでは教育歴

表 2　合成データの評価結果

	忠実性評価 ↓	提案手法 ↓
TVAE	0.062±0.033	0.491±0.031
CTGAN	0.077±0.057	0.611±0.062
TabDDPM	0.036±0.012	0.396±0.003

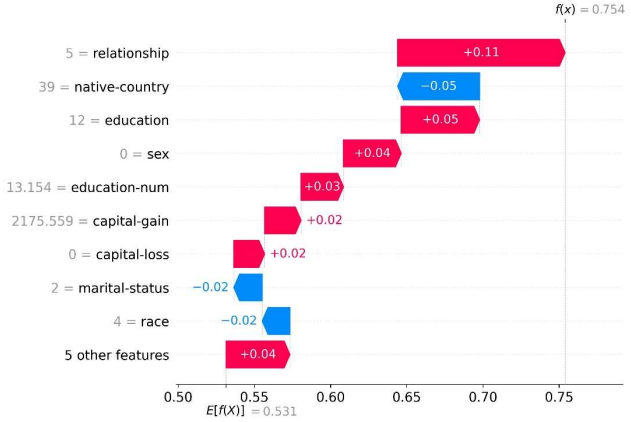

図 2　TabDDPM のあるデータに対する異常度根拠可視化結果

が 13.15 である一方，education（最終学歴）が 12（修士課程修了）となっていた．通常，修士課程修了は education-num（教育歴）が 14 となるはずであるため，属性の組み合わせに矛盾があることが確認できた．また，図 2 の影響度を確認すると教育歴と最終学歴は影響度が上位 5 項目に含まれており，この組み合わせの異常が異常度を高めた要因であることが示唆される．以上より，提案手法は属性の組み合わせに起因する異常を適切に捉えられる可能性が示唆された．

4. おわりに

既存の生成モデルの評価指標では，合成表形式データ自体の品質を直接評価できず，属性間の関係性や組み合わせを考慮できないという課題がある．この問題に対して，特徴抽出器と異常検知手法，Explainable AI（XAI：説明可能な AI）を組み合わせた新たな評価手法を構築し，従来手法との比較検証を行った．その結果，提案手法は合成データの組み合わせの異常を反映した評価指標であることが示唆された．

参考文献

[1] S. O. Arik, T. Pfister, Proceedings of the AAAI conference on artificial intelligence, Vol. 35, No. 8, pp.1-8, 2021.
[2] F. T. Liu, K. M. Ting, Z. Zhou, 2008 eighth ieee international conference on data mining, IEEE, pp.413-422, 2008.
[3] S. M. Lundberg and Lee Su-In, NeurIPS 2017, Vol. 30, 2017.
[4] Y.Du and N. Li, arXiv Preprint, 2402.06806, 2024.
[5] V. Borisov, et al., ICLR 2023, 2023.
[6] L. Xu, et al., NeurIPS 2019, 2019.
[7] A. Kotelnikov, et al., PMLR 202:17564-17579, 2023.
[8] D.P. Kingma, M. Welling, ICLR, 2014.
[9] Goodfellow, Ian, et al., Communications of the ACM, 2020.
[10] J. Sohl-Dickstein, et al., PMLR 37:2256-2265, 2015.