GANによる変数間相関の有無を仮定した欠測値補完の性能評価

西本尚樹[1)], 吉永 和美[1)], 袴田 遥[2)], 北山 香織理[1)], 菅原 亮平[1)], 阿部 涼[1)], 伊藤 陽一[1)]

1) 北海道大学病院 医療・ヘルスサイエンス研究開発機構 データサイエンスセンター

1．背景および目的

リアルワールドデータ（RWD）は、臨床現場の実態を反映する貴重な情報源として活用が進む一方、観測プロセスに由来するデータの欠損が不可避的に発生し、統計的推定の妥当性や解析精度を大きく損なう点が深刻な課題となっている。従来用いられてきた平均値補完や多重代入法などの補完手法は、変数間に存在する複雑な非線形関係や高次元な相関構造を十分に捉えきれない限界があった。

これに対し、Goodfellowら（2014）によって提唱された敵対的生成ネットワーク（GAN）[1-3]は、データ生成を担う生成器（Generator）と真偽判定を担う識別器（Discriminator）の minimax ゲームを介して、データ背後に潜む真の多変量確率分布を学習・再現できる枠組みである。

この対照学習メカニズムを欠損値補完に応用することで、多変量間の依存関係を保持した高度な補完が可能になると期待される。本研究では、PyTorchを用いてGANベースの欠損値補完プログラムを構築・実装し、データ内の変数間相関の強弱や欠損率が補完精度に与える影響および本手法の適用限界を定量的に評価した。

2．方法

1.シミュレーションデータセット

【データ生成条件】

基本構造: サンプル数 N=1,000、特徴量数 D=5（各値は[0,1] に正規化）

欠損条件: 20%の完全ランダム欠損（MCAR）を疑似生成

実験条件（2パターンのデータ構造）：

無相関データ: 各変数を独立した一様分布 U(0,1) から生成。

強相関データ: 共通の基盤傾向（Base）に対し、同一、線形（比例・反比例）、非線形（2乗）の依存関係と僅かなガウスノイズを加味して生成。

再帰性の担保: シード値を固定

2. 補完アルゴリズム

補完アルゴリズムには欠測パターンと変数間の相関関係を同時学習する手法を採用。1000人分5変数のデータ（相関なし／強い相関あり）に対し、5～20%の疑似欠損を生成した。GAN、ランダム補完、MissForestの性能をRMSEで比較した。

Pythonコード：

```
#  GANを使って欠損部のみを生成値で置き換える
# 1. 生成器（Generator）における入力合成と推論
class Generator(nn.Module):
    def forward(self, x, m):
        noise = torch.rand_like(x)
        # 観測値(m=1)は維持し、欠損箇所(m=0)にノイズを挿入してマスクと結合
        x_in = x * m + noise * (1 - m)
        inputs = torch.cat([x_in, m], dim=1)
        return self.net(inputs)  # 補完値の予測を出力

# 2. 補完データ（X_hat）の合成処理
# 観測値(batch_X * batch_M) と Gの予測値(G_sample * (1 - batch_M)) を合成
X_hat = batch_X * batch_M + G_sample * (1 - batch_M)
```

3. 解析環境

OS：Windows11 professional edition

プログラミング言語：

Python 3.13.14

PyTorch

Scala 3.3.3

sbt 1.12.1

Java 17.0.12(Oracle)

評価指標

$$RMSE = \sqrt{\frac{1}{N}(y-y)^2}$$

3．結果

欠損率20%・相関なしの場合、ランダム補完のRMSE（0.4021）に対しGANは0.4213と劣る結果となった。一方、同条件で強い相関がある場合、ランダム補完の0.4070に対しGANは0.2912とRMSEが小さい値となった。

表 1: データ構造（相関係数）および補完手法別の精度比較（RMSE）

データ条件	補完手法	RMSE	ベースライン (Random)
無相関データ (|r| ≈0)	GAN	0.4213	0.4021
	Diffusion	0.5350	0.4203
強相関データ (|r| ≈0.96~ 0.99)	GAN	0.2912(大幅改善)	0.4070

表 2: 欠損行における各変数のデータ値とGANによる補完結果の例

データ種別	変数 1	変数 2	変数 3	変数 4	変数 5 (欠損箇所)
元のデータ（欠損前）	0.0204	0.0126	0.0416	0.9724	0.0240
欠損データ（ NaN 化）	0.0204	0.0126	0.0416	0.9724	NaN
GAN による補完データ	0.0204	0.0126	0.0416	0.9724	0.0061

4．考察

本研究の結果は、GANを応用した欠損値補完モデルの有効性が「入力データ内に存在する変数間の相互依存関係の強さ」に大きく依存することを示している。相関がないデータでは、識別器（Discriminator）が真偽を判定する過程で本質的でないノイズを捉え、生成器（Generator）が過剰適合した不適切な値を補完したため精度が悪化したと考えられる。一方で、強相関データにおいては、生成器が多次元的な分布と相関構造を正確に学習し、欠損値を高精度に再現可能であることが証明された。

リアルワールドデータは一般に、同一患者の経時データや類似の生理学的指標など、強い共分散構造や相関関係を含んでいることが多い。そのため、本プログラムはRWDの解析精度を向上させる強力な手法となり得る。しかし、全変数が相関を持つとは限らず、一律の適用は逆効果となるリスクを孕む。したがって実務応用においては、事前探索（EDA）により相関構造の存在を十分に吟味した上でモデルを選択・調整するプロトコルの確立が不可欠である。今後は、多変量かつ複雑な非線形関係やカテゴリカル変数が混在するデータセットでの実証が課題となる。

5．結論

GANは変数間の相関を学習することで高い補完性能を発揮するが、その真価を引き出すにはデータ内の関係性を事前に吟味することが重要であると示唆された。

1. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in Advances in Neural Information Processing Systems 27 (NIPS 2014), 2014,. pp. 2672–2680.

2. Yoon, J., Jordon, J., & van der Schaar, M. (2018). GAIN: Missing Data Imputation using Generative Adversarial Nets. International Conference on Machine Learning (ICML), 5642-5651.

3.Wells, B. J., Chagin, K. M., Nowacki, A. S., & Kattan, M. W. (2013). Strategies for handling missing data in electronic health record-based research. Journal of the American Medical Informatics Association (JAMIA), 20(e2), e264-e270.