マスク言語モデルを参考にした表形式データの欠損値補完手法の提案

村岡　拓弥†　　木村　昌臣††

† †† 芝浦工業大学工学部情報工学科　〒 135-8548 東京都江東区豊洲 3-7-5
†† Universiti Malaysia Kelantan　Pengkalan Chepa,16100 Kota Bharu,Kelantan. Malaysia
E-mail: † al22027@shibaura-it.ac.jp, †† masaomi@shibaura-it.ac.jp

あらまし　近年，医療や金融など多岐にわたる分野において，表形式データの収集や活用が進んでいる．しかし，実社会で得られるデータには，機器の故障や入力ミスなどの理由により，欠損値が含まれることが極めて多く，これがデータ分析や機械学習の予測においてバイアスを招いている．従来の機械学習的手法は強力だが，逐次処理による誤差の伝播と高次の相互作用に課題がある．そのため，特徴量間の複雑な非線形な相互作用を捉えられる能力には限界がある．本研究では，全特徴量の相互作用を考慮した Transformer ベースの欠損値補完手法を提案する．本手法は，数値データの埋め込み時に単一の値だけでなく全特徴量の情報を統合する Global Input Tokenizer を導入し，マスク言語モデルを回帰問題に応用して学習を行う点に特徴がある．これにより，特徴量間の非線形的な関係を学習し，既存の統計手法や深層学習モデルより高い精度で補完を実現することを目指す．

キーワード　機械学習，欠損値補完，表形式データ，Transformer

1　研究背景と目的

近年，医療や金融，マーケティングなど多岐にわたる分野において，表形式データの収集と活用が進んでいる．しかし，実社会で得られるデータには，機器の故障や入力ミスなどの理由により，欠損値が含まれることが極めて多い．欠損値の存在はデータ分析や機械学習の予測においてバイアスを招く可能性がある．従来の欠損値補完には統計的手法や決定木系が広く用いられてきた．強力な手法ではあるが，逐次的に欠損値を埋めているため誤った情報が伝播し学習してしまう．また，複数の特徴量の相互作用を捉えるとき，木を深くする必要がある．しかし，木を深くすればするほど，分割される末端のノードに含まれるデータ数は激減し，統計的な信頼がなくなる．そのため，特徴量間の複雑な非線形な相互作用を捉えられる能力には限界がある．表形式データにおける欠損値補完は，観測済みの他の特徴量の関係性を用いて，欠損している値を予測するタスクと見なすことができる．これは文中の単語をその周辺単語から予測するマスク言語モデルのタスクに類似している．

そこで本研究では, マスク言語モデルを参考にした Transformer の欠損値補完手法の提案を行う．

2　先 行 研 究

欠損値補完に関する従来研究のうち，統計的アプローチや古典的な機械学習手法として，MICE [1] や MissForest [2]，KNN-Impute [3] がある．Van らによる MICE は，各特徴量を他の変数を用いた回帰モデルで連鎖的に予測・更新する多重代入法であるが，変数間の複雑な非線形関係に限界がある．Stekhoven らによる MissForest は，ランダムフォレストを用いることで非線形な相互作用を柔軟に扱え，Deep Learning 手法に匹敵する精度を示す．しかし，木を深くすればするほど，分割される末端のノードに含まれるデータ数は減少し，統計的な信頼がなくなる．また，Troyanskaya らによる KNN-Impute は，サンプル間の距離計算が必要なため，大規模データへの適用において計算効率が課題である．次に，Deep Learning を用いた生成モデルである．Yoon ら [4] が提案した GAIN は，GAN を応用し，Generator と Discriminator の敵対的学習を通して欠損値を生成する．しかし，観測値を使用しないため正しい分布に必ずしも近づくか分からないという課題が残る．Mattei ら [5] による MIWAE は，VAE は再構成誤差の最小化とガウス分布への近似を同時に行うため，生成されるデータや補完値が平均的な値によりやすく，表現力が制限されることが課題である．

近年では，表形式データへの Transformer の活用が進展している．Huang ら [6] による TabTransformer は数値データを含む表形式データへの活用をした手法である．しかし，カテゴリ変数は Transformer に入力するが，数値データは Attention 機構に入力されないため，データ全体の相互作用を捉えられない課題があった．これに対し, Gorishniy ら [7] が提案した FT-Transformer は，Feature Tokenizer を導入し数値データもトークン化することで，全特徴量を Transformer で処理することを可能にした．本研究の埋め込み手法はこの考え方を基礎としている．さらに，Somepalli ら [8] は行間の Attention や対照学習を取り入れ，Du ら [9] の ReMarker は，Masked Autoencoder の枠組みを表形式データに適用した手法である．入力の特徴量の一部をランダムにマスクし，残りの観測情報からマスクされた値を予測する自己教師あり学習を行うことで，特徴量間の複雑な依存関係を捉え，欠損値補完に応用した．

表 1: 実験に使用したデータセットの概要

データセット	N	F	クラス	領域	概要
Breast	569	30	2	医療	細胞核の特徴に基づく乳がん診断データ
Spam	4,601	57	2	テキスト	メールの単語頻度に基づくスパム検知
Letter	20,000	16	26	画像	画像から抽出されたアルファベットの統計的特徴
Credit	30,000	23	2	金融	顧客属性と支払い履歴に基づく債務不履行予測
Pima	768	8	2	医療	身体測定値に基づく糖尿病の発症予測

3 提案手法

3.1 提案手法の概要

提案モデルの全体像を図 1 に示す．本モデルは，数値データからなる表形式データを入力とする．従来の FTTransformer では，各特徴量を独立して埋め込みベクトルに変換したが，Global Input Tokenizer では全特徴量の情報を統合して特徴量の埋め込みを生成する点が特徴である．埋め込まれたトークン列は Transformer Encoder に入力され，Attention 機構によって特徴量間の複雑な相互作用が学習される．学習時には入力の一部を欠損させ，その値を予測することでモデルを最適化する．

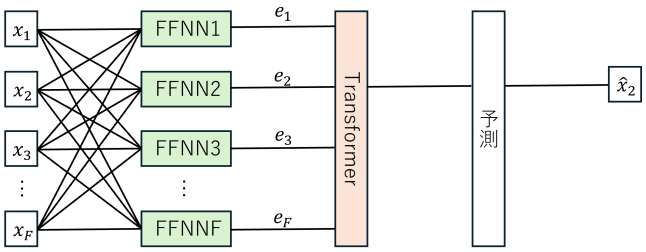

図 1: 提案モデルの全体構造

3.2 問題の定式化

データセット $\mathcal{D} = \{x^{(n)}\}_{n=1}^{N}$ を考える．ここで，N はサンプル数であり，各サンプル $x \in \mathbb{R}^F$ は F 個の数値特徴量からなるベクトル $x = [x_1, x_2, \ldots, x_F]$ である．欠損値補完において，入力ベクトル x の一部は欠損しており観測されない．本研究では，欠損箇所を示すマスクベクトル $m \in \{0,1\}^F$ を導入する ($m_i = 1$ ならば欠損，$m_i = 0$ ならば観測)．目的は，観測されたデータ x^{obs} を用いて，欠損している値 x^{miss} を正確に推定するモデル $\mathcal{M}$ を構築することである．

3.3 全特徴量を考慮した Global Input Tokenizer

従来手法である FT-Transformer は，第 i 番目の特徴量 x_i の埋め込みベクトル e_i は，自身の値のみに依存して生成される．つまり，以下の様に定義される．

$$e_i = f_i(x_i) \tag{1}$$

各特徴量を独立したトークンとして扱う点では自然言語処理の単語埋め込みに近いが，表形式データにおいて重要な変数の相互作用は，Transformer 層のみで学習されることになる．

本研究では，Global Input Tokenizer を提案する．これは，個々の特徴量の埋め込みを生成する際，単一の値ではなく，入力ベクトル全体 x を入力する．各特徴量 i に対応するフィードフォワードニューラルネットワーク $FFNN_i : \mathbb{R}^F \to \mathbb{R}^d$ を定義し，埋め込みベクトル $e_i \in \mathbb{R}^d$ を次のように定義する．

$$e_i = FFNN_i(x) \tag{2}$$

ここで，d は埋め込みの次元数である．この構造により，モデルは埋め込みの初期段階から他の変数の相互関係を利用することが可能となる．最終的に，全特徴量の埋め込みを結合し，埋め込み行列 $E \in \mathbb{R}^{F \times d}$ を得る．

$$E = [e_1, e_2, \ldots, e_F] \tag{3}$$

3.4 Self-Attention 機構

Self-Attention 機構は，特徴量間の重要度を動的に計算する機構である．入力 E に対して $Query(Q)$ と $Key(K)$，$Value(V)$ 行列を生成し，以下の式で Attention スコアを算出する．

$$Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d}})V \tag{4}$$

3.5 マスク言語モデルによる学習

学習時，入力サンプル x の各要素に対し，一定の確率 p_{mask} でマスク処理を行う．マスクされた要素の集合を M とする．マスクされた値 $\tilde{x}_i (i \in M)$ は，以下のルールに従って置換する．

- Fixed Value Replacement: 固定値に置換
- Random Replacement: データセット内のランダムな値に置換
- Original Value: 観測値を利用

3.6 損失関数

欠損値補完は真の値と予測値の差を最小化する回帰問題として定式化できる．Transformer Encoder の最終層の出力 $E^{(L)}$ から予測ヘッドを用いて予測値 $\hat{x}_i$ とする．

$$\hat{x}_i = F(E_i^{(L)}) \tag{5}$$

損失関数 $\mathcal{L}$ には，マスクされた集合 M に含まれる特徴量に対してのみ平均二乗誤差または平均絶対値誤差を計算し，モデルのパラメータを更新する．ここで，$|M|$ は集合 M の要素の総数のことである．

$$\mathcal{L} = \frac{1}{|M|}\sum_{i \in M}(x_i - \hat{x}_i) \quad \text{または} \quad \mathcal{L} = \frac{1}{|M|}\sum_{i \in M}|x_i - \hat{x}_i| \tag{6}$$

推論時にデータセット内で欠損している箇所を Fixed Value Replacement の値を入力し，予測値を補完値とする．

4 評価実験と結果

4.1 実験に使用したデータセット

本実験では，UCI Machine Learning Repository より収集した特性の異なる 5 つのデータセットを用いた．これらのデータセットは，サンプル数，特徴量数，データの領域が多岐にわたっており，多様な環境下での補完性能を評価するのに適している．各データセットの基本統計量を表 1 に示す．

4.2 目　　的

本章では，提案手法の有効性と汎用性を検証するために，数値データの欠損値補完実験を行う．欠損方法 [10] は MCAR と MAR，MNAR を仮定し，各データセットに対して人工的に欠損を生成した．補完精度の評価指標には，真の値と補完値の RMSE(二乗平均平方根誤差) を用いた．N_{miss} を欠損値の総数とすると，以下のように定義した．

$$RMSE = \sqrt{\frac{1}{N_{miss}} \sum_{n=1}^{N} \sum_{j=1}^{F} m_{n,j}(x_{n,j} - \hat{x}_{n,j})^2} \tag{7}$$

ここで，N はデータセットのサンプル数，F は特徴量の数を表す．$m_{n,j}$ は，n 番目のサンプルの j 番目の特徴量が欠損している場合に 1，そうでない場合に 0 をとるマスクである．また，$x_{n,j}$ は欠損前の真の値，$\hat{x}_{n,j}$ はモデルによって予測された補完値を表す．

実験の目的は，異なるデータの特性や欠損方法，欠損率において，提案手法が既存手法と比較して優れた RMSE を示すことを確認することである．

4.3 MCAR

欠損が変数の値や他の変数に依存せず，完全にランダムに発生する MCAR(Missing Completely At Random) を仮定した実験を行った．最も基本的な欠損パターンであり，モデルの基礎的な補完能力を測る指標となる．

本実験では，学習時に $p_{mask} = 0.2$ でマスク処理を行った．置換のそれぞれの比率を，以下の表 2 に示す．

表 2: 学習時におけるマスク置換の比率設定

データセット	Mask Replacement	Random Replacement	Original Value
Breast	0	0	10
Spam	0	0	10
Letter	7	2	1
Credit	0	0	10
Pima	0	0	10

各データセットに対して，MCAR による欠損を発生させた際の補完精度の比較結果を表 3 に示す．

表 3: 各手法における補完精度の比較 (MCAR)

手法	Breast	Spam	Letter	Credit	Pima
GAIN	5.46×10^{-2}	5.13×10^{-2}	$\mathbf{1.20 \times 10^{-1}}$	1.86×10^{-1}	–
MissForest	6.08×10^{-2}	5.53×10^{-2}	1.61×10^{-1}	1.98×10^{-1}	–
提案手法	$\mathbf{4.84 \times 10^{-2}}$	$\mathbf{3.04 \times 10^{-2}}$	2.46×10^{-1}	$\mathbf{3.17 \times 10^{-2}}$	2.09×10^{-2}

実験の結果，4 つのデータセットのうち Breast，Spam，Credit の 3 つにおいて，提案手法が最も低い RMSE を記録し，既存手法を上回る精度を達成した．一方で，Letter データセットにおいては，GAIN が最も良い精度を示し，提案手法は既存手法に及ばなかった．Letter データセットはアルファベットを 1 から 26 の整数値に変換したものである．本来はカテゴリデータであるが，提案手法は数値を連続値として扱い，ユークリッド距離に基づき誤差を最小化しようとする．そのため，整数間の等間隔の仮定がデータの特性と適合せず，生成モデルである GAIN と比較して精度が劣る結果となったと考える．

4.4 MAR

欠損がある観測されている他のデータに依存して発生する MAR(Missing At Random) の状況下での性能を検証した．これは，MCAR よりも補完が困難であり，現実データの欠損においても頻繁に観測される．

Pima データセットに対し，MAR による欠損を発生させた際の補完精度を確認した．欠損方法は年齢で昇順にソートし上位 10 %をテストデータとする．置換のそれぞれの比率を，以下の表 4 に示す．

表 4: Pima データセットにおけるマスク置換比率と補完精度の比較

マスク置換比率 (Mask : Random : Original)	RMSE (Mean)
0 : 0 : 10	2.10
7 : 2 : 1	5.37×10^{-1}

補完精度は 5.37×10^{-1} と MCAR と比較して精度が下がった．

4.5 MNAR

欠損が欠損している値そのものに依存する MNAR(Missing Not At Random) について検証した．統計的に最もバイアスが生じやすく補完が困難である．

Pima データセットに対し，MNAR による欠損を発生させた際の補完精度を確認した．欠損方法は測定器具の物理限界を想定し学習を行った．置換のそれぞれの比率を，以下の表 5 に示す．

表 5: Pima データセット (MNAR) におけるマスク置換比率と補完精度の比較

マスク置換比率 (Mask : Random : Original)	RMSE (Mean)
0 : 0 : 10	2.41
7 : 2 : 1	$\mathbf{7.77 \times 10^{-1}}$

補完精度は 7.77×10^{-1} と MCAR と比較して精度が下がった．

4.6 欠損率の変化による影響

モデルの堅牢性を評価するために，Credit データセットにおいて欠損率を 10% から 80% まで 10% 刻みで変化させた際の RMSE の推移を検証した．また，欠損方法は MCAR と仮定し，実験結果を図 2 に示す．図 2 より，提案手法は欠損率 10% から 80% の広範囲にわたり，既存手法よりも著しく低い RMSE を維持していることが確認できた．特に欠損率が 10% から 60% までは RMSE の上昇が極めて緩やかであり，利用可能な情報が半減していても補完精度が大きく損なわれないことが分かった．

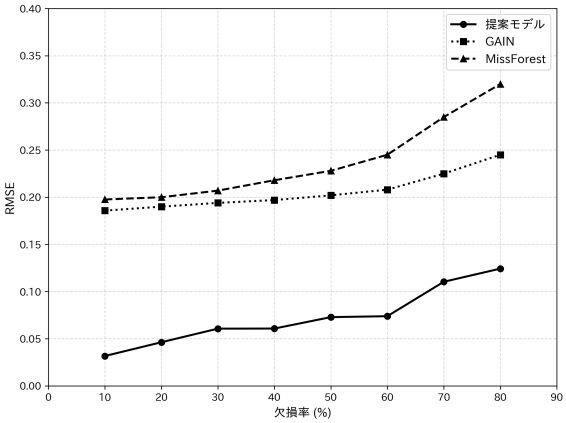

図 2: 欠損率の変化による精度の変化

4.7 Tokenizer の有効性の検証

Letter データセットを用いてトークン化の違いによる補完精度の変化についての実験を行った．実験の結果，Global Input Tokenizer の補完精度は 2.46×10^{-1} となり，既存の Feature Tokenizer の 2.69×10^{-1} と比較してより良い精度を達成した．

表 6: Tokenizer における補完精度の変化

Tokenizer	補完精度 (Mean)	標準偏差 (Std)
Global Input Tokenizer	$\mathbf{2.46 \times 10^{-1}}$	9.20×10^{-3}
既存手法	2.69×10^{-1}	7.40×10^{-3}

5 考　　察

本研究では，MCAR と MAR，MNAR という異なる 3 つの欠損メカニズムに対して提案手法の評価を行った．表 3 に示した MCAR の実験結果において，提案手法は Breast や Credit などのデータセットで GAIN や MissForest と比較して同等以上の精度を達成した．これは，Transformer の Self-Attention 機構が，観測された特徴量間の複雑な依存関係を効果的に捉えている可能性がある．もう 1 つの知見として，Pima データセットを用いた MAR と MNAR の実験結果から得られた．表 4 と表 5 に示した通り，学習時のマスクの置換比率を変えた場合，劇的な精度の向上が確認された．マスク時の Original Value の割合が 10，つまり全くマスクしない場合，欠損がない仮定で学習が行われてしまい，欠損値予測ができない．対して，適切に情報を隠し学習させることで，モデルは欠損していない他の特徴量から欠損値を推論する関係性を学習せざるを得なくなる．この学習プロセスが汎化性能を生み出していると考える．

6 まとめと展望

本研究では，表形式データにおける欠損値補完の精度向上を目的として，全特徴量の相互作用を埋め込みの初期段階から考慮する Global Input Tokenizer とマスク言語モデルを参考にした欠損値補完手法を提案した．提案手法は，マスク言語モデルを回帰タスクに応用することで，データセット全体に内在する列間の関係性を学習し，欠損値を高精度に推定することを可能にした．今後の展望としては，数値データに対し高い補完精度を達成できることが分かったため，カテゴリ値にも対応できる欠損値補完手法に改良する必要がある．また，MAR や MNAR のような偏った分布に対して損失関数の改良や大規模データセットに対する計算コストの削減に向けたモデルの軽量化についても検討を進める必要がある．

文　　献

[1] Stef van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, Vol. 45, No. 3, pp. 1–67, 2011.

[2] Daniel J and Peter. Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, Vol. 28, No. 1, pp. 112–118, 2012.

[3] Olga Troyanskaya and Michael Cantor. Missing value estimation methods for dna microarrays. *Bioinformatics*, Vol. 17, No. 6, pp. 520–525, 2001.

[4] Jinsung Yoon, James Jordon, and Mihaela van der Schaar. GAIN: Missing data imputation using generative adversarial nets. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, Vol. 80 of *Proceedings of Machine Learning Research*, pp. 5689–5698. PMLR, 10–15 Jul 2018.

[5] Pierre-Alexandre Mattei and Jes Frellsen. MIWAE: Deep generative modelling and imputation of incomplete data sets. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, Vol. 97 of *Proceedings of Machine Learning Research*, pp. 4413–4423. PMLR, 09–15 Jun 2019.

[6] Xin Huang, Ashish Khetan, Milan Cvitkovic, and Zohar Karnin. Tabtransformer: Tabular data modeling using contextual embeddings. *arXiv preprint arXiv:2012.06678*, 2020.

[7] Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting deep learning models for tabular data. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, Vol. 34, pp. 18932–18943. Curran Associates, Inc., 2021.

[8] Gowthami Somepalli, Micah Goldblum, Avi Schwarzschild, C. Bayan Bruss, and Tom Goldstein. Saint: Improved neural networks for tabular data via row attention and contrastive pre-training. *arXiv preprint arXiv:2106.01342*, 2021.

[9] Tianyu Du, Luca Melis, and Ting Wang. Remasker: Imputing tabular data with masked autoencoding. *arXiv preprint arXiv:2309.13793*, 2023.

[10] P. J. García-Laencina, J.-L. Sancho-Gómez, and A. R. Figueiras-Vidal. Pattern classification with missing data: a review. *Neural Computing and Applications*, Vol. 19, No. 2, pp. 263–282, 2010.