複数基準特徴選択とOOB重み付けアンサンブルを用いた欠損値補完手法

高良　流平†　　齊藤　史哲†

† 青山学院大学理工学研究科　〒 252–5258 神奈川県相模原市中央区淵野辺本町 5–10–1
E-mail: †c5625282@aoyama.jp, ††saitoh@ise.aoyama.ac.jp

あらまし　現実のデータ収集環境では，センサーの途切れや入力漏れ，システム障害などにより欠損値が頻繁に発生する．欠損値が存在すると，統計解析や機械学習モデルの推定に歪みが生じ，分析結果の信頼性が低下するため，適切な欠損処理が不可欠である．欠損処理には削除と補完があるが，削除は情報損失を伴う．一方，平均値補完などの統計的手法は，データの非線形性や特徴量間の関係を十分に捉えられない．この課題に対し，ランダムフォレストを用いた missForest が提案されているが，すべての特徴量を一律に用いるため，寄与の小さい特徴量の影響を受けやすい．そこで本研究では，複数の特徴選択基準に基づく特徴量部分集合を用い，OOB 指標に基づいて予測を統合するアンサンブル型欠損値補完手法を提案する．

キーワード　欠損値補完，ランダムフォレスト，アンサンブル学習，特徴選択

1　はじめに

1.1　研究背景

近年，センシング技術や情報通信の発達により，医療，アンケート調査，Iot センサ，顧客属性や販売実績といったビジネスデータ，さらには製造業におけるプロセスデータなど，様々な分野において大量かつ多量なデータを収集できるようになっている．これらのデータは，診断支援，品質推定，需要予測，プロセス監視など多様な意思決定を支える基盤として活用されている．一方で，現実の計測環境においては常に完全なデータが得られるとは限らず，計測機器の誤作動や通信エラー，人為的な誤入力などにより，データ中に欠損値が含まれる状況がしばしば発生する．例えば医療現場では，電子カルテ，検査データ，生体信号などが診断支援や予後予測のために広く利用されているが，患者ごとに実施される検査項目が異なることや，時間的制約などの理由からすべての指標を一律に取得することが困難である．その結果，測定タイミングの不規則性や長期追跡データにおける患者の離脱など，多様な要因に起因する欠損が生じることが指摘されている [3]．また，製造業におけるプロセスデータにおいても，データ駆動型ソフトセンサが製品品質の推定やプロセス監視のために広く利用されている一方で，すべての測定点にセンサを設置することは経済的に困難であり，高温・高圧環境におけるセンサの故障やドリフト，保守作業による停止などによって欠損が発生することが報告されている [4]．

このような欠損を含むデータをそのまま用いて分析や予測を行うと，解析精度の低下や推定値の偏り，さらには欠損を含むサンプルの除外による有効サンプルの減少といった問題を引き起こす．したがって，欠損値を適切に補完し，元の情報をできる限り保持した上でデータを活用することは，信頼性の高いデータ分析を実現する上で不可欠な前処理である．

1.2　既存の欠損値補完手法とその課題

欠損値補完の基本的な手法としては，平均値補完やホットデック法などの単一代入法が広く用いられてきた．これらの手法は実装が容易で計算コストも低い一方で，データの背後に隠れるクラスタ構造や非線形性を考慮できず，サンプルのばらつきに強く影響されるという課題を抱えている．その結果，補完後のデータにおいて分散の過小評価や推定値の歪みが生じる可能性がある．こうした単一代入法の限界を補う枠組みとして，多重代入法が提案されている．多重代入法は，欠損値の不確実性を複数の補完結果として反映できる点で理論的に優れた手法であるが，補完モデルの仮定や選択に強く依存するという課題が指摘されている．特に高次元データや複雑な非線形構造を持つデータに対しては，適切なモデル設計が困難である．

このような背景から，近年ではモデル依存性を低減しつつ，データ駆動的に欠損値を補完する機械学習に基づく手法が広く研究されている．代表的な手法として，局所的な距離関係に基づいて補完を行う kNN 補完や，ランダムフォレストを用いた missForest が挙げられる．kNN は比較的単純な構造のデータに対して有効である一方，高次元空間では距離尺度の信頼性が低下するという問題を抱えている．また，missForest は非線形性や特徴量間の相互作用をある程度捉えることが可能であり，混合型データにも適用できる汎用性の高い手法である．しかし，すべての特徴量を一律に利用して学習を行うため，ノイズ的な特徴量や寄与の小さい特徴量が分割に含まれ，予測精度が低下する場合がある．また，欠損率が高い場合や，サンプル数に比べて特徴量数が多い場合には，過学習や性能劣化が生じやすいという課題が指摘されている．

1.3　本研究の目的

本研究では，ランダムフォレストを用いた欠損値補完手法である missForest に着目し，その補完精度および安定性の向上を目的とする．missForest は非線形性や特徴量間の相互作用を

捉えられる有効な手法である一方で，すべての特徴量を一律に利用して学習を行うため，ノイズ的な特徴量や寄与の小さい特徴量の影響を受けやすいという課題を抱えている．特に，欠損率が高い場合や，サンプル数に対して特徴量数が多い状況では，予測精度の低下や過学習が生じやすい．

そこで本研究では，特徴量選択の観点を導入し，複数の基準に基づいて有効な特徴量部分集合を構築した上で，それぞれに対してランダムフォレストによる補完モデルを学習する．さらに，各補完モデルの汎化性能を反映する指標として，ランダムフォレストの外部誤差推定である OOB（Out-of-Bag）の決定係数 R^2 を用い，その値に応じて予測結果を加重平均することで，補完結果を統合するアンサンブル手法を提案する．

このように，特徴量選択によるノイズの抑制と，アンサンブル学習による安定化を同時に図ることで，既存の missForest に比べて，より頑健で信頼性の高い欠損値補完を実現することを目指す．本研究では，UCI Machine Learning Repository や OpenML にて公開されている複数の公開データセットを用いた数値実験を通じて，提案手法の有効性を評価する．

2 関 連 研 究

2.1 MissForest の概要

missForest は，ランダムフォレストを基盤とした反復型の欠損値補完手法であり，分布仮定を必要としない非パラメトリックな手法として提案されている [2]．各変数を目的変数，残りの変数を説明変数とみなしてランダムフォレストを学習し，欠損部分を予測する操作を変数ごとに繰り返すことで，データ行列全体の欠損値を補完する枠組みを採用している．この反復的な学習・更新により，特徴量間の非線形な関係や相互作用を考慮した補完が可能である．

missForest の特徴として，連続変数とカテゴリ変数が混在するデータに適用可能である点や，補完精度の推定にランダムフォレストの OOB 誤差を利用できる点が挙げられる．OOB 誤差を用いることで，外部の検証用データを用いることなく，補完結果の誤差を近似的に評価できる．また，多数の決定木の予測結果を平均化するランダムフォレストの構造は，多重代入的な性質を内包していると整理されている．

原著論文およびその後の研究において，missForest は医療データや生物データをはじめとする実データを用いた実験により，他の欠損値補完手法と比較して安定した補完性能を示すことが報告されており，機械学習に基づく欠損値補完手法の代表的な手法として位置づけられている．一方で，高次元データにおいては，補完に寄与しない特徴量の影響を受けやすいことが指摘されており，特徴量の扱いに関して改良の余地があるとされている．

3 提　　案

3.1 手法の概要

本研究で提案する手法は，欠損を含む各特徴量を目的変数とし，残りの特徴量を説明変数としてランダムフォレストを学習するという点では missForest と同様である．しかし，従来の missForest がすべての特徴量を一律に利用するのに対し，提案法では複数の基準に基づいて有効性の高い特徴量部分集合を構築し，各部分集合ごとに独立したランダムフォレストを学習させる．

このように，特徴量の選択基準を複数導入することで，不要あるいはノイズ的な特徴量の影響を低減し，多様性を持つ補完モデルを構築することを目的とする．以降では，まず特徴量部分集合の構築方法について述べ，その後，複数の補完モデルを統合する手法について説明する．

3.2 複数基準に基づく特徴量部分集合の構築

3.2.1 特徴選択導入の動機

missForest では，各特徴量の補完において，当該特徴量を目的変数とし，残りのすべての特徴量を説明変数としてランダムフォレストを学習する．この枠組みは，特徴量間の非線形な関係や相互作用を捉えられる一方で，補完に寄与しない特徴量やノイズ的な特徴量が説明変数に含まれる場合，予測結果のばらつきや過学習を引き起こす可能性がある．特に，サンプル数に対して特徴量数が多い場合や，特徴量間の相関構造が複雑な場合には，こうした影響が顕在化しやすいと考えられる．

また，各特徴量の補完に寄与する説明変数は，補完対象となる特徴量ごとに異なると考えられるが，missForest ではこれらを区別せず一律に利用している．そのため，目的変数に対する寄与が小さい特徴量が含まれることで，補完精度の低下や結果の不安定化を招く可能性がある．以上の点から，補完に有効な特徴量を選択的に用いることで，不要な情報の影響を低減し，より頑健な補完を行うことが期待される．

3.2.2 特徴量部分集合の構築方針

特徴量の重要度や補完への寄与は，評価基準やデータ構造に依存して変化するため，単一の基準に基づいて特徴量を選択した場合，特定の構造に偏った特徴量集合が得られる可能性がある．その結果，データによっては補完精度が十分に向上しない場合も考えられる．

そこで本研究では，複数の特徴選択基準に基づいて，異なる特徴量部分集合を構築する．各基準により選択された特徴量部分集合ごとに，独立したランダムフォレストを学習することで，補完に有効な特徴量の捉え方に多様性を持たせる．このようにして得られた複数の補完モデルを用いることで，特定の基準に依存した偏りを緩和し，より安定した補完結果を得ることを目指す．

3.3 各特徴選択基準

本節では，前節で述べた方針に基づき，特徴量部分集合を構築するために用いる各特徴選択基準について具体的に説明する．本研究では，複数の観点から有効な説明変数を抽出するため，以下に示す基準に基づいて特徴量部分集合を構築する．以降，補完対象となる特徴量を X_s，説明変数候補を $\{X_j \mid j \neq s\}$ とする．また，各特徴選択基準 m により得られる特徴量部分集

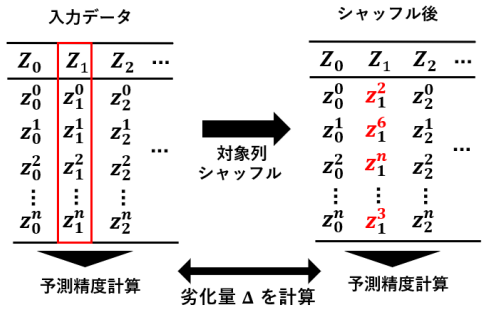

図 1　Permutation Importance の概念

合を $\mathcal{F}_s^{(m)}$ と表す．なお，本研究では連続変数のみを対象とするため，以下の重要度評価は回帰設定を前提とする．

3.3.1　全特徴量利用（ALL）

比較の基準として，補完対象 X_s に対し，残りのすべての特徴量を説明変数として用いる部分集合を構築する．すなわち，

$$\mathcal{F}_s^{(\mathrm{ALL})} = \{X_j \mid j \neq s\} \tag{1}$$

とする．この設定は，従来の missForest と同様に全特徴量を用いる場合に対応しており，提案法における内部的な比較の基準として位置づけられる．

3.3.2　Permutation Importance に基づく選択（PERM）

Permutation Importance（PI）は，学習済みモデルを固定したまま，各特徴量の値をサンプル間でランダムに入れ替えた際の予測性能の低下量に基づいて特徴量の重要度を評価する指標である．重要な特徴量ほど，当該特徴量の入れ替えによって予測性能が大きく劣化する．

図 1 に，PI の概念を示す．まず，元の入力データを用いて学習済みモデルによる予測性能を算出する．次に，対象となる特徴量の列のみをサンプル間でランダムにシャッフルし，他の特徴量およびモデルは固定したまま再度予測性能を算出する．このときの性能低下量 Δ を，当該特徴量の重要度として用いる．

本研究では，補完対象 X_s を目的変数としてランダムフォレストを学習し，各説明変数候補に対する PI を算出する．得られた重要度に基づき，上位の特徴量からなる部分集合 $\mathcal{F}_s^{(\mathrm{PERM})}$ を構築する．

3.3.3　Mean Decrease Impurity に基づく選択（MDI）

Mean Decrease Impurity（MDI）は，決定木の分岐における不純度の減少量を，各特徴量について集計することで重要度を評価する指標である．図 2 に，MDI の概念を示す．MDI は，ランダムフォレストを構成する各決定木において，特徴量が分割に用いられた際の不純度の減少量に基づいて算出される．具体的には，ある特徴量 X_j が分割に用いられたノードにおいて，分割前の不純度と分割後の子ノードにおける不純度との差を，当該特徴量による不純度減少量として算出する．図中左および中央は，各決定木において分割に用いられた特徴量と，対応する不純度減少量を記録し，特徴量ごとに累積する過程を示している．ランダムフォレストにおける MDI は，すべての決定木および分割ノードにおける不純度減少量を特徴量ごとに累積し，平均することで求められる．

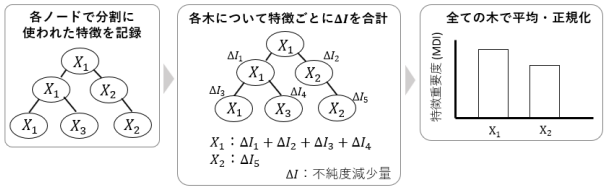

図 2　Mean Decrease Impurity の概念

本研究では，補完対象 X_s を目的変数として学習したランダムフォレストにおいて，説明変数候補の MDI を算出し，重要度が高い特徴量からなる部分集合 $\mathcal{F}_s^{(\mathrm{MDI})}$ を構築する．

3.3.4　相関係数に基づく選択（CORR）

単変量の関係に基づく基準として，補完対象 X_s と各説明変数候補 X_j の相関係数を用いる．具体的には，観測可能なサンプルに基づいて $|\mathrm{corr}(X_s, X_j)|$ を算出し，値の大きい特徴量からなる部分集合 $\mathcal{F}_s^{(\mathrm{CORR})}$ を構築する．

3.3.5　ランダム選択（RANDOM）

重要度に依存しない多様性確保のため，説明変数候補 $\{X_j \mid j \neq s\}$ からランダムに特徴量を抽出し，部分集合 $\mathcal{F}_s^{(\mathrm{RAND})}$ を構築する．この基準は，特定の重要度指標に基づく選択がデータ構造によって偏る可能性を緩和し，補完モデル群の多様性を確保することを目的とする．

3.3.6　部分集合サイズの扱い

PERM，MDI，CORR に基づく特徴選択では，特徴量部分集合のサイズ k をデータに基づいて自動的に決定する動的な選択方法を採用する．その手順を以下に示す．

Step 1: 補完対象となる特徴量 X_s に対し，当該特徴量を目的変数，$\{X_j \mid j \neq s\}$ を説明変数候補として，PERM，MDI，CORR の各基準に基づく特徴量重要度を算出する．

Step 2: 各基準において，算出した重要度に基づき，特徴量を重要度の高い順に順位付けする．この順位に従い，Top-1 から Top-d（d は全特徴量数）までの特徴量部分集合候補を段階的に生成する．

Step 3: 各 k に対して，対応する特徴量部分集合を用い，観測サンプルのみを用いてランダムフォレスト回帰モデルを学習する．

Step 4: 学習時に得られる OOB サンプルを用いて，各 k に対応する補完モデルの汎化性能指標を算出する．

Step 5: 得られた OOB 指標に基づき，汎化性能が最大となる k を選択し，当該特徴選択基準において採用する特徴量部分集合サイズとして決定する．

一方，RANDOM に基づく特徴選択では，特徴量の重要度順位が定義できないため，上記のような段階的な k の評価は行わない．その代わりに，全特徴量数 d に対する比率 k_{ratio} をあらかじめ設定し，$k = \lfloor d \times k_{\mathrm{ratio}} \rfloor$ として特徴量部分集合サイズを決定する．

3.4　特徴量部分集合ごとの補完モデル学習

本研究では，補完対象となる各特徴量を目的変数とし，前節までに構築した特徴量部分集合ごとに，独立したランダムフォ

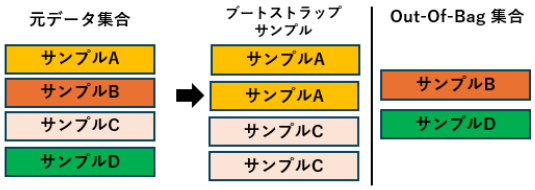

図 3　Out of Bag の概念

レスト回帰モデルを学習する．すなわち，各特徴選択基準 m により得られる特徴量部分集合 $\mathcal{F}_s^{(m)}$ に対して，補完モデル $f_s^{(m)}(\cdot)$ を構築する．

モデル学習には，当該特徴量が観測されているサンプルのみを用いる．一方，説明変数に含まれる欠損値については，各特徴量の平均値による一次補完を行った後，ランダムフォレストを学習する．これにより，欠損を含むデータに対しても各特徴量部分集合に基づく補完モデルの構築を可能とする．

また，ランダムフォレストの学習過程においては，ブートストラップ標本に基づく OOB サンプルが自然に得られる．本研究では，この OOB サンプルを用いて各補完モデルの汎化性能を評価し，次節におけるアンサンブル統合のための指標として利用する．

3.5　OOB 指標に基づくアンサンブル統合

本研究では，前節までに構築した特徴量部分集合ごとに学習された複数のランダムフォレスト補完モデルを統合するため，OOB 指標に基づく重み付きアンサンブルを用いる．ランダムフォレストはブートストラップ標本を用いて学習されるため，各モデルに対して学習データとは独立な OOB サンプルを用いた汎化性能の推定が可能である．

図 3 に，OOB サンプルの生成過程を示す．図は，単一の決定木に対するブートストラップ抽出の一例を模式的に表したものである．各決定木は元データ集合から復元抽出によって生成されたブートストラップ標本を用いて学習されるが，このとき，当該決定木の学習に一度も使用されなかったサンプルが存在する．これらのサンプルを OOB サンプルと呼ぶ．各決定木は異なるブートストラップ標本で学習されるため，OOB サンプルは木ごとに異なって生成される．本研究では，これらの OOB サンプルを用いて，各補完モデルの汎化性能を評価し，その結果をアンサンブル統合における重みとして利用する．

補完対象となる特徴量 X_s に対し，特徴選択基準 m に基づいて学習された補完モデルを $f_s^{(m)}(\cdot)$，対応する OOB 指標を $w_s^{(m)}$ とする．本研究では，OOB 指標として決定係数 R^2 を用い，$w_s^{(m)} < 0$ の場合には $w_s^{(m)} = 0$ とすることで，補完性能が低いモデルの影響を抑制する．

得られた重みは以下のように正規化される．

$$\tilde{w}_s^{(m)} = \frac{w_s^{(m)}}{\sum_{m'} w_s^{(m')}}. \tag{2}$$

ただし，$\sum_{m'} w_s^{(m')} = 0$ の場合には，すべてのモデルに等しい重みを割り当てる．

最終的な補完値は，各補完モデルの予測値の加重平均として求める．欠損しているサンプル i に対する X_s の補完値 $\hat{x}_{is}$ は，

$$\hat{x}_{is} = \sum_m \tilde{w}_s^{(m)} f_s^{(m)}(\mathbf{x}_{i,-s}) \tag{3}$$

と定義される．ここで，$\mathbf{x}_{i,-s}$ はサンプル i における補完対象以外の特徴量ベクトルを表す．

このように，OOB 指標を用いて各補完モデルを統合することで，特徴選択基準ごとの性能差を反映しつつ，単一の基準に依存しない安定した補完結果を得ることを目指す．

3.6　提案手法の全体像

提案手法における欠損値補完の処理手順を，以下に Step 1 から Step 5 として示す．

Step 1: 対象とするデータセット $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^N$ に対し，補完対象となる特徴量 X_s を順に選択する．

Step 2: X_s を目的変数，$\{X_j \mid j \neq s\}$ を説明変数候補として，ALL，PERM，MDI，CORR，RANDOM の各基準に基づき，特徴量部分集合 $\mathcal{F}_s^{(m)}$ を構築する．

Step 3: 各特徴量部分集合 $\mathcal{F}_s^{(m)}$ に対して，観測サンプルのみを用いてランダムフォレスト回帰モデル $f_s^{(m)}(\cdot)$ を学習する．

Step 4: 学習時に得られる OOB サンプルを用いて，各補完モデルの汎化性能指標 $w_s^{(m)}$ を算出する．

Step 5: OOB 指標に基づいて重みを正規化し，各補完モデルの予測値を加重平均することで，欠損値 $\hat{x}_{is}$ を推定する．

以上の手順を，補完対象となるすべての特徴量に対して繰り返すことで，データ行列全体の欠損値を補完する．

4　実　　験

4.1　評価方法の概要

本研究では，欠損値補完手法の性能を定量的に評価するため，正解値が既知である完全データを用い，人工的に欠損を導入する評価設定を採用した．図 4 に，本研究における欠損値補完手法の評価フローを示す．

図中左端の行列は，$n \times d$ 次元の完全データ行列 $X = (x_i^j)$ を表しており，x_i^j は i 番目のサンプルにおける j 番目の特徴量の正解値を示す．次に，完全データに対して人工的に欠損を導入し，一部のセルを欠損値（NA）として置き換える．図中では，欠損セルをグレー背景で表現している．欠損はセル単位で導入され，人工的に欠損させたセルの集合を Ω_{miss} と定義する．

生成した欠損データに対して，各補完手法を適用し，欠損セルに対応する値を推定する．補完によって得られた推定値は $\hat{x}_i^j$ と表し，図中では水色背景で示している．なお，観測済みセルの値は補完処理によって変更されない．最後に，人工的に欠損させたセル集合 Ω_{miss} のみに対して，推定値 $\hat{x}_i^j$ と正解値 x_i^j との差を用い，平均二乗誤差（MSE）により補完精度を評価する．

4.2　実 験 設 定

実験に先立ち，すべての特徴量に対して標準化を行い，以降の欠損導入，補完処理および評価は，この標準化空間において

実施した．標準化は訓練データの統計量に基づいて行い，同一の変換をテストデータに適用した．

欠損は，完全にランダムな欠損（MCAR: Missing Completely At Random）を仮定し，セル単位で導入した．欠損率は $r_{\text{miss}} \in \{0.1, 0.3, 0.5\}$ とし，欠損セル数は $[N \times d \times r_{\text{miss}}]$（$N$ はサンプル数，d は特徴量数）とした．欠損位置は一様分布に基づき，重複なしで無作為に選択した．

データはあらかじめ訓練データとテストデータに分割し，欠損は両者に対して独立に導入した．各反復においては，すべての補完手法に対して同一の欠損位置を用いることで，手法間の比較における公平性を担保した．各欠損率について 30 回の独立な反復実験を行い，反復ごとに欠損位置を再サンプリングした．

補完モデルは訓練データのみを用いて学習し，学習済みモデルをテストデータに適用した．評価は，テストデータに導入した欠損セルのみを対象として行い，以下の式で定義される平均二乗誤差（MSE）を補完精度の指標とした．結果は，30 回の反復における平均値および標準誤差（SE）として報告する．

4.3 実験に用いたデータ

本研究では，提案手法の有効性および汎用性を検証するため，UCI Machine Learning Repository にて公開されている代表的なベンチマークデータセットと，OpenML から取得した複数の実データセットを用いて実験を行った．

UCI のベンチマークデータとしては，“iris”，“wine”，“diabetes” の 3 つのデータセットを採用した．これらは欠損値補完に関する既存研究において広く用いられており，異なるサンプル数および特徴量数を有する標準的な評価用データセットである．

さらに，より現実的かつ多様なデータ構造に対する補完性能を評価するため，OpenML から取得した回帰データセットとして，“airfoil_self_noise”，“concrete”，“yacht_hydro”，“energy_efficiency” を用いた．これらのデータセットは，UCI のベンチマークデータと比較してサンプル数や特徴量数に多様性があり，実応用に近い条件下での欠損値補完性能を検証することが可能である．

本研究では，すべてのデータセットに対して連続値特徴量のみを対象とした．カテゴリカル変数やクラスラベルとして扱われる変数は，欠損補完の評価に直接寄与しないため，事前に除外した．これにより，補完対象を連続変数に統一し，手法間の比較を公平に行っている．

4.4 実 験 結 果

実験結果は表 1 に示すとおりである．表中の数値は，各手法および各欠損率における 30 回反復実験の平均 MSE を示しており，括弧内の数値は標準誤差である．評価はすべて標準化空間において，テストデータに導入した欠損セルのみを対象として行った．

UCI ベンチマークデータセットに関する結果について述べる．iris データセットでは，欠損率が低い条件において kNN や missForest が比較的良好な性能を示し，提案手法もこれらと同等の補完精度を達成している．一方，欠損率が高い条件では，ユークリッド距離に基づく近傍探索が有効に働く場合が見られ，低次元データにおける kNN の有効性が確認された．これらの結果から，提案手法は単純なデータ構造に対しても過剰適応することなく安定した補完性能を示すことが分かる．

wine データセットでは，すべての欠損率において提案手法が最小の補完誤差を示しており，既存手法に対する優位性が確認された．この結果は，特徴量間の相関が比較的強い中規模データにおいて，複数の特徴選択基準に基づく部分集合構築と，OOB 指標に基づくアンサンブル統合が有効に機能していることを示唆している．

diabetes データセットでは，欠損率が低い条件において missForest が良好な性能を示す一方で，欠損率が 0.3 および 0.5 の条件では提案手法が最小誤差を達成している．このことから，提案手法は中程度の相関構造を持つデータに対しても有効であることが示唆される．

次に，OpenML から取得した実データセットに関する結果について述べる．airfoil_self_noise および concrete データセットでは，すべての欠損率において提案手法が missForest や kNN を上回る，あるいは同等の補完精度を示した．特に，中程度から高い欠損率の条件では，提案手法が最小誤差を達成しており，特徴量数が多く，欠損率が高い条件においても安定した性能を示すことが確認された．

yacht_hydro データセットにおいても，提案手法はすべての欠損率で kNN および missForest より小さい誤差を示している．この結果は，サンプル数が限られたデータに対しても，複数の特徴選択基準に基づくモデル構築とアンサンブル統合が有効に機能していることを示唆している．

energy_efficiency データセットでは，欠損率が低い条件においても提案手法が最良の性能を示し，欠損率の増加に伴っても誤差の増大が比較的抑えられている．このことから，提案手法は，異なるデータ構造や規模を持つ複数の実データに対しても，一貫して高い補完精度を維持できることが分かる．

以上の結果より，提案手法は UCI のベンチマークデータセットに加え，OpenML の実データセットにおいても安定して良好な補完性能を示しており，より現実的なデータ環境に対しても有効である可能性が示された．

5 まとめと考察

5.1 実験結果のまとめ

本研究では，複数の公開データセットに対して人工的に欠損を導入し，既存の機械学習ベース補完手法および提案手法である RF-Ens の補完精度を比較した．評価は，テストデータに導入した欠損セルのみを対象とした平均二乗誤差（MSE）に基づいて行った．

実験の結果，提案手法 RF-Ens は，wine，airfoil_self_noise，concrete，yacht_hydro，energy_efficiency など多くのデータセットにおいて，安定して低い MSE を示した．特に，中〜高

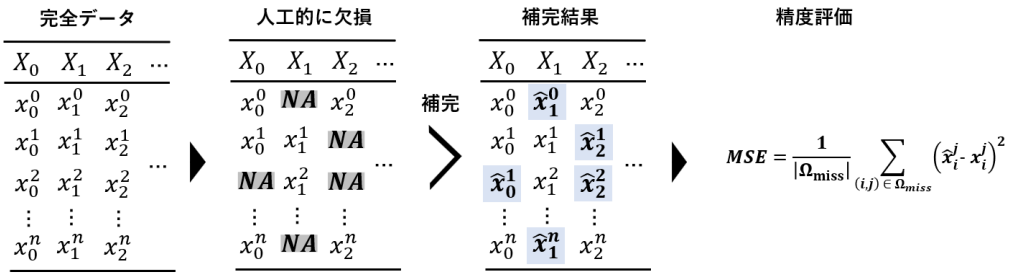

図 4　欠損値補完手法の評価フロー

表 1　提案法（RF-Ens）と既存補完手法の補完精度比較

data set	missing rate	mean	kNN(k=5)	missForest	RF-Ens (proposed)
iris	0.1	1.228(0.049)	0.309(0.030)	0.328(0.040)	0.335(0.044)
	0.3	1.149(0.022)	0.440(0.028)	0.416(0.034)	0.424(0.032)
	0.5	1.162(0.020)	0.606(0.020)	0.715(0.033)	0.705(0.034)
wine	0.1	0.940(0.028)	0.475(0.021)	0.471(0.016)	0.463(0.016)
	0.3	0.911(0.016)	0.587(0.014)	0.613(0.014)	0.582(0.013)
	0.5	0.943(0.009)	0.747(0.008)	0.752(0.014)	0.715(0.011)
diabetes	0.1	0.950(0.017)	0.621(0.015)	0.495(0.011)	0.500(0.010)
	0.3	0.991(0.013)	0.786(0.010)	0.635(0.008)	0.619(0.008)
	0.5	0.991(0.005)	0.959(0.009)	0.825(0.011)	0.772(0.007)
airfoil	0.1	0.959(0.013)	0.573(0.013)	0.276(0.012)	0.231(0.009)
	0.3	0.970(0.009)	0.876(0.010)	0.534(0.011)	0.396(0.008)
	0.5	0.972(0.006)	0.956(0.007)	0.831(0.011)	0.694(0.016)
concrete	0.1	0.992(0.023)	0.352(0.015)	0.275(0.017)	0.267(0.014)
	0.3	0.985(0.011)	0.688(0.008)	0.510(0.010)	0.466(0.007)
	0.5	0.979(0.006)	0.868(0.008)	0.792(0.009)	0.685(0.008)
yacht	0.1	0.899(0.026)	0.598(0.022)	0.548(0.026)	0.439(0.020)
	0.3	0.927(0.016)	0.769(0.018)	0.746(0.027)	0.569(0.016)
	0.5	0.932(0.009)	0.870(0.013)	0.955(0.021)	0.809(0.020)
energy	0.1	1.043(0.011)	0.554(0.012)	0.586(0.012)	0.406(0.011)
	0.3	1.032(0.007)	0.656(0.008)	0.572(0.011)	0.468(0.011)
	0.5	1.027(0.003)	0.798(0.006)	0.703(0.012)	0.589(0.009)

欠損率条件（$r_{\text{miss}} = 0.3, 0.5$）においても補完精度の大きな劣化が見られず，欠損率の増加に対するロバスト性が確認された．

また，データセットによっては，提案手法が既存手法と同等，あるいはそれ以上の補完精度を示しており，複数の特徴選択基準を統合するアプローチの有効性が確認された．

5.2 考　　察

提案手法 RF-Ens が安定した補完性能を示した要因として，複数の特徴選択基準に基づく部分集合の構築と，それらを統合するアンサンブル構造が有効に機能した点が挙げられる．単一の基準に基づく特徴選択では，データセットの特性によって重要な特徴量を十分に捉えられない可能性があるが，RF-Ens では PERM や MDI など異なる観点から得られた重要度情報を組み合わせることで，より多様な依存構造を反映できていると考えられる．さらに，各部分集合に基づく補完モデルを OOB 精度に応じて重み付けすることで，補完性能の高いモデルの寄与が強調され，結果として全体としての補完精度が向上したと考えられる．このような重み付き統合は，データセットや欠損率の違いに対して柔軟に適応できる点で有効である．

一方で，データ構造が比較的単純な場合や，特徴量間の依存関係が限定的な場合には，提案手法と既存手法との性能差が小さくなる傾向も見られた．これは，複雑な特徴選択や統合戦略が，すべての状況において常に大きな利点をもたらすわけではないことを示唆している．

本研究では，欠損が完全にランダムに発生する MCAR を仮定して評価を行ったが，実際のデータにおいては，観測されている他の変数に依存して欠損が生じる MAR（Missing At Random）や，欠損そのものが値に依存する MNAR（Missing Not At Random）が生じる場合も多い．これらの欠損メカニズムの違いは，特徴量間の依存構造や補完モデルの学習過程に影響を与える可能性があり，今後は欠損メカニズムの違いを考慮した評価が重要な課題となる．

また，本研究で提案した RF-Ens は，複数の特徴選択基準とアンサンブル学習を組み合わせた構成を取っているため，データ規模の増大に伴い計算コストが増加する可能性がある．今後は，大規模データへの適用を見据えた計算効率の改善や近似的手法の導入についても検討が必要である．

6 おわりに

本研究では，アンサンブルベースの欠損補完手法の代表例である missForest を基盤とし，複数の特徴選択基準を導入した欠損値補完手法を提案した．提案手法では，補完対象となる特徴量ごとに有効な説明変数の部分集合を構築し，各部分集合に基づいて学習したランダムフォレストを OOB 指標により統合することで，特徴選択の多様性と補完性能の安定化を図った．ベンチマークデータおよび OpenML データセットを用いた実験により，提案手法は missForest や kNN 補完と比較して，多くの条件下で同等以上の補完精度を示すことを確認した．特に，中程度以上の欠損率や特徴量間の相関構造を持つデータにおいて，特徴選択とアンサンブル統合の効果が有効に機能することが示唆された．一方で，本研究では連続変数のみを対象としており，カテゴリカル変数を含むデータへの適用や，欠損メカニズムが MCAR 以外の場合の検討は今後の課題である．今後は，より複雑な欠損構造や実データへの適用を通じて，提案手法の汎用性と実用性について検証を進めていく予定である．

謝辞 本研究は科学研究費（基盤 C）23K04275 による支援を受けたものです．

文　献

[1] 高橋 将宣，渡辺 美智子，欠測データ処理 R による単一代入法と多重代入法，共立出版，(2017).
[2] Stekhoven, D. J. and Buhlmann, P., MissForest: nonparametric missing value imputation for mixed-type data, Bioinformatics, Vol. 28, No. 1, pp. 112-118 (2012).
[3] Chen, Z., Tan, S., Chajewska, U., Rudin, C. and Caruana, R., Missing Values and Imputation in Healthcare Data: Can Interpretable Machine Learning Help?, Proceedings of Machine Learning Research, Vol. 209, pp. 86-99 (2023).
[4] Ma, L., Wang, M. and Peng, K., A missing manufacturing process data imputation framework for nonlinear dynamic soft sensor modeling and its application, Expert Systems with Applications, Vol. 237, Part A, Article 121428 (2024).
[5] Rubin, D. B., Inference and Missing Data, Biometrika, Vol. 63, No. 3, pp. 581–592 (1976).
[6] Little, R. J. A. and Rubin, D. B., Statistical Analysis with Missing Data (2nd ed.), Wiley, New York, NY (2002).
[7] Schafer, J. L. and Graham, J. W., Missing data: our view of the state of the art, Psychological Methods, Vol. 7, No. 2, pp. 147–177 (2002).
[8] 阿部 貴行，欠測データの統計解析，朝倉書店，(2016).
[9] Andridge, R. R. and Little, R. J. A., A Review of Hot Deck Imputation for Survey Non-response, International Statistical Review, Vol. 78, No. 1, pp. 40–64 (2010).
[10] Brick, J. M. and Kalton, G., Handling missing data in survey research, Statistical Methods in Medical Research, Vol. 5, No. 3, pp. 215–238 (1996).
[11] Rubin, D. B., Multiple Imputation for Nonresponse in Surveys, John Wiley & Sons, New York, NY (1987).
[12] Breiman, L., Random Forests, Machine Learning, Vol. 45, No. 1, pp. 5–32 (2001).
[13] Strobl, C., Boulesteix, A.-L., Zeileis, A. and Hothorn, T., Bias in random forest variable importance measures, BMC Bioinformatics, Vol. 8, Article 25 (2007).
[14] Breiman, L., Bagging predictors, Machine Learning, Vol. 24, No. 2, pp. 123–140 (1996).
[15] Guyon, I. and Elisseeff, A., An Introduction to Variable and Feature Selection, Journal of Machine Learning Research, Vol. 3, pp. 1157–1182 (2003).
[16] Hall, M. A., Correlation-based Feature Selection for Machine Learning, Ph.D. thesis, University of Waikato, Hamilton, New Zealand (1999).
[17] Díaz-Uriarte, R. and Alvarez de Andrés, S., Gene selection and classification of microarray data using random forest, BMC Bioinformatics, Vol. 7, Article 3 (2006).