アンサンブル学習に基づいた欠損値補完について

高良　流平†　　齊藤　史哲†

† 静岡大学情報学部　〒 432–8011 静岡県浜松市中区城北 3-5-1
E-mail: †{hanako,yamamoto}@design.inf.shizuoka.ac.jp

あらまし　これはアブストラクトです．300 字程度でまとめてください．

キーワード　情報検索，情報の信憑性，ヒューマンファクター

1　はじめに

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．しかし，現実の計測環境で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．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，サンプルサイズの減少による識別精度の低下といった問題を引き起こす．そのため，欠損値を適切に補完し，元の情報をできる限り維持した上で活用することは，信頼性のあるデータ分析を支える上で不可欠な前処理である．

基本的な欠損値補完手法としては，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いるが，これらは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ない，またはサンプルのばらつきに強く影響されるといった課題を抱えており，補完結果の頑健性や安定性に限界がある．一方，多重代入法は不確実性を考慮できる理論的に優れた手法であるが，モデル選択に強く依存するという課題がある．その他に，機械学習に基づく手法として k 近傍法やランダムフォレストを用いた missForest がある．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るが，すべての特徴量を一律に用いるため，ノイズ的な特徴量や寄与の小さい特徴量が学習に含まれ，分割の質が下がり，結果として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が多い場合，過学習や性能劣化が生じやすい．

そこで本研究では，複数の特徴選択基準に基づいて有効な特徴量を抽出し，各基準で構築したランダムフォレストの予測を OOB R^2 に応じて加重平均するアンサンブル補完手法を提案する．

2　提　　案

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補完手法である．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ことで，欠損セルを逐次的に予測・更新する．これをすべての欠損列に対して繰り返し，推定値が収束するまで反復を行うことで最終的な補完結果を得る．この手法はすべての特徴量を一律に利用するため，ノイズ的な特徴量や寄与の小さい特徴量が予測精度を阻害する可能性がある．また高欠損率や高次元の状況では，精度劣化や過学習が生じやすいという課題を抱えている．

2.2　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 が全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」を抽出し，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．これにより，不要あるいはノイズ的な特徴量の影響を低減し，より頑健な予測を得ることを目的とする．

特徴量選択の基準としては，五つの方法を用いた．従来法と同様にすべての特徴量をそのまま利用する方法，Permutation Importance に基づき寄与の大きい特徴量を抽出し，上位のみに限定する方法，ランダムフォレスト内部で得られる Mean Decrease Impurity に基づき，分岐に寄与した重要な特徴量を選択する方法，欠損を含む対象特徴量と他の特徴量との相関係数を計算し，その絶対値が大きい特徴量を優先的に採用する方法，特徴量をランダムに抽出して部分集合を構築し，モデル間の多様性を確保する方法である．

このように異なる基準で抽出した部分集合により欠損値を予測することで，missForest に比べてノイズ的特徴量の利用による精度低下を抑制し，多様な視点からの補完が可能になる．

2.3　アンサンブル統合

各基準に基づく方法から得られた複数の予測値は，それぞれのモデルが持つ汎化性能に応じて重みづけし，最終的に加重平均することで統合する．本研究では重みとして，ランダムフォレストの外部誤差推定である OOB R^2 を利用した．これは訓練に使用されなかったサンプルに対する決定係数であり，各部分モデルが未知データに対してどれだけ説明力を持つかを客観的に評価できる．

統合に際しては，OOB R^2 が高いモデルには大きな重みを，

低いモデルには小さな重みを与えることで，信頼性の高い予測を優先的に反映する．一方で，ランダムに特徴量を選ぶ方法のように，OOBR^2 が必ずしも高くないモデルも一定の重みをもつため，多様な特徴選択によるバリエーションを保持できる．

3 実　　験

3.1 実験設定

実験の前処理として，全特徴量を標準化し，この標準化空間において欠損の導入および補完を行った．完全データに対し，セル単位の MCAR（Missing Completely At Random）で欠損を導入し，欠損率 $r_{miss} \in \{0.1, 0.3, 0.5\}$ とした．欠損セル数は $[Ndr_{miss}]$（N はサンプル数，d は特徴量数）であり，位置は一様無作為に重複なしで選択した．また，各反復において全手法で同一の欠損位置を用いることで公平性を担保した．各欠損率につき 50 回の独立反復を実施し，毎反復で欠損位置を再サンプリングした．評価はすべて標準化空間において行い，欠損セルのみを対象とした平均二乗誤差（MSE）を指標とした．表 1 は各手法・各欠損率における 50 回反復の平均 (標準偏差) である．

3.2 実験に用いたデータ

実験には，UCI Machine Learning Repository にて公開されているベンチマークデータを利用した．採用したデータセットは，“ iris ”，“ wine ”，“ diabetes ”である．ここでは上記のデータに対してクラスとして扱われるカテゴリカル変数を削除し，特徴量のみを用いて欠損補完精度を評価した．

3.3 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・各欠損率における 50 回反復の MSE の平均を表している．また，括弧内の数値は試行毎の MSE の標準偏差を示している．

iris では，欠損率が 0.1 および 0.3 で missForest が最良な結果を示し，提案手法も同等の精度を示した．一方，欠損率が 0.5 では kNN が最小となり，低次元データではユークリッド距離に基づく近傍探索が有効に働いたと考えられる．このことから，単純なデータ構造に対しては kNN や missForest と同等の性能を発揮しつつ，提案手法は過剰適応することなく安定した補完が可能であることが確認された．wine では，全欠損率で提案手法が最小誤差となり，優位性が示された．これは特徴量間の相関が比較的強い中規模データでは，特徴選択と OOB 重みによるアンサンブル効果が有効に機能した結果と考えられる．

diabetes では欠損率 0.1 の際に missForest が優れたが，0.3 と 0.5 では提案手法が最小誤差を示し，中程度の相関を持つデータに対しても有効性を確認できた．

4 まとめと考察

iris データセットのような小規模なデータセットでは，提案手法は既存法と同等の精度を維持した．一方で，中規模以上のデータセットにおいては，特徴量選択を導入することで欠損補完の精度改善が確認された．これは，ノイズ的あるいは寄与の小さい特徴量を含めずに学習することで，欠損位置の違いによる推定値のばらつきが抑制され，より安定した補完が可能となったためである．さらに，相関の強い特徴量を優先的に利用することで，欠損値を説明力の高い情報に基づいて推定できるようになった．このように，特徴量選択はモデルの安定性と汎化性能を高め，欠損補完における頑健性の向上に寄与したと考えられる．

表 1　提案法（RF-Ens）と既存補完手法の精度比較

data set	missing rate	mean	kNN(k=5)	missForest	RF-Ens
iris	0.1	1.022(0.017)	0.261(0.012)	0.235(0.014)	0.243(0.010)
	0.3	1.018(0.008)	0.430(0.011)	0.365(0.014)	0.369(0.010)
	0.5	1.023(0.001)	0.545(0.009)	0.609(0.016)	0.548(0.012)
wine	0.1	1.003(0.013)	0.528(0.011)	0.491(0.010)	0.488(0.010)
	0.3	1.014(0.007)	0.634(0.006)	0.591(0.006)	0.583(0.006)
	0.5	1.021(0.004)	0.812(0.005)	0.786(0.009)	0.726(0.005)
diabetes	0.1	1.016(0.010)	0.643(0.007)	0.510(0.006)	0.518(0.006)
	0.3	1.010(0.004)	0.834(0.004)	0.649(0.005)	0.632(0.004)
	0.5	1.008(0.003)	1.007(0.003)	0.855(0.005)	0.779(0.004)

5 おわりに

本研究ではアンサンブルベースの欠損補完方法の代表的手法である missForest との比較において特徴量選択を導入した．ベンチマークデータによる性能評価により改善が確認できた．

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] E. F. Codd, “A Relational Model of Data for Large Shared Data Banks,” Communications of the ACM (CACM), Vol. 13, No. 6, pp. 377–387, 1970.
Stekhoven, D. J. and Buhlmann, P., MissForest: nonparametric missing value imputation for mixed-type data, Bioinformatics, Vol. 28, No. 1, pp. 112-118 (2012).