ソフトウェア開発実績データにおける欠損値補完への非負値行列因子分解の適用

Applying of Nonnegative Matrix Factorization for Missing Imputation in Software Development Data

川辺 裕貴 * 北村 大地 † 柿元 健 ‡

あらまし　本稿では，ソフトウェア開発実績データに含まれる欠損値の補完手法に非負値行列因子分解を適用した結果を報告する．

1　はじめに

ソフトウェア開発プロジェクトの初期段階において，様々な指標を正確に予測することはプロジェクトを成功させる大きな要因のひとつである．各種予測の手法として，開発実績データに基づいた定量的手法が数多く提案されている．しかし，開発実績データには欠損値が含まれていることが多く，定量的手法では欠損値が含まれていると，モデル構築を行えない，もしくは，モデルが構築できても精度が低下してしまう．欠損値を含む開発実績データから欠損値を取り除くために，欠損値を削除，あるいは何らかの値で補完する欠損値処理 [1] が適用される．欠損値処理において，削除の過程では情報量の低下，補完の過程では誤差の含有を招き，欠損値の割合（欠損率）が高いデータにおいて顕著化することが懸念される．

本稿では，欠損値の補完処理として，教師なし学習で入力データの潜在パターンを抽出可能な非負値行列因子分解 (NMF) [2] に基づく手法を適用する．

2　NMF を用いた欠損値補完

NMF は，全成分が 0 以上の行列 (非負行列) を，別の 2 つの非負行列（係数行列及び基底行列）の行列積に分解する低ランク近似法である．この分解で得られる係数行列と基底行列は，観測行列中の潜在的な非負パターンとその係数をそれぞれ表す．欠損値を含む非負行列に対しては，欠損要素を NMF のコスト関数から除外する最適化手法 [3] が適用でき，欠損値を無視しつつ係数行列及び基底行列を推定できる．従って，欠損値を含む開発実績データに上記の手法を適用することで，欠損値の影響を受けずに潜在パターンと係数を抽出でき，さらに係数行列と基底行列の行列積により，欠損値が補完された「開発実績データの低ランク近似行列」が得られる．NMF を用いた欠損値補完の手順は以下のとおりである．

1. 開発実績データのメトリクスごとに最小値が 0，最大値が 1 となる正規化を行う
2. 開発実績データの非欠損値を 1，欠損値を 0 としたインデクス行列を作成する
3. インデクス行列が 0 の要素を除外する NMF の最適化法 [3] により，係数行列及び基底行列を得る
4. 係数行列と基底行列の行列積（低ランク近似行列）から欠損値の補完結果を得る

3　評価実験

NMF による欠損値補完の精度を確認するために評価実験を行った．実験には，プロジェクト数 499 件，メトリクス数 16 個で欠損値を含まない China データセットを使用し，MCAR，MAR，NM の各欠損メカニズム [4] により 10%〜40%の欠損値

*Yuki Kawanabe, 香川高等専門学校電気情報工学科

†Daichi Kitamura, 香川高等専門学校電気情報工学科

‡Takeshi Kakimoto, 香川高等専門学校電気情報工学科

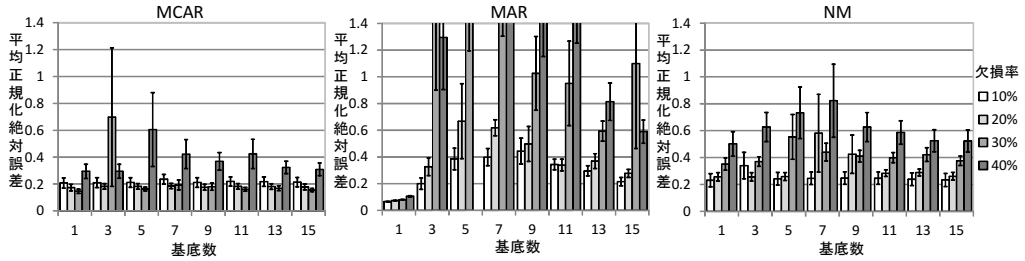

図 1　欠損メカニズムごとの欠損率と平均正規化絶対誤差の関係（基底数奇数のみ）

を与えた．欠損メカニズムにはランダム要素を含むため，MCAR と NM は各 10 回，MAR は欠損が依存するメトリクスを各メトリクスとした計 16 回与えたデータセットを作成した．欠損値を与えたデータセットに対して提案手法を適用し，補完値と実測値の誤差をメトリクスごとに正規化したうえで絶対誤差（MAE）の平均を算出した．提案手法の基底数（係数行列の列数かつ基底行列の行数）は 1〜15（メトリクス数 −1）で変化させた．

4　結果と考察

欠損メカニズム別に欠損率ごとの基底数と平均正規化絶対誤差の値を図 1 に示す．紙面の都合上，基底数が奇数の結果のみを示している．エラーバーは標準誤差を表す．図 1 より，欠損率 40%では欠損メカニズムに限らず精度が低下している．また，MCAR では欠損率 10%よりも 20%，30%の方が高い精度が得られているが，MCAR，NM では，欠損率が高くなるにつれて精度が低下する傾向がある．基底数に着目すると，MCAR は基底数に関わらず安定して精度が得られているが，MAR は欠損率 10%以外の精度は著しく低く，NM は基底数による差は小さいが欠損率増加による精度の低下が MCAR よりも大きくなっている．

MAR で，基底数が 1 の時のみ高い精度が得られている．基底数が 1 の場合は，各メトリクスの平均値とプロジェクトごとに求められた係数の積の行列となるため，平均値補完に近い補完となる．つまり，潜在パターンを求めない方が高い精度が得られており，MAR に対しては NMF を用いた欠損値補完は有効ではないことが考えられる．

5　まとめ

本稿では，非負値行列因子分解を用いてソフトウェア実績データの欠損値補完を行った．評価実験の結果，欠損メカニズムの特徴によって補完精度が大きく影響を受けることがわかった．他の欠損値処理との比較や，実際に工数等の予測を行って評価することが今後の課題である．

謝辞　本研究の一部は JSPS 科研費 JP19K11915，JP19K20306 の助成を受けた．

参考文献

[1] Myrtveit, I., et al. : Analyzing data sets with missing data: An Empirical Evaluation of Imputation Methods and Likelihood-Based Methods, IEEE Trans. Software Eng., vol. 27, no. 11, (2001), pp. 999–1013.
[2] Lee, D.D. and Seung, H.S., :Learning the parts of objects by non-negative matrix factorization, Nature, Vol. 401, (1999), pp. 788–791.
[3] Kitamura, D., et al. :Multichannel signal separation combining directional clustering and nonnegative matrix factorization with spectrogram restoration, IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 23, no. 4, (2015), pp. 654–669.
[4] Little, R.J.A., and Rubin, D.B., :Statistical analysis with missing data, 2nd edition, John Wiley and Sons, New York, (2002).