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

川辺裕貴　北村大地　〇柿元 健　（香川高専）

背景と問題点

コスト見積や欠陥予測等に用いられる定量的手法は，ソフトウェア開発実績データに基づいてモデルを構築する．ソフトウェア開発データには欠損値が含まれていることが多い．データに欠損値が含まれると，

- 定量的手法が適用できない
- 適用できても高い精度が得られない

という問題がある

欠損値を含むデータを定量的手法に利用するために欠損値を削除や補完する欠損値処理が適用される．

非負値行列因子分解を用いた欠損値補完

本研究では，欠損値を補完する手法として非負行列を2個の非負行列の行列積に分解する低ランク近似法である非負値行列因子分解（NMF）に基づく手法を適用する．具体的には欠損値除外NMF[1]を用いる．

NMFを用いることで，欠損値が補完された近似行列だけでなく，潜在パターンと係数を抽出できる．

図1にNMFを用いた欠損値補完の概要を示す．また，コスト関数と反復更新式には以下の式を用いた．

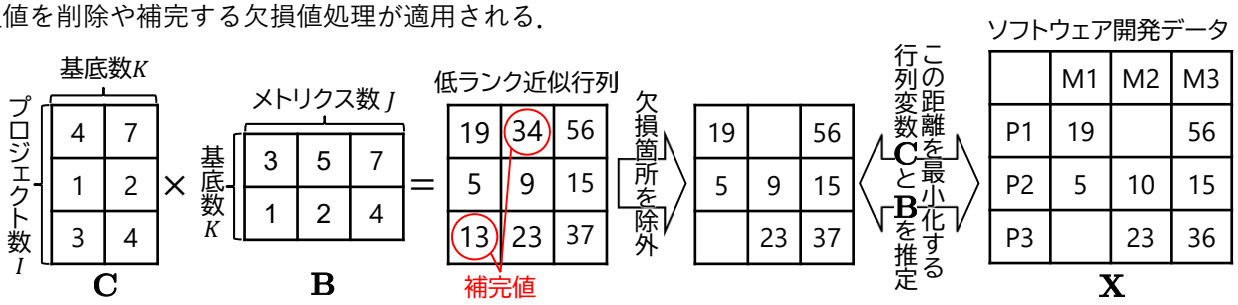

図1 NMFを用いた欠損値補完の概要（実際にはメトリクスごとに正規化を行っている）

最適化問題：$\mathrm{minimize}(\mathbf{C},\mathbf{B}) = \sum_{i=1}^{I}\sum_{j=1}^{J}\left(e_{ij}\left|x_{ij}-\sum_{k=1}^{K}(c_{ik}b_{kj})\right|^2\right)$

反復更新式：$c_{ik}^{(n+1)} = \dfrac{\sum_{j=1}^{J} e_{ij}x_{ij}b_{kj}^{(n)}}{\sum_{j=1}^{J}\left\{e_{ij}b_{kj}^{(n)}\left(\sum_{k'=1}^{K}c_{ik'}^{(n)}b_{k'j}^{(n)}\right)\right\}}c_{ik}^{(n)}$

$$b_{kj}^{(n+1)} = \frac{\sum_{i=1}^{I} e_{ij}x_{ij}c_{ik}^{(n)}}{\sum_{i=1}^{I}\left\{e_{ij}c_{ik}^{(n)}\left(\sum_{k'=1}^{K}c_{ik'}^{(n)}b_{k'j}^{(n)}\right)\right\}}b_{kj}^{(n)}$$

$\mathbf{X}$: 入力行列　$\mathbf{C}$: 係数行列　$\mathbf{B}$: 基底行列
I：プロジェクト数　J: メトリクス数　K: 基底数
e_{ij}: x_{ij}が欠損値なら0，非欠損値なら1
（$\mathbf{C},\mathbf{B}$の初期値は乱数）

[1] Kitamura, D., Saruwatari, H., Kameoka, H., Takahashi Y., Kondo, K., and Nakamura, S. :Multichannel signal separation combining directional clustering and nonnegative matrix factorization with spectrogram restoration, IEEE/ACM Transaction on Audio, Speech, and Language Processing, vol.23, no.4, (2015), pp.654-669.

欠損メカニズム

- MCAR (Missing Completely At Random)
  欠損する確率はデータ中のどの値にも依存しない
- MAR(Missing At Random)
  欠損する確率は他のメトリクスの値の大きさに依存する
- NM(Nonignorable Missingness)
  欠損する確率はその値自身の大きさに依存する

評価実験

Chinaデータセットからプロジェクト499件，メトリクス16個の欠損値を含まないデータセットを作成した．

作成したデータセットに対して，各欠損メカニズムごとに欠損率10, 20, 30, 40%で与えた．

評価指標には，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の平均値を算出した．

図3の棒グラフは精度，バーは標準偏差を表している．

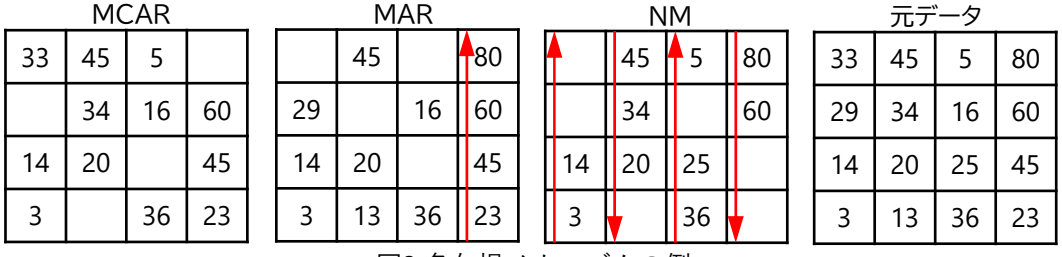

図2 各欠損メカニズムの例

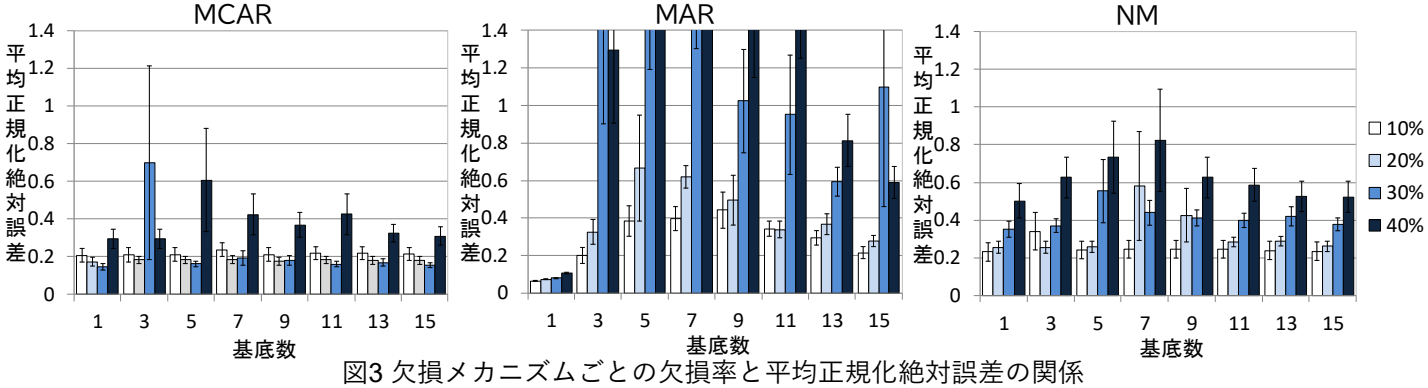

図3 欠損メカニズムごとの欠損率と平均正規化絶対誤差の関係