

カテゴリカル変数を含むソフトウェア開発実績データにおける 非負値行列因子分解による欠損値補完

川辺裕貴[†] 柿元健[‡]

香川高等専門学校 専攻科創造工学専攻[†] 香川高等専門学校 電気情報工学科[‡]

1. はじめに

ソフトウェアの開発プロジェクト計画作成は初期段階に正確なコスト見積を行うことが非常に重要である。正確にコストを管理し、プロジェクトを円滑に進めることでコスト超過を未然に防ぎプロジェクト完遂に必要な資源を確保することができる。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、欠損値が含まれていると、コスト見積が行えない場合がある。

また、開発実績データには欠損値の他にカテゴリカル変数を含むデータが得られることが多く数量的に扱うことが難しい。

これまでは、カテゴリカル変数を含まないメトリクスに対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 (nonnegative matrix factorization : NMF) [1]に基づく手法である欠損除外NMF[2]による欠損値補完により欠損値処理を行った[3]。

そこで本稿では、欠損除外NMFによる欠損値補完をカテゴリカル変数を含む開発実績データに適用させ欠損値処理を行い評価する。

2. 欠損除外 NMF による欠損値補完法

2.1 概要

本稿では、前稿と同様に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 NMF に基づく手法を用いる。具体的には欠損除外 NMF[1]による欠損値補完法である。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、メトリクスごとの特徴を表す基底行列とメトリクスごとの特徴とプロジェクトとの関係を表す係数行列に分解して潜在的パターンを抽出できる。

2.2 手順

欠損除外 NMF を用いた欠損値補完の手順は以下の通りである。

- ① ソフトウェア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 としたインデックス行列を作成する。
- ③ インデックス行列が 0 の要素を除外するNMFの最適化法により、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。
- ⑤ 欠損値補完した結果の正規化を復元することでレンジを元に戻す。

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式(1)、(2)及び(3)を用いる。

[コスト関数]

$$\min_{A,B} \|E \circ (X - AB)\|_{Fr}^2 \text{ s.t. } a_{ik}, b_{kj} \geq 0, \forall i,j,k \quad (1)$$

[反復更新式]

$$a_{ik}^{(n+1)} = a_{ik}^{(n)} \frac{\sum_{j=1}^J e_{ij} x_{ij} b_{kj}}{\sum_{j=1}^J e_{ij} b_{kj} \sum_{k=1}^K a_{ik}^{(n)} b_{kj}} \quad (2) \quad b_{kj}^{(n+1)} = b_{kj}^{(n)} \frac{\sum_{i=1}^I e_{ij} x_{ij} a_{ik}}{\sum_{i=1}^I e_{ij} a_{ik} \sum_{k=1}^K a_{ik}^{(n)} b_{kj}} \quad (3)$$

ここで、 X は実績データを行列で表した観測行列、 E はインデックス行列で、ともに $I \times J$ の非負行列 (I, J はプロジェクト数とメトリクス数)、 A は $I \times K$ の非負行列変数 (K は基底数)、 B は $K \times J$ の非負行列変数であり、それぞれ基底行列及び係数行列を表す。 e_{ij} 、 a_{ik} 及び b_{kj} は、それぞれ行列 E 、 A 及び B の非負の要素 (スカラー) である。

ソフトウェア開発実績データ X

	M1	M2
P1	443	
P2	323	1.0000
P3	527	1.0400

欠損要素の検出

インデックス行列 E

	1	2
1	0	
1	1	
1	1	

正規化

	1	2
1	0.079	
0	0	
0.136	1	

欠損除外NMF

	1	2	3	4
1	0.081	0.101	0.182	1.542
2	2.22e ⁻¹⁶	2.22e ⁻¹⁶	0.626	0.661
3	0.635	0.031		

A

B

	1	2
1	0.022	0.500
2	0.016	0.699
3	0.028	0.740

	M1	M2
P1	457.1	0.80
P2	332.0	0.99
P3	574.0	1.04

正規化の復元

	M1	M2
P1	457.1	0.80
P2	332.0	0.99
P3	574.0	1.04

補完値

近似行列 $\hat{X}$

図 1 NMFを用いた欠損値補完の概要

Missing Imputation Using Nonnegative Matrix Factorization in Software Development Data Including Categorical Variable
Kawanabe Yuki[†] Kakimoto Takeshi[‡]
Advanced Course in Industrial and Systems Engineering,
National Institute of Technology, Kagawa College[†]
Department of Electrical and Computer Engineering, National
Institute of Technology, Kagawa College[‡]

3. 評価実験

3.1 実験データセット

評価実験では、カテゴリカル変数を含んだ場合の精度の変化を基底数ごとに確認する。そのため、カテゴリ変数を含むデータセットと含まないデータセットの二種類のデータセットを作成する。

データセットは見積手法の評価に広く利用されており、ISBSGが収集した複数のソフトウェア開発企業の実績データから欠損値を含まない完全データセットを作成した。カテゴリカル変数を含まないデータセットは、プロジェクト数 755 件メトリクス数 17 件の完全データセットであり、カテゴリカル変数を含むデータセットは、カテゴリカル変数を含まないデータセットにカテゴリカル変数であるプログラミング言語を追加したプロジェクト数 755 件メトリクス数 18 件の完全データセットである。

3.2 実験方法

評価実験では、まず、カテゴリカル変数を含むデータセットにおいてカテゴリカル変数をダミー変数化させ数量的に扱えるようにした。そして、各データセットに欠損率 10%~40%の範囲で 10%ごとに欠損を与え、欠損除外NMFを用いて欠損値補完を行った。

評価指標には、式(4)で算出される平均正規化絶対誤差 (mean normalized absolute error : MNAE) を用いた。MNAEは値が小さいほど精度が高いことを示す。

$$MNAE = \frac{1}{m} \sum_{i=1}^m |V_i - M_i| \quad (4)$$

ここで、 V_i は正規化後の実測値、 M_i は正規化後の補完値、 m はプロジェクト数を示す。

4. 実験結果

カテゴリカル変数の有無ごとに基底数 K を 1~16 まで変化させたときのMNAEを図2、図3に示す。各グラフの縦軸はMNAE、横軸は基底数を表す。

図2、図3の結果から、カテゴリカル変数を含むデータセットでは基底数が2のときに最も精度が高く、含まないデータセットでは基底数が3のときに最も精度が高くなった。また、基底数を変化させたときの精度のばらつきも小さくなった。

5. 考察

カテゴリカル変数は離散値かつ取り得る値が限られているため、補完が連続値に比べ容易であると考えられる。そのため、パターン数である基底数は連続値よりも離散値でかつ取り得る

値が限られている方が少なくなるためカテゴリカル変数を含んでいるときよりも少ない基底数で最も精度が高くなったと考えられる。

6. まとめ

本論文ではカテゴリカル変数の有無ごとの精度の変化を確認した。評価実験の結果、カテゴリカル変数を含んだ際にも欠損除外NMFによる欠損値補完が適用できることが確認できた。

謝辞

本研究の一部はJSPS科研費JP19K11915 の助成を受けた。

参考文献

- [1] Lee, D.D. and Seung, H.S. : Learning the Parts of Objects by Non-negative Matrix Factorization, Nature, Vol.401, pp.788-791 (1999).
- [2] Kitamura, D. et al.: Multichannel Signal Separation Combining Directional Clustering and Nonnegative Matrix Factorization with Spectrogram Restoration, IEEE/ACM Trans. Audio, Speech, and Language Processing, Vol.23, No.4, pp.654-669 (2015).
- [3] 川辺裕貴, 北村大地, 柿元健: 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, 日本ソフトウェア科学会第 26 回ソフトウェア工学の基礎ワークショップ, pp.175~176 (2019).

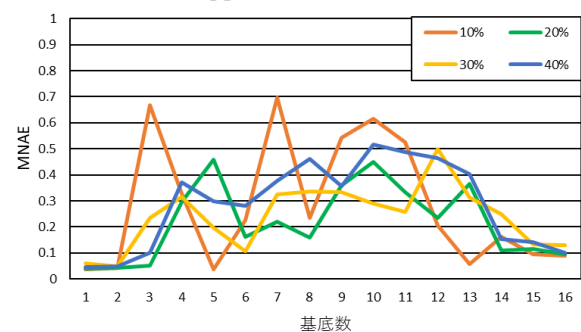


図2 カテゴリカル変数を含む場合のMNAE

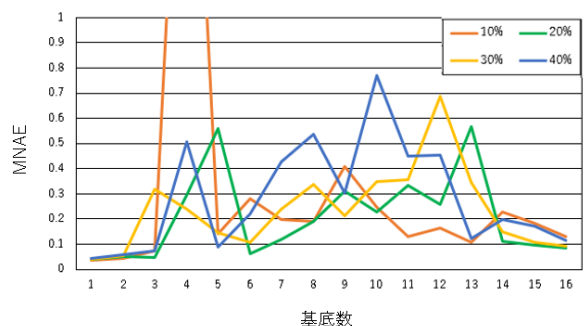


図3 カテゴリカル変数を含まない場合のMNAE