5M-08

最適化手法を用いた事例・変数統合型の欠損値補完

松本拓見 † 高野祐一 ‡

† 筑波大学大学院 システム情報工学研究群 ‡ 筑波大学 システム情報系

1 はじめに

近年は様々なデータセットを容易に入手できるようになったが，一方でデータの一部に欠損が含まれる場合が多い．統計的手法や機械学習のアルゴリズムは，完全なデータセットに依存しているため，欠損値を適切に補完する必要があり，多くの研究が行われている．

Bertsimas et al. [1] は，データセットにおいて事例間の距離を定義して近傍を決定し，その近傍間でより距離が近づくように欠損値を補完する最適化手法を提案した．この手法で欠損値を補完することで，多くのデータセットで従来の手法よりも優れた性能を示している．

欠損値補完の一種とみなせる推薦システムの文脈では Wang et al. [2] が，利用者近傍，アイテム近傍，利用者・アイテム近傍の 3 種類の近傍を利用することで，評価行列の推定精度を向上させている．

本研究では Bertsimas et al. [1] の事例間の近傍に基づいて欠損値を補完する最適化手法に，Wang et al. [2] で使われている近傍のアイデアを組み合わせる．つまり事例近傍間だけでなく変数間近傍も考慮し，最適化手法を用いることで欠損値を補完する．そして数値実験を実施して，提案手法の有効性を検証する．

2 既存研究

2.1 最適化手法を用いた欠損値補完

データセット $\boldsymbol{X} = \{\boldsymbol{x}_j\}_{j=1}^{n}$ が与えられているとする．変数は p 個あり，各変数 d は $d \in \{1, 2, \ldots, p_0\}$ のとき量的変数を表し，$d \in \{p_0+1, p_0+2, \ldots, p_0+p_1\}$ のとき質的変数を表す．また，$\boldsymbol{W} \in \mathbb{R}^{n \times p_0}$ を欠損値補完後の量的変数の行列，$\boldsymbol{V} \in \prod_{d=1}^{p_1}\{1, 2, \ldots, k_d\}$ を欠損値補完後の質的変数の行列とする．したがって，欠損値補完により（欠損していない成分も含めて）事例 $\boldsymbol{x}_j$ は $(\boldsymbol{w}_j, \boldsymbol{v}_j)$ に変換される．このとき以下の決定変数と集合を定義する．

$$z_{uj} = \begin{cases} 1 & \text{事例 } (\boldsymbol{w}_j, \boldsymbol{v}_j) \text{ が事例 } (\boldsymbol{w}_u, \boldsymbol{v}_u) \text{ の } K \text{ 近傍である} \\ 0 & \text{それ以外} \end{cases}$$

$$U = \{u : \text{事例 } \boldsymbol{x}_u \text{ に欠損成分がある }\},$$
$$N_0 = \{(u,d) : \text{成分 } x_{ud} \text{ が既知である}, 1 \le d \le p_0\},$$
$$N_1 = \{(u,d) : \text{成分 } x_{ud} \text{ が既知である}, p_0+1 \le d \le p_0+p_1\}$$

また，支持関数 $\mathbb{I}_{\{A\}}$ は命題 A が真の場合に 1，偽の場合に 0 をとる．このとき以下の目的関数を考える．

$$\sum_{u \in U}\sum_{j=1}^{n} z_{uj}\left[\sum_{d=1}^{p_0}(w_{ud}-w_{jd})^2 + \sum_{d=p_0+1}^{p_0+p_1}\mathbb{I}_{\{v_{ud} \neq v_{jd}\}}\right] \quad (1)$$

この目的関数は事例 $(\boldsymbol{w}_u, \boldsymbol{v}_u)$ と K 近傍にある事例との距離の総和を表しており，目的関数を最小化することで近傍間の事例がより近くなるように欠損値を補完する．式 (1) に複数の制約条件を加え，交互最適化を用いて最適化問題を解くことで多くのデータセットで優れた補完精度を示している [1].

2.2 評価行列の推定

Wang et al. [2] の手法では，利用者近傍とアイテム近傍を融合し，評価値を推定する際には互いに補完し合うことで，予測精度を向上させる．例を図 1 に示す．灰色の成分は欠損値を表す．値が既知でありかつ緑色の枠で囲まれた成分を利用者近傍，オレンジ色の枠で囲まれた成分をアイテム近傍，赤色の枠で囲まれた成分を利用者・アイテム近傍とする．欠損成分である評価値 x_{ud} を推定する際は，利用者近傍の評価値，アイテム近傍の評価値の成分を利用するだけでなく，利用者・アイテム近傍の評価値まで活用し，類似度で重みづけして x_{ud} の値を推定する．

3 提案手法

新たな決定変数と集合を以下のように定義する．ただし，変数 d の列ベクトルを $\boldsymbol{x}_{(d)}$ と表す．

$$z_{uj}^{(\mathrm{U})} = \begin{cases} 1 & \text{事例 } (\boldsymbol{w}_u, \boldsymbol{v}_u) \text{ が事例 } (\boldsymbol{w}_j, \boldsymbol{v}_j) \text{ の } K \text{ 近傍} \\ 0 & \text{それ例外} \end{cases}$$

$$z_{id}^{(\mathrm{I})} = \begin{cases} 1 & \text{変数 } (\boldsymbol{w}_i, \boldsymbol{v}_i) \text{ が変数 } (\boldsymbol{w}_d, \boldsymbol{v}_d) \text{ の } K \text{ 近傍である} \\ 0 & \text{それ以外} \end{cases}$$

Missing value imputation unifying instances and variables using optimization methods

Takumi MATSUMOTO†, Yuichi TAKANO†

†Degree Programs in Systems and Information Engineering, University of Tsukuba

‡Institute of Systems and Information Engineering, University of Tsukuba

$$
\begin{aligned}
I = \{&i : \text{変数 } \boldsymbol{x}_{(i)} \text{ に欠損成分がある},\\
&1 \le i \le p\},\\
I_0 = \{&i : \text{量的変数 } \boldsymbol{x}_{(i)} \text{ に欠損成分がある},\\
&1 \le i \le p_0\},\\
I_1 = \{&i : \text{質的変数 } \boldsymbol{x}_{(i)} \text{ に欠損成分がある},\\
&p_0 + 1 \le i \le p_0 + p_1\}
\end{aligned}
$$

$\boldsymbol{Z}^{(\mathrm{U})}$ は事例間の近傍変数の行列，$\boldsymbol{Z}^{(\mathrm{I})}$ は変数間の近傍変数の行列とする時，以下の目的関数を考える．

$$
\begin{aligned}
&c(\boldsymbol{Z}^{(\mathrm{U})}, \boldsymbol{Z}^{(\mathrm{I})}, \boldsymbol{W}, \boldsymbol{V}; \boldsymbol{X}) =\\
&\sum_{u \in U} \sum_{j=1}^{n} z_{uj}^{(\mathrm{U})} \left[\sum_{d=1}^{p_0} (w_{ud} - w_{jd})^2 + \sum_{d=p_0+1}^{p_0+p_1} \mathbb{I}_{\{v_{ud} \neq v_{jd}\}} \right]\\
&+ \sum_{i \in I_0} \sum_{d=1}^{p_0} z_{id}^{(\mathrm{I})} \left[\sum_{u=1}^{n} (w_{ui} - w_{ud})^2 \right]\\
&+ \sum_{i \in I_1} \sum_{d=p_0+1}^{p_0+p_1} z_{id}^{(\mathrm{I})} \left[\sum_{u=1}^{n} \mathbb{I}_{\{v_{ui} \neq v_{ud}\}} \right] \qquad (2)
\end{aligned}
$$

式 (2) の第 1 項は，事例間の近傍に基づく既存手法の目的関数 (1) に相当する．一方で式 (2) の第 2 項と第 3 項は変数間の近傍に基づいており，第 2 項は量的変数の近傍間の距離の総和を表し，第 3 項は質的変数の近傍間の距離の総和を表す．目的関数 (2) に制約条件を加え，以下の最適化問題を解くことで欠損値を補完する．

$$
\begin{aligned}
\text{minimize} \quad & c(\boldsymbol{Z}^{(\mathrm{U})}, \boldsymbol{Z}^{(\mathrm{I})}, \boldsymbol{W}, \boldsymbol{V}; \boldsymbol{X})\\
\text{subject to} \quad & w_{ud} = x_{ud} \quad ((u,d) \in N_0) && (3)\\
& v_{ud} = x_{ud} \quad ((u,d) \in N_1) && (4)\\
& z_{uu}^{(\mathrm{U})} = 0 \quad (u \in U) && (5)\\
& z_{ii}^{(\mathrm{I})} = 0 \quad (i \in I) && (6)\\
& \sum_{j=1}^{n} z_{uj}^{(\mathrm{U})} = K^{(\mathrm{U})} \quad (u \in U) && (7)\\
& \sum_{d=1}^{p_0+p_1} z_{id}^{(\mathrm{I})} = K^{(\mathrm{I})} \quad (i \in I) && (8)\\
& \boldsymbol{Z}^{(\mathrm{U})} \in \{0,1\}^{|U| \times n} && (9)\\
& \boldsymbol{Z}^{(\mathrm{I})} \in \{0,1\}^{|I| \times p} && (10)
\end{aligned}
$$

制約条件の式 (3)，(4) は既知の成分は固定することを表す．式 (5)，(6) は同一の事例や変数は近傍にはならないことを表す．式 (7)，(8) 式は近傍のサイズを指定している．なお $K^{(\mathrm{U})}, K^{(\mathrm{I})}$ は超パラメータであり，実験では値を変えながら精度を比較する．

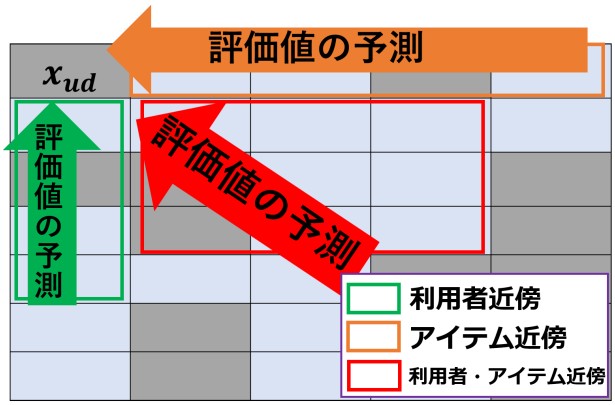

図 1: 評価行列の推定 [2] の概要

上式では決定変数が $\boldsymbol{Z}^{(\mathrm{U})}, \boldsymbol{Z}^{(\mathrm{I})}, \boldsymbol{W}, \boldsymbol{V}$ であり，非凸の最適化問題となる．そこで本研究では Bertsimas et al. [1] で扱われていた交互最適化を変数方向にも拡張して用いる．交互最適化は以下の 4 ステップで構成される．

ステップ 1（初期化）：欠損部分に初期値を代入する
ステップ 2（近傍の決定）：$\boldsymbol{W}$（量的変数）と $\boldsymbol{V}$（質的変数）を固定して $\boldsymbol{Z}^{(\mathrm{U})}, \boldsymbol{Z}^{(\mathrm{I})}$（近傍変数）を更新
ステップ 3（欠損値の補完）：$\boldsymbol{Z}^{(\mathrm{U})}, \boldsymbol{Z}^{(\mathrm{I})}$（近傍変数）を固定して $\boldsymbol{W}$（量的変数）と $\boldsymbol{V}$（質的変数）を更新
ステップ 4（収束判定）：目的関数の改善度をみて，収束条件を満たす場合は終了し，満たさない場合はステップ 2 に戻る．

4 数値実験

既存手法と比較することで提案手法の有効性を検証する．数値実験の詳細は当日述べる．

参考文献

[1] D. Bertsimas, C. Pawlowski & Y. D. Zhuo. From predictive methods to missing data imputation: An optimization approach. *Journal of Machine Learning Research*, 18(1), pp. 7133–7171, 2017.

[2] J. Wang, A. P. De Vries, & M. J. Reinders. Unifying user-based and item-based collaborative filtering approaches by similarity fusion. *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 501–508, 2006.