

L1 正則化非負値行列因子分解のための新たな階層的交互最小二乗法

A Novel Hierarchical Alternating Least Squares Algorithm for L1 Regularized Nonnegative Matrix Factorization

佐野 雄大[†] 右田 剛史[†] 高橋 規一[†]
Takehiro Sano Tsuyoshi Migita Norikazu Takahashi

1 はじめに

非負値行列因子分解 (NMF: Nonnegative Matrix Factorization) [1, 2] とは次元削減手法の 1 つであり, 顔画像処理, 音響信号処理, ネットワーク分析といった様々な分野に応用されている. 与えられた非負値行列と 2 つの非負値因子行列の積との誤差を最小化する制約付き最適化問題として定式化され, 誤差尺度としてフロベニウスノルムや一般化 KL ダイバージェンスがしばしば用いられる. 一方, 因子行列の疎性を制御するために L1 正則化項が用いられることもある. その場合, NMF は次の最適化問題として定式化される.

$$\begin{aligned} \text{minimize} \quad & f(\mathbf{U}, \mathbf{V}) = \frac{1}{2} \|\mathbf{X} - \mathbf{UV}^T\|_F^2 + \alpha \|\mathbf{U}\|_1 + \beta \|\mathbf{V}\|_1 \\ \text{subject to} \quad & \mathbf{U} \geq \mathbf{0}_{M \times K}, \mathbf{V} \geq \mathbf{0}_{N \times K} \end{aligned} \quad (1)$$

ここで $\mathbf{X} \in \mathbb{R}_+^{M \times N}$ は与えられた非負値行列であり, $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_K] \in \mathbb{R}_+^{M \times K}$ と $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_K] \in \mathbb{R}_+^{N \times K}$ は求める非負値因子行列, $\alpha, \beta \in \mathbb{R}_+$ は疎性を制御するためのパラメータである.

最適化問題 (1) のための反復計算法として次の乗法型更新式 (MUR) が広く知られている.

$$\begin{aligned} u_{ik} &\leftarrow \max \left\{ \epsilon, u_{ik} \frac{(\mathbf{XV})_{ik} - \alpha}{(\mathbf{UV}^T\mathbf{V})_{ik}} \right\} \\ v_{jk} &\leftarrow \max \left\{ \epsilon, v_{jk} \frac{(\mathbf{X}^T\mathbf{U})_{jk} - \beta}{(\mathbf{VU}^T\mathbf{U})_{jk}} \right\} \end{aligned} \quad (2)$$

ただし, $\epsilon \in \mathbb{R}_{++}$ は微小な正定数である. 更新式 (2) は勾配降下法の特殊な形とみなされるため, 大規模データに適した計算法である. 一方, 更新式 (2) は一反復あたりの目的関数値の減少が緩やかになることがあり, 目的関数の最小化に多くの反復を要する. その問題を解決するために Cichocki らは次の階層的交互最小二乗 (HALS) 法 [3] を提案した.

$$\begin{aligned} \mathbf{u}_k &\leftarrow \max \left\{ \epsilon, \frac{\mathbf{R}_k \mathbf{v}_k - \alpha \mathbf{1}_{M \times 1}}{\mathbf{v}_k^T \mathbf{v}_k} \right\} \\ \mathbf{v}_k &\leftarrow \max \left\{ \epsilon, \frac{\mathbf{R}_k^T \mathbf{u}_k - \beta \mathbf{1}_{N \times 1}}{\mathbf{u}_k^T \mathbf{u}_k} \right\} \end{aligned} \quad (3)$$

ただし, $\mathbf{R}_k = \mathbf{X} - \sum_{\tilde{k}=1, \tilde{k} \neq k}^K \mathbf{u}_{\tilde{k}} \mathbf{v}_{\tilde{k}}^T$ である.

どちらの計算法も目的関数値の単調減少性が保証されているが, 更新式 (2) と (3) から得られる因子行列の各要素は 0 の値をとることができないため, 疎行列とはならない. 本研究では, 更新式 (3) を改良して疎な因子行列を得るための計算法を提案する. 加えて, 提案手法の理論解析と数値実験もおこなう.

2 HALS 法の改良

BCD 法 [4] を用いて最適化問題 (1) を解くことは, 次の $2K$ 個の問題を繰り返し解くことに帰着される.

$$\begin{aligned} \text{minimize} \quad & g(\mathbf{u}_k) = \frac{1}{2} \|\mathbf{v}_k \mathbf{u}_k^T - \mathbf{R}_k^T\|_F^2 + \alpha \|\mathbf{u}_k\|_1 \\ \text{subject to} \quad & \mathbf{u}_k \geq \mathbf{0}_{M \times 1} \end{aligned} \quad (4)$$

$$\begin{aligned} \text{minimize} \quad & h(\mathbf{v}_k) = \frac{1}{2} \|\mathbf{u}_k \mathbf{v}_k^T - \mathbf{R}_k\|_F^2 + \beta \|\mathbf{v}_k\|_1 \\ \text{subject to} \quad & \mathbf{v}_k \geq \mathbf{0}_{N \times 1} \end{aligned} \quad (5)$$

$$(k = 1, 2, \dots, K)$$

そこで, 最適化問題 (4) と (5) を解くために次の反復計算法を提案する.

$$\begin{aligned} \mathbf{u}_k &\leftarrow \frac{\max \{0, \mathbf{R}_k \mathbf{v}_k - \alpha \mathbf{1}_{M \times 1} + \delta \mathbf{u}_k\}}{\mathbf{v}_k^T \mathbf{v}_k + \delta} \\ \mathbf{v}_k &\leftarrow \frac{\max \{0, \mathbf{R}_k^T \mathbf{u}_k - \beta \mathbf{1}_{N \times 1} + \delta \mathbf{v}_k\}}{\mathbf{u}_k^T \mathbf{u}_k + \delta} \end{aligned} \quad (6)$$

ただし, $\delta \in \mathbb{R}_{++}$ は微小な正定数である. 更新式 (6) は HALS 法の更新式 (3) と類似した計算法だが, 更新後の値が 0 となり得る点が (3) と異なる.

更新式 (6) に関して次の定理が成り立つ.

定理 1 更新式 (6) は目的関数値を単調に減少させる.

証明 最適化問題 (4) について考える. $\mathbf{u}_k = [x_1, x_2, \dots, x_M]^T$, $\mathbf{R}_k^T = [\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_M]$ とおくと, $g(\mathbf{u}_k)$ は

$$g(\mathbf{u}_k) = \frac{1}{2} \sum_{a=1}^M \|\mathbf{v}_k x_a - \mathbf{y}_a\|_2^2 + \alpha \sum_{a=1}^M x_a$$

と書き換えることができるので, 最適化問題 (4) を解くことは次の M 個の部分問題を解くことと等価である.

$$\begin{aligned} \text{minimize} \quad & k(x_a) = \frac{1}{2} \|\mathbf{v}_k x_a - \mathbf{y}_a\|_2^2 + \alpha x_a \quad (a = 1, 2, \dots, M) \\ \text{subject to} \quad & x_a \geq 0 \end{aligned}$$

目的関数が $k(x_a) = \frac{1}{2} \mathbf{v}_k^T \mathbf{v}_k x_a^2 + (\alpha - \mathbf{y}_a^T \mathbf{v}_k) x_a + \frac{1}{2} \mathbf{y}_a^T \mathbf{y}_a$ と書けることに注意し,

$$l(x, x_a) = k(x_a) + k'(x_a)(x - x_a) + (k''(x_a)/2! + \delta/2)(x - x_a)^2$$

と定義すると, これは $k(x_a)$ の補助関数 [2] である. 補助関数 $l(x, x_a)$ の x に関する最小点を求め, それを用いて x_a の値を更新すると $k(x_a)$ の値は減少する [2]. そこで方程式 $\partial l(x, x_a) / \partial x = k'(x_a) + (k''(x_a) + \delta)(x - x_a) = 0$ を x について解くと

$$x = \frac{\mathbf{y}_a^T \mathbf{v}_k - \alpha + \delta x_a}{\mathbf{v}_k^T \mathbf{v}_k + \delta} \quad (7)$$

が得られる. 非負制約に注意すると, 式 (7) の右辺が正ならば $l(x, x_a)$ の最小点は (7) で与えられ, そうでなければ

[†] 岡山大学大学院自然科学研究科 Graduate School of Natural Science and Technology, Okayama University

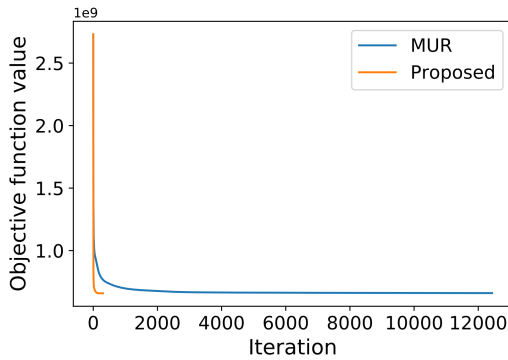


図1 COIL20 データセット

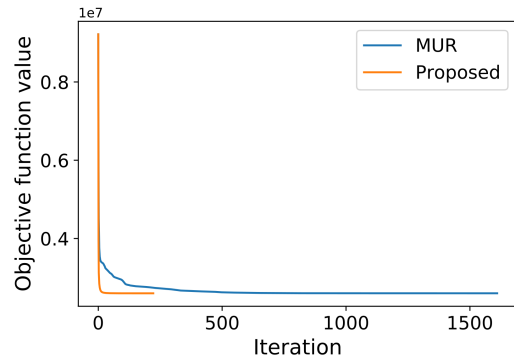


図2 CLUTO (tr23) データセット

ば $x = 0$ が最小点となる. 更新式 (6) の第一式は以上の解析によって導かれたものであり, 補助関数の性質から, この更新によって目的関数値は減少する. 更新式 (6) の第二式も同様の解析によって導かれるので, この更新によって目的関数値は減少する. \square

定理1より, $C = f(\mathbf{U}^{(0)}, \mathbf{V}^{(0)})$ とおくと, 任意の反復回数 t に対して $0 \leq au_{ik}^{(t)} \leq C$ および $0 \leq \beta v_{jk}^{(t)} \leq C$ が成り立つ. したがって, Zangwill の大域収束定理 [5] より次の定理が成り立つ.

定理2 $\alpha > 0, \beta > 0$ とすると, 任意の初期値に対して更新式 (6) によって生成される点列 $\{(\mathbf{U}^{(t)}, \mathbf{V}^{(t)})\}_{t=0}^{\infty}$ は少なくとも1つ収束部分列をもち, かつ任意の収束部分列の極限は最適化問題 (1) の停留点である.

3 数値実験

本章では2つの数値実験を実施する. 実験1は更新式 (2), (3) および (6) における目的関数値の推移の比較である. 実験2は更新式 (3) と (6) によって得られる因子行列の疎性の比較である. どちらの実験でも $\epsilon = \delta = 10^{-8}$ とし, COIL20 データセット¹⁾ と CLUTO (tr23) データセット²⁾ を用いた. また, 初期値は特異値分解に基づく初期化法 [6] で設定し, 終了条件には目的関数値の単調減少性を考慮して次の不等式を用いた.

$$\frac{f(\mathbf{U}^{(t)}, \mathbf{V}^{(t)}) - f(\mathbf{U}^{(t+1)}, \mathbf{V}^{(t+1)})}{f(\mathbf{U}^{(0)}, \mathbf{V}^{(0)}) - f(\mathbf{U}^{(t+1)}, \mathbf{V}^{(t+1)})} \leq 10^{-7}$$

3.1 実験1

$\alpha = 1.0, \beta = 1.0$ とした場合の実験結果を図1と図2に示す. どちらのデータセットにおいても更新式 (6) は更新式 (2) より少ない反復回数で終了条件を満たした. 一方, 図1と図2には示していないが, 更新式 (3) は更新式 (6) と極めて類似した収束曲線となった.

3.2 実験2

疎性を評価するために, 次の指標 [7] を用いる.

$$\text{sparseness}(\mathbf{x}) = \frac{\sqrt{N} - \left(\sum_{i=1}^N |x_i| \right) / \sqrt{\sum_{i=1}^N x_i^2}}{\sqrt{N} - 1}$$

ただし, $\mathbf{x} \in \mathbb{R}^N, \mathbf{x} \neq \mathbf{0}$ である. この指標は0から1までの値をとり, 疎であるほど1に近い値となる.

1) <https://www.cs.columbia.edu/CAVE/software/softlib/coil-20.php>
2) <http://glaros.dtc.umn.edu/gkhome/views/cluto>

表1 疎性の評価

	COIL20	CLUTO (tr23)
特徴量の次元数 M	1024	5833
データ数 N	1440	204
クラス数 K	20	6
HALS 法	0.572393792	0.657563256
提案手法	0.572393794	0.657563753

低次元データの疎性はクラスタリングのような応用例において重要であるため, 低次元データ, すなわち, 行列 \mathbf{V}^T の列 $\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_N$ を用いて $\frac{1}{N} \sum_{i=1}^N \text{sparseness}(\mathbf{z}_i)$ を算出した. ただし, $\alpha = 0.0, \beta = 10.0$ とした. 実験結果を表1に示す. 更新式 (6) の方がより疎な結果が得られている. さらに, 更新式 (6) で得られた行列 \mathbf{V} は COIL20 データセットの場合 42.4%, CLUTO (tr23) データセットの場合 38.6% が0となることも確認した.

4 おわりに

本研究では, L1 正則化項をもつ NMF のための階層的交互最小二乗法を改良し, その理論解析をおこなった. また, 数値実験によって既存手法との比較もおこない, 疎な因子行列が高速に得られることも確認した. 一方, L1 正則化非負値テンソル因子分解に対して本研究を応用することが今後の課題として挙げられる.

参考文献

- [1] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791 (1999).
- [2] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," *Advances in Neural Information Processing Systems*, pp. 556–562 (2001).
- [3] A. Cichocki, R. Zdunek, and S. Amari, "Hierarchical als algorithms for nonnegative matrix and 3d tensor factorization," *International Conference on Independent Component Analysis and Signal Separation*, pp. 169–176 (2007).
- [4] J. Kim, Y. He, and H. Park, "Algorithms for nonnegative matrix and tensor factorizations: A unified view based on block coordinate descent framework," *Journal of Global Optimization*, vol. 58, no. 2, pp. 285–319 (2014).
- [5] W. I. Zangwill, *Nonlinear Programming: A Unified Approach*, Englewood Cliffs, NJ, Prentice-Hall (1969).
- [6] C. Boutsidis and E. Gallopoulos, "SVD based initialization: A head start for nonnegative matrix factorization," *Pattern Recognition*, vol. 41, no. 4, pp. 1350–1362 (2008).
- [7] P. O. Hoyer, "Non-negative matrix factorization with sparseness constraints," *Journal of Machine Learning Research*, vol. 5, pp. 1457–1469 (2004).