One Class SVM に基づく Kernel NMF の幾何学的構造解釈手法に関する研究

1X18C001-1　飯塚悠太郎
指導教員　後藤正幸

1. 研究背景と目的

次元削減に基づくクラスタリングや情報圧縮は，画像や文書などの高次元データの分析における解釈性向上や計算コスト削減などに有用である．このような次元削減手法の1つに，非負値データを対象とした行列分解である Non-negative Matrix Factorization（以下，NMF）がある．NMF は，圧縮する次元（以下，基底）数の事前設定が必要な上，解が初期値に依存し一意に定まらない（解の非一意性）という問題点を持つ．これに対し Klingberg ら [1] は，解の非一意性を解決するため NMF を「最小の凸錐体を形成する基底を見つける問題」として幾何学的に解釈する手法を提案した．さらに Essid[2] は，時間計算量の削減および基底数の決定に One Class Support Vector Machine（以下，OC-SVM）を用いる Support Vector NMF（以下，SV-NMF）を提案している．これにより，得られたサポートベクトルから基底を選択することで，基底数を自動で決定することが可能となった．

OC-SVM は従来，各データ点との距離が近くデータが密に分布するデータ群である標準データに対して，各データ点から距離が遠いデータである外れ値データを検出する異常検知手法である．そのため，画像などの複雑な構造をもつデータに SV-NMF を適用する場合，基底に外れ値データが含まれてしまい，データ全体の様相を解釈することが困難となる．

そこで，本研究では，基底から外れ値データを除去することで，データに対する解釈性を向上させるような基底の決定手法を提案する．最後に，提案手法を外れ値データを含んだ実データセットに適用し，得られた基底や基底数の変化から，提案手法の有効性を検証する．

2. 提案手法への準備

2.1. Kernel NMF

NMF は，J 次元の非負値をもつベクトルであるデータ $\boldsymbol{v}_i (i \in \{1,\ldots,I\})$ に対して，行列 $\boldsymbol{V} = (\boldsymbol{v}_1,\ldots,\boldsymbol{v}_I)$ を基底行列 $\boldsymbol{W}$ と各データの特徴を表す重み行列 $\boldsymbol{H} = (\boldsymbol{h}_1,\ldots,\boldsymbol{h}_I)$ の積に分解する手法である．このように非負値のデータを加法的な構成成分に分解することで，より低次元の特徴量で各パターンを表現する．これに対し，NMF を拡張し，非線形データに対応可能な形にしたカーネル NMF のモデル式は以下の式 (1) で表される．

$$\boldsymbol{V}_\phi = \boldsymbol{W}_\phi \boldsymbol{H} \tag{1}$$

ここで，$\boldsymbol{V}_\phi = (\phi(\boldsymbol{v}_1),\ldots,\phi(\boldsymbol{v}_I))$ とし，$\phi(\cdot)$ は，カーネル空間 $\mathcal{H}$ 上への写像を意味する．また，$\boldsymbol{W}_\phi$ はカーネル空間 $\mathcal{H}$ 上の基底行列を表す．

2.2. NMF の幾何学的解釈と SV-NMF

NMF の解の非一意性という問題に対し，Klingberg らは，$\boldsymbol{W}$ によって生成される凸錐体を用いた幾何学的解釈の方法を提案している [1]．$\boldsymbol{V}$ が 2 次元の場合における，NMF の幾何学的解釈のイメージを図 1 に示す．

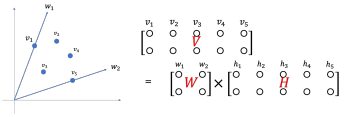

図 1: NMF の幾何学的解釈

$\boldsymbol{V}$ の各列を 2 次元空間上のデータ点，$\boldsymbol{W}$ の各行を原点から伸びる基底とすると，$\boldsymbol{W}$ の各行はデータ点の集合する領域を包含する凸錐体の外縁を表す．この幾何学的解釈により，NMF は凸錐体の外縁を求める問題に帰着できる．さらに，最小の凸錐体を求めることで，データの特徴をうまく捉えた行列分解が可能となり，多次元の場合も同様に解釈できる．

しかし，この幾何学的解釈による最小の凸錐体の探索は，時間計算量が大きくなる．そこで，これを容易に実現するために，Essid は OC-SVM を利用し，低計算コストで基底を探索する SV-NMF を提案した [2]．OC-SVM は，分離超平面を境に正例である標準データと負例である外れ値データを識別する教師なし手法である．OC-SVM の任意の 2 点のデータ $\boldsymbol{x}_i$，$\boldsymbol{x}_\ell$ の距離に関する RBF カーネル $\kappa(\boldsymbol{x}_i,\boldsymbol{x}_\ell)$ を $\kappa'(\boldsymbol{x}_i,\boldsymbol{x}_\ell) = \kappa(\boldsymbol{x}_i,\boldsymbol{x}_\ell)/\sqrt{\kappa(\boldsymbol{x}_i,\boldsymbol{x}_i)\kappa(\boldsymbol{x}_\ell,\boldsymbol{x}_\ell)}$ に変換すると，特徴空間内のすべてのデータ点が単位超球面上に存在するよう写像される．

このことから，最小の凸錐体を求めることは，データを原点から最大のマージンで分離する超平面を求めることと同等となる．この性質を明らかにした SV-NMF では，OC-SVM を利用することで最小の凸錐体を低計算コストで求めることができ，分離超平面を支えるサポートベクトルによって基底ベクトルを一意に選択可能となる．

3. 提案手法

3.1. 提案への着想

NMF の幾何学的解釈の方法は，外れ値データが存在すると適切な基底が決定できないと指摘されている [1]．つまり，SV-NMF は「実データなどの複雑なデータに適用する場合，基底にデータ全体の重要な特徴をよく表すグループを形成できないような外れ値データが含まれてしまい，基底の解釈が困難となる」という問題がある．

これは SV-NMF の基底の選択法に原因がある．分離超平面を支えるサポートベクトルは SVM，OC-SVM 共にラグランジュ乗数 $\alpha_i \neq 0$ となる i 番目のデータ点のことである．ただし，OC-SVM には外れ値データの割合を決定する事前パラメータ ν がある．この ν は分離超平面の決定に影響するため，サポートベクトル数は ν に対して単調増加する．そこで SV-NMF では，OC-SVM が Margin SVM であることに着目し，全サポートベクトル (以下，All SVs) の中から $0 < \alpha_i < 1$ となる Margin SVs を基底として取得し

ている．しかし，この方法では，負例側のサポートベクトルも基底に選択するため，実データなどの外れ値データを含むデータに適用する場合，基底に外れ値データを含んでしまう．そこで，本研究では SV-NMF をもとに，外れ値データに影響されにくい少数の基底を選択可能で，標準データの特徴を捉えた解釈性の高い基底の取得方法を提案する．

3.2. 基底の選択法

データ点 $\boldsymbol{x}_\ell$ に対する OC-SVM の判別関数は式 (2) で表される．

$$f(\boldsymbol{x}_\ell) = \sum_{i=1}^{I} \alpha_i \kappa(\boldsymbol{x}_i, \boldsymbol{x}_\ell) - \rho \tag{2}$$

式 (2) に対し，負例である外れ値データは $f(\boldsymbol{x}_\ell) \leq 0$ となる $\boldsymbol{x}_\ell$ である．そこで本提案では，Margin SVs のうち，$f(\boldsymbol{x}_\ell) > 0$ となる標準データの $\boldsymbol{x}_\ell$(以下，Margin SVs and Positive) をサポートベクトルとして選択する．提案手法により，外れ値データを含まないデータの重要な特徴をよく表した基底を獲得することが可能になる．

4. 再構成誤差

SV-NMF および提案手法は Kernel NMF と同様の構造をもつ．そのため，基底ベクトルが選択されたのちに，式 (1) に基づき，$\mathcal{H}$ 上での再構成誤差を最小とする $\boldsymbol{H}$ を求める．このとき，全データに対する再構成誤差 C_ϕ は以下のように定義される．

$$C_\phi = \frac{1}{I} \sum_{i=1}^{I} \|\phi(\boldsymbol{v}_i) - \boldsymbol{W}_\phi \boldsymbol{h}_i\|_{\mathcal{H}}^2 \tag{3}$$

ただし，$\boldsymbol{W}_\phi = (\phi(\boldsymbol{w}_1), \ldots, \phi(\boldsymbol{w}_k), \ldots, \phi(\boldsymbol{w}_K))$ とし，$\boldsymbol{w}_k$ は $\boldsymbol{v}_i$ から選択され，K は選択された基底数である．カーネル NMF は，なるべく K を小さくしつつ，C_ϕ の増加を抑えることを達成できれば，同程度の圧縮性能で解決性を高めた基底ベクトルを選択できたと言える．

5. 評価実験

5.1. 実験条件

2 つの実データセットを組み合わせ，外れ値データを意図的に含んだデータに対して提案手法を適用し，得られた基底の特徴と基底数の推移を観察することで，提案手法の有効性を確認する．対象データは，0〜9 の手書き数字に関する画像 [3]（標準データ）各 100 枚，合計枚数 $N = 1,000$ とする．これに外れ値データとしてファッション画像 [4] を νN 枚（$\nu > 0$）加えた，計 $I = N + \nu N$ 枚の画像データ (8bit グレースケール/28×28 次元) を用いる．なお，$\nu \in \{0.01, 0.02, 0.04, 0.08, 0.10\}$ とした．また，評価指標として用いる C_ϕ は，$f(\boldsymbol{x}_\ell) > 0$ となるデータのみから算出する．

5.2. 結果と考察

標準データに対する再構成誤差 C_ϕ と ν を変化させたときの基底数の関係を図 2 に示す．図 2 より，従来手法では ν を変化させても基底数が約 20 から下がらないのに対し，提案はその下限を押し下げ，従来よりも常に少ない基底数とすることができる．つまり，提案手法では少ない基底で解釈を行うことが可能となる．これは分析対象とする標準データのみが基底として選択されたためである．さらに，提案手法の中でも提案手法の C_ϕ が最小となった点は，周囲の他点と比べ，基底数を少なくしながら，再構成誤差をある程度小さくできている．つまり，データを表す基底ベクトルを選択できたと考えられる．

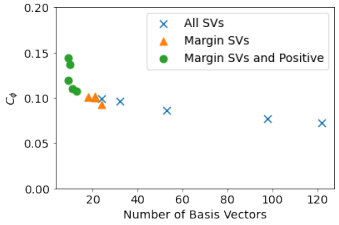

図 2: 再構成誤差

ここで，$\boldsymbol{H}$ の各行は各基底ベクトルへの重みを意味するため，各行で値が大きいデータを読み解くことでデータ全体の解釈が可能であると考えられる．そこで，図 3 に SV-NMF，図 4 に提案手法における $\nu = 0.01$ の場合の各手法により得られた k 番目の基底ベクトル $\boldsymbol{w}_k$ に対して，$\boldsymbol{H}$ の各行の重みの大きい上位 10 件のデータの一部を示す．

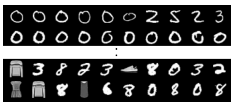

図 3: SV-NMF

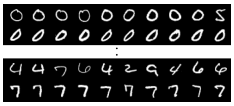

図 4: 提案手法

図 3 より，SV-NMF にはファッション画像（外れ値データ）が含まれていることがわかる．これに対し，図 4 より，提案手法によって得られる解釈に標準データである数字画像のみが抽出されている．また，図 4 は行ごとに類似した画像抽出がなされている．つまり，提案手法は外れ値データを除外し，データ全体の重要な特徴をよく表す標準データのみを基底として獲得することに成功している．

以上より，提案手法ではより正確にデータの特徴を捉えた少数の基底が獲得でき，従来手法に比べて解釈性が向上することが示された．

6. まとめと今後の課題

本研究では，SV-NMF の問題を解決し，解釈性を向上させる基底の選択手法を提案した．また外れ値データを意図的に含んだ実データセットに適用することで，提案手法の有効性を確認した．今後の課題としては，他の実データへの適用などが挙げられる．

参考文献

[1] Bradley Klingenberg, James Curry, and Anne Dougherty. Non-negative matrix factorization: Ill-posedness and a geometric algorithm. *Pattern Recognition*, Vol. 42, No. 5, pp. 918–928, 2009.

[2] Slim Essid. A single-class svm based algorithm for computing an identifiable nmf. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2053–2056. IEEE, 2012.

[3] Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Processing Magazine*, Vol. 29, No. 6, pp. 141–142, 2012.

[4] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.