Adaptive Projected Matrix Factorization Method for Data Clustering

Mulin Chen[a], Qi Wang[a,b,*], Xuelong Li[c,d]

[a]*School of Computer Science and Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an 710072, Shaanxi, P. R. China*
[b]*Unmanned System Research Institute (USRI), Northwestern Polytechnical University, Xi'an 710072, Shaanxi, P. R. China*
[c]*Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an, 710119, Shaanxi, P. R. China*
[d]*University of Chinese Academy of Sciences, Beijing, 100049, P. R. China*

Abstract

Data clustering aims to group the data samples into clusters, and has attracted many researchers in a variety of multidisciplinary fields, such as machine learning and data mining. In order to capture the geometry structure, many methods perform clustering according to a predefined affinity graph. So the clustering performance is largely determined by the graph quality. Unfortunately, the graph quality cannot be guaranteed in various real-world applications. In this paper, an Adaptive Projected Matrix Factorization (APMF) method is proposed for data clustering. Our contributions are threefold: (1) instead of keeping the graph fixed, graph learning is taken as a part of the clustering procedure; (2) the clustering is performed in the projected subspace, so the noise in the input data space is alleviated; (3) an efficient and effective algorithm is developed to solve the proposed problem, and its convergence is proved. Extend experiments on nine real-world benchmarks validate the effectiveness of the proposed method, and verify its superiority against the state-of-the-art competitors.

Keywords: Clustering, Graph Learning, Subspace Learning, Matrix Factorization

*Corresponding author
Email address: crabwq@gmail.com (Qi Wang)

1. Introduction

Clustering is a fundamentally important task in the field of machine learning, and involves various practical applications, such as image classification [1], saliency detection [2] and signal processing [3]. Given the desired cluster number, the objective of clustering is to partition the data into different clusters, such that the samples within the same cluster have high similarity. In the past decades, a great number of techniques have proposed on clustering, such as k-means clustering [4], hierarchical clustering [5], multiview clustering [6, 7, 8, 9], spectral clustering [10], non-negative matrix factorization [11], support vector clustering [12] and maximum margin clustering [13].

One of the most popular clustering method is non-negative matrix factorization (NMF). NMF approximates the data matrix with the product of two non-negative matrices, which can be considered to be the clustering centroid and indicator matrix respectively. And the clustering result is indicated by the index of the maximal element within each indicator vector. Since the non-negative constraints allow only additive operation (no subtractive and combination), the factorization leads to a part-based representation, which implies the perception mechanism of human brain [14]. So NMF is suitable to learn the parts of objects, and has shown good performance in image annotation [15], face recognition [16], crowd analysis [17, 18, 19] and document clustering [20].

Though NMF achieved prominent performance, it suffers from some problems. The main drawback of NMF is that it neglects the local correlation of data points, so it fails to perceive the data manifold. To tackle this problem, some graph-based NMF methods [21, 22, 16, 23, 14, 24, 25] have been proposed. By assuming that the points with close distances should be grouped into the same cluster, these approaches first construct a weighted affinity graph based on the data points' distances, and then perform matrix factorization according to the spectrum of the graph. So their performance highly depends on the quality of the affinity graph. However, in practical applications, the graph quality cannot be guaranteed because: (1) the points' distances in the input data space

may be influenced by the noise; (2) the construction approach (0-1 weighting, Gaussian weighting) may be inappropriate. So the clustering performance tends to be adversely affected by the graph quality, and it's necessary to optimize the affinity graph in the clustering procedure.

Another major limitation of the existing NMF-based methods is that the factorization is accomplished in the input data space. According to Li et al. [26], data with high dimensionality is always lying within a low dimensional manifold. And in the desired low dimensional subspace, the noise within the original data is alleviated and more discriminative features are preserved. However, most existing methods process the input data directly. So they cannot capture the intrinsic data geometry accurately, and are sensitive to the data noise. In recent years, the success of subspace clustering methods [27, 28, 29, 30, 31] has proved that the exploration of subspace improves the robustness. However, most matrix factorization methods neglect this aspect.

To mitigate the above issues, we propose an Adaptive Projected Matrix Factorization (APMF) method in this paper, which inherits the merits of both spectral clustering and NMF. The data points are projected into a low dimensional subspace, so the noise in the input space is alleviated and more valuable features are captured. And the data graph is updated during the clustering procedure, so the graph quality is improved. In addition, the proposed method learns the indicator matrix directly without post-processing, so it achieves stable performance.

The main contributions of this paper are summarized as follows.

(1) The graph learning is jointly incorporated in the NMF framework. By optimizing the data graph adaptively during the matrix factorization procedure, the geometric structure is explored and the graph quality is improved.

(2) The data is projected to a low dimensional subspace, which preserves more discriminative features and contains less noise. Both the graph learning and clustering are accomplished in the projected subspace, so the data

noise in the input space is avoided.

(3) The proposed problem is solved with an efficient alternative algorithm, and no post-processing is involved. Experimental results demonstrate that the proposed method is effective and outperforms the representative competitors.

The rest of this paper is organized as follows. Section 2 reviews the related works on non-negative matrix factorization. Section 3 proposes the Adaptive Projective Matrix Factorization method, and introduces an efficient optimization algorithm. Experimental results are presented and theoretically discussed in Section 4. Finally Section 5 summarizes the conclusion.

2. Related Work

In this section, we briefly review the original non-negative matrix factorization (NMF) and its graph-based variants, and discuss their advantages and limitations.

2.1. Non-negative matrix factorization

Given a data matrix $X = [x_1, x_2, ..x_n] \in \mathbb{R}^{d\times n}$ (d is the dimensionality and n is the number of samples), NMF aims to approximate X with two non-negative matrices :

$$X \approx FG^T, \tag{1}$$

where $F \in \mathbb{R}^{d\times k}$ and $G \in \mathbb{R}^{n\times k}$ are the clustering centroid and cluster indicator respectively, and k is the desired cluster number. So the objective function of NMF is defined as a closed form problem:

$$\min_{F\geq 0, G\geq 0} ||X - FG^T||_F^2, \tag{2}$$

where $||\cdot||_F$ is the Frobenius norm. Benefited from the non-negative constraints, NMF obtains a part-based representation. This property provides a physiological and psychological interpretation for non-negative data [21], such as face

images and documents. However, from Eq. (2) we can see that NMF just focus on the global data geometry, and ignore the local aspect. So some graph-based variants are proposed.

2.2. Graph-based variants of NMF

The investigation of data relationship is essential to handle the data with complicated structures. To this end, Cai et al. [21] proposed to impose the graph regularization term on NMF by solving the following problem:

$$\min_{F\geq 0, G\geq 0} ||X - FG^T||_F^2 + \lambda tr(G^T LG), \tag{3}$$

where $tr()$ is the trace operator, $L = D - \frac{S^T+S}{2}$ is the Laplacian matrix, S is the affinity graph, and D is a diagonal matrix with $D_{ii} = \sum_j \frac{S_{ij}+S_{ji}}{2}$. The second term enforces that the samples with small distance should have similar indicator vectors. Thus, the local relationship between the points is exploited such that the clustering performance can be improved.

The residue error of each data point is squared in the objective function (3), so problem (3) is sensitive to outliers. By introducing the $\ell_{2,1}$ norm, Nie et al. [22] proposed a more robust version:

$$\min_{G\geq 0, G^T G=I} ||X - FG^T||_{2,1} + \lambda tr(G^T LG), \tag{4}$$

where I is the identity matrix. The $\ell_{2,1}$ norm makes the objective function robust to outliers, and the constraint $G^T G = I$ ensures the optimal solution to be unique. However, the cluster result may be inaccurate if the affinity graph is constructed with low quality.

Instead of defining a simple graph, Zeng et al. [32] utilized the hyper graph L_{hyper}, where each edge connects more than two data points. And the objective function is denoted as

$$\min_{F\geq 0, G\geq 0} ||X - FG^T||_F^2 + \lambda tr(G^T L_{hyper} G), \tag{5}$$

There are also some other methods [16, 23, 14, 24, 25] that introduced manifold regularization terms. However, the predefined graph may be with low quality,

and affect the clustering performance adversely. Recently, Zhang et al. [33] proposed to update the graph in the clustering phase, and validated the rationality of performing graph learning.

The shortcoming shared by all the above methods is that the clustering is performed in the input data space, where the valuable features may be corrupted by noise. So both the graph construction and matrix factorization tend to be affected. Li et al. [26] proved that the influence of data noise can be reduced in the subspace. So it's necessary to combine the subspace learning strategy into matrix factorization methods.

3. Methodology

In this section, we introduce the proposed Adaptive Projected Matrix Factorization (APMF), and design an efficient optimization algorithm. To capture the local data structure while improving the robustness, the proposed method jointly incorporates graph learning and subspace clustering into the matrix factorization framework.

3.1. Adaptive Projected Matrix Factorization Method

In this part, the Adptive Projected Matrix Factorization (APMF) method is presented. To find the optimal graph for clustering, we propose to update the graph when performing matrix factorization. So we have the following objective function:

$$\begin{aligned} \min_{F,G,S} & ||X - FG^T||_{2,1} + \lambda tr(G^{\mathrm{T}} LG) \\ & +\alpha \sum_{i,j} S_{ij}^2 ||(x_i - x_j)||_2^2, \\ s.t. & \sum_j S_{ij} = 1, S_{ij} \geq 0, G^T G = I, G \geq 0, \end{aligned} \tag{6}$$

where α is a parameter. The third term enforces that the S_{ij} should be large if x_i and x_j are close to each other. So the affinity graph S can be updated according to the points' distances, and the local relationship within the data manifold can be captured. However, the clustering is still accomplished in the input data space. According to Chen et al. [26], more discriminative features

can be captured if the points are transformed into the desired subspace. So we project the data into subspace with a transformation matrix $W \in \mathbb{R}^{d\times m}$ ($m \ll d$). And the objective function is rewritten as

$$\begin{aligned}\min_{F,G,S,W} &||W^T X - FG^T||_{2,1} + \lambda tr(G^{\mathrm{T}} LG) \\ &+\alpha \sum_{i,j} S_{ij}^2 ||W^T(x_i - x_j)||_2^2, \\ s.t. &\sum_j S_{ij} = 1, S_{ij} \geq 0, G^T G = I, G \geq 0, W^T W = I,\end{aligned} \tag{7}$$

where the orthogonal constraint on W guarantees the uniqueness of W. In the optimization, we propose an alternative method to update each variable. Thus, both the matrix factorization and the graph learning are achieved in the learned subspace, and the merit of subspace clustering is inherited. Finally, we can obtain the cluster indicator G in the desired subspace, and the noise within the input space is avoided.

3.2. Optimization algorithm

Problem (7) is not convex, so we propose to solve it with the Augmented Lagrangian Multiplier (ALM) [34]. Because both the $\ell_{2,1}$ norm and Laplacian regularization depend on G, we introduce two auxiliary variables $E = W^T X - FG^T$ and $Z = G$. So problem (7) is rewritten as

$$\begin{aligned}\min_{E,Z,F,G,S,W,\Lambda_1,\Lambda_2,\mu} &||E||_{2,1} + \lambda tr(Z^{\mathrm{T}} LG) \\ &+\alpha \sum_{i,j} S_{ij}^2 ||W^T(x_i - x_j)||_2^2 \\ s.t. &E = W^T X - FG^T, Z = G, G^T G = I, \\ &\textstyle\sum_j S_{ij} = 1, S_{ij} \geq 0, W^T W = I,\end{aligned} \tag{8}$$

which is equivalent to the following ALM problem

$$\begin{aligned}\min_{E,Z,F,G,S,W,\Lambda_1,\Lambda_2,\mu} &||E||_{2,1} + \lambda tr(Z^{\mathrm{T}} LG) \\ &+\alpha \sum_{i,j} S_{ij}^2 ||W^T(x_i - x_j)||_2^2 \\ &+\tfrac{\mu}{2} ||E - W^T X + FG^T + \tfrac{1}{\mu}\Lambda_1||_F^2 \\ &+\tfrac{\mu}{2} ||Z - G + \tfrac{1}{\mu}\Lambda_2||_F^2, \\ s.t. &Z \geq 0, G^T G = I, \textstyle\sum_j S_{ij} = 1, S_{ij} \geq 0, W^T W = I,\end{aligned} \tag{9}$$

where $\Lambda_1 \in \mathbb{R}^{m\times n}$ and $\Lambda_2 \in \mathbb{R}^{k\times d}$ are ALM multipliers, and μ is a scalar parameter. Since problem (9) involves several variables to be optimized, we propose to solve it with an alternative algorithm. Firstly, F and G are initialized with the method in [22], and W and S are initialized with the strategy in [35]. Then we update one variable while keeping the others fixed.

Update E: to update E, we fix all the other variables, and then problem (9) becomes

$$\min_{E} ||E||_{2,1} + \frac{\mu}{2}||E - M||_F^2, \tag{10}$$

where $M = W^T X - FG^T - \frac{1}{\mu}\Lambda_1$. As proved in [36], the optimal solution of problem (10) is

$$e_i = \begin{cases} (1 - \frac{1}{\mu||m_i||_2})m_i, \text{if } ||m_i||_2 \geq \frac{1}{\mu} \\ 0, \qquad\qquad \text{else} \end{cases}, \tag{11}$$

where e_i and m_i are the i-th column of E and M respectively.

Update Z: when solving Z, problem (9) is reduced to

$$\min_{Z\geq 0} \lambda tr(Z^{\mathrm{T}} LG) + \frac{\mu}{2}||Z - G + \frac{1}{\mu}\Lambda_2||_F^2. \tag{12}$$

Spanning the Frobenius norm and removing the irrelevant terms, we have

$$\min_{Z\geq 0} ||Z - P||_F^2, \tag{13}$$

where $P = G - \frac{1}{\mu}\Lambda_2 - \frac{\lambda}{\mu}LG$. It's obvious that the optimal Z can be updated as

$$Z_{ij} = max(P_{ij}, 0). \tag{14}$$

Update F: optimizing F yields the following problem

$$\min_{F} ||F - HG||_F^2, \tag{15}$$

where $H = W^T X - E - \frac{1}{\mu}\Lambda_1$. Since there is no constraint on F, we can compute the optimal solution as $F = HG$.

Update G: for updating G, problem (9) is transformed to the following form:

$$\begin{aligned} \min_{G^T G=I} \lambda tr(Z^{\mathrm{T}} LG) + \frac{\mu}{2}||E - W^T X + FG^T + \frac{1}{\mu}\Lambda_1||_F^2 \\ + \frac{\mu}{2}||Z - G + \frac{1}{\mu}\Lambda_2||_F^2, \end{aligned} \tag{16}$$

which can be further simplified to

$$\min_{G^TG=I} ||G-K||_F^2, \tag{17}$$

where $K = (W^TX - E - \frac{1}{\mu}\Lambda_1)^TF + Z - \frac{\lambda}{\mu}LZ + \frac{1}{\mu}\Lambda_2$. According to [22], the optimal solution of the above problem is

$$G = UV^T, \tag{18}$$

where $U \in \mathbb{R}^{k\times d}$ and $V \in \mathbb{R}^{d\times d}$ are the left and right singular vectors of the compact SVD decomposition of G.

Update W: when updating W, we have

$$\min_{W^TW=I} \frac{\mu}{2}||W^TX - Q||_F^2 + \alpha\sum_{i,j} s_{ij}^2||W^T(x_i - x_j)||_2^2, \tag{19}$$

where $Q = FG^T + E + \frac{1}{\mu}\Lambda_1$. Denoting a scatter matrix $\widetilde{S} \in \mathbb{R}^{d\times d}$ as $\widetilde{S} = \sum_{i,j} S_{ij}^2(x_i - x_j)(x_i - x_j)^T$, then problem (20) becomes

$$\min_{W^TW=I} tr(W^T\widetilde{S}W) - \frac{\mu}{\alpha}tr(W^TXQ), \tag{20}$$

which can be solved with the Generalized Power Iteration method [37].

Update S: in order to optimize S, we transform the objective into the following problem:

$$\min_{S_{ij}=1, S_{ij}\geq 0} \frac{\lambda}{\alpha}\sum_{i,j} S_{ij}||z_i - g_j||_2^2 + \sum_{i,j} S_{ij}^2||W^T(x_i - x_j)||_2^2, \tag{21}$$

where z_i is the i-th column of Z, and g_j is the j-th column of G. Problem (23) is independent for different i, so we optimize the following problem separately for each i:

$$\min_{S_{ij}=1, S_{ij}\geq 0} \frac{\lambda}{\alpha}\sum_{j} S_{ij}||z_i - g_j||_2^2 + \sum_{j} S_{ij}^2||W^T(x_i - x_j)||_2^2. \tag{22}$$

Denoting a diagonal matrix $V \in \mathbb{R}^{n\times n}$ with $V_{jj} = ||W^T(x_i - x_j)||_2^2$ and defining a column vector $u_i \in \mathbb{R}^{n\times 1}$ with its j-th element as $\frac{\lambda}{2\alpha}||z_i - g_i||_2^2$, the above problem can be rewritten as

$$\min_{s_i^T\mathbf{1}=1, s_i\geq 0} \frac{1}{2}s_i^TVs_i + s_i^Tu_i, \tag{23}$$

where $s_i \in \mathbb{R}^{n\times 1}$ is a column vector with its j-th element as S_{ij}, and $\mathbf{1} \in \mathbb{R}^{n\times 1}$ is a column vector with all its elements as 1. Problem (23) can be readily solved with an efficient method [38].

Update Λ_1, Λ_2 and μ: according to [34], the ALM parameters Λ_1, Λ_2 and μ are updated as

$$\begin{aligned} \Lambda_1 &= \Lambda_1 + \mu(W^T X - FG^T - E), \\ \Lambda_2 &= \Lambda_2 + \mu(Z - G), \\ \mu &= \rho\mu, \end{aligned} \tag{24}$$

where the parameter ρ is set as 1.5 in the experiments.

The details of the optimization algorithm are exhibited in Algorithm 1.

Algorithm 1 Optimization algorithm

Input: Data matrix X, cluster number c, projection dimension m, parameters λ, α.

1: Initialize data graph S, matrices F and G, and transformation matrix W.
2: **repeat**
3: Update E with Eq. (11).
4: Update Z with Eq. (14)
5: Update F by solving problem (15).
6: Update G according to Eq. (18).
7: Update W by solving problem (20).
8: Update S by solving problem (23).
9: Update Λ_1, Λ_2 and μ according to Eq. (24).
10: **until** Converge

Output: The learned graph S.

4. Experiments

In this section, experiments are conducted on real-world datasets to demonstrate the effectiveness of the proposed APMF. The clustering accuracy (ACC),

Table 1: Description of datasets.

Datasets	Samples	Dimensionality	Classes
BA	1400	320	36
Jaffe	213	676	10
Yale	165	256	15
Dermatology	366	34	6
Ecoli	336	7	8
Ionosphere	351	34	2
Movement	360	90	15
Semeion	1593	256	10

Table 2: ACC of different methods on eight real-world datasets. The best results are in bold face.

Datasets	k-means	RCut	NCut	NMF	GNMF	RMNMF	AMF	APMF
BA	0.41	0.15	0.42	0.17	0.20	0.40	0.44	**0.46**
Jaffe	0.75	0.77	0.80	0.67	0.47	0.80	0.77	**0.92**
Yale	0.43	0.24	0.51	0.35	0.32	0.38	0.48	**0.57**
Dermatology	0.71	0.65	0.92	0.73	0.61	0.72	0.95	**0.96**
Ecoli	0.64	0.63	0.52	0.35	0.44	0.70	0.67	**0.76**
Ionoshpere	0.70	0.52	0.69	0.63	0.64	0.72	0.71	**0.83**
Movement	0.44	0.40	0.46	0.37	0.26	0.39	**0.49**	**0.49**
Semeion	0.57	0.25	0.52	0.37	0.36	0.44	0.67	**0.68**

normalized mutual information (NMI) [22] and Purity [19] are taken as evaluation measurements.

Datasets: Eight real world benchmark datasets are employed, including one handwritten digit dataset, i.e., Binary Alphabet (BA) [39], two face image dataset, i.e., Jaffe [40] and Yale [41], and five datasets from UCI Machine Learning Repository [42], i.e., Dermatology, Ecoli, Ionosphere, Movement and Semeion. The details of the datasets are exhibited in Table 1.

Competitors: In the experiments, the proposed APMF is compared with six state-of-the-arts, including k-means, Ratio Cut (RCut) [43] Normalized Cut (NCut) [44], NMF [11], Graph-regularized NMF (GNMF) [21] and Robust Manifold NMF (RMNMF) [22]. For k-means, RCut, NCut and RMNMF, the self-

Table 3: NMI of different methods on eight real-world datasets. The best results are in bold face.

Datasets	k-means	RCut	NCut	NMF	GNMF	RMNMF	AMF	APMF
BA	0.57	0.24	0.57	0.29	0.35	0.52	0.58	**0.59**
Jaffe	0.82	0.81	0.85	0.69	0.52	0.77	0.78	**0.90**
Yale	0.52	0.30	0.56	0.42	0.38	0.47	0.55	**0.61**
Dermatology	0.81	0.68	0.88	0.64	0.50	0.85	0.90	**0.91**
Ecoli	0.58	0.56	0.48	0.21	0.57	0.51	0.58	**0.62**
Ionoshpere	0.12	0.12	0.20	0.15	0.16	0.13	0.23	**0.29**
Movement	0.56	0.51	0.60	0.40	0.27	0.49	0.59	**0.62**
Semeion	0.52	0.22	0.48	0.32	0.34	0.34	0.58	**0.59**

Table 4: Purity of different methods on eight real-world datasets. The best results are in bold face.

Datasets	k-means	RCut	NCut	NMF	GNMF	RMNMF	AMF	APMF
BA	0.44	0.18	0.45	0.17	0.37	0.46	0.46	**0.48**
Jaffe	0.77	0.80	0.84	0.69	0.48	0.80	0.78	**0.92**
Yale	0.45	0.27	0.52	0.35	0.28	0.39	0.50	**0.58**
Dermatology	0.8	0.72	0.93	0.75	0.61	0.73	0.94	**0.96**
Ecoli	0.81	0.84	0.79	0.51	0.54	0.80	0.83	**0.86**
Ionoshpere	0.70	0.64	0.72	0.64	0.64	0.72	0.71	**0.83**
Movement	0.47	0.43	0.49	0.38	0.26	0.42	**0.50**	**0.50**
Semeion	0.60	0.31	0.56	0.39	0.40	0.46	0.67	**0.69**

tune Gaussian method [45] is used to construct the graph. For GNMF, the 0-1 weighting graph is used since it's more suitable [21]. Because k-means, RCut and NCut are sensitive to initialization, we perform them for 200 repetitions and report the average performance. In order to evaluate whether the data projection is effective, we further denote the method with objective (6) as Adaptive Matrix Factorization (AMF), and compare it with the proposed APMF. For a fair comparison, all the competitors utilize their optimal parameters. And for our APMF, the parameters λ and α are all set as 0.01, and the projection dimensionality m is determined empirically according to the data dimensionality d ($m \ll d$).

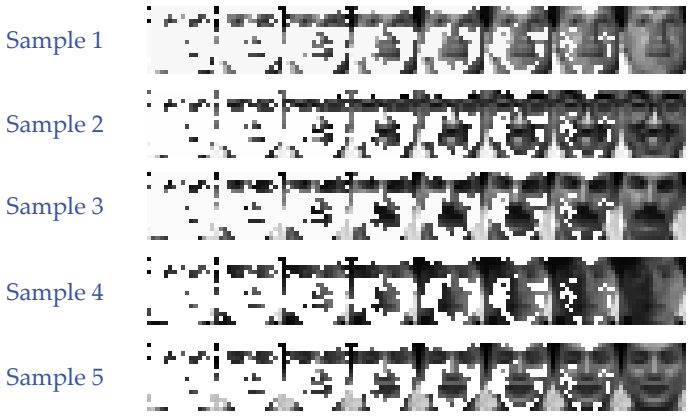
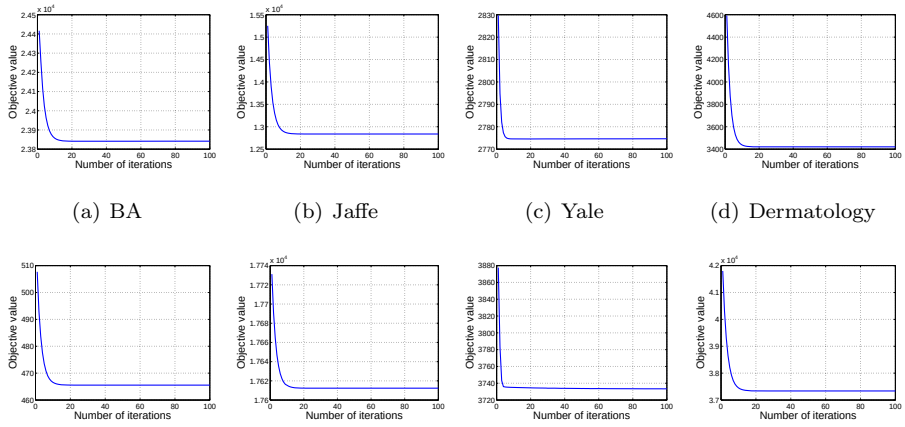


Figure 1: Features captured by the transformation matrix W. From to right, the selected feature number is $\{32, 64, 96, 128, 160, 192, 224, 256\}$. Most discriminative features, such as eyes, nose and mouth, are preserved.

Performance: Table 2, 3 and 4 exhibits the clustering results of different methods. It can be seen that APMF has the highest ACC, NMI and Purity on all datasets, which implies the best performance. In particular, APMF outperforms the competitors a lot on Jaffe and Ionosphere. k-means and NMF fails in most cases, since they can not perceive the local data structure. RCut, NCut, GNMF and RMNMF show better performance because they exploit the manifold structure. However, they are limited to learn the optimal affinity graph for clustering, and easily affected by the noise in the input data space. Moveover, k-means, RCut and NCut are sensitive to the initialization. AMF shows good performance on some datasets because it learns the data graph adaptively. But it fails to find the low dimensional subspace, so it's easily affected by data noise. The comparison between AMF and APMF demonstrates that the learning of subspace does improve the clustering performance. The proposed APMF takes graph learning as a part of the clustering stage, and learns the local data structure in the desired subspace. So it achieves the best results.

In addition, we evaluate whether the projected subspace preserves discriminant features. We randomly select five samples from the Yale dataset. With the



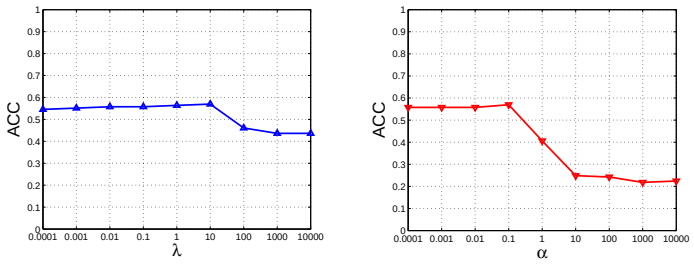
(a) BA (b) Jaffe (c) Yale (d) Dermatology

(e) Ecoli (f) Ionosphere (g) Movement (h) Semeion

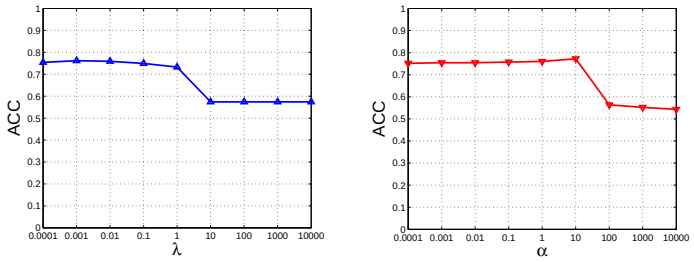
Figure 2: Convergence curves on different datasets.

learned optimal transformation matrix W, the larger value of $||W_{i,:}||_2$ indicates that the i-th feature is more valuable [35]. So we select the most important $\{32, 64, 96, ..., 256\}$ features by sorting $||W_{i,:}||_2$, and show the selected features in Figure 1. Each pixel is a feature, the selected ones maintain their original pixel value, and the unselected ones are shown in white. From left to right, the feature number is $\{32, 64, 96, ..., 256\}$ respectively. And we can see in Figure 1 that the transformation matrix W captures the most discriminative features, such as eyes, mouth and nose. And the skin pixels, which can be considered as the background noise, are dropped in most cases. So the data projection operation helps to capture the valuable features while reducing the influence of noise.

Convergence: We further discuss the convergence behavior of APMF. During the optimization of each parameter, a global or local optimal solution is obtained. So the ALM objective value in problem (9) decreases monotonously in each iteration, and finally converges to a local optima. In addition, as μ grows exponentially with the number of iterations, the third and fourth terms in the ALM objective function will be close to 0. So problem (9) converges to the



(a) Yale



(b) Ecoli

Figure 3: The influence of λ and α on (a) Yale and (b) Ecoli datasets.

original objective function in problem (7) asymptotically. Therefore, the convergence of problem (7) is guaranteed. Fig. 2 visualizes the convergence curves of the original objective value on different datasets. As shown in the figure, the optimization algorithm converges within ten iterations, which verifies the effectiveness of the optimization.

Parameter sensitivity: There are two regularization parameters, λ and α, in the proposed method. λ controls the weight of the manifold learning term, and α balances the graph learning term. We investigate their impacts on the final clustering performance. Figure 3 shows the ACC curves of APMF with varying λ and α respectively. It can be seen that the performance of APMF is very stable when λ and α are within the range of $\{0.0001, 0.001, ..., 0.1\}$. When the values of λ and α are larger than 1, the approximate error in problem (7) becomes large, so the performance decreases. Thus, we set λ and α as 0.01 throughout the experiments.

5. Conclusions

In this paper, a new clustering method called Adaptive Projected Matrix Factorization (APMF) has been put forward. The proposed method jointly combines the graph learning method into the NMF framework, so it is able to learn the optimal affinity graph for clustering. In addition, the proposed method takes advantage of subspace clustering by projecting the input data into the desired subspace, so the noise in the input data space is alleviated and more discriminative features are captured. Moreover, an efficient algorithm is designed to optimize the proposed problem with proved convergence. Extensive experiments on several datasets validates the superior performance of APMF.

Acknowledge

This work was supported by the National Key R&D Program of China under Grant 2017YFB1002202, National Natural Science Foundation of China under Grant 61773316, Fundamental Research Funds for the Central Universities under Grant 3102017AX010, and the Open Research Fund of Key Laboratory of Spectral Imaging Technology, Chinese Academy of Sciences.

References

[1] Q. Wang, J. Lin, Y. Yuan, Salient band selection for hyperspectral image classification via manifold ranking, IEEE Transactions on Neural Network and Learning System 27 (6) (2016) 1279–1289.

[2] X. Yao, J. Han, D. Zhang, F. Nie, Revisiting co-saliency detection: a novel approach based on two-stage multi-view spectral rotation co-clustering, IEEE Transactions on Image Processing 26 (7) (2017) 3196–3209.

[3] W. Liu, N. Zheng, X. Li, Nonnegative matrix factorization for EEG signal classification, in: International Symposium on Neural Networks, 2004, pp. 470–475.

[4] J. Macqueen, Some methods for classification and analysis of multivariate observations, in: Berkeley Symposium on Mathematical Statistics and Probability, 1967, pp. 281–297.

[5] J. Rohlf, Adaptive hierarchical clustering schemes, Systematic Zoology 19 (1) (1970) 58.

[6] X. Li, M. Chen, F. Nie, Q. Wang, A multiview-based parameter free framework for group detection, in: AAAI Conference on Artificial Intelligence, 2017, pp. 4147–4153.

[7] C. Deng, Z. Lv, W. Liu, J. Huang, D. Tao, X. Gao, Multi-view matrix decomposition: a new scheme for exploring discriminative information, in: International Joint Conference on Artificial Intelligence, 2015, pp. 3438–3444.

[8] C. Deng, X. Liu, C. Li, D. Tao, Active multi-kernel domain adaptation for hyperspectral image classification, Pattern Recognition 77 (2018) 306–315.

[9] X. Yao, J. Han, D. Zhang, F. Nie, Revisiting co-saliency detection: A novel approach based on two-stage multi-view spectral rotation co-clustering, IEEE Transactions on Image Processing 26 (7) (2017) 3196–3209.

[10] A. Ng, M. Jordan, Y. Weiss, On spectral clustering: Analysis and an algorithm, in: Advances in Neural Information Processing System, 2001, pp. 849–856.

[11] D. Li, H. Seung, Algorithms for nonnegative matrix factorization, Advances in Neural Information Processing Systems 13 (6) (2000) 556–562.

[12] A. Ben-Hur, D. Horn, H. Siegelmann, V. Vapnik, Support vector clustering, Journal of Machine Learning Research 2 (2001) 125–137.

[13] L. Xu, J. Neufeld, B. Larson, D. Schuurmans, Maximum margin clustering, in: Advances in Neural Information Processing Systems, 2004, pp. 1537–1544.

[14] X. Li, G. Cui, Y. Dong, Graph regularized non-negative low-rank matrix factorization for image clustering, IEEE Transactions on Cybernetics 47 (11) (2017) 3840–3853.

[15] D. Tao, D. Tao, X. Li, X. Gao, Large sparse cone non-negative matrix factorization for image annotation, ACM Transactions on Intelligent Systems and Technology 8 (3) (2017) 37:1–37:21.

[16] S. Z. Li, X. Hou, H. Zhang, Q. Cheng, Learning spatially localized, parts-based representation, in: Computer Vision and Pattern Recognition, 2001, pp. 207–212.

[17] M. Chen, Q. Wang, X. Li, Anchor-based group detection in crowd scenes, in: IEEE International Conference on Acoustics, Speech and Signal Processing, 2017, pp. 1378–1382.

[18] M. Chen, Q. Wang, X. Li, Patch-based topic model for group detection, SCIENCE CHINA Information Sciences 60 (11) (2017) 113101:1–113101:7.

[19] Q. Wang, M. Chen, X. Li, Quantifying and detecting collective motion by manifold learning, in: AAAI Conference on Artificial Intelligence, 2017, pp. 4292–4298.

[20] Z. Xiong, Y. Zang, X. Jiang, X. Hu, Document clustering with an augmented nonnegative matrix factorization model, in: Advances in Knowledge Discovery and Data Mining, 2014, pp. 348–359.

[21] D. Cai, X. He, J. Han, T. Huang, Graph regularized nonnegative matrix factorization for data representation, IEEE Transactions on Pattern Analysis and Machine Intelligence 33 (8) (2011) 1548–1560.

[22] J. Huang, F. Nie, H. Huang, C. Ding, Robust manifold nonnegative matrix factorization, ACM Transactions on Knowledge Discovery from Data 8 (3) (2013) 11:1–11:21.

[23] D. Tao, J. Cheng, M. Song, X. Lin, Manifold ranking-based matrix factorization for saliency detection, IEEE Transactions on Neural Network and Learning System 27 (6) (2016) 1122–1134.

[24] X. Zhang, L. Zhao, L. Zong, X. Liu, H. Yu, Multi-view clustering via multi-manifold regularized nonnegative matrix factorization, in: IEEE International Conference on Data Mining, 2014, pp. 1103–1108.

[25] C. P. et al., Nonnegative matrix factorization with integrated graph and feature learning, ACM Transactions on Intelligent Systems and Technology 8 (3) (2017) 42:1–42:29.

[26] X. Li, M. Chen, F. Nie, Q. Wang, Locality adaptive discriminant analysis, in: International Joint Conference on Artificial Intelligence, 2017, pp. 2201–2207.

[27] J. Chen, H. Zhang, H. Mao, Y. Sang, Z. Yi, Symmetric low-rank representation for subspace clustering, Neurocomputing 173 (2016) 1192–1202.

[28] X. Zhang, C. Xu, X. Sun, G. Baciu, Schatten-q regularizer constrained low rank subspace clustering model, Neurocomputing 182 (2016) 36–47.

[29] J. Wang, D. Shi, D. Cheng, Y. Zhang, J. Gao, LRSR: low-rank-sparse representation for subspace clustering, Neurocomputing 214 (2016) 1026–1037.

[30] W. Chen, E. Zhang, Z. Zhang, A laplacian structured representation model in subspace clustering for enhanced motion segmentation, Neurocomputing 208 (2016) 174–182.

[31] Q. Yin, S. Wu, R. He, L. Wang, Multi-view clustering via pairwise sparse subspace representation, Neurocomputing 156 (2015) 12–21.

[32] K. Zeng, J. Yu, C. Li, J. You, T. Jin, Image clustering by hyper-graph regularized non-negative matrix factorization, Neurocomputing 138 (2014) 209–217.

[33] L. Zhang, Q. Zhang, B. Du, J. You, D. Tao, Adaptive manifold regularized matrix factorization for data clustering, in: International Joint Conference on Artificial Intelligence,, 2017, pp. 3399–3405.

[34] F. Nie, H. Wang, H. Huang, C. Ding, Joint schatten p -norm and ℓ_p -norm robust matrix completion for missing value recovery, Knowledge and Information Systems 42 (3) (2015) 525–544.

[35] F. Nie, W. Zhu, X. Li, Unsupervised feature selection with structured graph optimization, in: AAAI Conference on Artificial Intelligence, 2016, pp. 1302–1308.

[36] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, Y. Ma, Robust recovery of subspace structures by low-rank representation, IEEE Transactions on Pattern Analysis and Machine Intelligence 35 (1) (2013) 171–184.

[37] F. Nie, R. Zhang, X. Li, A generalized power iteration method for solving quadratic problem on the stiefel manifold, Science China Information Sciences 60 (11) (2017) 112101.

[38] F. Nie, X. Wang, M. Jordan, H. Huang, The constrained laplacian rank algorithm for graph-based clustering, in: AAAI Conference on Artificial Intelligence, 2016, pp. 1969–1976.

[39] P. Belhumeur, J. Hespanha, D. Kriegman, Eigenfaces vs. fisherfaces: Recognition using class specific linear projection, IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (7) (1997) 711–720.

[40] M. Lyons, J. Budynek, S. Akamatsu, Automatic classification of single facial images, IEEE Transactions on Pattern Analysis and Machine Intelligence 21 (12) (1999) 1357–1362.

[41] X. He, S. Yan, Y. Hu, P. Niyogi, H. Zhang, Face recognition using laplacianfaces, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (3) (2005) 328–340.

[42] M. Lichman, UCI machine learning repository (2013).

[43] P. Chan, M. Schlag, J. Zien, Spectral k-way ratio-cut partitioning and clustering, IEEE Transactions on CAD of Integrated Circuits and Systems 13 (9) (1994) 1088–1096.

[44] J. Shi, J. Malik, Normalized cuts and image segmentation, IEEE Transactions on Pattern Analysis on Machine Intelligence 22 (8) (2000) 888–905.

[45] L. Manor, P. Perona, Self-tuning spectral clustering, in: Advances in Neural Information Processing Systems, 2004, pp. 1601–1608.