Pattern Recognition 111 (2021) 107676

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Discriminative subspace matrix factorization for multiview data clustering

Jiaqi Ma[a], Yipeng Zhang[a], Lefei Zhang[a,b,*]

[a] School of Computer Science, Wuhan University, Wuhan 430072, China
[b] State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan 430072, China

ARTICLE INFO

Article history:
Received 19 April 2020
Revised 22 July 2020
Accepted 22 September 2020
Available online 28 September 2020

Keywords:
Dimension reduction
Multiview
Clustering
Machine learning

ABSTRACT

In a real-world scenario, an object is easily considered as features combined by multiple views in reality. Thus, multiview features can be encoded into a unified and discriminative framework to achieve satisfactory clustering performance. An increasing number of algorithms have been proposed for multiview data clustering. However, existing multiview methods have several drawbacks. First, most multiview algorithms focus only on origin data in high dimension directly without the intrinsic structure in the relative low-dimensional subspace. Spectral and manifold-based methods ignore pseudo-information that can be extracted from the optimization process. Thus, we design an unsupervised nonnegative matrix factorization (NMF)-based method called discriminative multiview subspace matrix factorization (DMSMF) for clustering. We provide the following contributions. (1) We extend linear discriminant analysis and NMF to a multiview version and connect them to a unified framework to learn in the discriminant subspace. (2) We propose a multiview manifold regularization term and discriminant multiview manifold regularization term that instruct the regularization term to discriminate different classes and obtain the geometry st ructure from the low-dimensional subspace. (3) We design an effective optimization algorithm with proven convergence to obtain an optimal solution procedure for the complex model. Adequate experiments are conducted on multiple benchmark datasets. Finally, we demonstrate that our model is superior to other comparable multiview data clustering algorithms.

1. Introduction

Clustering is a classic task that is widely examined in various domains. Nonnegative matrix factorization (NMF) and spectral clustering (SC) are two widely used techniques for clustering. NMF aims to learn two nonnegative matrices with a product that can approximate the original data matrix [1]. According to [2], the two matrices represent cluster centroid attributes and cluster indicator information. The clustering result can be obtained directly from the cluster indicator matrix without conducting extra post-processing. SC is a classic tool that is applied extensively to nonconvex patterns and linearly inseparable clusters. Spectral-based methods optimize the process of learning an adjacency matrix and obtain final results by performing Eigen decomposition on the Laplacian matrix [3]. However, most methods are designed for single-view features in the data clustering task.

Increasing amounts of unlabeled data are received and collected daily from diverse sources in multiple views. Unsupervised multiview learning is employed to tackle unlabeled data, which are clustered from the perspective of multiview learning [4,5]. Multiview clustering aims to segment points into clusters based on representations of an object from different views. We classify this method into three main categories. The first category incorporates multiview integration into the clustering process by optimizing its objective function, such as the method in Kumar et al. [6]. The second category projects multiview data to a shared subspace in low dimension and then conducts post-processing to obtain the final result, such as the approach in Chaudhuri et al. [7]. The last category is called late integration, and this type of algorithms performs clustering on each individual view and fuses the results into one on consensus [8].

In the past few years, many multiview algorithms based on SC and NMF have been designed to improve multiview clustering quality. For SC algorithms, Wang [9] proposed constrained SC with the first view considered as a similarity matrix and the other view as a constraint. However, this method is restricted to data with only two views. Cai [10] introduced a way to unify modals. This

[*] Corresponding author at: School of Computer Science, Wuhan University, Wuhan 430072, China.
E-mail address: zhanglefei@whu.edu.cn (L. Zhang).

https://doi.org/10.1016/j.patcog.2020.107676
0031-3203/© 2020 Elsevier Ltd. All rights reserved.

method can incorporate diverse data attributes by learning a common Laplacian matrix. Although Kumar raised two spectral-based algorithms [6,11] to ensure the consistence of eigenvectors on all views, he pursued an implicit clustering consistency among different views [12]. For NMF-based algorithms, Tolic [13] designed a nonlinear orthogonal NMF approach for subspace clustering. Lu [14] attempted to establish a connection between linear discriminant analysis (LDA) and NMF in a supervised or semi-supervised manner, but this approach cannot be applied to clustering. Ma [15] established an unsupervised framework, but the technique is still problematic.

Numerous other multiview clustering algorithms have been proposed and demonstrated excellent performance on multiple benchmark datasets [16]. Wang [17] introduced MultiCC to discover multiple independent ways of organizing a dataset into clusters. Nie [18] suggested a new multiview clustering method that is completely self-weighted. Zhang [19] explored underlying complementary information from multiple views. Nie [20] introduced an auto-weighted way for fast matrix factorization, while Huang [21] tried it from the perspective of deep matrix decomposition. By conducting graph structure fusion on different views, Zhan [22] illustrated a global graph with exactly n_c connected components that reflect cluster indicators, thereby making post-processing unnecessary. Zhu [23] suggested a one-step multiview clustering method to solve the previous two-step problem. Hu [24] designed a dynamic way to assign auto weights to different views. Moreover, Huang [25] proposed an auto-weighted multiview clustering method via kernelized graph learning. For an incomplete multiview clustering task, Liu [26] proposed a late fusion approach to simultaneously clustering and imputing the incomplete base clustering matrices via kernel learning. Furthermore, Huang [27] tried a multiview method from intact space to address the view insufficiency issue associated with multiview clustering.

To summarize, the advantages of existing multiview algorithms are as follows. First, such methods consider and balance relationships among views. Second, these methods address the connection among several traditional techniques, such as K-means and SC. Third, constraints on objective functions are well designed to speed up convergence. However, disadvantages remain. Intrinsic data information in the low-dimensional subspace is lacking. Moreover, information obtained from the optimization process cannot be used in most SC and manifold regularization methods. Finally, unavoidable outliers result in residue errors owing to the l_2-norm.

The manifold regularization term was combined with clustering models to improve performance [28–30]. Cai [31] captured local manifold geometry and proposed a graph model. Meanwhile, Zhang [32] adopted adaptive manifold regularization for matrix factorization to learn a satisfactory affinity matrix. A joint framework for integrating sparsity NMF, representation, and adaptive weight was proposed by Zhang et al. [33]. Wang [34] learned a relatively low-dimensional discriminative mapping through a Grassmann manifold. Allab [35] proposed a multi-manifold matrix decomposition method for data co-clustering. Zhang [36] applied manifold regularization terms to the low-dimensional subspace and cluster indicators and learned considerable local geometrical information from raw data. These models show that the manifold regularization term improves the clustering performance and can be expanded to other frameworks.

However, the existing methods have several drawbacks. First, the multiview framework based on NMF factorizes only matrices in high-dimensional space and ignores intrinsic data information in the low-dimensional subspace. Thus, ordinary NMF requires a large number of constraints to capture complex structures. Second, local relationships among the data points obtained from the optimization process, such as pseudo-information, cannot be used in most SC and manifold regularization methods. The geometric structure of data distribution lacks an effective capturing technique. Third, squaring matrix factorization enlarges residue errors caused by the l_2-norm framework. Thus, an improved norm should be chosen to avoid the effect of outliers.

In this study, an unsupervised multiview NMF-based framework called discriminative multiview subspace matrix factorization (DMSMF) is proposed for clustering. The structural block diagram of the proposed DMSMF method is shown in Fig. 1. Briefly, the original dataset can be viewed as several matrices according to the different views in Fig. 1. We consider each view as a concatenated matrix and several view-specific matrices. For the concatenated matrix, we extend the proposed unified framework to a multiview version that combines NMF and LDA. For the view-specific matrices, we design two regularization terms by manifold learning. These constraints instruct themselves to learn from their intrinsic structure and avoid outliers. Finally, by incorporating them together, we obtain our proposed DMSMF method. The primary work and contributions of this study are summarized below.

1. A unified framework for clustering in the discriminant subspace that combines multiview LDA with multiview NMF is proposed to find and utilize the intrinsic low-dimensional structure in the projection subspace.
2. A pseudo-supervised multiview manifold regularization term is proposed to discover the subspace that distinguishes different classes. This regularization term utilizes pseudo-information to instruct itself and refine clustering results.
3. An augmented Lagrangian multiplier (ALM)-based optimization algorithm based on the DMSMF model is proposed to effectively seek an optimal solution.

The remainder of the paper is described as follows. The derivation of our model is shown in detail in Section 2, and the optimization process and corresponding algorithm are presented in Section 3. Several theoretical connections between our framework and other classic methods are discussed in Section 4, and adequate experiments on multiple benchmark datasets and further research are presented in Section 5. Finally, Section 6 elaborates on the conclusions.

2. Problem formulation

Single-view data are important in clustering tasks, and several classic algorithms are proven to be useful for these types of data. However, multiview data can be observed from different points of views. We assume that the data matrix is denoted as $\mathbf{X} = \{\mathbf{X}^{\{1\}}, \mathbf{X}^{\{2\}}, \dots, \mathbf{X}^{\{i\}}, \dots, \mathbf{X}^{\{V\}}\}$, where $\mathbf{V}$ refers to the number of views. In the remaining parts of this section, first, we extend NMF and LDA to a multiview version then form a connection between them to deduce an unsupervised multiview framework. In addition, two multiview manifold regularizations are combined in the framework to improve performance.

2.1. Multiview NMF

We assume the existence of a data matrix $\mathbf{X} = [\mathbf{x_1}, \mathbf{x_2}, \dots, \mathbf{x_n}], \mathbf{x_i} \in \mathbb{R}^{d\times 1}$, and d and n represent the numbers of dimension and sample, respectively. NMF tries to approximate $\mathbf{X}$ with the product of $\mathbf{F}$ and $\mathbf{G^T}$ ($\mathbf{F}$ and $\mathbf{G^T}$ are all non-negative matrices), as follows:

$$\min_{\mathbf{F}\geq 0, \mathbf{G}\geq 0} \|\mathbf{X} - \mathbf{F}\mathbf{G^T}\|_F^2, \tag{1}$$

where $\mathbf{F} \in \mathbb{R}^{d\times c}$ refers to a centroid matrix and $\mathbf{G} \in \mathbb{R}^{n\times c}$ means a indicator matrix in clustering. The product of $\delta\mathbf{F}$ and $\mathbf{G^T}/\delta$ results in the residue error when the scalar $\delta > 0$.

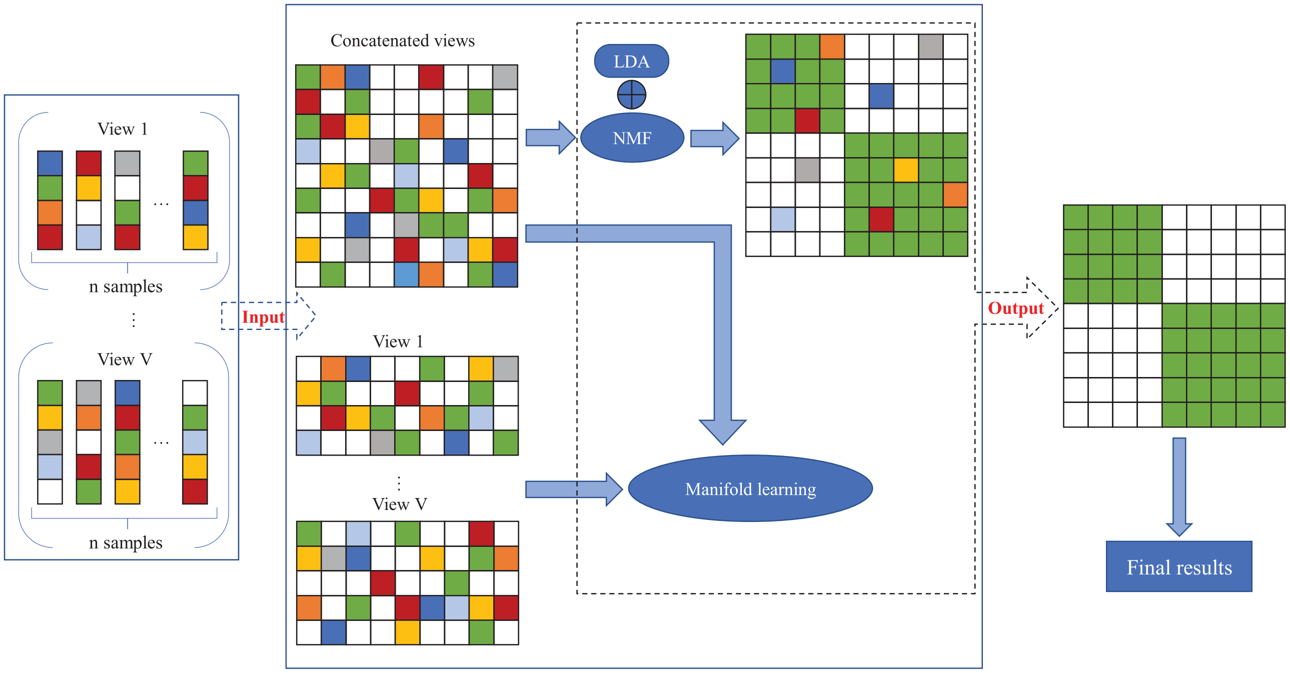

Fig. 1. The structural block diagram of the proposed DMSMF method.

To obtain the unique solution of Problem (1), Huang [37] imposed the orthogonal constraint on $\mathbf{G}$. Problem (1) can be reformulated as:

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

where $\mathbf{I_c} \in \mathbb{R}^{c\times c}$ denotes an identity matrix. Thus we can achieve its optimal solution and keep its uniqueness by these constraints simultaneously.

In this part, we pay close attention to its extension, that is, multiview NMF. We suppose that a total of $\mathbf{V}$ views are observed for each data point, that is, $\mathbf{x_i} = \{\mathbf{x_i^{(1)}} \in \mathbb{R}^{d_1}, \ldots, \mathbf{x_i^{(V)}} \in \mathbb{R}^{d_V}\}$. To complete feature concatenation in a multiview manner, we denote another form of points as $\mathbf{x_i} = [\mathbf{x_i^{(1)}}, \mathbf{x_i^{(2)}}, \ldots, \mathbf{x_i^{(V)}}]^\mathbf{T} \in \mathbb{R}^d, d = \sum_{k=1}^{\mathbf{V}} d_i$. Thus the original feature matrix in the multiview version is also extended to $\mathbf{X} = [\mathbf{X^{(1)}}, \mathbf{X^{(2)}}, \ldots, \mathbf{X^{(V)}}]^\mathbf{T} \in \mathbb{R}^{d\times n}$.

We assume that the unique cluster indicator matrix $\mathbf{G}$ is shared by all views. Accordingly, we minimize the following problem for the kth view:

$$\min_{\mathbf{F}\geq 0,\mathbf{G}\geq 0,\mathbf{G}^T\mathbf{G}=\mathbf{I_c}} \|\mathbf{X^{(k)}}-\mathbf{F^{(k)}}\mathbf{G}^\mathbf{T}\|_F^2. \tag{3}$$

Under the kth view, the cluster centroid $\mathbf{F^{(k)}}$ is combined as a whole and the multiview cluster centroid is represented as $\mathbf{F} = [\mathbf{F^{(1)}}, \mathbf{F^{(2)}}, \ldots, \mathbf{F^{(V)}}]^\mathbf{T} \in \mathbb{R}^{d\times c}$. According to the aforementioned definitions, Problem (3) can be unified simply to:

$$\min_{\mathbf{F}\geq 0,\mathbf{G}\geq 0,\mathbf{G}^T\mathbf{G}=\mathbf{I_c}} \|\mathbf{X}-\mathbf{F}\mathbf{G}^\mathbf{T}\|_F^2, \tag{4}$$

where $\mathbf{X} = [\mathbf{X^{(1)}}, \mathbf{X^{(2)}}, \ldots, \mathbf{X^{(V)}}]^\mathbf{T} \in \mathbb{R}^{d\times n}$ and $\mathbf{F} = [\mathbf{F^{(1)}}, \mathbf{F^{(2)}}, \ldots, \mathbf{F^{(V)}}]^\mathbf{T} \in \mathbb{R}^{d\times c}$ are the multiview feature matrix and multiview cluster centroid, respectively.

2.2. Multiview LDA

We suppose that we have V views and a corresponding data matrix $\mathbf{X} = [\mathbf{X^{(1)}}, \mathbf{X^{(2)}}, \ldots, \mathbf{X^{(V)}}]^\mathbf{T} \in \mathbb{R}^{d\times n}$. For the kth view, LDA tries to learn a matrix $\mathbf{W^{(k)}} \in \mathbb{R}^{d_k\times m}$ to project the d_k-dimensional representation into the m-dimensional subspace.

With a label matrix $\mathbf{L} = [\mathbf{l_1}, \mathbf{l_2}, \ldots, \mathbf{l_n}]^\mathbf{T} \in \{0, 1\}^{n\times c}$, Yang [38] defined three scatter matrices, namely, the total-class matrix $\mathbf{S_t^{(k)}}$, the between-class matrix $\mathbf{S_b^{(k)}}$ and the within-class matrix $\mathbf{S_w^{(k)}}$, as follows:

$$\begin{aligned}
\mathbf{S_t^{(k)}} &= \sum_{i=1}^{n} (\mathbf{x_i^{(k)}} - \mu^{(\mathbf{k})})(\mathbf{x_i^{(k)}} - \mu^{(\mathbf{k})})^\mathbf{T} = \mathbf{X^{(k)}HHX^{(k)}}^\mathbf{T},\\
\mathbf{S_b^{(k)}} &= \sum_{i=1}^{c} \mathbf{n_i}(\mu_\mathbf{i}^{(\mathbf{k})} - \mu^{(\mathbf{k})})(\mu_\mathbf{i}^{(\mathbf{k})} - \mu^{(\mathbf{k})})^\mathbf{T} = \mathbf{X^{(k)}HSS^THX^{(k)}}^\mathbf{T},\\
\mathbf{S_w^{(k)}} &= \mathbf{S_t^{(k)}} - \mathbf{S_b^{(k)}},
\end{aligned} \tag{5}$$

where $\mu^{(k)}$ represents the average of all points, $\mu_i^{(k)}$ denotes the average of the ith class points and n_i expresses the amount of the ith class points. $\mathbf{S} = \mathbf{L}(\mathbf{L^TL})^{-\mathbf{1/2}}$ indicates a scaled label matrix and $\mathbf{H} \in \mathbb{R}^{n\times n}$ is $\mathbf{I_n} - \frac{1}{\mathbf{n}}\mathbf{1_n1_n^T}$. ($\mathbf{I_n}$ is a n-dimensional identity matrix and $\mathbf{1_n} \in \mathbb{R}^{n\times 1}$ is a all 1 column vector)

To learn the optimal $\mathbf{W^{(k)}}$, LDA tries to make points from same class closer to one another. Thus, we rewrite its objective function as:

$$\begin{aligned}
&\min_{\mathbf{W^{(k)}}} \mathbf{tr}(\mathbf{W^{(k)}}^\mathbf{T}\mathbf{S_w^{(k)}}\mathbf{W^{(k)}})\\
&s.t.\quad \mathbf{W^{(k)}}^\mathbf{T}\mathbf{S_w^{(k)}}\mathbf{W^{(k)}} = \mathbf{I_m},
\end{aligned} \tag{6}$$

where $\mathbf{tr}(\,\cdot\,)$ is the trace operator. In consideration of Eq. (5), Problem (6) is equivalent to:

$$\begin{aligned}
&\min_{\mathbf{W^{(k)}}} \mathbf{tr}(\mathbf{W^{(k)}}^\mathbf{T}(\mathbf{S_t^{(k)}} - \mathbf{S_b^{(k)}})\mathbf{W^{(k)}})\\
=&\min_{\mathbf{W^{(k)}}} \|\mathbf{W^{(k)}}^\mathbf{T}\mathbf{X^{(k)}H}(\mathbf{I_n} - \mathbf{SS^T})\|_F^2\\
&s.t.\quad \mathbf{W^{(k)}}^\mathbf{T}\mathbf{S_w^{(k)}}\mathbf{W^{(k)}} = \mathbf{I_m}.
\end{aligned} \tag{7}$$

To unify the different views, we optimize this problem to

$$\begin{aligned}
&\min_{\mathbf{W^{(k)}}} \sum_{k=1}^{\mathbf{V}} \|\mathbf{W^{(k)}}^\mathbf{T}\mathbf{X^{(k)}H}(\mathbf{I_n} - \mathbf{SS^T})\|_F^2\\
=&\min_{\mathbf{W}} \|\mathbf{W^TXH}(\mathbf{I_n} - \mathbf{SS^T})\|_F^2\\
&s.t.\quad \mathbf{W^TS_wW} = \mathbf{I_m}.
\end{aligned} \tag{8}$$

where $\mathbf{X} = [\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(V)}]^{\mathbf{T}} \in \mathbb{R}^{d\times n}$ and $\mathbf{W} = [\mathbf{W}^{(1)^{\mathbf{T}}}, \dots, \mathbf{W}^{(V)^{\mathbf{T}}}]^{\mathbf{T}} \in \mathbb{R}^{d\times m}$ are the multiview feature and projection matrices, respectively.

2.3. Multiview matrix factorization in discriminative subspace

Most current methods cannot capture the intrinsic structure in high-dimensional data and thus rely on it in the low-dimensional subspace. In the remaining parts of this subsection, we discover a connection between multiview NMF and multiview LDA.

By considering the framework of multiview NMF, the optimal **F** is easily equal to **XG** by fixing **G**. By replacing **F** with **XG**, Problem (4) becomes:

$$\|\mathbf{X} - \mathbf{X}\mathbf{G}\mathbf{G}^{\mathbf{T}}\|_F^2 = \|\mathbf{X}(\mathbf{I_n} - \mathbf{G}\mathbf{G}^{\mathbf{T}})\|_F^2, \tag{9}$$

which is similar to Problem (8). The difference between Problem (8) and Problem (9) is that $\mathbf{W^TXH}$ and **S** correspond **X** and **G**, respectively. Moreover, the optimal solution of **G** for Problem (4) actually denotes the unique cluster indicator in the NMF framework, and **S** in Problem (8) is the scaled label matrix. In this way, **G** and **S** have similar meanings in practice. Therefore, the unsupervised multiview matrix factorization framework is connected to multiview LDA, as follows:

$$\begin{aligned} &\min_{\mathbf{F},\mathbf{G},\mathbf{W}} \|\mathbf{W^TXH} - \mathbf{FG^T}\|_F^2, \\ &s.t. \quad \mathbf{F} \in \mathbb{R}^{m\times c}, \mathbf{G} \in \mathbb{R}^{n\times c}, \mathbf{W} \in \mathbb{R}^{d\times m}, \\ &\mathbf{G} \geq 0, \mathbf{G^TG} = \mathbf{I_c}, \mathbf{W^TS_tW} = \mathbf{I_m}, \end{aligned} \tag{10}$$

where **G** can be considered as the pseudo-information matrix or the unique cluster indicator. **W** is the multiview projection matrix that can find the discriminant subspace in the multiview matrix factorization framework.

In accordance with [37], we replace the Frobenius norm with the $l_{2,1}$-norm to improve the robustness. In this way, we obtain a robust unsupervised multiview framework as follows:

$$\begin{aligned} &\min_{\mathbf{F},\mathbf{G},\mathbf{W}} \|\mathbf{W^TXH} - \mathbf{FG^T}\|_{2,1}, \\ &s.t. \quad \mathbf{F} \in \mathbb{R}^{m\times c}, \mathbf{G} \in \mathbb{R}^{n\times c}, \mathbf{W} \in \mathbb{R}^{d\times m}, \\ &\mathbf{G} \geq 0, \mathbf{G^TG} = \mathbf{I_c}, \mathbf{W^TS_tW} = \mathbf{I_m}, \end{aligned} \tag{11}$$

2.4. Pseudo supervised multiview manifold regularization

Considering the original data from the perspective of multiview $\mathbf{X} = [\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \dots, \mathbf{X}^{(V)}]^{\mathbf{T}} \in \mathbb{R}^{d\times n}$ and its projected low-dimensional multiview matrix $\mathbf{Y} = [\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)}, \dots, \mathbf{Y}^{(V)}]^{\mathbf{T}} \in \mathbb{R}^{m\times n}$, $\mathbf{W} = [\mathbf{W}^{(1)^T}, \mathbf{W}^{(2)^T}, \dots, \mathbf{W}^{(V)^{\mathbf{T}}}]^{\mathbf{T}} \in \mathbb{R}^{d\times m}$ is the multiview matrix that projects a data matrix from d dimension into m dimension, that is, $\mathbf{Y} = \mathbf{W^TX}$.

As discussed previously, multiview LDA learns to discover a subspace that can discriminate different classes by optimizing $\mathbf{tr}(\mathbf{S_W^{(k)}})$ and $\mathbf{tr}(\mathbf{S_B^{(k)}})$. Following [39], the following problem is solved to minimize $\mathbf{S_W^{(k)}}$:

$$\begin{aligned} &\min \mathbf{tr}(\mathbf{S_W^{(k)}}) \\ &= \min_{\vec{y_i^j}^{(k)}} tr\left(\sum_{i=1}^{C}\sum_{j=1}^{N_i}(\vec{y_i^j}^{(k)} - y_i^{\vec{m}^{(k)}})(\vec{y_i^j}^{(k)} - y_i^{\vec{m}^{(k)}})^T\right), \end{aligned} \tag{12}$$

where C denotes the class number, N_i represents the number of points in the ith class, $\vec{y_i^j}^{(k)}$ indicates the jth sample in the ith class under the kth view, and $y_i^{\vec{m}^{(k)}}$ expresses the centroid of the ith class under the kth view. Problem (12) can be reduced into

$$\min_{\mathbf{Y_i^{(k)}}} \sum_{i=1}^{N} \mathbf{tr}(\mathbf{Y_i^{(k)}} \mathbf{L_i^{W^{(k)}}} \mathbf{Y_i^{(k)^T}}), \tag{13}$$

where

$$\mathbf{L_i^{W^{(k)}}} = \frac{\mathbf{1}}{\mathbf{N_i^2}} \begin{bmatrix} \mathbf{N_i - 1} \\ -\vec{\mathbf{e}}_{\mathbf{N_i-1}} \end{bmatrix} [\mathbf{N_i - 1} \quad -\vec{\mathbf{e}}^{\mathbf{T}}_{\mathbf{N_i-1}}],$$

$\mathbf{Y_i^{(k)}} = \mathbf{YS_i^{(k)}} = [\vec{\mathbf{y_i}}^{(\mathbf{k})}, \mathbf{y_i^{\vec{1}}}^{(\mathbf{k})}, \dots, \mathbf{y_i^{\vec{N_i-1}}}^{(\mathbf{k})}]$ and $\vec{\mathbf{e}}_{\mathbf{N_i-1}} = [\mathbf{1}, \dots, \mathbf{1}]^{\mathbf{T}} \in \mathbb{R}^{N_i-1}$.

$\mathbf{L_W^{(k)}}$ is the alignment matrix and $\mathbf{S_i}$ is a selection matrix. The pseudo supervised multiview manifold regularization originates from a multiview version of pseudo supervised regularization and LDA. However, as a multiview objective function, we sum up all the views and calculate the final loss in a unified framework. According to [40], we share the same $\mathbf{L_i^W}$ and $\mathbf{S_i}$ given that we share the same cluster indicator matrix **G** under different views. Accordingly, we can obtain $\mathbf{L_W^{(k)}}$

$$\mathbf{L_W} = \sum_{\mathbf{i=1}}^{\mathbf{N}} \mathbf{S_i L_i^W S_i}^{\mathbf{T}}. \tag{14}$$

In this way, we actually obtain a concatenated-like feature matrix that uses view-specific information implicitly and will not cause loss of information in the entire framework.

With Problem (13) and Eq. (14) [39], we obtain:

$$\begin{aligned} &\min_{\mathbf{Y_i^{(k)}}} \sum_{i=1}^{N} \mathbf{tr}(\mathbf{Y_i^{(k)}} \mathbf{L_i^{W^{(k)}}} \mathbf{Y^{(k)}{}_i}^{\mathbf{T}}) \\ &= \min_{\mathbf{Y}} \sum_{i=1}^{N} \mathbf{tr}(\mathbf{YS_i^{(k)}} \mathbf{L_i^{W^{(k)}}} \mathbf{S_i^{(k)}}^{\mathbf{T}} \mathbf{Y^T}) \\ &= \min_{\mathbf{Y}} \mathbf{tr}(\mathbf{YL_WY^T}). \end{aligned} \tag{15}$$

Following the definition of multiview LDA, we replace **Y** with the form of $\mathbf{W^TX}$. Thus, Problem (15) is equivalent to the following problem:

$$\begin{aligned} &\min_{\mathbf{W}} \mathbf{tr}(\mathbf{W_W^{TXL}X^{TW}}), \\ &s.t. \quad \mathbf{WW^T} = \mathbf{I_d}. \end{aligned} \tag{16}$$

Problem (16) is a manifold regularizer that is developed from multiview LDA. This regularization term is inspired by the within-class scatter from LDA. In contrast to other SC and manifold regularization methods, it considers the pseudo-label information. Although the pseudo-label information is not the real label information, it can instruct and improve clustering results during iterations. Thus, it needs data structure information to instruct itself.

2.5. Discriminant multiview manifold regularization

From the perspective of manifold learning, data samples in the region of the low-dimensional manifold with a high density can be assumed to be from the same class. For the kth view, the low-dimensional data graph $\mathbf{S^{(k)}} \in \mathbb{R}^{n\times n}$ is built by $\mathbf{W^{(k)T}X^{(k)}}$ when $\mathbf{W^{(k)}}$ is fixed. Naturally, we minimize the following problem to obtain its geometry structure from low-dimensional subspace, that is,

$$\min_{\mathbf{G}} \mathbf{tr}(\mathbf{G^TL_m^{(k)}G}), \tag{17}$$

where $\mathbf{L_m^{(k)}}$ is the Laplacian matrix of the data graph $\mathbf{S^{(k)}}$.

We extend Problem (17) to a multiview version and obtain:

$$\min_{\mathbf{G}} \sum_{k=1}^{V} (\alpha^{(k)}) \mathbf{tr}(\mathbf{G^TL_m^{(k)}G}). \tag{18}$$

In Problem (18), $\alpha = [\alpha^{(1)}, \alpha^{(2)}, \dots, \alpha^{(V)}]$ denotes a set of non-negative weight coefficients. The constraints are $\alpha > 0$ and $\sum_{k=1}^{V} \alpha^{(k)} = 1$. $\alpha^{(i)}$ decides the importance of the ith view to the entire framework. If we assign $\alpha^{(i)} = 1$ corresponding to the ith view with all the other j having $\alpha^{(j)} = 0$ in its views, we can easily obtain its minimum $\mathbf{tr}(\mathbf{G^T L_m^{(i)} G})$. This procedure will result in the failure to discover their multiview complementary property. To avoid this unusual solution, we perform the same trick used in Wang et al. [41]. $\alpha^{(k)}$ in Problem (18) is set as $(\alpha^{(k)})^r$ and $r > 1$ is a scale parameter that controls feature weight. By adopting this trick, $\sum_{k=1}^{V}(\alpha^{(k)})^r$ reaches its minimum value when $\alpha^{(k)} = \frac{1}{V}$.

Therefore, we replace $\alpha^{(k)}$ in Problem (18) with $(\alpha^{(k)})^r$, and define the multiview manifold regularization formula as:

$$\min_{\mathbf{G}} \sum_{k=1}^{V} (\alpha^{(k)})^r \mathbf{tr}(\mathbf{G^T L_m^{(k)} G}). \tag{19}$$

By adjusting the weights of the multiview features, we can assign the proper contribution of each view. This term encodes the complementary information and redundant properties appropriately and improves clustering ability [42] [43].

2.6. Objective function

By combining Problems (11), (16), and (19), we form the framework of the proposed DMSMF method, as follows:

$$\begin{aligned}
&\min_{\mathbf{F},\mathbf{G},\mathbf{W},\alpha} \|\mathbf{W^T X H} - \mathbf{F G^T}\|_{2,1} + \\
&\lambda_1 \sum_{k=1}^{V} (\alpha^{(k)})^r \mathbf{tr}(\mathbf{G^T L_m^{(k)} G}) + \lambda_2 \mathbf{tr}(\mathbf{W^T X L_W X^T W}), \\
&s.t. \quad \mathbf{F} \in \mathbb{R}^{m\times c}, \mathbf{G} \in \mathbb{R}^{n\times c}, \mathbf{W} \in \mathbb{R}^{d\times m}, \mathbf{G} \ge 0, \\
&\mathbf{G^T G} = \mathbf{I_c}, \mathbf{W^T S_t W} = \mathbf{I_m}, \mathbf{W W^T} = \mathbf{I_d}, \sum_{k=1}^{V} \alpha^{(k)} = 1,
\end{aligned} \tag{20}$$

where λ_1 and λ_2 are two non-negative manifold regularization parameters. When $\mathbf{W}$ and $\mathbf{G}$ are updated, we reconstruct $\mathbf{L_m}$ and $\mathbf{L_W}$ which are a Laplacian matrix and an alignment matrix for further updates. The proposed method is inspired by LDA and is an unsupervised method that do not use any true label information. Instead, we use the pseudo-label information to instruct and improve the clustering results during iterations. The pseudo-label information can be estimated roughly with clustering methods such as CAN [44].

3. Solution

With the constraint $\mathbf{W^T S_t W} = \mathbf{I_m}$, Problem (20) is difficult to solve. Following [15], we need to let $\mathbf{S_t}$ be positive definite to decompose it by Cholesky decomposition. According to [45], if we disturb the diagonal elements of $\mathbf{S_t}$ by a scalar which is small enough but beyond zero, the matrix $\mathbf{S_t}$ will still be positive definite. So we first provide the diagonal elements of $\mathbf{S_t}$ with a sufficiently small disturbance, such as a scalar $\theta > 0$. Then, we decompose it in the form of $\mathbf{S_t} = \mathbf{R^T R}$. Thereafter, if we denote $\mathbf{RW}$, $(\mathbf{R^{-1}})^\mathbf{T}\mathbf{XH}$ and $(\mathbf{R^{-1}})^\mathbf{T}\mathbf{X}$ as $\mathbf{P}$, $\mathbf{A}$, and $\mathbf{B}$, respectively, then Problem (20) is reduced into:

$$\begin{aligned}
&\min_{\mathbf{F},\mathbf{G},\mathbf{P},\alpha} \|\mathbf{P^T A} - \mathbf{F G^T}\|_{2,1} + \lambda_1 \sum_{k=1}^{V} (\alpha^{(k)})^r \mathbf{tr}(\mathbf{Gbf^T L_m^{(k)} G}) + \\
&\lambda_2 \mathbf{tr}(\mathbf{P^T B L_W B^T P}), \\
&s.t. \quad \mathbf{F} \in \mathbb{R}^{m\times c}, \mathbf{G} \in \mathbb{R}^{n\times c}, \mathbf{P} \in \mathbb{R}^{d\times m}, \mathbf{G} \ge 0, \\
&\mathbf{G^T G} = \mathbf{I_c}, \mathbf{P^T P} = \mathbf{I_m}, \sum_{k=1}^{V} \alpha^{(k)} = 1.
\end{aligned} \tag{21}$$

However, this non-convex problem cannot be solved easily. An ALM-based method is proposed here to overcome this problem [46]. Problem (21) depends on $\mathbf{P}$ and $\mathbf{G}$. Thus, we choose to import three auxiliary variables, namely, $\mathbf{E} = \mathbf{P^T A} - \mathbf{F G^T}$, $\mathbf{Z_1} = \mathbf{G}$ and $\mathbf{Z_2} = \mathbf{P}$ to simplify the problem. Accordingly, Problem (21) is easily rewritten as the following ALM problem:

$$\begin{aligned}
&\min_{\mathbf{E}}, \mathbf{G}, \mathbf{Z_1}, \mathbf{P}, \mathbf{Z_2}, \mathbf{F}, \alpha \|\mathbf{E}\|_{2,1} + \lambda_1 \textstyle\sum_{k=1}^{V} \left(\alpha^{(k)}\right)^r \mathbf{tr}\left(\mathbf{G^T L_m^{(k)} Z_1}\right) \\
&+\lambda_2 tr\left(P^T B L_W B^T Z_2\right) + \tfrac{\mu}{2} \|\mathbf{P^T A} - \mathbf{F G^T} - \mathbf{E} + \tfrac{\mathbf{\Lambda_1}}{\boldsymbol{\mu}}\|_F^2 \\
&+\tfrac{\mu}{2} \|\mathbf{G} - \mathbf{Z_1} + \tfrac{\mathbf{\Lambda_2}}{\boldsymbol{\mu}}\|_F^2 + \tfrac{\mu}{2} \|\mathbf{P} - \mathbf{Z_2} + \tfrac{\mathbf{\Lambda_3}}{\boldsymbol{\mu}}\|_F^2, \\
&s.t. \quad \mathbf{Z_1} \ge 0, \mathbf{G} \ge 0, \mathbf{G^T G} = \mathbf{I_c}, \mathbf{P^T P} = \mathbf{I_m}, \textstyle\sum_{k=1}^{V} \alpha^{(k)} = 1,
\end{aligned} \tag{22}$$

where $\mu \in \mathbb{R}^{1\times 1}$ represents the ALM parameter, and Λ_1, Λ_2, and Λ_3 denote the ALM multipliers. Next, we can then divide the problem into several sub-problems and then optimize each variable in every iteration.

3.1. Rules for updating *E*

For the sub-problem regarding $\mathbf{E}$, we shall fix all the other variables. All the terms regarding those fixed variables are considered as constants. Thus, the problem can be expressed as:

$$\min_{\mathbf{E}} \|\mathbf{E}\|_{\mathbf{2,1}} + \frac{\mu}{\mathbf{2}} \|\mathbf{E} - \mathbf{M}\|_F^2, \tag{23}$$

where $\mathbf{M} = \mathbf{P^T} - \mathbf{F G^T} + \frac{\mathbf{\Lambda_1}}{\mu}$. We use the method in Huang et al. [37] to compute the optimal $\mathbf{E}$, as follows:

$$\mathbf{E}_{:,\mathbf{q}} = \begin{cases} (1 - \frac{1}{\mu \|\mathbf{M}_{:,\mathbf{i}}\|_2})\mathbf{M}_{:,\mathbf{i}}, & if \ \|\mathbf{M}_{:,\mathbf{i}}\|_2 \ge \frac{1}{\mu}. \\ 0, & \text{else.} \end{cases} \tag{24}$$

3.2. Rules for updating *Z₁*

When updating $\mathbf{Z_1}$, Problem (22) becomes:

$$\begin{aligned}
&\min_{Z_1} \lambda_1 \sum_{k=1}^{V} (\alpha^{(k)})^r \mathbf{tr}(\mathbf{G^T L_m^{(k)} Z_1}) + \frac{\mu}{2} \|\mathbf{G} - \mathbf{Z_1} + \frac{\mathbf{\Lambda_2}}{\mu}\|_F^2. \\
&s.t. \quad \mathbf{Z_1} \ge 0, \sum_{k=1}^{V} \alpha^{(k)} = 1
\end{aligned} \tag{25}$$

The sub-problem is simple and can be further transformed into the following closed-from version:

$$\begin{aligned}
&\min_{\mathbf{Z_1}} \|\mathbf{Z_1} - \mathbf{T}\|_\mathbf{F}^\mathbf{2}, \\
&s.t. \quad \mathbf{Z_1} \ge 0,
\end{aligned} \tag{26}$$

where $\mathbf{T} = \mathbf{G} + \frac{\Lambda_2}{\mu} - \frac{\lambda_1}{\mu} \sum_{k=1}^{V} (\alpha^{(k)})^r \mathbf{L_m^{(k)} G}$. Therefore, the optimal $\mathbf{Z_1}$ is:

$$\mathbf{Z_{1_{ij}}} = \max(\mathbf{T_{ij}}, 0). \tag{27}$$

3.3. Rules for updating *Z₂*

Similar to solving this sub-problem regarding Z_2, Problem (22) can be simplified to:

$$\begin{aligned}
&\min_{\mathbf{Z}_2 \lambda_2 \mathbf{tr}} (\mathbf{P^{TBL_W} B^T Z_2}) + \frac{\mu}{2} \|\mathbf{P} - \mathbf{Z_2} + \frac{\Lambda_3}{\mu}\|_F^2, \\
&s.t. \quad \mathbf{Z_2^T Z_2} = \mathbf{I_m}.
\end{aligned} \tag{28}$$

If we remove all the irrelevant terms regarding $\mathbf{Z_2}$, we can rewrite this sub-problem into:

$$\min_{\mathbf{Z_2}} \|\mathbf{Z_2} - \mathbf{R}\|_{\mathbf{F}}^2, \quad s.t. \quad \mathbf{Z_2^T Z_2} = \mathbf{I_m}, \tag{29}$$

where $\mathbf{R} = -\frac{\lambda_2}{\mu}\mathbf{BL_W B^T P} + \mathbf{P} + \frac{\mathbf{\Lambda_3}}{\mu}$. This condition is equivalent to:

$$\max_{\mathbf{Z_2}} \mathbf{tr}(\mathbf{Z_2^{TR}}), \quad s.t. \quad \mathbf{Z_2^T Z_2} = \mathbf{I_m}. \tag{30}$$

A proof was given in Huang et al. [37] to ensure the optimal solution, that is,

$$\mathbf{Z_2} = \mathbf{U_3 V_3^T}, \tag{31}$$

where $\mathbf{U_3} \in \mathbb{R}^{d\times m}$ and $\mathbf{V_3} \in \mathbb{R}^{m\times m}$ are computed as the singular vectors of SVD on $\mathbf{R}$.

3.4. Rules for updating *G*

Problem (22) regarding $\mathbf{G}$ is reduced to the following sub-problem:

$$\min_{\mathbf{G}} \lambda_1 \sum_{k=1}^{V} (\alpha^{(k)})^r \mathbf{tr}(\mathbf{G^T L_m^{(k)} Z_1}) + \frac{\mu}{2}\|\mathbf{P^T A} - \mathbf{FG^T} - \mathbf{E} + \frac{\Lambda_1}{\mu}\|_F^2 + \frac{\mu}{2}\|\mathbf{G} - \mathbf{Z_1} + \frac{\mathbf{\Lambda_2}}{\mu}\|_F^2, \quad s.t. \quad \mathbf{G} \geq 0, \mathbf{G^T G} = \mathbf{I_c}. \tag{32}$$

Similar to $\mathbf{Z_2}$, the above-mentioned problem reaches its optimization as:

$$\mathbf{G} = \mathbf{U_1 V_1^T}, \tag{33}$$

where $\mathbf{U_1} \in \mathbb{R}^{n\times c}$ and $\mathbf{V_1} \in \mathbb{R}^{c\times c}$ are two singular vectors of SVD on $\mathbf{K}$ and $\mathbf{K} = (-\frac{\lambda_1}{\mu}\sum_{k=1}^{V}(\alpha^{(k)})^r \mathbf{L_m^{(k)} Z_1}) + (\mathbf{P^T A} - \mathbf{E} + \frac{\mathbf{\Lambda_1}}{\mu})^{\mathbf{T}}\mathbf{F} + (\mathbf{Z_1} - \frac{\mathbf{\Lambda_2}}{\mu})$.

3.5. Rules for updating *P*

By updating $\mathbf{P}$, Problem (22) is simplified to:

$$\min_{\mathbf{P}} \lambda_2 \mathbf{tr}(\mathbf{P^T BL_W B^T Z_2}) + \frac{\mu}{2}\|\mathbf{P^T A} - \mathbf{FG^T} - \mathbf{E} + \frac{\mathbf{\Lambda_1}}{\mu}\|_{\mathbf{F}}^2 + \frac{\mu}{\mathbf{2}}\|\mathbf{P} - \mathbf{Z_2} + \frac{\mathbf{\Lambda_3}}{\mu}\|_{\mathbf{F}}^2, \quad s.t. \quad \mathbf{P^T P} = \mathbf{I_m}. \tag{34}$$

This sub-problem reaches its optimal value according to:

$$\mathbf{P} = \mathbf{U_2 V_2^T}, \tag{35}$$

where $\mathbf{U_2} \in \mathbb{R}^{d\times m}$ and $\mathbf{V_2} \in \mathbb{R}^{m\times m}$ result from the same SVD operation on $\mathbf{D}$ and $\mathbf{D} = -\frac{\lambda_2}{\mu}\mathbf{BL_W B^T Z_2} + \mathbf{A}(\mathbf{FG^T} + \mathbf{E} - \frac{\Lambda_1}{\mu})^{\mathbf{T}} + (\mathbf{Z_2} - \frac{\Lambda_3}{\mu})$.

3.6. Rules for updating *F*

Optimizing problem (22) with regard to $\mathbf{F}$ results in:

$$\min_{\mathbf{F}} \frac{\mu}{2}\|\mathbf{P^T A} - \mathbf{FG^T} - \mathbf{E} + \frac{\mathbf{\Lambda_1}}{\mu}\|_F^2. \tag{36}$$

With the constraint $\mathbf{G^T G} = \mathbf{I_c}$, this sub-problem is rewritten into:

$$\min_{\mathbf{F}} \|\mathbf{F} - (\mathbf{P^T A} - \mathbf{E} + \frac{\mathbf{\Lambda_1}}{\mu}\mathbf{G})\|_F^2. \tag{37}$$

The optimal $\mathbf{F}$ can be computed as:

$$\mathbf{F} = (\mathbf{P^T A} - \mathbf{E} + \frac{\mathbf{\Lambda_3}}{\mu})\mathbf{G}. \tag{38}$$

3.7. Rules for updating *α*

The weight parameter α can also be considered a variable. Thus, we formulate the sub-problem as follows:

$$\min_{\alpha} \sum_{k=1}^{V} (a^{(k)})^r \mathbf{tr}(\mathbf{G^T L_m^{(k)} Z_1}) \quad s.t. \quad \mathbf{Z_1} \geq 0, \mathbf{G} \geq 0, \mathbf{G^T G} = \mathbf{I_c}, \sum_{k=1}^{V} a^{(k)} = 1. \tag{39}$$

To solve this sub-problem, we apply a Lagrangian multiplier η. Thus, the Lagrangian function of Eq. (39) is deduced as:

$$L(\alpha^{(k)}, \eta) = \sum_{k=1}^{V} (\alpha^{(k)})^r p^{(k)} - \eta(\sum_{k=1}^{V} \alpha^{(k)} - 1), \tag{40}$$

where $p^{(k)} = \mathbf{tr}(\mathbf{G^T L_m^{(k)} Z_1})$. If we take the derivative with regard to $\alpha^{(k)}$ to 0, then we have

$$r(\alpha^{(k)})^{r-1}p^{(k)} - \eta = 0 \Rightarrow \alpha^{(k)} = (\frac{\eta}{rp^{(k)}})^{\frac{1}{r-1}}. \tag{41}$$

Substituting $\sum_{k=1}^{V} \alpha^{(k)} = 1$ into Eq. (41) transforms the updating rule of $\alpha^{(k)}$ into

$$\alpha^{(k)} = (rp^{(k)})^{\frac{1}{1-r}} \Big/ \sum_{k=1}^{V} (rp^{(k)})^{\frac{1}{1-r}}. \tag{42}$$

3.8. Rules for updating *μ*, *Λ₁*, *Λ₂* and *Λ₃*

The following rules are used for the ALM parameters:

$$\begin{aligned} \Lambda_1 &= \Lambda_1 + \mu(\mathbf{P^T A} - \mathbf{FG^T} - \mathbf{E}), \\ \Lambda_1 &= \Lambda_1 + \mu(\mathbf{G} - \mathbf{Z_1}), \\ \Lambda_1 &= \Lambda_1 + \mu(\mathbf{P} - \mathbf{Z_2}), \\ \mu &= \rho\mu, \end{aligned} \tag{43}$$

where ρ is a parameter for deciding the time it takes for convergence.

The optimization procedures of the DMSMF method are described in Algorithm 1.

Algorithm 1 DMSMF.

```
Input:
Original data matrix X as $\mathbf{X} = [\mathbf{X^{(1)}}, \mathbf{X^{(2)}}, \ldots, \mathbf{X^{(V)}}]^{\mathbf{T}} \in \mathbb{R}^{d\times n}$;
Cluster number $k$;
Regularization parameters $\lambda$ and $\mu$;
Scale parameter $r$
Output:
Cluster indicator G.
Initialize $\mathbf{G} \in \mathbb{R}^{n\times c}$;
Initialize $\mathbf{W} = [\mathbf{W^{(1)}}^{\mathbf{T}}, \mathbf{W^{(2)}}^{\mathbf{T}}, \ldots, \mathbf{W^{(V)}}^{\mathbf{T}}]^{\mathbf{T}} \in \mathbb{R}^{d\times m}$;
Initialize $\mathbf{F} = \mathbf{W^T XG}$ and $\mathbf{P} = \mathbf{RW}$ and compute $\mathbf{H}, \mathbf{A}, \mathbf{B}, \mathbf{S_t}, \mathbf{S_b}, \mathbf{S_w}$;
Initialize $\alpha^{(k)}$ for each view;
while not converge do
    Calculate and update E by Eq. (24);
    Calculate and update $\mathbf{Z_1}$ by Eq. (27);
    Calculate and update $\mathbf{Z_2}$ by Eq. (31);
    Calculate and update G by Eq. (33);
    Calculate and update P by Eq. (35);
    Calculate and update F by Eq. (38);
    Calculate and update $\alpha$ by Eq. (42);
    Calculate and update ALM parameters by Eq. (43);
end while
```

4. Connections with other methods

In this section, we show that our framework has connections to other classic methods theoretically. Specifically, when we choose a proper combination of λ_1 and λ_2, our method takes advantage of their merits.

4.1. Connection to NMF

The relationship between Problem (20) and NMF is discussed in this section. The objective function of NMF clustering is:

$$\begin{aligned}&\min_{\mathbf{F},\mathbf{G}} \|\mathbf{X}-\mathbf{F}\mathbf{G}^{\mathbf{T}}\|_F^2,\\ &s.t\quad \mathbf{G}\geq 0, \mathbf{G}^{\mathbf{T}}\mathbf{G}=\mathbf{I_c}.\end{aligned} \tag{44}$$

For Problem (44), we can apply $l_{2,1}$-norm for robustness. As a result, we obtain the following robust version of NMF:

$$\begin{aligned}&\min_{\mathbf{F},\mathbf{G}} \|\mathbf{X}-\mathbf{F}\mathbf{G}^{\mathbf{T}}\|_{2,1},\\ &s.t\quad \mathbf{G}\geq 0, \mathbf{G}^{\mathbf{T}}\mathbf{G}=\mathbf{I_c}.\end{aligned} \tag{45}$$

When $\lambda_1 \to 0$ and $\lambda_2 \to 0$ in Problem (20), we can obtain it in the form of Problem (46), as follows:

$$\begin{aligned}&\min_{\mathbf{F},\mathbf{G},\mathbf{W}} \|\mathbf{W}^{\mathbf{T}}\mathbf{X}\mathbf{H}-\mathbf{F}\mathbf{G}^{\mathbf{T}}\|_{2,1},\\ &s.t.\quad \mathbf{G}\geq 0, \mathbf{G}^{\mathbf{T}}\mathbf{G}=\mathbf{I_c}, \mathbf{W}\mathbf{W}^{\mathbf{T}}=\mathbf{I_d}.\end{aligned} \tag{46}$$

For Problem (46), if we replace $\mathbf{W}^{\mathbf{T}}\mathbf{X}$ with low-dimensional matrix $\mathbf{Y}$, then we can extend it into the following form:

$$\min_{\mathbf{F},\mathbf{G},\mathbf{Y}} \left\|\mathbf{Y}(\mathbf{I_n}-\frac{\mathbf{1}}{\mathbf{n}}\mathbf{1_n})-\mathbf{F}\mathbf{G}^{\mathbf{T}}\right\|_{2,1} \tag{47}$$

Theorem 4.1. *Zero-centered operation on origin data matrix affects its cluster centroid matrix. The clustering results remain unchanged.*

We consider the following form of input data matrix:

$$\begin{aligned}\mathbf{Z}&=\mathbf{Y}(\mathbf{I_n}-\frac{\mathbf{1}}{\mathbf{n}}\mathbf{1_n})\\ &=\mathbf{Y}-\frac{\mathbf{Y}}{\mathbf{n}}\mathbf{1_n},\end{aligned} \tag{48}$$

and the following problem:

$$\begin{aligned}&\min_{\mathbf{F},\mathbf{G}}\|\mathbf{Z}-\tilde{\mathbf{F}}\tilde{\mathbf{G}}^{\mathbf{T}}\|_{2,1}\\ &\quad=\left\|\mathbf{Y}(\mathbf{I_n}-\frac{\mathbf{1}}{\mathbf{n}}\mathbf{1_n})-\tilde{\mathbf{F}}\tilde{\mathbf{G}}^{\mathbf{T}}\right\|_{2,1},\end{aligned} \tag{49}$$

it actually performs zero-centered operation on matrix Y by row.

The origin data **Y** *can be denoted in the following form of NMF:*

$$\begin{aligned}&\min_{\mathbf{F},\mathbf{G}} \|\mathbf{Y}-\mathbf{F}\mathbf{G}^{\mathbf{T}}\|_{2,1},\\ &s.t\quad \mathbf{G}\geq 0, \mathbf{G}^{\mathbf{T}}\mathbf{G}=\mathbf{I_c}.\end{aligned} \tag{50}$$

In Problems (49) and (50), the two cluster centroid matrices differ, as centroids are changed. Meanwhile, the final clustering results are the same, because post-processing is needed on $\tilde{\mathbf{G}}$ *and* **G** *for results.*

Proof. We can transform Eq. (48) into an element form as follows:

$$\begin{aligned}z_{ij}&=y_{ij}-\frac{\sum_{k=1}^{n}y_{ik}}{n}\\ &=y_{ij}-y_i^{mean},\end{aligned} \tag{51}$$

where y_i^{mean} represents the mean of the ith row.

For every z_{ij} in the raw data matrix, we minus y_{ij} with mean of each row. Thus, we perform zero-centered operation on matrix Y by row.

By optimizing Problem (49) and (50), we obtain $\tilde{\mathbf{G}}$ and **G** with their $\tilde{\mathbf{F}}$ and **F**, respectively. Given the zero-centered operation on matrix **Y**, centroids in **F** are shifted to centroids in $\tilde{\mathbf{F}}$. Thus, the zero-centered operation on the origin data matrix affects its cluster centroid matrix.

We must conduct post-processing on $\tilde{\mathbf{G}}$ and **G** to obtain the final clustering results. The data distributions of **Z** and **Y** remain the same. Thus, we will obtain the same clustering results from the original data and their zero-centred operation version. □

According to Theorem 4.1, Problems (46) and (45) have similar forms. Thus, our method implicitly originates from NMF clustering with dimensionality reduction and zero-centred operation.

4.2. Connection to SC

Classic SC is based on the Laplacian graph. We must learn a graph partition to reduce the weights among different groups.

When $\lambda_1 \to \infty$ and $\lambda_2 \to 0$ in Problem (20), we can approximate it as Problem (52):

$$\begin{aligned}&\min_{G,\alpha}\sum_{k=1}^{V}(\alpha^{(k)})^r\mathbf{tr}(\mathbf{G}^{\mathbf{T}}\mathbf{L}_{\mathbf{m}}^{(\mathbf{k})}\mathbf{G}),\\ &s.t.\quad \mathbf{G}\geq 0, \mathbf{G}^{\mathbf{T}}\mathbf{G}=\mathbf{I_c}, \sum_{k=1}^{V}\alpha^{(k)}=1.\end{aligned} \tag{52}$$

Problem (52) can achieve its minimum with $a^{(k)}=\frac{1}{V}$. Accordingly, Problem (52) turns into:

$$\begin{aligned}&\min_{\mathbf{G}}\sum_{k=1}^{V}\mathbf{tr}(\mathbf{G}^{\mathbf{T}}\mathbf{L}_{\mathbf{m}}^{(\mathbf{k})}\mathbf{G}),\\ &s.t.\quad \mathbf{G}\geq 0, \mathbf{G}^{\mathbf{T}}\mathbf{G}=\mathbf{I_c},\end{aligned} \tag{53}$$

For every **k**, we compute $\mathbf{tr}(\mathbf{G}^{\mathbf{T}}\mathbf{L}_{\mathbf{m}}^{(\mathbf{k})}\mathbf{G})$ and then sum them up for optimization. The resultant is the form of multiview SC. According to **L**, we can induce three different methods from graph cut and perturbation theories [47].

4.3. Connection to LDA

In this section, we show the relationship between Problem (20) and multiview LDA. When $\lambda_1 \to 0$ and $\lambda_2 \to \infty$ in Problem (20), we can approximate it into Problem (54), as follows:

$$\begin{aligned}&\min_{\mathbf{W}}\mathbf{tr}(\mathbf{W}^{\mathbf{T}}\mathbf{X}\mathbf{L_W}\mathbf{X}^{\mathbf{T}}\mathbf{W}),\\ &s.t.\quad \mathbf{W}^{\mathbf{T}}\mathbf{S_t}\mathbf{W}=\mathbf{I_m}, \mathbf{W}\mathbf{W}^{\mathbf{T}}=\mathbf{I_d},\end{aligned} \tag{54}$$

According to Section 2.4, we can deduce that:

$$\min_{\mathbf{Y}_{\mathbf{i}}^{(\mathbf{k})}}\sum_{i=1}^{N}\mathbf{tr}(\mathbf{Y}_{\mathbf{i}}^{(\mathbf{k})}\mathbf{L}^{\mathbf{W}_{\mathbf{i}}^{(\mathbf{k})}}\mathbf{Y}^{(\mathbf{k})}{}_{\mathbf{i}}^{\mathbf{T}}). \tag{55}$$

Problem (55) is exactly in the form of multiview LDA. It performs LDA on every view of original data. With this term, we can find its low-dimensional data structure.

5. Experiments

In this section, we evaluate the performance of our algorithm and that of several comparable methods on multiple datasets. The experimental results illustrate the effectiveness of the DMSMF method for multiview clustering tasks. The parameter sensitivity and convergence studies for the DMSMF method are also discussed in this section.

Table 1
Description of datasets.

Datasets	View 1	View 2	View 3	View 4	View 5	Size	Classes
Cora	1433	2708	–	–	–	2708	7
Citeseer	3703	3312	–	–	–	3312	6
Wiki	128	10	–	–	–	2866	10
Digits	216	76	240	47	6	2000	10
USPS	100	100	100	100	100	300	10

5.1. Datasets

We select five representative datasets to test performance. Table 1 explains the detailed features of datasets used in our experiments.

1. **Cora**: Cora is a dataset that consists of 2708 articles on machine learning. This dataset can be divided into two individual views. One view indicates the existence of a word in a paper, and the other view reflects how the papers cite one another and their relationships.
2. **Citeseer**: This dataset contains 4732 citations with regard to 3312 publications. All the publications have unique labels. The six labels are AI, DB, HC, ML, Agents, and IR. Citeseer and Cora have a similar structure. One view indicates the existence of a word in a paper, and the other view reflects how the papers cite one another and their relationships.
3. **Wiki text-image data**: This dataset is composed of 2866 image text pairs with two different views. The first view contains 10-dimensional latent Dirichlet allocation model-based text features, and the second view includes 128-dimensional SIFT histogram image features.
4. **UCI Digits**: This dataset is an original UCI handwritten digit dataset and contains 2000 digital data points, from 0 to 9. Each class includes 200 data samples. It is extracted from six views: 76 Fourier coefficients of character shapes, 216 profile correlations, 64 Karhunen Loeve coefficients, 240 pixel averages in 2 3 windows, 47 Zernike moments, and 6 morphological features.
5. **USPS**: Similar to Xie [48], we design synthesis data based on the USPS digit dataset [49]. This dataset contains handwritten digit images iwth a resolution of 16 16 pixels. Thus, the original feature dimensionality of each sample is 256. We generate five different views by mapping the data into different subspaces via PCA. PCA maps the data into a low-dimensional subspace. Thus, the maximum variance of the input data matrix is preserved. For the first view, digits labeled 1 and 2 are selected to construct a sub-feature matrix and fed into the PCA to obtain the first view. The second view is generated similar to the first view, but only digits labeled 3 and 4 are used, and so on. In the end, we obtain five synthesis views for the USPS digit dataset.

5.2. Evaluation metrics

Following [50], we adopt two common evaluation metrics to measure the clustering performance of the models by quantity.

1. **Clustering accuracy (ACC)** [31] reveals the overall connection of clusters and their original classes. Through ACC, we can learn whether one cluster includes samples belonging to its corresponding class. Its definition is given as follows:

$$Acc = \frac{\sum_{i=1}^{n} \delta(map(r_i), l_i)}{n}, \tag{56}$$

where r_i represents a predicted cluster label, l_i refers to its original class and n is viewed as the total point number of the dataset. $\delta(x,\ y)$ defines a delta function that equals 1 when $x = y$ and 0 otherwise. In addition, $map(r_i)$ is considered a permutation mapping function that links each r_i to its true label from the dataset.
2. **Purity** measures whether data points from one cluster belong to the primarily class. From the perspective of cluster analysis, purity evaluates the cluster quality of methods as an external evaluation criterion. It refers to the percentage of data samples that are classified properly in the range of 0 and 1. Thus, the purity of a clustering task is defined as follows:

$$Purity = \sum_{i=1}^{K} \frac{n_i}{n} P(S_i), \quad P(S_i) = \frac{1}{n_i} \max_{j}(n_i^j), \tag{57}$$

where S_i refers to the cluster size of n_i, and n_i^j denotes the number of the ith cluster assigned to the jth cluster. K represents the total cluster number in this dataset, and n is considered the total number of all data points.

Both metrics are generally positively correlated. Therefore, a large value corresponds to satisfactory clustering performance.

5.3. Compared algorithms

We compare the proposed method with the following clustering algorithms to illustrate the efficiency of DMSMF intuitionally. For readers convenience, we introduce the algorithms briefly.

1. **Stack K-means**: Known as a classic single-view clustering algorithm, all feature vectors are stacked together for further K-means
2. **ConNMF**: NMF [51] is also a single-view method, and we concatenate all the feature representations as one view to run the basic NMF algorithm.
3. **ConRMNMF**: RMNMF [37] is a single-view NMF-based clustering algorithm that considers outliers and introduces geometric information. We run this algorithm on the concatenated feature representation.
4. **Co-trainSC**: Co-trainSC [11] adopts a cotraining framework for SC and is a parameter-free method used for multiview tasks.
5. **Co-regSC**: Co-regSC [6] regularizes the spectral embeddings of every view to ensure their similarity and adopts a pairwise co-regularization framework. Co-regSC performs K-means based on a unified embedding to achieve the final results. Thus, Co-regSC is a centroid co-regularization SC method.
6. **MultiNMF**: MultiNMF [52] regularizes the coefficient matrices explored from various views to the same degree. It can effectively dig the latent structure embedded in multiple views.
7. **MLAN**: MLAN [53] is a parameter-free algorithm that performs clustering while learning its local structure. We highlight its ability to allocate an optimal weight for each view automatically and avoid unnecessary weight and parameters.

The number of clusters k is acknowledged as given information in the aforementioned algorithms.

5.4. Experimental setting

In the experiments, some algorithms perform K-means after processing multiview features or are sensitive to the initialization.

Table 2
Clustering results of ACC in percentage.

Datasets	Cora	Citeseer	Wiki	Digits	USPS
Stack	20.90	26.36	52.02	51.46	64.87
ConNMF	17.04	21.43	53.98	50.03	55.53
ConRMNMF	21.43	30.07	54.29	54.44	65.60
Co-trainSC	18.91	20.20	52.58	53.65	67.67
Co-regSC	22.38	20.86	51.26	63.80	69.00
MultiNMF	25.78	19.87	51.87	70.95	18.67
MLAN	25.55	35.99	16.68	97.30	68.33
DMSMF	**26.18**	**36.26**	**56.80**	**97.55**	**70.00**

Table 3
Clustering results of Purity in percentage.

Datasets	Cora	Citeseer	Wiki	Digits	USPS
Stack	26.11	36.48	58.16	56.86	70.80
ConNMF	26.14	36.49	**61.06**	51.26	61.53
ConRMNMF	26.18	36.47	59.25	57.91	71.13
Co-trainSC	26.11	36.47	59.63	59.25	73.00
Co-regSC	26.26	36.47	58.09	68.55	76.00
MultiNMF	26.29	36.47	54.95	70.95	19.67
MLAN	26.14	36.50	19.09	97.30	76.67
DMSMF	**26.29**	**36.56**	61.03	**97.55**	**78.67**

For all the methods, we repeat each algorithm 10 times with random initialization then calculate their average results. For K-means, we apply a fast MATLAB method [54]. For RMNMF, the graph is constructed by finding the five nearest neighbors, in which each edge is weighted from 0 to 1. All the initial values of μ and Λ are set empirically, because they slightly influence clustering performance. For our proposed algorithm, the convergence condition is set empirically to 10^{-4}. In other words, we check the difference between the updated and old variables in each iteration. The algorithm will stop the iterations when the difference is smaller than 10^{-4}. Given the rapid convergence ability of our proposed algorithm, the iteration number can be set to 50 to rapidly evaluate its performance.

To compare the methods fairly, we select several parameter combinations for the experiments then choose the best results for comparison. For Co-trainSC, Co-regSC, and MultiNMF, we perform experiments as the default set. For RMNMF and the proposed DMSMF method, we set the regularization parameters by searching the grid of $\{10^{-5}, 10^{-4}, \dots, 10^{4}, 10^{5}\}$ and the iteration number to 100. MLAN is a parameter-free method, and its cluster number is the only parameter given as the input.

As a matrix factorization-based approach, we need to initialize matrix **G** and **W. G** is a cluster indicator matrix; thus, we initialize it with certain existing methods, such as K-means. **W** is a projection matrix; thus, we initialize it randomly.

5.5. Clustering results

Tables 2 and 3 present the ACC and purity of several compared clustering algorithms on five datasets. The highest ACC and purity of each dataset are presented in bold. The two tables indicate that except for the purity of the Wiki dataset, our method consistently achieves the highest ACC and purity. Compared with other methods, our method can effectively offer extra information from multiview data. The following detailed points are derived from the results:

1. Compared with methods using manifold regularization terms, such as RMNMF, the proposed algorithm demonstrates better performance ithe ACC and purity measurement. Therefore, our method can fully capture the local geometrical structure hidden in high-dimensional space. Specifically, the DMSMF method performs well in exploring the intrinsic discriminative data structure.
2. Our DMSMF method consistently outperforms the SC-based algorithms, namely, Co-trainSC and Co-regSC, in the ACC and purity measurement on the five datasets. Although these SC methods make eigenvectors agree on all views, they lack prior knowledge of each view s weight. Considering the weight set for each view in our algorithm is suitable and convinced by the experimental results.
3. Compared with the multiview algorithms, the single-view clustering methods that concatenate only feature representations, such as ConNMF and ConRMNMF, perform poorly in the ACC and purity measurement on most of the datasets. Therefore, merely concatenating all features is impractical. Each view cannot be considered equal owing to its complementary and redundant information.
4. Our algorithm outperforms the other NMF-based multiview algorithms, such as MultiNMF, especially on the Digits and USPS datasets. The two multiview manifold regularization terms demonstrate their effectiveness owing to the intrinsic geometry structure captured in the discriminative subspace and by utilizing pseudo-information for iterative optimization.
5. Our method is superior to MLAN. The Wiki dataset presents an interesting case. Notably, compared with the other datasets, the Wiki dataset is small, with only two views. Under such a condition, ConNMF can achieve close performance by merely concatenating them together. MLAN performs very poorly in the ACC and purity measurement. According to the Introduction, the dimensional numbers of two views are relatively low. This phenomenon may be due to the idea of adaptive neighbors. When the dimension of one dataset is much lower than the sample number, the origin graph is unsuitable.

5.6. Parameter sensitivity

The aforementioned DMSMF method contains two manifold regularization parameters, namely, λ_1 and λ_2, which determine the weight of the two manifold regularization terms. When one parameter changes, the clustering result may change. Learning whether the model is sensitive and finding an optimal parameter combination are challenging and significant tasks. Thus, to explore the effect of λ_1 and λ_2 on the algorithm performance, we tune the parameters to the same range of $\{10^{-5}, 10^{-4}, \dots, 10^{4}, 10^{5}\}$ and demonstrate the ACC of the DMSMF method via 3D visualization, as shown in Fig. 2.

Fig. 2 shows that the performance of our method deteriorates as the scale of λ_1 increases. The two manifold regularization parameters affect the final clustering performance substantially. Thus, the proposed model is sensitive to the parameters and should be tuned carefully. An important future objective is to seek a wide range of parameters and develop a regularization parameter setting rule to obtain the optimal parameter combination.

5.7. Convergence study

The convergence study is also described in this section. Problem (22) can be decomposed into seven sub-problems, which are reduced to closed-form problems. Theoretically, the objective function will decrease by updating each variable iteratively in the seven sub-problems and finally converge to a local optimal value owing to its proven convergence.

The iteration number and corresponding objective function value are shown in Fig. 3 as a convergence curve. The X-axis denotes the number of iterations performed by the DMSMF method on the datasets, and the Y-axis is the objective function value in

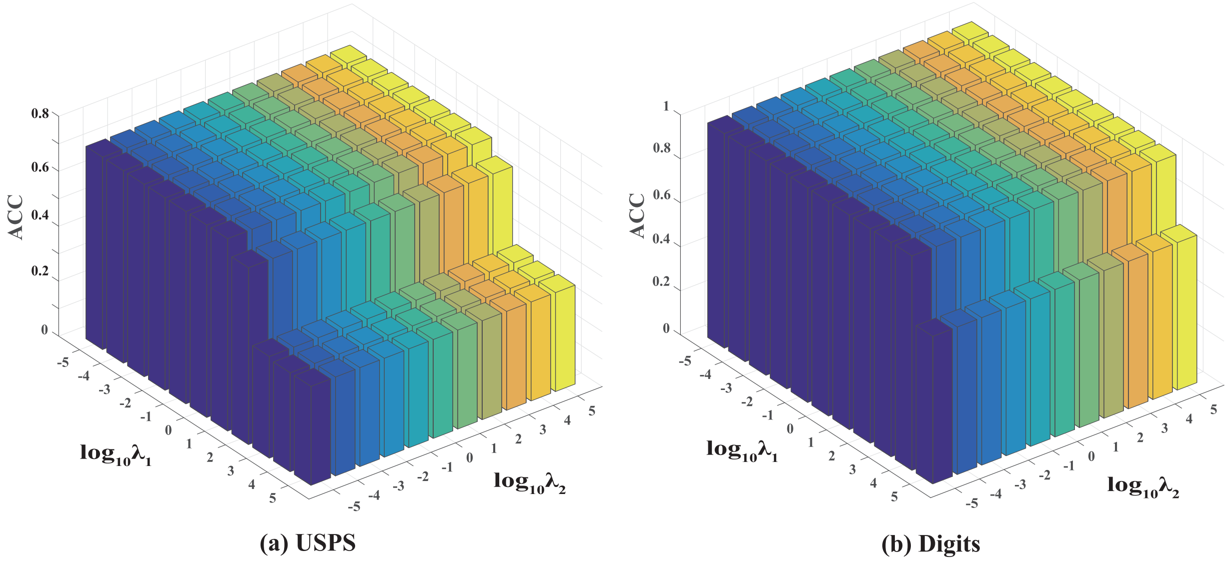

Fig. 2. Clustering Accuracy (ACC) of DMSMF on (a) USPS and (b) Digits with varying λ_1 and λ_2.

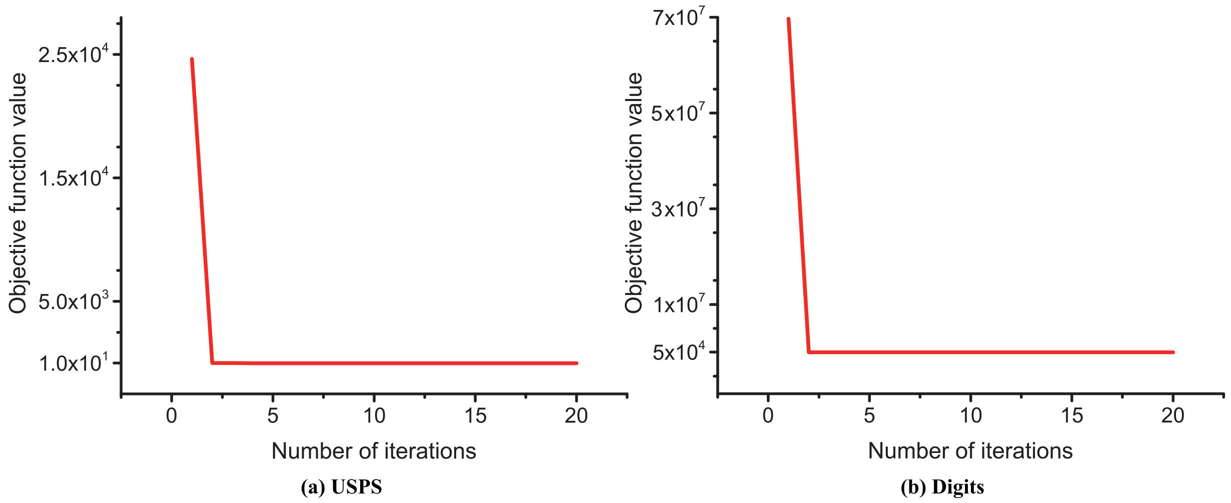

Fig. 3. Convergence curves on (a) USPS and (b) digits.

each iteration. The red solid lines remain flat after several iterations. This finding indicates that the proposed DMSMF method converges after only a few iterations. The convergence study confirms that the proposed DMSMF method is effective and efficient and has satisfactory convergence ability.

6. Conclusion

In this study, a method called DMSMF is proposed for multiview data clustering. In contrast to other manifold regularized clustering methods, our method utilizes pseudo-information to optimize several sub-problems via iterations. The intrinsic geometry structure can be captured in the discriminative subspace by the DMSMF method. The Frobenius norm is changed to the $l_{2,1}$-norm in our model to avoid outliers and enhance robustness. The proposed DMSMF method is proven to be satisfactorily optimized by the suggested ALM-based method. We conduct adequate experiments on multiple benchmark datasets and illustrate that our model is superior to other comparable multiview data clustering algorithms. In future research, we can fit our model for increasingly challenging tasks, such as feature embedding. Given the sensitivity of the model, we must develop a rule for selecting parameters to obtain the optimal combination.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under grants 62076188 and 61771349 and in part by the Science and Technology Major Project of Hubei Province (Next-Generation AI Technologies) under Grant 2019AEA170. The numerical calculations in this paper had been supported by the supercomputing system in the Supercomputing Center of Wuhan University.

References

[1] Y. Meng, R. Shang, F. Shang, L. Jiao, S. Yang, R. Stolkin, Semi-supervised graph regularized deep NMF with bi-orthogonal constraints for data representation, IEEE Trans. Neural Netw. Learn. Syst. PP (2019) 1–14.

[2] C.H.Q. Ding, T. Li, M.I. Jordan, Convex and semi-nonnegative matrix factorizations, IEEE Trans. Pattern Anal. Mach. Intell. 32 (1) (2010) 45–55.
[3] F.R. Chung, F.C. Graham, Spectral Graph Theory, American Mathematical Soc., 1997.
[4] S. Bickel, T. Scheffer, Multi-view clustering, in: Proc. Int. Conf. Data Min., 2004, pp. 19–26.
[5] L. Song, C. Wang, L. Zhang, B. Du, Q. Zhang, C. Huang, X. Wang, Unsupervised domain adaptive re-identification: theory and practice, Pattern Recognit. 102 (2020) 107173.
[6] A. Kumar, P. Rai, H.D. III, Co-regularized multi-view spectral clustering, in: Proc. Adv. Neural Inf. Process. Syst., 2011, pp. 1413–1421.
[7] K. Chaudhuri, S.M. Kakade, K. Livescu, K. Sridharan, Multi-view clustering via canonical correlation analysis, in: Proc. Int. Conf. Mach. Learn., 2009, pp. 129–136.
[8] E. Bruno, S. Marchand-Maillet, Multiview clustering: a late fusion approach using latent models, in: Proc. ACM Int. Conf. Res. Develop. Inf. Retr., 2009, pp. 736–737.
[9] X. Wang, B. Qian, I. Davidson, Improving document clustering using automated machine translation, in: Proc. ACM Int. Conf. Inf. Knowl. Manag., 2012, pp. 645–653.
[10] X. Cai, F. Nie, H. Huang, F. Kamangar, Heterogeneous image feature integration via multi-modal spectral clustering, in: Proc. IEEE Conf. Comput. Vis. Pattern Recog., 2011, pp. 1977–1984.
[11] A. Kumar, H.D. III, A co-training approach for multi-view spectral clustering, in: Proc. Int. Conf. Mach. Learn., 2011, pp. 393–400.
[12] C. Wang, J. Lai, P.S. Yu, Multi-view clustering based on belief propagation, IEEE Trans. Knowl. Data Eng. 28 (4) (2016) 1007–1021.
[13] D. Tolic, N. Antulov-Fantulin, I. Kopriva, A nonlinear orthogonal non-negative matrix factorization approach to subspace clustering, Pattern Recognit. 82 (2018) 40–55.
[14] Y. Lu, Z. Lai, X. Yong, X. Li, D. Zhang, C. Yuan, Nonnegative discriminant matrix factorization, IEEE Trans. Circuits Syst. Video Tech. 27 (7) (2017) 1392–1405.
[15] J. Ma, Y. Zhang, L. Zhang, B. Du, D. Tao, Pseudo supervised matrix factorization in discriminative subspace, in: Proc. Int. Joint Conf. Artif. Intell., 2019, pp. 4554–4560.
[16] B.-Y. Liu, L. Huang, C.-D. Wang, S. FAN, P. Yu, Adaptively weighted multiview proximity learning for clustering, IEEE Trans. Cybern. PP (2019).
[17] X. Wang, G. Yu, C. Domeniconi, J. Wang, Z. Yu, Z. Zhang, Multiple co-clusterings, in: Proc. Int. Conf. Data Min., 2018, pp. 1308–1313.
[18] F. Nie, J. Li, X. Li, Self-weighted multiview clustering with multiple graphs, in: Proc. Int. Joint Conf. Artif. Intell., 2017, pp. 2564–2570.
[19] C. Zhang, Q. Hu, H. Fu, P. Zhu, X. Cao, Latent multi-view subspace clustering, in: Proc. IEEE Conf. Comput. Vis. Pattern Recog., 2017, pp. 4333–4341.
[20] F. Nie, S. Shi, X. Li, Auto-weighted multi-view co-clustering via fast matrix factorization, Pattern Recognit. 102 (2020) 107207.
[21] S. Huang, Z. Kang, Z. Xu, Auto-weighted multi-view clustering via deep matrix decomposition, Pattern Recognit. 97 (2019) 107015.
[22] K. Zhan, C. Niu, C. Chen, F. Nie, C. Zhang, Y. Yang, Graph structure fusion for multiview clustering, IEEE Trans. Knowl. Data Eng. 31 (10) (2019) 1984–1993.
[23] X. Zhu, S. Zhang, W. He, R. Hu, C. Lei, P. Zhu, One-step multi-view spectral clustering, IEEE Trans. Knowl. Data Eng. 31 (10) (2019) 2022–2034.
[24] S. Hu, X. Yan, Y. Ye, Dynamic auto-weighted multi-view co-clustering, Pattern Recognit. 99 (2019) 107101.
[25] S. Huang, Z. Kang, I.W. Tsang, Z. Xu, Auto-weighted multi-view clustering via kernelized graph learning, Pattern Recognit. 88 (2019) 174–184.
[26] X. Liu, X. Zhu, M. Li, L. Wang, C. Tang, J. Yin, D. Shen, H. Wang, W. Gao, Late fusion incomplete multi-view clustering, IEEE Trans. Pattern Anal. Mach. Intell. 41 (10) (2019) 2410–2423.
[27] L. Huang, H. Chao, C. Wang, Multi-view intact space clustering, Pattern Recognit. 86 (2019) 344–353.
[28] R. Shang, Y. Meng, W. Wang, F. Shang, L. Jiao, Local discriminative based sparse subspace learning for feature selection, Pattern Recognit. 92 (2019) 219–230.
[29] R. Shang, Z. Zhang, L. Jiao, C. Liu, Y. Li, Self-representation based dual–graph regularized feature selection clustering, Neurocomputing 171 (2016) 1242–1253.
[30] N. Wang, S. Ma, J. Li, Y. Zhang, L. Zhang, Multistage attention network for image inpainting, Pattern Recognit. 106 (2020) 107448.
[31] D. Cai, X. He, J. Han, Locally consistent concept factorization for document clustering, IEEE Trans. Knowl. Data Eng. 23 (6) (2011) 902–913.
[32] L. Zhang, Q. Zhang, B. Du, X. Huang, Y.Y. Tang, D. Tao, Simultaneous spectral-spatial feature selection and extraction for hyperspectral images, IEEE Trans. Cybern. 48 (1) (2018) 16–28.
[33] Y. Zhang, Z. Zhang, S. Li, J. Qin, G. Liu, M. Wang, S. Yan, Unsupervised nonnegative adaptive feature extraction for data representation, IEEE Trans. Knowl. Data Eng. 31 (12) (2019) 2423–2440.
[34] B. Wang, Y. Hu, J. Gao, Y. Sun, H. Chen, M. Ali, B. Yin, Locality preserving projections for grassmann manifold, in: Proc. Int. Joint Conf. Artif. Intell., 2017, pp. 2893–2900.
[35] K. Allab, L. Labiod, M. Nadif, Multi-manifold matrix decomposition for data co–clustering, Pattern Recognit. 64 (2017) 386–398.
[36] L. Zhang, Q. Zhang, L. Zhang, D. Tao, X. Huang, B. Du, Ensemble manifold regularized sparse low-rank approximation for multiview feature embedding, Pattern Recognit. 48 (10) (2015) 3102–3112.
[37] J. Huang, F. Nie, H. Huang, C.H.Q. Ding, Robust manifold nonnegative matrix factorization, ACM Trans. Knowl. Discov. Data. 8 (3) (2013) 11:1–11:21.
[38] Y. Yang, H.T. Shen, Z. Ma, Z. Huang, X. Zhou, L2,1-norm regularized discriminative feature selection for unsupervised learning, in: Proc. Int. Joint Conf. Artif. Intell., 2011, pp. 1589–1594.
[39] T. Zhang, D. Tao, X. Li, Y. Jie, Patch alignment for dimensionality reduction, IEEE Trans. Knowl. Data Eng. 21 (9) (2009) 1299–1313.
[40] Z. Zhang, H. Zha, Principal manifolds and nonlinear dimensionality reduction via tangent space alignment, J. Shanghai Univ. 8 (4) (2004) 406–424.
[41] M. Wang, X. Hua, X. Yuan, Y. Song, L. Dai, Optimizing multi-graph learning: towards a unified video annotation scheme, in: Proc. ACM Int. Conf. Multimedia., 2007, pp. 862–871.
[42] B. Geng, D. Tao, C. Xu, L. Yang, X. Hua, Ensemble manifold regularization, IEEE Trans. Pattern Anal. Mach. Intell. 34 (6) (2012) 1227–1233.
[43] J. Gui, D. Tao, Z. Sun, Y. Luo, X. You, Y.Y. Tang, Group sparse multiview patch alignment framework with view consistency for image classification, IEEE Trans. Image Process. 23 (7) (2014) 3126–3137.
[44] F. Nie, X. Wang, H. Huang, Clustering and projected clustering with adaptive neighbors, in: Proc. ACM SIGKDD, 2014, pp. 977–986.
[45] M. Belkin, Problems of Learning on Manifolds, Springer Berlin Heidelberg, 2003.
[46] F. Nie, H. Wang, H. Huang, C. Ding, Joint schatten p-norm and ℓ_p-norm robust matrix completion for missing value recovery, Knowl. Inf. Syst. 42 (3) (2015) 525–544.
[47] U. von Luxburg, A tutorial on spectral clustering, Stat. Comput. 17 (4) (2007) 395–416.
[48] B. Xie, Y. Mu, D. Tao, K. Huang, M-SNE: multiview stochastic neighbor embedding, IEEE Trans. Syst. Man Cybern. Part B 41 (4) (2011) 1088–1096.
[49] J.J. Hull, A database for handwritten text recognition research, IEEE Trans. Pattern Anal. Mach. Intell. 16 (5) (1994) 550–554.
[50] Y. Liu, Q. Gao, Z. Yang, S. Wang, Learning with adaptive neighbors for image clustering, in: Proc. Int. Joint Conf. Artif. Intell., 2018, pp. 2483–2489.
[51] D.D. Lee, H.S. Seung, Algorithms for non-negative matrix factorization, in: Proc. Adv. Neural Inf. Process. Syst., 2000, pp. 556–562.
[52] J. Gao, J. Han, J. Liu, C. Wang, Multi-view clustering via joint nonnegative matrix factorization, in: Proc. SIAM Int. Conf. Data Min., 2013, pp. 252–260.
[53] F. Nie, G. Cai, X. Li, Multi-view clustering and semi-supervised classification with adaptive neighbours, in: Proc. AAAI Conf. Artif. Intell., 2017, pp. 2408–2414.
[54] D. Cai, Litekmeans: the fastest matlab implementation of kmeans, Available at:http://www.zjucadcg.cn/dengcai/Data/Clustering.html (2011).

Jiaqi Ma, received the B.S. degree in School of Information Engineering, Zhengzhou University, China, in 2018. He is currently pursuing the M.S. degree in the School of Computer Science, Wuhan University, China. His research interests include machine learning and pattern recognition.

Yipeng Zhang received the B.S. degree in electronic engineering from the School of Electronic Information, Wuhan University, Wuhan, China, and the master's degree in electrical engineering from Syracuse University, Syracuse, NY, USA. He is currently pursuing the Ph.D. degree in the School of Computer Science, Wuhan University. His current research interests include deep learning, architecture-optimization on the FPGA, as well as the machine learning-oriented processor and accelerator.

Lefei Zhang, received the B.S. and Ph.D. degrees from Wuhan University, Wuhan, China, in 2008 and 2013, respectively. He is currently a professor with the School of Computer Science, Wuhan University. His research interests include pattern recognition, image processing, and remote sensing. Dr. Zhang is a reviewer of more than 30 international journals, including the IEEE TPAMI, TIP and TGRS.