

Embedded Feature Selection on Graph-Based Multi-View Clustering

Wenhui Zhao¹, Guangfei Li¹, Haizhou Yang¹, Quanxue Gao^{1*}, Qianqian Wang¹

¹ School of Telecommunication Engineering, Xidian University, Shaanxi 710071, China

whzhao@stu.xidian.edu.cn, liguangfei_dream@hotmail.com, leoyhz@qq.com, qxgao@xidian.edu.cn, qqwang@xidian.edu.cn

Abstract

Recently, anchor graph-based multi-view clustering has been proven to be highly efficient for large-scale data processing. However, most existing anchor graph-based clustering methods necessitate post-processing to obtain clustering labels and are unable to effectively utilize the information within anchor graphs. To solve these problems, we propose an Embedded Feature Selection on Graph-Based Multi-View Clustering (EFS-GMC) approach to improve the clustering performance. Our method decomposes anchor graphs, taking advantage of memory efficiency, to obtain clustering labels in a single step without the need for post-processing. Furthermore, we introduce the $\ell_{2,p}$ -norm for graph-based feature selection, which selects the most relevant data for efficient graph factorization. Lastly, we employ the tensor Schatten p -norm as a tensor rank approximation function to capture the complementary information between different views, ensuring similarity between cluster assignment matrices. Experimental results on five real-world datasets demonstrate that our proposed method outperforms state-of-the-art approaches.

Introduction

Over the past few decades, there has been immense interest in developing numerous exceptional clustering algorithms, including subspace-based clustering (Luo et al. 2018; Xie et al. 2020), non-negative matrix factorization clustering (Gao et al. 2013; Salah, Ailem, and Nadif 2018), and graph-based clustering (Hu et al. 2020; Nie, Li, and Li 2017). Notably, graph-based clustering methods have been widely developed due to their excellent performance in capturing the spatial structure of nonlinear data.

The key step in graph-based clustering methods is to construct an $N \times N$ affinity graph matrix to represent the similarity between different N data points. However, this operation can be time-consuming and memory-intensive. To address this issue, anchor graph-based methods (Li et al. 2020) have been proposed to construct an $N \times M$ ($M \ll N$) anchor graph, where anchor graphs are used to measure the relationship between N data points and M anchors. However, post-processing (e.g., K -means) is required in most anchor graph-based methods to obtain final clustering labels, which not only increases the computational time but

also leads to the clustering performance being limited by K -means. To this end, SFMC (Li et al. 2020) manipulates the joint graph by a connectivity constraint, so that the connected components can indicate clusters directly. MSC-BG (Yang et al. 2022) proposed imposing constraints on the rank of the Laplacian matrix to obtain an affinity graph matrix with K connected components. Nevertheless, imposing constraints on connected components may result in a smaller number of connected components than K , leading to a significant decrease in clustering performance.

Moreover, most anchor graph-based clustering algorithms take advantage of all the data points, but the anchor points corresponding to noise and redundant data in data points are useless. Therefore, LAPIN (Nie et al. 2023) method obtains a better coefficient matrix by applying sparse constraints to the data matrix, thus to alleviate the impact of noise to some extent. However, the distribution of noise is difficult to estimate, and the sparse representation of the noise term is hard to guarantee. Furthermore, the quality differences between different data views can also significantly affect clustering performance. Accordingly, AMGL (Nie, Li, and Li 2016) automatically learns optimal weights for each view by minimizing the squared-root trace. Although these methods have achieved good results, they cannot fully utilize the complementary information in the adjacency matrix of different views.

To address these issues, we propose an Embedded Feature Selection on Graph-Based Multi-View Clustering (EFS-GMC) method, which can obtain the final cluster label in one step. Specifically, we adapt non-negative matrix factorization directly to the anchor graph to get the final cluster indicator matrix in one step, thus avoiding post-processing. Besides, we draw inspiration from feature selection for raw data points and apply feature selection to the anchor graph. Specifically, we minimize the $\ell_{2,p}$ -norm to make the learned anchor map representation more sparse to filter out the anchor points corresponding to noise and redundant data, which significantly reduces the effect of noise. In addition, we refer to the weighted tensor Schatten p -norm minimization (WTSNM) (Xia et al. 2022) and propose to employ the tensor Schatten p -norm minimization to explore the low-rank structure embedded in inter-view graphs. The main contributions of our method are as follows:

- Our method performs non-negative matrix decomposi-

*Corresponding author

Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

tion of the learned anchor graphs to obtain a discrete label matrix, allowing us to obtain clustering results directly in one step without the need for post-processing.

- We propose a method that minimizes the $\ell_{2,p}$ -norm to ensure the sparsity of the learned anchor graph, thereby achieving the selection of representative anchor points while eliminating the redundant ones and present a novel and efficient algorithm with a closed-form solution.
- We employ LPP (Lu et al. 2016) manifold learning to ensure label consistency among adjacent sample points, and explore the low-rank structure of inter-view graphs using Schatten p -norm, which fully leverages the complementary information embedded in the graphs.
- We propose an efficient algorithm to solve the model via ALM, and we carry out experiments on real multi-view datasets to demonstrate the effectiveness of our proposed method.

Methodology

Notations and Definitions: In this paper, we use bold calligraphy letters for 3rd-order tensors, e.g., $\mathcal{A} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$, and bold upper case letters for matrices, e.g., \mathbf{A} . $\mathbf{A}_{i:}$ and $\mathbf{A}_{:j}$ are the i -th row and j -th column of matrix \mathbf{A} , separately. The v -th frontal slice of \mathcal{A} is \mathcal{A}_v . $\overline{\mathcal{A}}$ is the discrete Fast Fourier Transform (FFT) of \mathcal{A} along the third dimension, i.e., $\overline{\mathcal{A}} = \text{fft}(\mathcal{A}, :, 3)$. The trace of matrix \mathbf{A} is denoted by $\text{tr}(\mathbf{A})$. \mathbf{I} is an identity matrix.

Definition 1 (Gao et al. 2021) Given $\mathcal{G} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$, the tensor Schatten p -norm of \mathcal{G} is defined as

$$\|\mathcal{G}\|_{\mathbb{S}^p} = \left(\sum_{i=1}^{n_3} \|\overline{\mathcal{G}}_i\|_{\mathbb{S}^p}^p \right)^{\frac{1}{p}} = \left(\sum_{i=1}^{n_3} \sum_{j=1}^h \sigma_j(\overline{\mathcal{G}}_i)^p \right)^{\frac{1}{p}} \quad (1)$$

where $h = \min(n_1, n_2)$, $p \in (0, 1]$, $\sigma_j(\overline{\mathcal{G}}_i)$ is the j -th singular value of $\overline{\mathcal{G}}_i$. The Schatten p -norm can approximate the rank function more tightly when p is chosen appropriately.

Definition 2 (Wang et al. 2018; Liao et al. 2018) Given $\mathbf{H} \in \mathbb{R}^{n_1 \times n_2}$, the $\ell_{2,p}$ -norm is defined as

$$\|\mathbf{H}\|_{2,p} = \sum_{i=1}^{n_1} \|\mathbf{H}_{i:}\|_2^p = \sum_{i=1}^{n_1} \left(\sum_{j=1}^{n_2} \mathbf{H}_{ij}^2 \right)^{\frac{p}{2}} \quad (2)$$

where $p \in (0, 1]$. Specially, when $p = 1$, $\ell_{2,p}$ -norm becomes $\ell_{2,1}$ -norm, i.e., $\|\mathbf{H}\|_{2,1} = \sum_{i=1}^{n_1} \sqrt{\sum_{j=1}^{n_2} \mathbf{H}_{ij}^2}$.

Definition 3 (Dong et al. 2016) Given $\mathbf{Z} \in \mathbb{R}^{N \times M}$, weight matrix $\mathbf{W} \in \mathbb{R}^{N \times N}$, the Sparse Gradient Pursuit is defined as

$$\|\nabla \mathbf{Z}\|_{\ell_1} = \sum_{i=1}^N \sum_{j=1}^N \mathbf{W}_{ij} \|\mathbf{Z}_{:i} - \mathbf{Z}_{:j}\|_1 = \|\mathbf{KZ}\|_1 \quad (3)$$

where $\nabla \mathbf{Z}$ represents the gradient of \mathbf{Z} and \mathbf{K} denotes the gradient matrix of the adjacency KNN graph (Yang et al. 2014).

Problem Formulation and Objective

Anchor graph-based methods typically require learning a shared graph using predefined graphs $\mathbf{S}^v \in \mathbb{R}^{N \times M}$, which capture the relationships between N data points and M anchor points. However, there are some drawbacks to these methods: (1) The cluster labels must be obtained via post-processing, which can limit clustering performance; (2) All data points in the anchor graphs are used, which can introduce redundant data and lead to inefficiencies; (3) These methods process each view separately, which prevents them from fully leveraging the complementary information in the adjacency matrices of different views.

In response to the above-mentioned disadvantages, we use non-negative matrix factorization (Ding et al. 2006) to obtain the final global cluster assignment matrix by factorizing the anchor graph in one step, thus avoiding post-processing. To ensure that results before and after matrix factorization are close, $\ell_{2,1}$ -norm is used for non-negative matrix factorization to avoid the increasing error caused by the square of F -norm. Thus, we have

$$\begin{aligned} \min_{\mathbf{G}^v, \mathbf{H}^v} & \sum_{v=1}^V \frac{1}{\alpha^v} \left\{ \|\mathbf{S}^v - \mathbf{G}^v \mathbf{H}^{v\top}\|_{2,1} \right\} \\ \text{s.t. } & \mathbf{G}^{v\top} \mathbf{G}^v = \mathbf{I}, \mathbf{G} \geq \mathbf{0}, \sum_{v=1}^V \alpha^v = 1, \alpha^v \geq 0 \end{aligned} \quad (4)$$

where α^v is the non-negative normalized weight factor, $\mathbf{S}^v \in \mathbb{R}^{N \times M}$ is pre-defined anchor graph (Li et al. 2020), $\mathbf{G}^v \in \mathbb{R}^{N \times C}$ is the cluster assignment matrix, $\mathbf{H}^v \in \mathbb{R}^{M \times C}$ is the latent feature matrix, C is the number of clusters.

To better construct anchor points, we consider selecting the most representative data points from the anchor map. When reconstructing the anchor graph \mathbf{S}^v , we can add a diagonal matrix $\text{diag}(\mathbf{f})$ with $\mathbf{f}_i = \{0, 1\}$ corresponding to the matrix \mathbf{S}^v for feature selection. At this point, the reconstruction matrix corresponding to anchor graph \mathbf{S}^v is $\tilde{\mathbf{S}}^v = \mathbf{S}^v \text{diag}(\mathbf{f})$. We observe that the matrix \mathbf{S}^v can be reconstructed by \mathbf{G}^v and \mathbf{H}^v , thus the i -th column vectors of the reconstruction matrix $\tilde{\mathbf{S}}^v$ can be represented as $\tilde{\mathbf{S}}_{:i}^v = \mathbf{G}^v \mathbf{H}_{i:}^{v\top}$. Considering $\|\tilde{\mathbf{S}}_{:i}^v\|_2 = \|\mathbf{G}^v \mathbf{H}_{i:}^{v\top}\|_2 = \|\mathbf{H}_{i:}^v\|_2$, we can see that the reconstruction for \mathbf{S}^v is heavily dependent on the matrix \mathbf{H}^v , and when $\|\tilde{\mathbf{S}}_{:i}^v\|_2$ is close to 0, it means that the corresponding anchor point is not representative and should be excluded. Therefore, ensuring the row sparsity of \mathbf{H}^v can achieve feature selection on the graph easily. The corresponding model in this case is:

$$\begin{aligned} \min_{\mathbf{G}^v, \mathbf{H}^v} & \sum_{v=1}^V \frac{1}{\alpha^v} \left\{ \|\mathbf{S}^v - \mathbf{G}^v (\text{diag}(\mathbf{f}) \mathbf{H}^v)^\top\|_{2,1} \right\} \\ \text{s.t. } & \mathbf{G}^{v\top} \mathbf{G}^v = \mathbf{I}, \mathbf{G} \geq \mathbf{0}, \sum_{v=1}^V \alpha^v = 1, \alpha^v \geq 0, \mathbf{f} \in \{0, 1\}^M \end{aligned} \quad (5)$$

where $\text{diag}(\mathbf{f}) \in \mathbb{R}^{M \times M}$ with $\mathbf{f}_i = \{0, 1\}$. When $\mathbf{f}_i = 0$, the corresponding i -th row of $\text{diag}(\mathbf{f}) \mathbf{H}^v$ is $\mathbf{0}^\top$, and the i -th column of the reconstructed anchor graph $\tilde{\mathbf{S}}^v$ also tends toward

0. Otherwise, when $\mathbf{f}_i = 1$, it indicates that the feature associated with graph \mathbf{S}^v is useful and should be retained. This allows for efficient feature selection on the anchor graph \mathbf{S}^v . However, directly imposing a constraint on a specific row in matrix \mathbf{H}^v to be $\mathbf{0}^T$ is too strict and difficult to solve. Therefore, we propose a new row sparsity norm, termed as the $\ell_{2,p}$ -norm (see Definition 2). By using this norm, the resulting \mathbf{H}^v matrix can be made even sparser, leading to further enhancement of the performance of the algorithm. Besides, using the $\ell_{2,p}$ -norm constraint for non-negative matrix factorization can reduce the reconstruction error. Therefore, we can obtain:

$$\begin{aligned} & \min_{\mathbf{G}^v, \mathbf{H}^v} \sum_{v=1}^V \frac{1}{\alpha^v} \left\{ \|\mathbf{S}^v - \mathbf{G}^v \mathbf{H}^{vT}\|_{2,p} + \lambda \|\mathbf{H}^v\|_{2,p} \right\} \\ \text{s.t. } & \mathbf{G}^{vT} \mathbf{G}^v = \mathbf{I}, \mathbf{G}^v \geq \mathbf{0}, \sum_{v=1}^V \alpha^v = 1, \alpha^v \geq 0 \end{aligned} \quad (6)$$

In equation (6), the sparsity of the rows in matrix \mathbf{H}^v controls the column sparsity of the anchor graph \mathbf{S}^v , thereby enabling the \mathbf{H}^v matrix to realize feature selection on the anchor graph. Furthermore, the matrix \mathbf{G}^v serves as the corresponding label embedding matrix. To effectively learn the feature selection matrix \mathbf{H}^v , it is necessary to ensure that adjacent samples in the high-dimensional manifold remain adjacent after dimension reduction. Inspired by the local preserving projection (LPP) (Lu et al. 2016) algorithm, we consider adding a regularization term to the matrix \mathbf{G}^v to preserve the label consistency between adjacent sample points by employing the idea of LPP manifold learning. This leads to our new model formulation:

$$\begin{aligned} & \min_{\mathbf{G}^v, \mathbf{H}^v} \sum_{v=1}^V \frac{1}{\alpha^v} \left\{ \|\mathbf{S}^v - \mathbf{G}^v \mathbf{H}^{vT}\|_{2,p} \right. \\ & \quad \left. + \gamma \text{tr}(\mathbf{G}^{vT} \tilde{\mathbf{L}}^v \mathbf{G}^v) + \lambda \|\mathbf{H}^v\|_{2,p} \right\} \\ \text{s.t. } & \mathbf{G}^{vT} \mathbf{G}^v = \mathbf{I}, \mathbf{G}^v \geq \mathbf{0}, \sum_{v=1}^V \alpha^v = 1, \alpha^v \geq 0 \end{aligned} \quad (7)$$

where the normalized Laplacian matrix $\tilde{\mathbf{L}}^v$ can be calculated by $\tilde{\mathbf{L}}^v = \mathbf{I} - \mathbf{S}^v (\Delta^v)^{-1} \mathbf{S}^{vT}$ and the diagonal elements of diagonal matrix Δ^v are $\Delta_{ii}^v = \sum_{i=1}^N \mathbf{S}_{ii}^v$.

During the optimization process of the model, $\text{tr}(\mathbf{G}^{vT} \tilde{\mathbf{L}}^v \mathbf{G}^v)$ needs to be transformed into the square of F -norm for computation, with the corresponding expression being:

$$\text{tr}(\mathbf{G}^{vT} \tilde{\mathbf{L}}^v \mathbf{G}^v) = \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \mathbf{W}_{ij}^v \|\mathbf{G}_{i:}^v - \mathbf{G}_{j:}^v\|_F^2 \quad (8)$$

where \mathbf{W}^v is the adjacency matrix and \mathbf{G}^v is the cluster assignment matrix. Minimizing the ℓ_1 -norm optimization problem tends to set some elements to 0, i.e., only the part of the data that fits well is selected for estimating the matrix \mathbf{G}^v to ensure the sparsity. Therefore, we propose to use

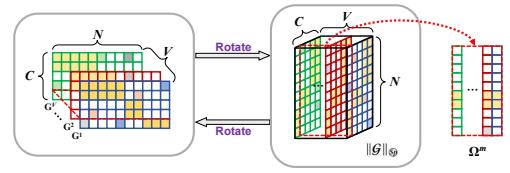


Figure 1: Construction of $\mathbf{G} \in \mathbb{R}^{N \times V \times C}$.

the ℓ_1 -norm instead of the square of the F -norm. Inspired by Definition 3, we convert the row operation of matrix \mathbf{G}^v into the column operation of \mathbf{G}^{vT} to obtain the corresponding expression:

$$\frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \mathbf{W}_{ij}^v \|\mathbf{G}_{i:}^v - \mathbf{G}_{j:}^v\|_1 = \frac{1}{2} \|\mathbf{G}^{vT} \mathbf{T}^{vT}\|_1 \quad (9)$$

where $\mathbf{T}^v \in \mathbb{R}^{O \times N}$ is the corresponding gradient matrix of adjacent K -nearest neighbor (KNN) graph (the k -th row satisfies $\mathbf{T}_{ki}^v = -\mathbf{T}_{kj}^v = \mathbf{W}_{ij}^v$), $O = K * N$ is the number of edges in the KNN graph. Combining (9) with (7), we have:

$$\begin{aligned} & \min_{\mathbf{G}^v, \mathbf{H}^v} \sum_{v=1}^V \frac{1}{\alpha^v} \left\{ \|\mathbf{S}^v - \mathbf{G}^v \mathbf{H}^{vT}\|_{2,p} + \right. \\ & \quad \left. \frac{\gamma}{2} \|\mathbf{G}^{vT} \mathbf{T}^{vT}\|_1 + \lambda \|\mathbf{H}^v\|_{2,p} \right\} \\ \text{s.t. } & \mathbf{G}^{vT} \mathbf{G}^v = \mathbf{I}, \mathbf{G}^v \geq \mathbf{0}, \sum_{v=1}^V \alpha^v = 1, \alpha^v \geq 0 \end{aligned} \quad (10)$$

However, equation (10) does not fully exploit the complementary information in different views. Hence, we use the tensor Schatten p -norm (defined in Definition 1) to measure the similarity between different \mathbf{G}^v and obtain the final global cluster assignment matrix $\mathbf{C} = \sum_{v=1}^V \frac{\mathbf{G}^v}{\alpha^v}$, incorporating weight information. Specifically, we construct a 3rd-order tensor \mathbf{G} from \mathbf{G}^v (as illustrated in Figure 1) and consider the corresponding Schatten p -norm after rotation. Notably, Ω^m ensures that the relationship between the N data points and the c -th cluster is consistent across views. Therefore, $\|\mathbf{G}\|_{\mathfrak{S}}$ allows for a comprehensive exploration of information hidden between different views. Combining (10) with the Schatten p -norm, our final model can be expressed as:

$$\begin{aligned} & \min_{\mathbf{G}^v, \mathbf{H}^v} \sum_{v=1}^V \frac{1}{\alpha^v} \left\{ \|\mathbf{S}^v - \mathbf{G}^v \mathbf{H}^{vT}\|_{2,p} + \right. \\ & \quad \left. \frac{\gamma}{2} \|\mathbf{G}^{vT} \mathbf{T}^{vT}\|_1 + \lambda \|\mathbf{H}^v\|_{2,p} \right\} + \beta \|\mathbf{G}\|_{\mathfrak{S}}^p \\ \text{s.t. } & \mathbf{G}^{vT} \mathbf{G}^v = \mathbf{I}, \mathbf{G}^v \geq \mathbf{0}, \sum_{v=1}^V \alpha^v = 1, \alpha^v \geq 0 \end{aligned} \quad (11)$$

Optimization

We propose an efficient optimization method based on the Augmented Lagrange Multiplier (ALM) method. This

method involves introducing auxiliary variables \mathcal{J} , \mathbf{P}^v , and \mathbf{Q}^v , which allows us to rewrite (11) as:

$$\begin{aligned} & \min \sum_{v=1}^V \frac{1}{\alpha^v} \left\{ \|\mathbf{P}^v\|_{2,p} + \frac{\rho_0}{2} \|\mathbf{S}^v - \mathbf{G}^v \mathbf{H}^{v\top} - \mathbf{P}^v + \frac{\mathbf{K}^v}{\rho_0}\|_F^2 + \right. \\ & \quad \left. \frac{\gamma}{2} \|\mathbf{Q}^v\|_1 + \frac{\rho_1}{2} \|\mathbf{G}^{v\top} \mathbf{T}^{v\top} - \mathbf{Q}^v + \frac{\mathbf{M}^v}{\rho_1}\|_F^2 + \right. \\ & \quad \left. \lambda \|\mathbf{H}^v\|_{2,p} \right\} + \beta \|\mathcal{J}\|_{\mathbb{Q}}^p + \frac{\rho_2}{2} \|\mathcal{G} - \mathcal{J} + \frac{\mathcal{W}}{\rho_2}\|_F^2 \\ \text{s.t. } & \mathbf{G}^{v\top} \mathbf{G}^v = \mathbf{I}, \mathbf{Q}^v \geq \mathbf{0}, \sum_{v=1}^V \alpha^v = 1, \alpha^v \geq 0 \end{aligned} \quad (12)$$

where \mathcal{W} , \mathbf{K}^v and \mathbf{M}^v are Lagrange multiplier, ρ_0 , ρ_1 and ρ_2 are the penalty parameters. The optimization process could be separated into the following steps:

• \mathbf{Q}^v sub-problem:

$$\begin{aligned} & \arg \min_{\mathbf{Q}^v} \sum_{v=1}^V \frac{1}{\alpha^v} \left\{ \frac{\gamma}{2} \|\mathbf{Q}^v\|_1 + \frac{\rho_1}{2} \|\mathbf{G}^{v\top} \mathbf{T}^{v\top} - \mathbf{Q}^v + \frac{\mathbf{M}^v}{\rho_1}\|_F^2 \right\} \\ \text{s.t. } & \mathbf{G}^{v\top} \mathbf{G}^v = \mathbf{I}, \mathbf{Q}^v \geq \mathbf{0}, \sum_{v=1}^V \alpha^v = 1, \alpha^v \geq 0 \end{aligned} \quad (13)$$

Considering every single view individually, it follows that

$$\arg \min_{\mathbf{Q}^v} \frac{\gamma}{2\rho_1} \|\mathbf{Q}^v\|_1 + \frac{1}{2} \|\mathbf{Q}^v - \mathbf{C}^v\|_F^2 \quad \text{s.t. } \mathbf{G}^{v\top} \mathbf{G}^v = \mathbf{I}, \mathbf{Q}^v \geq \mathbf{0} \quad (14)$$

where $\mathbf{C}^v = \mathbf{G}^{v\top} \mathbf{T}^{v\top} + \frac{\mathbf{M}^v}{\rho_1}$. Inspired by (Hale, Yin, and Zhang 2008), we have

$$\mathbf{Q}^{v*} = \Theta_{\frac{\gamma}{2\rho_1}}(\mathbf{C}^v) \quad (15)$$

where the i, j -th element of $\Theta_{\frac{\gamma}{2\rho_1}}(\mathbf{C}^v)$ is defined as

$$\Theta_{\frac{\gamma}{2\rho_1}}(\mathbf{C}^v)_{ij} = \text{sgn}(\mathbf{C}^v_{ij}) \times \max \left(\left| \mathbf{C}^v_{ij} - \frac{\gamma}{2\rho_1} \right|, 0 \right) \quad (16)$$

• \mathbf{H}^v sub-problem:

$$\begin{aligned} & \arg \min_{\mathbf{H}^v} \sum_{v=1}^V \frac{1}{\alpha^v} \left\{ \frac{1}{2} \|\mathbf{A}^v - \mathbf{H}^v\|_F^2 + \frac{\lambda}{\rho_0} \|\mathbf{H}^v\|_{2,p} \right\} \\ \text{s.t. } & \mathbf{G}^{v\top} \mathbf{G}^v = \mathbf{I}, \sum_{v=1}^V \alpha^v = 1, \alpha^v \geq 0 \end{aligned} \quad (17)$$

where $\mathbf{A}^v = \left(\mathbf{S}^v - \mathbf{P}^v + \frac{\mathbf{K}^v}{\rho_0} \right)^\top \mathbf{G}^v$. In order to solve (17), we need the following Lemma 1 2 and Theorem 1.

Lemma 1 (Gao et al. 2021) Considering

$$\min_{\delta \geq 0} f(\delta) = \frac{1}{2} (\delta - \omega)^2 + \lambda \delta^p \quad \text{s.t. } 0 < p < 1 \quad (18)$$

by using the Generalized Soft-Thresholding (GST):

$$\tau_p^{GST}(\lambda) = (2\lambda(1-p))^{\frac{1}{2-p}} + \lambda p(2\lambda(1-p))^{\frac{p-1}{2-p}}, \quad (19)$$

Thus, δ^* can be obtained by

$$\begin{cases} \delta^* = 0, & \omega \leq \tau_p^{GST}(\lambda) \\ \delta^* = \text{sign}(\omega) S_p^{GST}(\omega, \lambda), & \text{otherwise} \end{cases} \quad (20)$$

where $S_p^{GST}(\omega, \lambda)$ can be solved by $S_p^{GST}(\omega, \lambda) - \omega + \lambda p(S_p^{GST}(\omega, \lambda))^{p-1} = 0$.

Lemma 2 (Yang et al. 2020) Considering $\mathbf{Q}, \mathbf{P} \in \mathbb{C}^{N \times M}$, and $F : \mathbb{C}^{N \times M} \rightarrow \mathbb{C}$ is represented as $F(\mathbf{Q}) = f \circ \vec{\sigma}_{\mathbf{Q}} = f(\sigma_1(\mathbf{Q}), \dots, \sigma_K(\mathbf{Q}))$, where $\vec{\sigma}_{\mathbf{Q}}$ is the vector with components of the non-increasing singular values of \mathbf{Q} . If $F(\mathbf{Q})$ is a complex invariant function and consider the SVD $\mathbf{P} = \mathbf{U} \sum \mathbf{V}^H$, then the optimal solution to

$$\min_{\mathbf{Q}} \left\{ \frac{1}{2} \|\mathbf{P} - \mathbf{Q}\|_F^2 + F(\mathbf{Q}) \right\} \quad (21)$$

is $\mathbf{Q}^* = \mathbf{U} \sum_{\mathbf{Q}}^* \mathbf{V}^H$, where $\sum_{\mathbf{Q}}^* = \text{diag}(\vec{\sigma}_{\mathbf{Q}}^*)$ and

$$\vec{\sigma}_{\mathbf{Q}}^* = \arg \min_{\vec{\sigma}} \left\{ \frac{1}{2} \|\vec{\sigma} - \vec{\sigma}_{\mathbf{P}}\|_2^2 + f(\vec{\sigma}) \right\} \quad (22)$$

Theorem 1 Suppose $\mathbf{H} \in \mathbb{R}^{N \times M}$, the solution of

$$\arg \min_{\mathbf{H}} \frac{1}{2} \|\mathbf{A} - \mathbf{H}\|_F^2 + \mu \|\mathbf{H}\|_{2,p} \quad (23)$$

is $\mathbf{H}^* = [\mathbf{H}_{1:}^*; \dots; \mathbf{H}_{N:}^*]^T$, where the i -th row element is

$$\mathbf{H}_{i:}^* = \sigma^* \frac{\mathbf{A}_{i:}}{\|\mathbf{A}_{i:}\|_2} \quad (24)$$

where σ^* can be obtained by

$$\sigma^* = \arg \min_{x \geq 0} \frac{1}{2} (x - \|\mathbf{A}_{i:}\|_2)^2 + \mu x^p \quad (25)$$

which can be obtained by the General Shrinkage Thresholding (GST) algorithm(Gao et al. 2021) (see Lemma 1).

Proof 1 We rewrite (23) as a row-wise manner to get

$$\arg \min_{\mathbf{H}_{i:}^*} \sum_{i=1}^N \left\{ \frac{1}{2} \|\mathbf{A}_{i:} - \mathbf{H}_{i:}\|_2^2 + \mu \|\mathbf{H}_{i:}\|_2^p \right\} \quad (26)$$

Considering each row separately, (26) can be rewritten as

$$\arg \min_{\mathbf{H}_{i:}^*} \frac{1}{2} \|\mathbf{A}_{i:} - \mathbf{H}_{i:}\|_2^2 + \mu \|\mathbf{H}_{i:}\|_2^p \quad (27)$$

We perform economy SVD to $\mathbf{H}_{i:}$ to get $\sigma(\mathbf{H}_{i:}) = \sqrt{\mathbf{H}_{i:} \mathbf{H}_{i:}^T} = \|\mathbf{H}_{i:}\|_2$. Referring to Lemma 2, we have $F(\mathbf{H}_{i:}) = \mu \|\mathbf{H}_{i:}\|_2^p = \mu (\sigma(\mathbf{H}_{i:}))^p = f(\sigma(\mathbf{H}_{i:}))$, i.e., $f(x) = \mu(x)^p$. Thus, the optimal result of (27) is

$$\mathbf{H}_{i:}^* = \mathbf{u}_i \sum_{\mathbf{H}_{i:}}^* \mathbf{v}_i^T = [1] \sigma^*(\mathbf{H}_{i:}) \frac{\mathbf{A}_{i:}}{\|\mathbf{A}_{i:}\|_2} = \sigma^*(\mathbf{H}_{i:}) \frac{\mathbf{A}_{i:}}{\|\mathbf{A}_{i:}\|_2} \quad (28)$$

where $\mathbf{u}_i = \mathbf{I}^T = [1]$ and $\mathbf{v}_i^T = \frac{\mathbf{A}_{i:}}{\|\mathbf{A}_{i:}\|_2}$ can be decomposed by economy SVD of $\mathbf{A}_{i:}$. Since $\sigma(\mathbf{A}_{i:}) = \sqrt{\mathbf{A}_{i:} \mathbf{A}_{i:}^T} = \|\mathbf{A}_{i:}\|_2$ is the only singular value of $\mathbf{A}_{i:}$, $\sigma^*(\mathbf{H}_{i:})$ can be obtained by

$$\begin{aligned}\sigma^*(\mathbf{H}_{i:}) &= \arg \min_{x \geq 0} \frac{1}{2} \|x - \sigma(\mathbf{A}_{i:})\|_2^2 + f(x) \\ &= \arg \min_{x \geq 0} \frac{1}{2} (x - \|\mathbf{A}_{i:}\|_2)^2 + \mu x^p\end{aligned}\quad (29)$$

Thus, taking each view into account, the optimal solution of (17) can be easily obtained by Theorem 1 as

$$\mathbf{H}^{v*} = [\mathbf{H}_{1:}^{v*}; \dots; \mathbf{H}_{N:}^{v*}]^T \quad (30)$$

where the i -th row element is $\mathbf{H}_{i:}^{v*} = \sigma^* \frac{\mathbf{A}_{i:}^v}{\|\mathbf{A}_{i:}^v\|_2}$.

• \mathcal{J} sub-problem:

$$\arg \min_{\mathcal{J}} \beta \|\mathcal{J}\|_{\mathcal{B}}^p + \frac{\rho_2}{2} \|\mathcal{G} - \mathcal{J} + \frac{\mathbf{W}}{\rho_2}\|_F^2 \quad \text{s.t. } \mathbf{G}^{vT} \mathbf{G}^v = \mathbf{I} \quad (31)$$

Now, the idea is to use the following Theorem 2 (Gao et al. 2021):

Theorem 2 Let $\mathcal{S} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$, $h = \min(n_1, n_2)$ have the t -SVD $\mathcal{S} = \mathcal{U} * \mathcal{A} * \mathcal{V}^T$. Then the solution of

$$\arg \min_{\mathcal{X}} \frac{1}{2} \|\mathcal{X} - \mathcal{S}\|_F^2 + \tau \|\mathcal{X}\|_{\mathcal{B}}^p \quad (32)$$

is the following

$$\mathcal{X}^* = \Gamma_{\tau}(\mathcal{S}) = \mathcal{U} * \text{iftt}(\mathcal{C}_{\tau}(\bar{\mathcal{S}})) * \mathcal{V}^T \quad (33)$$

where $\mathcal{C}_{\tau}(\bar{\mathcal{S}})$ is a 3rd-order tensor, whose diagonal elements can be obtained by the General Shrinkage Thresholding (GST) algorithm(Gao et al. 2021).

Accordingly, the solution of (31) is

$$\mathcal{J}^* = \Gamma_{\frac{\beta}{\rho_2}}(\mathcal{G} + \frac{\mathbf{W}}{\rho_2}) \quad (34)$$

• \mathbf{G}^v sub-problem:

$$\begin{aligned}\arg \min_{\mathbf{G}^v} \sum_{v=1}^V \frac{1}{\alpha^v} \left\{ \frac{\rho_0}{2} \|\mathbf{S}^v - \mathbf{G}^v \mathbf{H}^{vT} - \mathbf{P}^v + \frac{\mathbf{K}^v}{\rho_0}\|_F^2 + \right. \\ \left. \frac{\rho_1}{2} \|\mathbf{G}^{vT} \mathbf{T}^{vT} - \mathbf{Q}^v + \frac{\mathbf{M}^v}{\rho_1}\|_F^2 \right\} + \sum_{v=1}^V \frac{\rho_2}{2} \|\mathbf{G}^v - \mathbf{J}^v + \frac{\mathbf{W}^v}{\rho_2}\|_F^2 \\ \text{s.t. } \mathbf{G}^{vT} \mathbf{G}^v = \mathbf{I}, \sum_{v=1}^V \alpha^v = 1, \alpha^v \geq 0\end{aligned}\quad (35)$$

By simple calculation, (35) is equivalent to

$$\begin{aligned}\arg \max_{\mathbf{G}^v, \mathbf{H}^v} \sum_{v=1}^V \frac{1}{\alpha^v} \left\{ \text{tr}(\mathbf{G}^{vT} \mathbf{D}^v \mathbf{G}^v) + 2 \text{tr}(\mathbf{G}^{vT} \mathbf{E}^v) \right\} \\ \text{s.t. } \mathbf{G}^{vT} \mathbf{G}^v = \mathbf{I}, \sum_{v=1}^V \alpha^v = 1, \alpha^v \geq 0\end{aligned}\quad (36)$$

where $\mathbf{E}^v = \frac{\rho_0}{2\alpha^v} \mathbf{A}^v \mathbf{H}^v + \frac{\rho_1}{2\alpha^v} \mathbf{T}^{vT} \mathbf{B}^{vT} + \frac{\rho_2 \mathbf{C}^v}{2}$ and $\mathbf{D}^v = -\frac{\rho_1}{2\alpha^v} \mathbf{T}^{vT} \mathbf{T}^v$. Update $\mathbf{F}^v = \mathbf{D}^v \mathbf{G}^v + \mathbf{E}^v$ and perform SVD decomposition on \mathbf{F}^v to get $\mathbf{U} \mathbf{\Sigma} \mathbf{V}^T = \mathbf{F}^v$. Update $\mathbf{G}^{v*} = \mathbf{U} \mathbf{V}^T$. Repeat until convergence and get new \mathbf{G}^v .

• \mathbf{P}^v sub-problem:

$$\begin{aligned}\arg \min_{\mathbf{P}^v} \sum_{v=1}^V \frac{1}{\alpha^v} \left\{ \frac{1}{2} \|\mathbf{P}^v - \mathbf{N}^v\|_F^2 + \frac{1}{\rho_0} \|\mathbf{P}^v\|_{2,p} \right\} \\ \text{s.t. } \mathbf{G}^{vT} \mathbf{G}^v = \mathbf{I}, \sum_{v=1}^V \alpha^v = 1, \alpha^v \geq 0\end{aligned}\quad (37)$$

where $\mathbf{N}^v = \mathbf{S}^v + \mathbf{G}^v \mathbf{H}^{vT} - \frac{\mathbf{K}^v}{\rho_0}$. Inspired by Theorem 1, taking each view into account, the optimal solution of (37) can be easily obtained as

$$\mathbf{P}^{v*} = [\mathbf{P}_{1:}^{v*}; \dots; \mathbf{P}_{N:}^{v*}]^T \quad (38)$$

where the i -th row element is $\mathbf{P}_{i:}^{v*} = \sigma^* \frac{\mathbf{N}_{i:}^v}{\|\mathbf{N}_{i:}^v\|_2}$.

• α^v sub-problem:

$$\arg \min_{\alpha^v} \sum_{v=1}^V \frac{\tau_v}{\alpha^v} \quad \text{s.t. } \sum_{v=1}^V \alpha^v = 1, \alpha^v \geq 0 \quad (39)$$

where $\tau_v = \|\mathbf{P}^v\|_{2,p} + \frac{\rho_1}{2} \|\mathbf{G}^{vT} \mathbf{T}^{vT} - \mathbf{Q}^v + \frac{\mathbf{M}^v}{\rho_1}\|_F^2 + \frac{\gamma}{2} \|\mathbf{Q}^v\|_1 + \frac{\rho_0}{2} \|\mathbf{S}^v - \mathbf{G}^v \mathbf{H}^{vT} - \mathbf{P}^v + \frac{\mathbf{K}^v}{\rho_0}\|_F^2 + \lambda \|\mathbf{H}^v\|_{2,p}$.

Due to the fact that $\sum_{v=1}^V \alpha^v = 1$ and $\alpha^v \geq 0$, it follows that (39) is equivalent to

$$\arg \min_{\alpha^v, \eta} \sum_{v=1}^V \frac{\tau_v}{\alpha^v} - \eta \left(\sum_{v=1}^V \alpha^v - 1 \right) \quad (40)$$

where η is the Lagrange multiplier. Setting the partial derivatives with respect to α^v and η in (40) to be zero gives $\alpha^v = \sqrt{\frac{\tau_v}{\eta}}$ and $\eta = \left(\sum_{v=1}^V \sqrt{\tau_v} \right)^2$. Then it is not hard to see

$$\alpha^v = \sqrt{\tau_v} \left/ \sum_{v=1}^V \sqrt{\tau_v} \right. \quad (41)$$

The remaining variables are updated as follows

$$\mathbf{K}^v = \mathbf{K}^v + \rho_0 \left(\mathbf{S}^v - \mathbf{G}^v \mathbf{H}^{vT} - \mathbf{P}^v \right) \quad (42)$$

$$\mathbf{M}^v = \mathbf{M}^v + \rho_1 (\mathbf{G}^v - \mathbf{Q}^v) \quad (43)$$

$$\mathbf{W} = \mathbf{W} + \rho_2 (\mathcal{G} - \mathcal{J}) \quad (44)$$

$$\rho_i = \min(pho_rho \times \rho_i, max_rho_i) \quad (45)$$

where $i = 0, 1, 2$, max_rho_i and pho_rho are constants.

• **Clustering labels sub-problem:** The final clustering assignment matrix $\mathbf{C} = \sum_{v=1}^V \frac{\mathbf{G}^v}{\alpha^v}$ is constructed using $\mathbf{G}^v \in \mathbb{R}^{N \times C}$, where N represents the number of data points and C denotes the number of categories. The whole algorithm is summarized in Algorithm 1.

Complexity

The main computational complexity of this model is $\mathcal{O}(VN M d + VN^2 C t)$, where $d = \sum_{v=1}^V d_v$, V , M , N , C and t are the number of views, anchors samples, number of clusters and number of iterations, respectively.

Algorithm 1: EFSGMC

Input: Data matrices: $\{\mathbf{X}^v\}_{v=1}^V \in \mathbb{R}^{N \times d_v}$, anchors number M , and cluster number K .

Output: Cluster assignment matrix $\tilde{\mathbf{G}}$ with K classes.

- 1: Initialize $\mathcal{W} = \mathcal{J} = \mathbf{0}$, $\mathbf{K}^v = \mathbf{M}^v = \mathbf{0}$, $\rho_0, \rho_1, \rho_2, \text{pho_}\rho = 1.1, \text{max_}\rho_i = 10^{10}, \alpha^v = \frac{1}{V}, \gamma, \lambda, \beta$.
- 2: **while** not converge **do**
- 3: Update \mathbf{Q}^v by using (15)
- 4: Update \mathbf{H}^v by using (30)
- 5: Update \mathcal{J} by using (34)
- 6: Update \mathbf{G}^v by solving (36)
- 7: Update \mathbf{P}^v by solving (37)
- 8: Update α^v by using (41)
- 9: Update $\mathbf{K}^v, \mathbf{M}^v, \mathcal{W}$ and ρ_i by using (42), (43), (44) and (45), respectively;
- 10: Directly achieve the K clusters based on the cluster assignment matrix $\tilde{\mathbf{G}} = \sum_{v=1}^V \frac{\mathbf{G}^v}{\alpha^v}$;
- 11: **end while**
- 12: **return** Clustering results.

Scale	Normal				Large
Dataset	MSRC	HW	Mnist	Cal101	Reuters
Size	210	2000	4000	2386	18758
Sample	1622	345	69	3766	107727
Views	5	4	3	6	5
Clusters	7	10	4	20	6

Table 1: Statistics of Real Benchmark Datasets.

Experiments

Experimental Setup

Datasets As shown in Table 1, MSRC (Winn and Jojic 2005), HW (Dua and Graff 2017), Mnist4 (Deng 2012), Cal101-20 (Fei-Fei, Fergus, and Perona 2007) and Reuters (Apté, Damerau, and Weiss 1994) are selected.

Settings The hyperparameters are selected in the range of $[0.0001, 0.001, 0.01, 0.1, 0.5, 1, 5, 10, 50, 100, 1000, 10000]$ to obtain the optimal results. Adjusting K within the range of 2-10, the optimal value of 5 was chosen.

Comparative algorithms A total of 11 state-of-the-art algorithms were selected as comparison algorithms, including Co-reg (Kumar and Rai 2011), SwMC (Nie, Li, and Li 2017), MVSC (Li et al. 2015), SMSC (Hu et al. 2020), AMGL (Nie, Li, and Li 2016), MLAN (Nie et al. 2018), SFMC (Li et al. 2020), RMSC (Xia et al. 2014), CSMSC (Luo et al. 2018), MSC-BG (Yang et al. 2022), FPMVSCAG (Wang et al. 2021).

Evaluation metrics ACC, NMI and Purity indexes are used to evaluate the clustering performance of all methods.

Experimental Results

Comparisons with State-of-the-art Methods The clustering performance of different algorithms on 5 data sets is

shown in Table 2. Specifically, our method and MSC-BG algorithm both obtain optimal and sub-optimal clustering performance. On the largest dataset Reuters, the proposed method and MSC-BG and SFMC algorithms achieve satisfactory clustering results, but our method is superior. All three algorithms use the anchor selection method, significantly reducing memory consumption. Our algorithm significantly improves performance compared to MSC-BG, which uses anchor selection and tensor Schatten p -norm. The primary reason may be that our method uses feature selection to filter out redundant information. Moreover, compared with the SMSC algorithm, our method considers both inter-view and intra-view potential information, substantially improving clustering performance.

Effect of parameter p of Schatten p -norm We vary the p from 0.1 to 1.0 to investigate the impact of the tensor Schatten p -norm on two datasets. Experiments show that the proposed approach achieves the optimal performance at 0.5, 0.8, which indicates that different p values can better capture the latent space distribution of different datasets. Thus, adjusting p on different datasets can improve performance in practical applications .

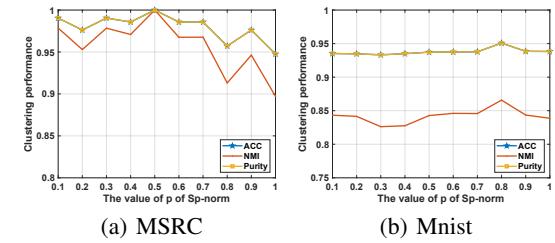
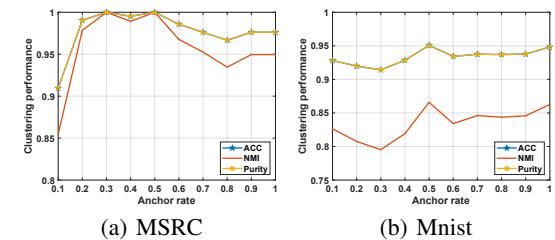
Figure 2: The clustering result with varying p of Schatten p -norm.

Figure 3: The clustering result with varying anchor rates.

Effect of anchor rate We conducted ten experiments by incrementally increasing the anchor rate from 0.1 to 1.0, as shown in Figure 3. Experiments show that the algorithm does not necessarily perform optimally when the anchor rate is set to 1.0. This could be attributed to redundant and irrelevant data in the actual dataset, which can be removed using the anchor selection method. Additionally, an anchor selection strategy reduces memory usage, enabling effective clustering of large-scale data.

Dataset	MSRC			HW			Mnist			Cal101			Reuters		
Metric	ACC	NMI	Purity												
Co-reg	0.635	0.578	0.659	0.784	0.758	0.795	0.785	0.602	0.786	0.412	0.587	0.754	0.563	0.326	0.552
SwMC	0.776	0.774	0.805	0.758	0.833	0.792	0.914	0.799	0.912	0.599	0.493	0.700	OM	OM	OM
MVSC	0.794	0.672	0.756	0.796	0.820	0.808	0.733	0.651	0.780	0.595	0.613	0.717	0.596	0.347	0.574
SMSC	0.766	0.717	0.804	0.742	0.781	0.759	0.913	0.789	0.913	0.582	0.590	0.748	OM	OM	OM
AMGL	0.751	0.704	0.789	0.704	0.762	0.732	0.910	0.785	0.910	0.557	0.552	0.677	OM	OM	OM
MLAN	0.681	0.630	0.733	0.778	0.832	0.812	0.744	0.659	0.744	0.526	0.474	0.666	OM	OM	OM
SFMC	0.810	0.721	0.810	0.853	0.871	0.873	0.917	0.801	0.917	0.642	0.595	0.748	0.602	0.354	0.552
RMSC	0.762	0.663	0.769	0.681	0.661	0.713	0.705	0.486	0.705	0.385	0.512	0.742	OM	OM	OM
CSMSC	0.758	0.735	0.793	0.806	0.793	0.867	0.643	0.645	0.832	0.474	0.648	0.563	OM	OM	OM
MSC-BG	0.981	0.960	0.981	0.889	0.922	0.889	0.938	0.861	0.938	0.667	0.727	0.794	0.640	0.484	0.686
FPMVS-CAG	0.843	0.738	0.843	0.85	0.787	0.850	0.887	0.719	0.887	0.635	0.611	0.723	0.526	0.323	0.603
EFSGMC	1.000	1.000	1.000	0.994	0.984	0.994	0.951	0.866	0.951	0.741	0.725	0.839	0.618	0.518	0.739

Table 2: The clustering result of the selected methods. ("OM" means out-of-memory.)

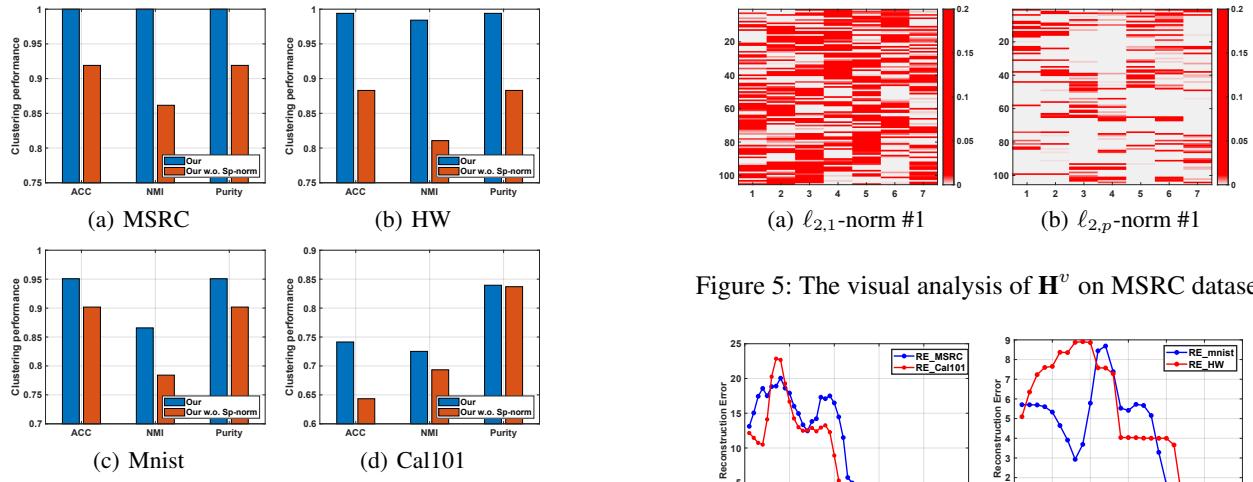


Figure 4: The ablation studies on selected datasets.

Ablation studies about Schatten p -norm To evaluate the effectiveness of utilizing the tensor Schatten p -norm to extract complementary information from inter-view low-rank space, we conducted corresponding experiments on four datasets. An analysis of Figure 4 shows that utilizing this norm can help reveal complementary information that may be hidden between different views.

Effect of minimizing $\ell_{2,p}$ -norm of \mathbf{H}^v To investigate the role of the $\ell_{2,p}$ -norm, we set $p = 1$ to obtain the $\ell_{2,1}$ -norm and present the corresponding visual analysis of the feature selection matrix \mathbf{H}^v in Figure 5. The usage of the $\ell_{2,p}$ -norm results in sparser \mathbf{H}^v . This indicates that the $\ell_{2,p}$ -norm can lead to sparser feature selection matrices and facilitate feature selection on the graph, enabling the removal of redundant and noisy anchor points in the anchor graph.

Analysis of convergence curves We analyzed the relationship between the reconstruction error of anchor graph \mathbf{S}^v and the number of iterations, where the Reconstruction Error (RE) is defined as $RE = \|\mathbf{S}^v - \mathbf{G}^v \mathbf{H}^{vT} - \mathbf{P}^v\|_\infty$ on

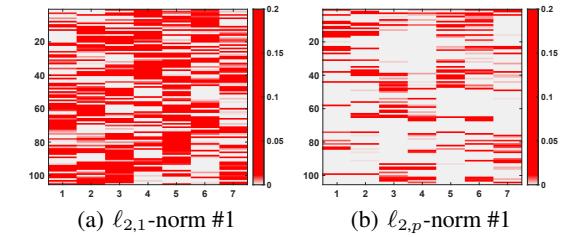
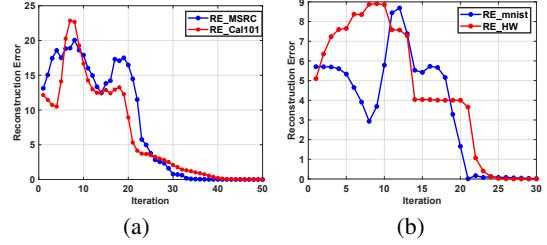
Figure 5: The visual analysis of \mathbf{H}^v on MSRC dataset.

Figure 6: The convergence curves.

the dataset MSRC, HW, Mnist4, and Cal101. Experiments show that the models converge within 50 iterations on the four datasets.

Conclusion

In this paper, we utilize the non-negative matrix decomposition on the anchor graph to obtain the cluster label in one step. Additionally, we introduce a novel $\ell_{2,p}$ -norm for feature selection on the anchor graph and provide an effective solution, significantly improving clustering efficiency. We include the minimization tensor Schatten p -norm of cluster assignment matrices to enhance clustering performance, which helps explore the complementary information and representation space structure between different views. We introduce our algorithm and provide an efficient solution. Extensive experiments verify the validity of EFSGMC.

Acknowledgements

This work is supported by the Science and technology project of Xi'an (Grant 2022JH-JSYF-0009), the Natural Science Basic Research Program of Shaanxi Province (Grant 2023-JC-YB-534), the National Natural Science Foundation of China under Grants 62176203, Natural Science Foundation of Shandong Province under Grant ZR202102180986, the Fundamental Research Funds for the Central Universities and the Innovation Fund of Xidian University, Guangxi Key Laboratory of Digital Infrastructure under Grant GXDIOP2023010 Natural Science Foundation of Guangdong Province, 2023A1515011845.

References

Apté, C.; Damerau, F.; and Weiss, S. M. 1994. Automated Learning of Decision Rules for Text Categorization. *ACM Trans. Inf. Syst.*, 12(3): 233–251.

Deng, L. 2012. The MNIST Database of Handwritten Digit Images for Machine Learning Research. *IEEE Signal Process. Mag.*, 29(6): 141–142.

Ding, C.; Li, T.; Peng, W.; and Park, H. 2006. Orthogonal nonnegative matrix t-factorizations for clustering. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 126–135.

Dong, J.; Liu, R.; Tang, K.; Wang, Y.; Zhang, X.; and Su, Z. 2016. Sparse Gradient Pursuit for Robust Visual Analysis. In *ACCV*, 369–384.

Dua, D.; and Graff, C. 2017. UCI Machine Learning Repository.

Fei-Fei, L.; Fergus, R.; and Perona, P. 2007. Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories. *Comput. Vis. Image Understand.*, 106(1): 59–70.

Gao, J.; Han, J.; Liu, J.; and Wang, C. 2013. Multi-View Clustering via Joint Nonnegative Matrix Factorization. 252–260. SIAM.

Gao, Q.; Zhang, P.; Xia, W.; Xie, D.; Gao, X.; and Tao, D. 2021. Enhanced Tensor RPCA and its Application. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(6): 2133–2140.

Hale, E. T.; Yin, W.; and Zhang, Y. 2008. Fixed-Point Continuation for l_1 -Minimization: Methodology and Convergence. *SIAM J. Optim.*, 19(3): 1107–1130.

Hu, Z.; Nie, F.; Wang, R.; and Li, X. 2020. Multi-view spectral clustering via integrating nonnegative embedding and spectral embedding. *Inf. Fusion*, 55: 251–259.

Kumar, A.; and Rai, P. 2011. Co-regularized multi-view spectral clustering. In *NeurIPS*, 1413–1421.

Li, X.; Zhang, H.; Wang, R.; and Nie, F. 2020. Multiview clustering: A scalable and parameter-free bipartite graph fusion method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1): 330–344.

Li, Y.; Nie, F.; Huang, H.; and Huang, J. 2015. Large-Scale Multi-View Spectral Clustering via Bipartite Graph. In Bonet, B.; and Koenig, S., eds., *AAAI*, 2750–2756.

Liao, S.; Li, J.; Liu, Y.; Gao, Q.; and Gao, X. 2018. Robust Formulation for PCA: Avoiding Mean Calculation With $\ell_{2,p}$ -norm Maximization. In *AAAI*, 3604–3610.

Lu, Y.; Lai, Z.; Xu, Y.; Li, X.; Zhang, D.; and Yuan, C. 2016. Low-Rank Preserving Projections. *IEEE Trans. Cybern.*, 46(8): 1900–1913.

Luo, S.; Zhang, C.; Zhang, W.; and Cao, X. 2018. Consistent and Specific Multi-View Subspace Clustering. In *AAAI*, 3730–3737.

Nie, F.; Cai, G.; Li, J.; and Li, X. 2018. Auto-Weighted Multi-View Learning for Image Clustering and Semi-Supervised Classification. *IEEE Trans. Image Process.*, 27(3): 1501–1511.

Nie, F.; Chang, W.; Wang, R.; and Li, X. 2023. Learning an Optimal Bipartite Graph for Subspace Clustering via Constrained Laplacian Rank. *IEEE Trans. Cybern.*, 53(2): 1235–1247.

Nie, F.; Li, J.; and Li, X. 2016. Parameter-Free Auto-Weighted Multiple Graph Learning: A Framework for Multiview Clustering and Semi-Supervised Classification. In *IJCAI*, 1881–1887.

Nie, F.; Li, J.; and Li, X. 2017. Self-weighted Multiview Clustering with Multiple Graphs. In *IJCAI*, 2564–2570.

Salah, A.; Ailem, M.; and Nadif, M. 2018. Word Co-Occurrence Regularized Non-Negative Matrix Tri-Factorization for Text Data Co-Clustering. In *AAAI*, 3992–3999.

Wang, Q.; Gao, Q.; Gao, X.; and Nie, F. 2018. $\ell_{2,p}$ -Norm Based PCA for Image Recognition. *IEEE Trans. Image Process.*, 27(3): 1336–1346.

Wang, S.; Liu, X.; Zhu, X.; Zhang, P.; Zhang, Y.; Gao, F.; and Zhu, E. 2021. Fast parameter-free multi-view subspace clustering with consensus anchor guidance. *IEEE Transactions on Image Processing*, 31: 556–568.

Winn, J. M.; and Jojic, N. 2005. LOCUS: Learning Object Classes with Unsupervised Segmentation. In *ICCV*, 756–763.

Xia, R.; Pan, Y.; Du, L.; and Yin, J. 2014. Robust Multi-View Spectral Clustering via Low-Rank and Sparse Decomposition. In *AAAI*, 2149–2155.

Xia, W.; Zhang, X.; Gao, Q.; Shu, X.; Han, J.; and Gao, X. 2022. Multiview Subspace Clustering by an Enhanced Tensor Nuclear Norm. *IEEE Trans. Cybern.*, 52(9): 8962–8975.

Xie, D.; Zhang, X.; Gao, Q.; Han, J.; Xiao, S.; and Gao, X. 2020. Multiview Clustering by Joint Latent Representation and Similarity Learning. *IEEE Trans. Cybern.*, 50(11): 4848–4854.

Yang, H.; Gao, Q.; Xia, W.; Yang, M.; and Gao, X. 2022. Multiview Spectral Clustering With Bipartite Graph. *IEEE Trans. Image Process.*, 31: 3591–3605.

Yang, M.; Luo, Q.; Li, W.; and Xiao, M. 2020. Multiview Clustering of Images with Tensor Rank Minimization via Nonconvex Approach. *SIAM J. Imaging Sci.*, 13(4): 2361–2392.

Yang, Y.; Wang, Z.; Yang, J.; Han, J.; and Huang, T. S. 2014. Regularized l_1 -Graph for Data Clustering. In *BMVC*.