Efficient Multi-view Unsupervised Feature Selection with Adaptive Structure Learning and Inference

Chenglong Zhang1, Yang Fang2, Xinyan Liang3, Han Zhang4, Peng Zhou5, Xingyu Wu6, Jie Yang7, Bingbing Jiang1,*, and Weiguo Sheng1
1Hangzhou Normal University, Hangzhou, China
2Chongqing University of Posts and Telecommunications, Chongqing, China
3Shanxi University, Taiyuan, China
4Northwestern Polytechnical University, Xi’an, China
5Anhui University, Hefei, China
6Hong Kong Polytechnical University, Hong Kong SAR, China
7University of Technology Sydney, NSW, Australia
{clzhang123,liangxinyan48}@163.com, fangyang@cqupt.edu.cn, zhanghan9937@gmail.com, zhoupeng@ahu.edu.cn, xingy.wu@polyu.edu.hk, jie.yang-1@uts.edu.au, jiangbb@hznu.edu.cn, w.sheng@ieee.org

Abstract

As data with diverse representations become high-dimensional, multi-view unsupervised feature selection has been an important learning paradigm. Generally, existing methods encounter the following challenges: (i) traditional solutions either concatenate different views or introduce extra parameters to weight them, affecting the performance and applicability; (ii) emphasis is typically placed on graph construction, yet disregarding the clustering information of data; (iii) exploring the similarity structure of all samples from the original features is suboptimal and extremely time-consuming. To solve this dilemma, we propose an efficient multi-view unsupervised feature selection (EMUFS) to construct bipartite graphs between samples and anchors. Specifically, a parameter-free manner is devised to collaboratively fuse the membership matrices and graphs to learn the compatible structure information across all views, naturally balancing different views. Moreover, EMUFS leverages the similarity relations of data in the feature subspace induced by $l_{2,0}$-norm to dynamically update the graph. Accordingly, the cluster information of anchors can be accurately propagated to samples via the graph structure and further guide feature selection, enhancing the quality of selected features and the computational costs in solution processes. A convergent optimization is developed to solve the formulated problem, and experiments demonstrate the effectiveness and efficiency of EMUFS.

1 Introduction

As information technology develops rapidly, data collected from heterogeneous sources often contains multiple representations [Hu et al., 2019; Zhong and Pun, 2020; Liang et al., 2022b; Liang et al., 2022a; Zhao et al., 2023; Peng et al., 2023]. To process this kind of data, multi-view learning has been proposed in recent years [Hu et al., 2021; Liu et al., 2022; Yang et al., 2024; Xu et al., 2024; Liang et al., 2024]. As a special case, multi-view feature selection that aims to select representative features from the original feature space has become a fundamental task. Depending on the availability of data class labels, current methods can be realized in supervised, semi-supervised, and unsupervised ways [Bai et al., 2021; Jiang et al., 2022; Zhang et al., 2023b; Li et al., 2024b; Zhao et al., 2024]. Considering the expensive cost of manually labeling data, several researchers have devoted to the multi-view unsupervised feature selection, which utilizes the intrinsic data structure to select informative features without the guidance of label information.

To select a feature subset from multi-view data, existing methods can be mainly categorized into two manners. The first kind of method directly concatenates the features from different views and then invokes single-view models on the concatenated features. Typical methods include feature selection with graph learning [Nie et al., 2016; Chen et al., 2023; Tang et al., 2023] and spectral feature selection [Zhao and Liu, 2007; Li and Tang, 2015; Zhou et al., 2023]. Considering that different views have specific properties and contribute variably to final models, this kind of method treats different views equally and neglects the difference between them, causing performance deterioration in practical domains [Tang et al., 2019; Zhong and Pun, 2021]. Instead of simply concatenating different views, another kind of method exploits the underlying correlations among views and introduces weights to balance the contributions of different views. Representative methods include feature selection with adap-
tive similarity and view weight (ASVW) [Hou et al., 2017], multilevel projections with adaptive neighbor graph for multi-view feature selection (MAMFS) [Zhang et al., 2021], and robust feature selection via multi-group adaptive graph representation (MGAGR) [You et al., 2023]. Commonly, these methods follow two separate steps, i.e., constructing a graph on each view to explore view-specific structures and performing feature selection based on the weighted combination of multiple graphs. To avoid the trivial solution of view weights, an extra weight-related parameter that needs to be manually determined is introduced, weakening their applicability [Li et al., 2022]. Moreover, they suffer from a higher computation due to constructing graphs on all samples. Recently, [Chen et al., 2023] proposed to learn a bipartite graph between samples and anchors to reduce computational costs. However, they focus on learning graphs to characterize similarity structures, yet are incapable of capturing the inherent cluster structure of data that can positively guide feature selection.

To explore the cluster structure information, [Feng et al., 2013] proposed to learn the feature projections and the pseudo cluster labels of data (i.e., the cluster indicator matrix) by a linear regression model. Inheriting from [Feng et al., 2013], many variants have been developed in recent years. For example, [Dong et al., 2018] incorporated graph learning into the framework of [Feng et al., 2013], such that the similarity graph can be updated in the feature selection process. Instead of the orthogonal constraint, [Shi et al., 2023] imposed a binary hash constraint on the cluster indicator matrix to improve the discrimination. Despite achieving some progress, these methods still face the following limitations: i) the similarity graph directly derived from the original features is susceptible to poor-quality features since the original data usually contains irrelative features, impairing the graph reliability and finally affecting the effectiveness of selected features; ii) their focus lies in learning the cluster indicator matrix of all samples, such that the optimization procedure involves the decomposition or inverse operations of high-order matrices, leading to expensive computational complexity.

Motivated by the aforementioned issues, we propose an efficient multi-view unsupervised feature selection with adaptive structure learning and inference (EMUFS). Fig. 1 illustrates the basic framework of EMUFS. The main contributions of this paper are summarized as follows:

- We propose an efficient multi-view unsupervised feature selection method that simultaneously leverages membership matrices and bipartite graphs to capture the cluster information of anchors and the similarity structure between samples and anchors, facilitating the ultimate feature selection and reducing computational costs.
- We design a collaborative fusion manner for membership matrices and graphs to learn the compatible structures, so that the cluster information of samples can be inferred via the similarity between samples and anchors, balancing different views without extra parameters.
- We utilize the neighbor relations of data in the selected feature subspace to adaptive learn a unified bipartite graph, reducing the impact of poor-quality features.

## 2 Our Proposed Methodology

### 2.1 Notations

Throughout the paper, vectors are written in bold lowercase letters, and matrices are written in bold uppercase letters. Besides, $m_i$ represents the $i$-th row of a given matrix $M$, $\text{Tr}(M)$
denotes the trace of $M$ and $\|M\|_F = \sqrt{\text{Tr}(M^T M)}$ denotes the Frobenius norm. $\|M\|_{2,1} = \sum_i \|m_i\|_2$ and $\|M\|_{2,0} = \sum_i \|m_i\|_0$ denote the $l_{2,1}$-norm and $l_{2,0}$-norm of $M$, respectively. Table 1 lists the frequently used notations.

### 2.2 Similarity Structure Learning and Fusion

Most methods construct $n$-order graphs to mine the similarity relations among samples, leading to high computational costs [Hu et al., 2022; Li et al., 2024a]. To this end, our proposed EMUFs learns bipartite graphs between samples and anchors to improve the efficiency of graph construction. Specifically, the alternate sampling strategy proposed by [Li et al., 2022] is used to generate a consistent anchor set on different views. Then, the view-specific similarity structures between samples and generated anchors can be captured by solving:

$$\min_{S_{v,1}=1, S_{v,2} \geq 0} \sum_{i \in V} \sum_{j \in V} \|x_i^v - z_j^v\|^2 s_{ij}^v + \tau \|S_v\|_F^2,$$  

(1)

where $S_v \in \mathbb{R}^{n \times m}$ is the bipartite graph on the $v$-th view, and $s_{ij}^v$ measures the similarity between sample $x_i$ and anchor $z_j$ in $v$-th view. Eq.(1) can be directly solved by the adaptive neighbor strategy [Nie et al., 2014].

To learn a consensus similarity graph, the view-specific similarity matrices on different views are first generated by the FCM [Wang et al., 2017] that applied an $\text{min}$ function to mine the view-specific cluster structures of data, which avoids the decomposition or clustering information propagating from anchors in the selected feature subspace:

$$\min_{S_{1:V}} \sum_{v=1}^V \sum_{i,j=1}^m \|x_i^v - z_j^v\|^2 s_{ij}^v + \tau \|S_v\|_F^2,$$  

(2)

where $S_v$ denotes a unified graph that compatibly crosses multiple views, and the exponential parameter $\eta$ controls the distribution of the view weights $\{\phi_v\}_{v=1}^V$. To further reveal the underlying relation between $S$ and $\{S_v\}_{v=1}^V$, we fix $\phi$ and set the derivation of Eq. (2) w.r.t. $S$ to zero:

$$\sum_{v=1}^V \phi_v s_{ij}^v (S - S_v) = 0 \iff S = \sum_{v=1}^V \alpha_v S_v; \quad \alpha_v = \phi_v \sum_{v=1}^V \phi_v,$$  

(3)

where $\alpha_v$ can be regarded as the weight since $\alpha_v \geq 0$ and $\sum_{v=1}^V \alpha_v = 1$. In Eq. (3), the equality relation of $S$ and $\sum_{v=1}^V \alpha_v S_v$ can be further relaxed by introducing a flexible regression residue (i.e., $S - \sum_{v=1}^V \alpha_v S_v$) to measure the mismatch between them, achieving a new fusion paradigm as:

$$\min_{S_{1:V}, S \geq 0, \alpha \geq 0} \sum_{v=1}^V \alpha_v s_{ij}^v \|S - \sum_{v=1}^V \alpha_v S_v\|_1.$$  

(4)

Eq. (4) can fuse multiple graphs without the exponential parameter $\eta$. Further, considering that the poor-quality features in the original space can undermine graph structures, we propose to dynamically update the graph $S$ according to the relations of samples and anchors in the selected feature subspace:

$$\min_{W, S, \alpha} \sum_{i=1}^n \sum_{j=1}^m \|W^T x_i - W^T z_j\|_F^2 s_{ij} + \lambda \|S - \sum_{v=1}^V \alpha_v S_v\|_2 \quad \text{s.t.} \quad S_1 = 1, S \geq 0, \alpha \geq 0, \alpha^T 1 = 1, \|W\|_{2,0} = k.$$  

(5)

In contrary to previous methods [Nie et al., 2016; Hou et al., 2017] that applied an $l_{2,1}$-norm on the feature projection matrix $W$, the $l_{2,0}$-norm constraint is imposed on $W$ to automatically select salient features without sorting features in advance. Moreover, the graph learned from the selected features can accurately explore the similarity structure and alleviate the impacts of low-quality features.

### 2.3 Cluster Information Inference on Graph

A membership matrix that contains the cluster structure of data is widely used since it can be regarded as the latent discriminative information in unsupervised scenarios [Zhang et al., 2023a; Lu et al., 2023]. For convenience, existing models often generate membership matrices on each view independently and coalesce them for further analyses. Nevertheless, the clusters preserved in the membership matrices might be mismatched between different views, which means that different membership matrices are not aligned column-widely [Wang et al., 2022]. As a result, the direct fusion of membership matrices might suffer from inconsistent cluster information across views. To solve this limitation, the membership matrices on different views are first generated by the FCM [Bezdek et al., 1984] to mine the view-specific cluster structures, then different membership matrices can be aligned by solving the following problem:

$$\min_{T_v} \|U_1 - U_v T_v\|_F^2 \quad \text{s.t.} \quad T_v^T T_v = I,$$  

(6)

where the permutation matrix $T_v$ makes $U_1$ and $U_v$ consistent at the level of clusters and can be directly solved by performing the SVD on $U_1^T U_v$. We exploit membership matrices on anchors instead of all samples to explore the latent cluster structure of data, which avoids the decomposition or
inverse operations of high-order matrices in the subsequent solution process. To capture the consensus cluster structure of anchors, the same fusion approach as Eq. (4) can be employed to merge the aligned membership matrices:

\[
\min_{U_{1:1}, U_{2:0}, \alpha \geq 0, \alpha' \\ 1:1} \left\| U - \sum_{v=1}^{V} \alpha_v \tilde{U}_v \right\|_F^2, \tag{7}
\]

where \( \tilde{U}_v \) (i.e., \( U_v \mathbf{T}_v \)) denotes the aligned membership matrix, serving as the label guidance for feature selection. With the learned bipartite graph \( S \), the label guidance (i.e., cluster information) of samples can be directly inferred from anchors, as shown in Fig. 2. Therefore, the final optimization objective of the proposed EMDFS is obtained as:

\[
\min_{W \in \mathbb{R}^{(2,0)=k, F \geq 0, S \geq 0, U \geq 0, \alpha \geq 0}} \left\| X^T W - F \right\|_F^2 + \lambda \left\| F - SU \right\|_F^2 + \beta \left( \left\| S - \sum_{v=1}^{V} \alpha_v S_v \right\|_F^2 + \left\| U - \sum_{v=1}^{V} \alpha_v U_v \right\|_F^2 \right) + \gamma \sum_{i=1}^{n} \sum_{j=1}^{m} \left\| W^T x_i - W^T z_j \right\|_2^2 s_{ij},
\]

s.t. \( F^1 = 1, S^1 = 1, U^1 = 1, \alpha^T 1 = 1, \) \( \) where \( F \) denotes the cluster labels of samples and simultaneously leverages the similarity and cluster information. In Eq. (8), the membership matrix and the bipartite graph are collaboratively fused to facilitate the cluster information inference on the graph. Moreover, the similarity structure is also guided by the relations between samples and anchors in the selected feature subspace, enhancing the graph quality. Consequently, similarity structure learning, cluster information inference and feature selection are incorporated into a unified framework and can benefit from each other, such that the discriminative features can be selected under the guidance of the learned structure compatible across all views.

### 3 Optimization and Analyses

To solve the problem in Eq. (8), an iterative optimization is devised to solve one variable by fixing other variables. Considering that the \( l_2,0 \)-norm optimization problem is complicated, an auxiliary variable \( E = W \) is introduced to transform Eq. (8) into the following equivalent problem:

\[
\min_{\|E\|_{2,0} = k, F \geq 0, S \geq 0, U \geq 0, \alpha \geq 0} \left\| X^T W - F \right\|_F^2 + \lambda \left\| F - SU \right\|_F^2 + \beta \left( \left\| S - \sum_{v=1}^{V} \alpha_v S_v \right\|_F^2 + \left\| U - \sum_{v=1}^{V} \alpha_v U_v \right\|_F^2 \right) + \gamma \sum_{i=1}^{n} \sum_{j=1}^{m} \left\| W^T x_i - W^T z_j \right\|_2^2 s_{ij}, \tag{9}
\]

where \( \mu = \mathbb{R}^{1 \times 1} \) and \( \Pi = \mathbb{R}^{n \times c} \) denote the penalty parameter and Lagrange multipliers, respectively. The solution procedures are given as follows:

- **U subproblem**: When fixing other variables, the problem of Eq. (9) is simplified into:

\[
\min_{U_{1:1}, U_{2:0}, \alpha \geq 0} \lambda \left\| F - SU \right\|_F^2 + \beta \left\| U - \sum_{v=1}^{V} \alpha_v \tilde{U}_v \right\|_F^2, \tag{10}
\]

For computational efficiency, we first ignore the constraints to calculate the latent solution \( U^* \) and then project \( U^* \) into the constrained space. Specifically, setting the derivative of Eq. (10) w.r.t. \( U \) to zero, the latent solution of \( U \) is obtained as:

\[
U^* = (\lambda s^T S + \beta I)^{-1}(s^T F + \beta \sum_{v=1}^{V} \alpha_v \tilde{U}_v). \tag{11}
\]

Subsequently, the optimal solution of \( U \) can be derived by solving:

\[
\min_{U_{1:1}, U_{2:0}, \alpha \geq 0} \left\| U - U^* \right\|_F^2. \tag{12}
\]

Eq. (11) can be solved with a closed-form solution [Huang et al., 2015]. Unlike previous methods that directly learn the cluster information of all samples and need the \( O(n^2) \) computational complexity at least, EMDFS explores the cluster structure of anchors first and then uses the learned bipartite graph to propagate the cluster information, avoiding the decomposition or inverse operations of \( n \)-order dense matrices.

- **α subproblem**: By fixing the other variables, the optimization subproblem w.r.t. \( \alpha \) is:

\[
\min_{\alpha \geq 0} \left\| s - \sum_{v=1}^{V} \alpha_v s_v \right\|_F^2 + \left\| U - \sum_{v=1}^{V} \alpha_v \tilde{U}_v \right\|_F^2. \tag{13}
\]

where \( Q = \psi^T \psi + \Omega^T \Omega \), \( p = \psi^T \psi + \Omega^T \omega \). Since Q is semi-definite, Eq. (13) is a quadratic convex problem and can be solved efficiently [Jiang et al., 2023].

- **W subproblem**: By fixing other variables except for \( W \), Eq. (9) is simplified into:

\[
\min_{W} \left\| X^T W - F \right\|_F^2 + \gamma \sum_{i=1}^{n} \sum_{j=1}^{m} \left\| W x_i - W z_j \right\|_2^2 s_{ij} + \frac{\mu}{2} \left\| E - W + \frac{\Pi}{\mu} \right\|_F^2. \tag{14}
\]

Setting the derivative of Eq. (14) w.r.t. \( W \) to zero, we have:

---

**Algorithm 1 Optimization procedures for EMDFS**

**Input**: Multi-view data \( X = [X_1, \cdots, X_V]^T \), the cluster number \( c \), and the parameters \( \lambda, \beta \) and \( \gamma \);

1. Initialize the view-weights \( \alpha_v = 1/V \), a random feature selection matrix \( W, E = \Pi = 0, \mu = 1 \) and \( \rho = 1.1 \);
2. Generate the anchor set \( \{Z_v\}_{v=1}^{V} \), and initialize the membership matrices \( \{U_v\}_{v=1}^{V} \) by FCM;
3. Learn the permutation matrix \( \mathbf{T}_v \) by solving Eq. (6);
4. **repeat**
   5. Update \( U \) by Eq. (10);
   6. Update \( \alpha \) by Eq. (13);
   7. Update \( W \) by Eq. (15);
   8. Update \( E \) by Eq. (16);
   9. Update \( S \) by Eq. (19);
   10. Update \( F \) by Eq. (19);
   11. Update \( \Pi \) and \( \mu \) by Eq. (20);
5. until Eq. (9) converges;

**Output**: Feature selection matrix \( W \) with \( k \) nonzero rows.
\[ W = (XX^T + \gamma M + \frac{\mu I}{2})^{-1} (XF + \frac{\mu}{2} E + \frac{\Pi}{2}), \]

where \( M = XX^T + ZAZ^T - XSZ^T - ZS^2 X^T \).

\* **E subproblem:** By fixing the other variables except for \( E \), we have the following subproblem:

\[
\min_{\|E\|_{2,1} = k} \|E - W + \frac{\Pi}{\mu} \|_F^2.
\]

\* **S subproblem:** By fixing the other variables except for \( S \), we have the following subproblem:

\[
\min_S \lambda \|F - SU\|_F^2 + \gamma \sum_{i=1}^n \sum_{j=1}^m \|W^T x_i - W^T z_j\|^2_{S_{ij}} + \beta \|S - \sum_{v=1}^V \alpha_v S_v\|_F^2 \quad \text{s.t.} \quad S1 = 1, S \geq 0.
\]

Noting that the optimization problem in Eq. (17) is independent for each row (i.e., \( s_i \)), we can calculate \( S \) by rows:

\[
\min_{s_1, \ldots, s_{V}, \lambda} \lambda \|f_i - s_i U\|_2^2 + \beta \|s_i - a_i\|^2 + \gamma \|d_i\|^2.
\]

where \( f_i \) and \( a_i \) denote the \( i \)-th row of \( F \) and \( A = \sum_{v=1}^V \alpha_v S_v \), and \( d_i \) is a row vector with \( d_{ij} = \|W^T x_i - W^T z_j\|^2 \). Setting the derivative of Eq. (18) w.r.t. \( s_i \) to zero, the latent solution of \( s_i \) is obtained as: \( s_i^* = (\lambda FU^T + \beta a_i - \gamma d_i)/(\lambda UU^T + \beta I)^{-1} \). Thus, the optimal solution of \( s_i \) can be obtained by solving:

\[
\min_{s_1, \ldots, s_{V}, \lambda} \|s_i - s_i^*\|^2.
\]

which can be effectively solved by [Huang et al., 2015].

\* **F subproblem:** Fixing other variables, we can optimize \( F \) by rows:

\[
\min_{f_1, \ldots, f_{V}, \lambda} \|x_i^T W - f_i\|_2^2 + \lambda \|f_i - s_i U\|_2^2.
\]

Eq. (20) can be solved with a closed-form solution.

\* **Update ALM parameters:** In each iteration, the penalty parameter \( \mu \) and the Lagrange multipliers \( \Pi \) are updated as:

\[
\Pi = \Pi + \mu (E - W) \quad \mu = \rho \mu.
\]

where \( \rho \) is a constant updated rate. During the optimization, EMUFs separately optimizes each subproblem, decreasing the objective value monotonically until convergence.

**Computational Complexity**

The steps for solving the problem (9) are summarized in Algorithm 1. Specifically, selecting \( m \) anchors takes \( O(nm) \), calculating \( \{U_v\}_{v=1}^V \) by FCM requires \( O(mdcV) \), and learning the permutation matrices needs \( O(mc^2V) \). Afterward, updating \( U, W \), and \( S \) involve the inverses of matrices, requiring \( O(nd^2 + ndm + \frac{d^3}{m}) \) and \( O(nm^2 + ndc + m^3) \). Besides, calculating \( \alpha \) takes \( O(nmV^2) \), and the optimization of \( E \) and \( F \) respectively take \( O(dk + dc) \) and \( O(nc) \). Since \( c \) and \( V \) are small constants, the computational complexity of EMUFs in each iteration is \( O(nm^2 + nd^2 + m^3 + d^3) \), which is linearly related to the sample scale.

**4 Experiments**

**4.1 Experimental Settings**

In this section, six real-word datasets are employed, including flower-17\(^1\), Leaves\(^2\), NUS\(^3\), Scene\(^4\), ALOI\(^5\) and Youtube\(^6\). The details of each dataset are listed in Table 2. To comprehensively verify the superiority and effectiveness of EMUFs, we conduct experiments with six state-of-the-art competitors, including (1) Unsupervised Feature Selection with Structured Graph Optimization (SOGFS) [Nie et al., 2016]; (2) Multi-View Clustering and Feature Learning via Structured Sparsity (MVCSS) [Wang et al., 2013]; (3) Multi-view Unsupervised Feature Selection with Adaptive Similarity and View Weight (ASVV) [Hou et al., 2017]; (4) Multi-view Feature Selection via Nonsignature Structured Graph Learning (NSGL) [Bai et al., 2020]; (5) Multilevel Projections with Adaptive Neighbor Graph for Unsupervised Multi-View Feature Selection (MAMFS) [Zhang et al., 2021]; (6) Robust Unsupervised Feature Selection via Multi-Group Adaptive Graph Representation (MGAGR) [You et al., 2023]. To ensure comparison fairness, the parameters of all competitors are tuned following their respective works. The regularization parameters for EMUFs are searched in a grid of \( \{10^{-3}, 10^{-2}, \ldots, 10^{3}\} \), with the number of anchors set as \( m = 10\% \times n \). The K-means clustering is independently executed 20 times on the selected feature subsets, and the average results, including the clustering accuracy (ACC) and the normalized mutual information (NMI), are reported to evaluate the performance.

**4.2 Comparison Results**

The means and standard deviations of ACC and NMI are presented in Tables 4 and 5, where the optimal and second-best results are prominently indicated in bold and underlined formatting, respectively. The following conclusions can be drawn:

\[ \text{https://www.robots.ox.ac.uk/~vgg/data/flowers/} \]
\[ \text{https://archive.ics.uci.edu/dataset/} \]
\[ \text{http://people.csail.mit.edu/torralba/code/spatialenvelope/} \]
\[ \text{https://aloi.science.uva.nl/} \]
\[ \text{https://archive.ics.uci.edu/dataset/269/} \]
Table 4: ACC of different methods with different numbers of features. The best and second results are in bold and underlined.

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Feature ratio</th>
<th>10%</th>
<th>15%</th>
<th>20%</th>
<th>25%</th>
<th>30%</th>
<th>35%</th>
</tr>
</thead>
<tbody>
<tr>
<td>SOGFS</td>
<td>18.12 ± 0.59</td>
<td>18.35 ± 0.78</td>
<td>18.08 ± 0.35</td>
<td>18.49 ± 0.59</td>
<td>19.76 ± 0.70</td>
<td>19.22 ± 0.36</td>
<td></td>
</tr>
<tr>
<td>MVCSS</td>
<td>23.20 ± 0.98</td>
<td>23.34 ± 1.21</td>
<td>23.51 ± 1.30</td>
<td>24.50 ± 1.17</td>
<td>25.33 ± 1.55</td>
<td>24.99 ± 0.83</td>
<td></td>
</tr>
<tr>
<td>ASV</td>
<td>22.21 ± 1.04</td>
<td>22.49 ± 0.91</td>
<td>22.28 ± 0.71</td>
<td>22.37 ± 0.81</td>
<td>21.40 ± 0.90</td>
<td>21.68 ± 0.69</td>
<td></td>
</tr>
<tr>
<td>flower-17</td>
<td>NSGL</td>
<td>26.62 ± 0.79</td>
<td>25.81 ± 2.04</td>
<td>26.40 ± 1.56</td>
<td>26.54 ± 1.04</td>
<td>26.76 ± 1.53</td>
<td>27.65 ± 1.21</td>
</tr>
<tr>
<td></td>
<td>MAMFS</td>
<td>23.18 ± 1.07</td>
<td>24.70 ± 0.95</td>
<td>25.32 ± 1.09</td>
<td>26.36 ± 1.25</td>
<td>25.89 ± 1.20</td>
<td>27.98 ± 1.12</td>
</tr>
<tr>
<td></td>
<td>MGAGR</td>
<td>25.82 ± 1.33</td>
<td>25.29 ± 1.15</td>
<td>25.09 ± 0.89</td>
<td>25.36 ± 1.35</td>
<td>25.61 ± 0.79</td>
<td>25.47 ± 1.02</td>
</tr>
<tr>
<td></td>
<td>EMUFS</td>
<td>30.49 ± 1.30</td>
<td>31.51 ± 1.87</td>
<td>30.82 ± 1.59</td>
<td>30.05 ± 1.15</td>
<td>31.20 ± 1.14</td>
<td>29.15 ± 0.99</td>
</tr>
</tbody>
</table>

Leaves

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Feature ratio</th>
<th>10%</th>
<th>15%</th>
<th>20%</th>
<th>25%</th>
<th>30%</th>
<th>35%</th>
</tr>
</thead>
<tbody>
<tr>
<td>SOGFS</td>
<td>43.91 ± 1.57</td>
<td>50.64 ± 1.29</td>
<td>52.84 ± 1.88</td>
<td>53.29 ± 1.33</td>
<td>60.39 ± 1.58</td>
<td>60.31 ± 1.98</td>
<td></td>
</tr>
<tr>
<td>MVCSS</td>
<td>36.51 ± 0.84</td>
<td>42.84 ± 0.75</td>
<td>49.13 ± 1.54</td>
<td>53.09 ± 0.85</td>
<td>54.12 ± 1.34</td>
<td>57.98 ± 1.72</td>
<td></td>
</tr>
<tr>
<td>ASV</td>
<td>46.85 ± 0.94</td>
<td>53.11 ± 1.25</td>
<td>60.11 ± 1.37</td>
<td>62.70 ± 1.59</td>
<td>65.98 ± 1.45</td>
<td>68.22 ± 1.51</td>
<td></td>
</tr>
<tr>
<td>flower-17</td>
<td>NSGL</td>
<td>35.41 ± 0.86</td>
<td>44.63 ± 1.26</td>
<td>48.07 ± 1.45</td>
<td>52.22 ± 1.60</td>
<td>54.41 ± 1.54</td>
<td>58.22 ± 1.78</td>
</tr>
<tr>
<td></td>
<td>MAMFS</td>
<td>54.53 ± 1.25</td>
<td>60.05 ± 1.63</td>
<td>63.57 ± 1.58</td>
<td>64.59 ± 1.27</td>
<td>66.45 ± 1.14</td>
<td>67.83 ± 2.19</td>
</tr>
<tr>
<td></td>
<td>MGAGR</td>
<td>53.16 ± 1.61</td>
<td>59.31 ± 1.49</td>
<td>61.79 ± 1.72</td>
<td>63.58 ± 2.17</td>
<td>66.96 ± 1.65</td>
<td>68.71 ± 1.45</td>
</tr>
<tr>
<td></td>
<td>EMUFS</td>
<td>56.31 ± 1.08</td>
<td>62.92 ± 1.34</td>
<td>67.06 ± 1.76</td>
<td>69.95 ± 1.47</td>
<td>71.45 ± 1.64</td>
<td>72.34 ± 1.50</td>
</tr>
</tbody>
</table>

NUS

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Feature ratio</th>
<th>10%</th>
<th>15%</th>
<th>20%</th>
<th>25%</th>
<th>30%</th>
<th>35%</th>
</tr>
</thead>
<tbody>
<tr>
<td>SOGFS</td>
<td>19.04 ± 0.24</td>
<td>19.22 ± 0.56</td>
<td>20.53 ± 0.27</td>
<td>20.63 ± 0.24</td>
<td>19.35 ± 0.25</td>
<td>20.32 ± 0.43</td>
<td></td>
</tr>
<tr>
<td>MVCSS</td>
<td>23.93 ± 0.56</td>
<td>24.62 ± 0.80</td>
<td>25.44 ± 0.72</td>
<td>25.43 ± 0.98</td>
<td>25.98 ± 0.84</td>
<td>26.21 ± 0.58</td>
<td></td>
</tr>
<tr>
<td>ASV</td>
<td>23.08 ± 0.63</td>
<td>24.21 ± 0.79</td>
<td>23.75 ± 0.85</td>
<td>23.62 ± 0.41</td>
<td>23.33 ± 0.46</td>
<td>23.94 ± 0.59</td>
<td></td>
</tr>
<tr>
<td>flower-17</td>
<td>NSGL</td>
<td>23.16 ± 1.09</td>
<td>24.58 ± 1.47</td>
<td>25.51 ± 0.76</td>
<td>25.58 ± 0.66</td>
<td>26.00 ± 0.87</td>
<td>25.76 ± 0.90</td>
</tr>
<tr>
<td></td>
<td>MAMFS</td>
<td>23.65 ± 0.56</td>
<td>26.20 ± 0.74</td>
<td>23.18 ± 0.86</td>
<td>23.69 ± 0.22</td>
<td>33.62 ± 0.42</td>
<td>33.36 ± 0.67</td>
</tr>
<tr>
<td></td>
<td>MGAGR</td>
<td>22.35 ± 0.35</td>
<td>24.63 ± 1.05</td>
<td>24.81 ± 0.88</td>
<td>25.04 ± 0.73</td>
<td>24.81 ± 0.61</td>
<td>25.79 ± 0.51</td>
</tr>
<tr>
<td></td>
<td>EMUFS</td>
<td>25.70 ± 0.78</td>
<td>26.30 ± 0.92</td>
<td>27.31 ± 0.94</td>
<td>27.29 ± 0.69</td>
<td>27.68 ± 1.18</td>
<td>27.62 ± 1.02</td>
</tr>
</tbody>
</table>

Figure 3: Running time versus the number of samples.

be drawn: (1) Across varying numbers of selected features, EMUFS consistently exhibits competitive or superior results, fully demonstrating its effectiveness in multi-view feature selection. (2) The ACC and NMI scores achieved through the EMUFS outperform those of the single-view methodology. This underscores the assertion that harnessing distinct information from multiple views enhances the selection of informative features. (3) Compared with other multi-view feature selection methods, EMUFS consistently attains superior performance, highlighting the effectiveness of collaborative utilization of the cluster structure and the similarity structure. Meanwhile, to assess the efficiency of EMUFS, Table 3 summarizes the computation complexity of each method and Fig. 3 shows the running times versus the training sample scale on the ALOI and Youtube datasets. We find that the running times of EMUFS exhibit a linear increase, while other methods emerge an exponentially increasing trend with the increased sample scale, fully validating the efficiency of propagating cluster information on the bipartite graph.

4.3 Ablation Study

To investigate the significance of adaptive structure learning and inference, an ablation study is conducted to design three variants of EMUFS: EMUFS$_1$, wherein the similarity structure is only derived from original data, overlooking the information from projection space; EMUFS$_2$, a variation of EMUFS that excludes the collaborative fusion model, guiding feature selection by fixed graphs and membership matrices; and EMUFS$_3$ employs the unaligned membership matrices to learn the cluster structure. The results of EMUFS and its distilled versions are depicted in Fig. 4. We can conclude that: (1) The ACC of EMUFS$_1$ is inferior to EMUFS, indicating that the dynamic graph learning on selected feature space has an effective influence on the overall performance; (2) EMUFS
achieves superior results than EMUFS$_2$, validating that the collaborative fusion of similarity structure and cluster information can facilitate feature selection; (3) Comparative analyses between EMUFS and EMUFS$_3$ demonstrate that simply concatenating different membership matrices impairs the consistency of clustering centers among views.

### 4.4 Visualization

To visually evaluate the quality of selected features, the T-SNE is employed to project the selected features into a two-dimensional space [Van der Maaten and Hinton, 2008]. In this context, we select 320 samples from 20 clusters of the Leaves dataset for visualization, in which each sample has three views (i.e. View #1, #2 and #3), and each view contains 64 features. Figs. 5 (a)-(c) show the results of the original three views, while Figs. 5 (d)-(f) depict the visualizations

![Figure 4: ACC of EMUFS and its simplified versions.](image)

**Table 5:** NMI of different methods with different numbers of features. The best and second results are in bold and underlined.
of feature subsets selected by ASVW, MAMFS, and EMUFS, respectively. To ensure a fair evaluation, the number of selected features is set to 64 for each method. As depicted in Fig. 5, the visualizations of the original views exhibit significant overlaps between different clusters. In contrast, ASVW, MAMFS, and EMUFS achieve separable results, highlighting the significance of multi-view feature selection. Moreover, the inter-cluster distance obtained by EMUFS is larger than those of ASVW and MAMFS, demonstrating that EMUFS can select a compact subset of discriminative features.

4.5 Parameter Sensitivity and Convergence

EMUFS encompasses three intrinsic parameters, namely $\lambda$, $\beta$ and $\gamma$. To assess the impacts of these parameters, we present the clustering accuracies under distinct parameter settings, alongside varying numbers of selected features, in Fig. 6. This depiction shows that EMUFS can perform better, particularly when $\lambda$ assumes a value smaller than 1. This observation effectively indicates that the adaptive structure learning and fusion indeed contribute to identifying informative features. Meanwhile, to illustrate the convergence behavior of EMUFS, Fig. 7 provides the variations of the objective function values, which shows that the objective function exhibits a rapid decrease and converges to a stable value quickly.

Figure 6: ACC with different parameters on the NUS and SCENE datasets.

Figure 7: Variation curves of objective function values.

5 Conclusion

In this paper, we propose an efficient multi-view unsupervised feature selection with adaptive structure learning and inference (EMUFS). Unlike existing methods that exploit the unreliable similarity structures as well as overlook the cluster structures of data, EMUFS exploits the similarity structure from the selected feature space to alleviate the impacts of low-quality features and explores the cluster structure by membership matrices, so that the feature selection process can be effectively guided by the collaborative cooperation of structures. Moreover, EMUFS learns the cluster structure of samples from anchors via the information propagation on the bipartite graph, reducing the computational cost and strengthening the scalability for large-scale tasks. Extensive experiments demonstrate the effectiveness of EMUFS and its superiority against the state-of-the-art competitors.
Acknowledgments

This work was supported by the "Pioneer" and "Leading Goose" R&D Program of Zhejiang Province, China (Grant No. 2023C01022), the National Natural Science Foundation of China (Nos. 62306171 and 62176001), the Science and Technology Major Project of Shanxi (No. 202201020101006), the Natural Science Foundation of Chongqing (No. CSTB2022NSCQ-MSX1202).

References


