Pattern Recognition 159 (2025) 111159

Contents lists available at ScienceDirect

Pattern Recognition

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

Semi-supervised multi-view feature selection with adaptive similarity fusion and learning

Bingbing Jiang [a], Jun Liu [a], Zidong Wang [b], Chenglong Zhang [a,*], Jie Yang [c], Yadi Wang [d], Weiguo Sheng [a], Weiping Ding [e,*]

[a] School of Information Science and Technology, Hangzhou Normal University, No.2318, Yuhangtang Rd, Hangzhou 311121, Zhejiang, China
[b] Department of Computer Science, Brunel University London, Uxbridge, Middlesex UB8 3PH, Kingston Lane, London, UK
[c] Australian AI Institute, University of Technology Sydney, 15 Broadway Ultimo, Sydney, NSW 2007, Australia
[d] School of Computer and Information Engineering, Henan University, Jinming Campus, Kaifeng 475004, Henan, China
[e] School of Artificial Intelligence and Computer Science, Nantong University, No.9, Seyuan Rd, Nantong 226019, Jiangsu, China

ARTICLE INFO

Keywords:
Multi-view feature selection
Semi-supervised learning
Similarity fusion
Graph learning

ABSTRACT

Existing multi-view semi-supervised feature selection methods typically need to calculate the inversion of high-order dense matrices, rendering them impractical for large-scale applications. Meanwhile, traditional works construct similarity graphs on different views and directly fuse these graphs from the view level, ignoring the differences among samples in various views and the interplay between graph learning and feature selection. Consequently, both the reliability of graphs and the discrimination of selected features are compromised. To address these issues, we propose a novel multi-view semi-supervised feature selection with Adaptive Similarity Fusion and Learning (ASFL) for large-scale tasks. Specifically, ASFL constructs bipartite graphs for each view and then leverages the relationships between samples and anchors to align anchors and graphs across different views, preserving the complementarity and consistency among views. Moreover, an effective view-to-sample fusion manner is designed to coalesce the aligned graphs while simultaneously exploiting the neighbor structures in projection subspaces to construct the joint graph compatible across views, reducing the adverse effects of noisy features and outliers. By incorporating bipartite graph fusion and learning, label propagation, and $l_{2,0}$-norm multi-view feature selection into a unified framework, ASFL not only avoids the expensive computation in the solution procedures but also enhances the quality of selected features. An effective optimization strategy with fast convergence is developed to solve the objective function, and experimental results validate its efficiency and effectiveness over state-of-the-art methods.

1. Introduction

As information collection technologies rapidly develop, data with diverse feature representations become more and more popular in practical applications, referred to as multi-view data. Compared to single-view data, multi-view data offers a more comprehensive depiction of research objects, as different views are complementary or partly independent [1–6]. Considering the instability of external environments, data collected from diverse types of equipment inevitably contain noisy features, such that the direct use of multi-view data not only requires expensive costs but also impairs the performance of later tasks. To this end, multi-view feature selection aims to select a subset of relevant features from heterogeneous feature spaces, becoming a fundamental yet challenging paradigm in multi-view learning [7–10].

Depending on the availability of the class labels in the training procedures, existing multi-view feature selection can be conducted in three manners: supervised, unsupervised, and semi-supervised [11–14]. Among them, supervised methods need enough labeled information to guide feature selection [15–17], while unsupervised methods commonly utilize the intrinsic structure of data to select features without the guidance of label information [18–21]. In many real-world domains, there exist scarce labeled samples because of the laborious process of labeling data, whereas large amounts of unlabeled samples are easily available. Therefore, the semi-supervised manner that leverages both labeled and unlabeled samples to select discriminative features from the original feature space, has attracted widespread attention [22,23]. To select features from multi-view data with limited

* Corresponding authors.
E-mail addresses: jiangbb@hznu.edu.cn (B. Jiang), 15570471996@163.com (J. Liu), zidong.wang@brunel.ac.uk (Z. Wang), clzhang123@163.com (C. Zhang), yadiwang@henu.edu.cn (Y. Wang), w.sheng@ieee.org (W. Sheng), dwp9988@163.com (W. Ding).

https://doi.org/10.1016/j.patcog.2024.111159
Received 23 July 2024; Received in revised form 10 October 2024; Accepted 3 November 2024
Available online 12 November 2024

labeled information, the direct way is to use label propagation to predict the pseudo labels of unlabeled samples and then invoke single-view models on the concatenated features of different views. Typical methods generally learn graphs to capture the underlying similarity structures within the data and learn a sparse feature projection matrix to identify informative features [24]. Specifically, Lai et al. [25] proposed to combine label propagation and sparse projection learning for feature selection. Shi et al. [26] introduced binary label learning to enrich the label information of unlabeled samples, so as to supervise the feature selection process. However, these methods are not specifically devised for multi-view tasks, since they equally treat multiple views and discard the distinction and correlation among views, affecting the effectiveness of ultimate feature selection.

Accordingly, researchers introduced the view weights to distinguish the contributions of different views instead of simply concatenating them [27]. Specifically, Multi-view Laplacian Sparse Feature Selection (MLSFS) [28] had been proposed to construct similarity graphs independently on each view, and then introduce view weight parameters to integrate the results of label propagation on different views. To avoid the trivial solution of view weights, an exponent hyper-parameter $\rho(\rho > 1)$ is required for MLSFS to control the distribution of view weights. MLSFS imposes the $l_{2,1/2}$-norm constraint on the feature projection matrix, to learn a sparse feature projection for feature selection. Thereafter, Shi et al. [29] replaced the graph Laplacian matrix in MLSFS with the Hessian matrix to further exploit the structure information of data. To update similarity structures during feature selection, Multi-view Adaptive Semi-supervised Feature Selection (MASFS) has been proposed to exploit the distances among training samples in the original space to learn view-specific similarity graphs [30]. However, these multi-view semi-supervised feature selection methods directly derive the similarity structures of data from the original feature space that inevitably contains irrelative dimensions, such that the constructed graphs might be affected by noises and often suboptimal, ultimately leading to performance degradation. To this end, Jiang et al. [31] proposed to adaptively learn graphs from both the original feature space and the projection subspace induced by the $l_{2,1}$-norm constraint. Generally, the methods mentioned above construct the n-order graphs to mine the relations among training samples (n is the number of training samples) and apply the $l_{2,p}$-norm ($0 < p \leq 1$) regularization to feature projections for feature selection. Despite achieving some progress, they suffer from the high computation costs and the suboptimal performance of feature selection. Specifically, it takes the computational complexities of $\mathcal{O}(n^3)$ and $\mathcal{O}(n^2)$ when solving the prediction label matrix of training samples and constructing n-order similarity graphs, respectively. Therefore, these methods are challenging to tackle large-scale tasks, heavily limiting their applicability in practice. Meanwhile, the $l_{2,p}$-norm regularization cannot ensure that certain rows of feature projection matrices are 0, which means that these $l_{2,p}$-norm feature selection methods need the sparsity-induced parameters to balance the projection losses and the row-sparsity of projection matrices, increasing the burden of parameter tuning. To select important features, they have to calculate and sort the feature scores (i.e., the l_2-norm of each row for feature projection matrices), which separates the procedures of selecting features from learning projection matrices, such that subtle variations of the feature scores might result in different features, increasing the instability of feature selection.

To improve the computational efficiency, recent methods attempted to construct bipartite graphs between training samples and a small number of anchors for predicting labels [32]. For example, Chen et al. [33] proposed to pick out features and simultaneously exploit selected anchors to construct bipartite graphs in the projected subspace. However, this method is an unsupervised single-view model, which not only fails to utilize the labeled information in the training samples but also ignores the consistency and complementarity among views, making it difficult to select a discriminative subset of features. To match multi-view data, Zhang et al. [34] adaptively combined the similarity structures of different views to construct a bipartite graph for label propagation and adopted the $l_{2,1}$-norm regularization to achieve the row-sparsity of the learned feature projection. Unfortunately, the method in [34] generates anchors from the concatenated feature space of multiple views, such that the obtained anchors cannot take into account the differences among views. Moreover, the aforementioned semi-supervised multi-view methods discriminate and fuse similarity structures/graphs directly from the view level, which ignores the impacts of samples from different views on the graph construction, making the learned graphs susceptible to outliers within each view.

To break such inherent limitations, this paper proposes a novel semi-supervised multi-view feature selection method with adaptive similarity fusion and learning (ASFL), whose basic framework is illustrated in Fig. 1. The main contributions of ASFL are summarized as follows:

- By incorporating the view-to-sample similarity fusion, label propagation with bipartite graph learning and the $l_{2,0}$-norm constraint into a unified framework, we propose a novel semi-supervised multi-view feature selection method, which tactfully avoids calculating the inverse of n-order dense matrices, so as to efficiently handle large-scale multi-view data.
- Unlike previous methods, ASFL introduces an adaptive view-to-sample weight to fuse the similarities among samples from multiple views and simultaneously leverages the distances in the projection subspace induced by $l_{2,0}$-norm to learn an optimal graph, not only properly balancing different views but also effectively alleviating the adverse impacts of noisy features and outliers on the graph.
- We employ the $l_{2,0}$-norm to automatically identify salient features during learning procedures, which releases ASFL from the sparsity-induced parameters as well as the operations of sorting features, enhancing the quality and reliability of selected features.
- We develop an efficient optimization to alternately solve the formulated objective function, and extensive experiments demonstrate the effectiveness and superiority of the ASFL when tackling semi-supervised multi-view feature selection tasks.

The rest of this paper is organized as follows. Section 2 introduces the details and optimization procedures of the proposed ASFL. Section 3 conducts experiments to validate the proposed ASFL from different aspects, and Section 4 draws the conclusion and discusses possible directions in future research.

2. The proposed method

2.1. Notations

Throughout the paper, matrices and vectors are written as boldface uppercase letters and lowercase letters, respectively. For a matrix $\mathbf{M}$, $\mathrm{Tr}(\mathbf{M})$ and $\|\mathbf{M}\|_F$ denote the trace and the Frobenius norm of $\mathbf{M}$. $\|\mathbf{M}\|_{2,1} = \sum_i \|\mathbf{m}_i\|_2$ and $\|\mathbf{M}\|_{2,0} = \sum_i \|\mathbf{m}_i\|_2^0$ represent $l_{2,1}$-norm and $l_{2,0}$-norm of $\mathbf{M}$, in which $\mathbf{m}_i$ denotes the ith row of $\mathbf{M}$. Table 1 summaries the main notations in this paper.

2.2. Problem formulation

In this section, we elaborate on the crucial details of the proposed ASFL, which comprises three parts: feature selection with the $l_{2,0}$-norm constraint, adaptive view-to-sample similarity fusion and learning, and label propagation with the learned bipartite graph. Specifically, to achieve feature selection, the sparse linear regression [35,36] is widely adopted, which can be formulated as:

$$\min_{\mathbf{W},\mathbf{b}} \|\mathbf{X}^T\mathbf{W} + \mathbf{1}_n\mathbf{b}^T - \mathbf{F}\|_F^2 + \gamma R(\mathbf{W}), \tag{1}$$

where $\mathbf{W} \in \mathbb{R}^{d\times c}$ is the feature projection matrix, and $\mathbf{b} \in \mathbb{R}^{c\times 1}$ denotes the bias vector. $R(\mathbf{W})$ represents a sparse regularization term on $\mathbf{W}$ and

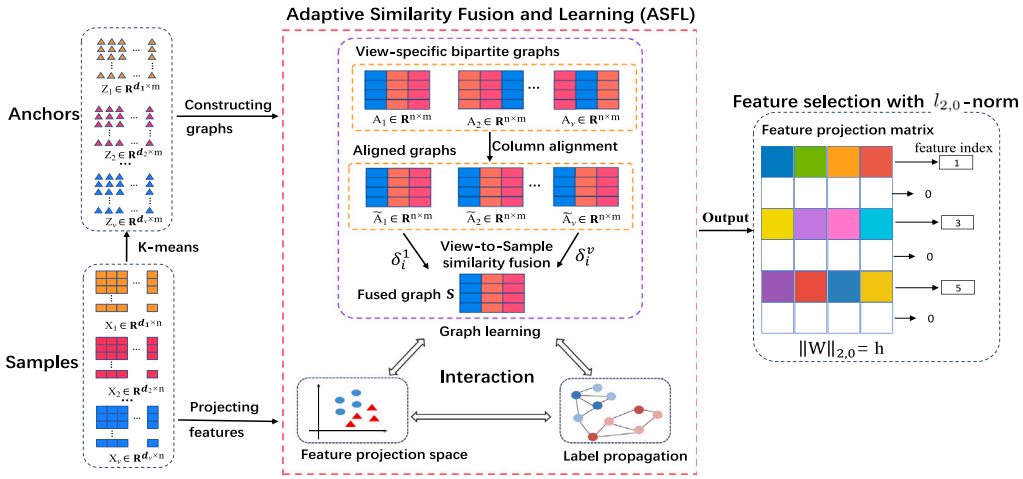


Fig. 1. Schematic illustration of ASFL. Concretely, by learning the view-specific bipartite graphs and aligning them from the column, ASFL properly ensures the complementary and consensus among views. After that, ASFL introduces an adaptive view-to-sample weight to fuse the similarities from multiple views, which effectively alleviates the undesirable influence of outliers. By incorporating bipartite graph fusion and learning, label propagation and $l_{2,0}$-norm feature selection into a unified framework, ASFL avoids the inverses of high-order dense matrices and enhances the quality of selected features ultimately.

Table 1
Description of notations.

Notation	Description
n, l	The number of training data and labeled data
c, v	The number of classes and views
$d_r, d = \sum_{r=1}^{v} d_r$	The dimension of rth view and all views
$\mathbf{x}_i^r \in \mathbb{R}^{d_r \times 1}$	The ith sample in rth view
$\mathbf{x}_i = [\mathbf{x}_i^1, \ldots, \mathbf{x}_i^v] \in \mathbb{R}^{d \times 1}$	The ith sample
$\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{d \times n}$	The concatenated feature matrix
$\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_m] \in \mathbb{R}^{d \times m}$	The generated anchors
$\mathbf{Y}_l \in \mathbb{R}^{l \times c}$	The label matrix of labeled samples
$\mathbf{Y}_n = [\mathbf{Y}_l; \mathbf{0}] \in \mathbb{R}^{n \times c}$	The label matrix
$\mathbf{F} \in \mathbb{R}^{n \times c}$	The predicted label matrix
$\mathbf{W} \in \mathbb{R}^{d \times c}$	The feature projection matrix
$\mathbf{I}_n \in \mathbb{R}^{n \times n}$	The n-order identity matrix
$\mathbf{b} \in \mathbb{R}^{c \times 1}$	The bias vector
$\mathbf{1}_c \in \mathbb{R}^{c \times 1}$	The vector with c elements being 1

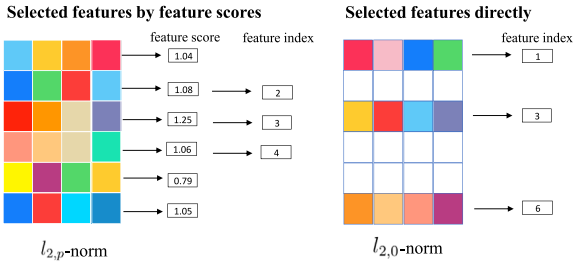


Fig. 2. Feature projection matrices obtained by $l_{2,p}$-norm and $l_{2,0}$-norm.

can be generally realized by the $l_{2,p}$-norm ($0 < p \leq 1$), thereby making $\mathbf{W}$ sparse in rows. Although achieving a certain degree of sparsity, the $l_{2,p}$-norm fails to enable several rows of $\mathbf{W}$ to be 0. Therefore, the feature selection methods based on the $l_{2,p}$-norm need to calculate the scores $\{\|\mathbf{w}_i\|_2\}_{i=1}^{d}$ ($\mathbf{w}_i$ is the ith row of $\mathbf{W}$) of all features first and then sort the feature scores by descending order, so as to select the first few features. However, the procedure of sorting features is separated from the feature projection learning, such that subtle changes of $\{\|\mathbf{w}_i\|_2\}_{i=1}^{d}$ can lead to different results, thereby selecting diverse subsets of features. Considering these factors, we directly introduce the $l_{2,0}$-norm on the $\mathbf{W}$, reformulating Eq. (1) as follows:

$$\min_{\mathbf{W},\mathbf{b}} \|\mathbf{X}^T\mathbf{W} + \mathbf{1}_n\mathbf{b}^T - \mathbf{F}\|_F^2 \quad \text{s.t.} \quad \|\mathbf{W}\|_{2,0} = h, \tag{2}$$

where h is the number of selected features. Benefiting from the $l_{2,0}$-norm, Eq. (2) can select important features automatically during training procedures, without the feature sorting operation and the sparsity-induced parameter (i.e., γ). Fig. 2 shows the feature projection matrices obtained by the $l_{2,p}$-norm regularization and the $l_{2,0}$-norm regularization, respectively. Obviously, the $l_{2,0}$-norm constraint on $\mathbf{W}$ (i.e., $\|\mathbf{W}\|_{2,0} = h$) is the optimal solution for feature selection, whereas it is difficult to solve due to the non-convex property. To solve this problem, several efforts have been made to develop primarily two optimization methods: projection gradient descent (PGD) [37] and augmented Lagrangian method (ALM) [38]. Generally, the PGD-based methods update the solution of $\mathbf{W}$ by gradient information first and then project it to the feasible region in each iteration, which is complex and time-consuming. In contrast, the ALM optimization manner is more efficient with stable solutions by introducing auxiliary variables to equivalently convert the $l_{2,0}$-norm problem into a simple problem. Therefore, we adopt the ALM method to solve the problem of $\|\mathbf{W}\|_{2,0} = h$, whose optimization details are provided in the next subsection.

In semi-supervised scenarios, the samples with true class labels are relatively limited, which motivates researchers to leverage similarity structures to enrich the labeled information of training samples [39]. However, existing methods construct graphs to capture the similarity structures and learn the predicted labels for training samples by graph-based label propagation, which takes the computational complexities of $O(n^2)$ and $O(n^3)$, respectively [40]. To tackle large-scale data, Zhang et al. [41] proposed to generate anchors on the concatenated space from multiple views and construct bipartite graphs between training samples and anchors to reduce computational complexity. However, considering that each view has specific statistical characteristics, such feature concatenation manner neglects the distinctions between different views, potentially affecting the quality of anchors [42]. To preserve the complementarity among views, we propose to generate m anchors $\{\mathbf{z}_j^r\}_{j=1}^{m}$ on each view separately and construct bipartite graphs $\{\mathbf{A}^r\}_{r=1}^{v} \in \mathbb{R}^{n \times m}$ to explore the view-specific similarity structures. We note that the anchor sets on different views might be mismatched, making view-specific graphs not aligned in columns [42]. To solve this limitation, the connection relations between training samples and anchors on the bipartite graphs $\{\mathbf{A}^r\}_{r=1}^{v}$ are effectively used. Specifically, each column of bipartite graphs records the similarities between its corresponding anchor and all samples, which can be regarded as

an n-dimensional representation of this anchor. Considering that the similarity relations between anchors and samples on different views will not be drifting with views, thus each view can be selected as the aligned standard in practice. For convenience, the first view is employed as the standard to align others by solving the following problem:

$$\min_{\mathbf{O}^r} \|\mathbf{A}^1 - \mathbf{A}^r\mathbf{O}^r\|_F^2 \quad \text{s.t.} \quad (\mathbf{O}^r)^T\mathbf{O}^r = \mathbf{I}, \mathbf{O}^r(\mathbf{O}^r)^T = \mathbf{I}, \tag{3}$$

where $\mathbf{A}^1$ and $\mathbf{A}^r$ denote the bipartite graphs for the first and the rth view, $\mathbf{O}^r$ represents the permutation matrix that aligns the rth view to the first view. Eq. (3) can be efficiently solved by the singular value decomposition (i.e., $(\mathbf{A}^1)^T\mathbf{A}^v = \boldsymbol{\Phi}\boldsymbol{\Sigma}\boldsymbol{\Psi}$), and the solution is calculated as: $\mathbf{O}^r = \boldsymbol{\Psi}\boldsymbol{\Phi}^T$. Denoting $\widetilde{\mathbf{A}}^r := \mathbf{A}^r\mathbf{O}^r$ and $\widetilde{\mathbf{Z}}^r := \mathbf{Z}^r\mathbf{O}^r$ as the aligned graphs and anchors, we propose an adaptive view-to-sample manner to fuse the cross-view similarity structures:

$$\min_{\mathbf{S}\mathbf{1}=\mathbf{1},\mathbf{S}\geq 0,\boldsymbol{\delta}\mathbf{1}=\mathbf{1},\boldsymbol{\delta}\geq 0} \sum_{i=1}^{n} \|\mathbf{s}_i - \sum_{r=1}^{v} \delta_i^r \widetilde{\mathbf{a}}_i^r\|_2^2, \tag{4}$$

where $\mathbf{S} \in \mathbb{R}^{n\times m}$ represents the learned unified similarity graph, and $\boldsymbol{\delta} = [\boldsymbol{\delta}_1; ...; \boldsymbol{\delta}_n] \in \mathbb{R}^{n\times v}$ is the view-to-sample weight matrix. It is worth emphasizing that the ith row of $\boldsymbol{\delta}$, i.e., $\boldsymbol{\delta}_i = [\delta_i^1, \ldots, \delta_i^v] \in \mathbb{R}^{1\times v}$, denotes the weights of the ith training sample in v different views, and the rth column of $\boldsymbol{\delta}$ measures the importance of n training samples in the rth view. Different from the previous studies that solely discriminate multiple views at the view level, Eq. (4) introduces the view-to-sample weights to adaptively discriminate and coalesce the similarity structures of training samples across views, such that the contribution diversity of different views and samples is taken into account, enhancing the robustness of the fused similarity structures against poor-quality views and outliers. Considering that the feature selection(projection) matrix $\mathbf{W}$ can identify discriminative features and effectively reduce the influence of noisy features, ASFL further leverages the similarity relationships in the projected feature subspace (i.e., $\mathbf{X}^T\mathbf{W}$) among samples and anchors to learn a more reliable bipartite graph $\mathbf{S}$, finally formulating the adaptive view-to-sample similarity fusion and learning model as follow:

$$\begin{aligned}\min_{\mathbf{S},\boldsymbol{\delta}} &\sum_{i=1}^{n}\sum_{j=1}^{m} \|\mathbf{W}^T\mathbf{x}_i - \mathbf{W}^T\widetilde{\mathbf{z}}_j\|_2^2 s_{ij} + \theta \sum_{i=1}^{n} \|\mathbf{s}_i - \sum_{r=1}^{v} \delta_i^r \widetilde{\mathbf{a}}_i^r\|_2^2 \\ \text{s.t.}\ \ &\mathbf{S}\mathbf{1} = \mathbf{1}, \mathbf{S} \geq 0, \boldsymbol{\delta}\mathbf{1} = \mathbf{1}, \boldsymbol{\delta} \geq 0,\end{aligned} \tag{5}$$

where θ is the regularization parameter. In Eq. (5), the aligned anchors are denoted as unlabeled samples, with the corresponding prediction label matrix represented as $\mathbf{G} \in \mathbb{R}^{m\times c}$. To obtain the predicted labels of unlabeled samples and anchors, we define an augmented graph $\mathbf{R}$ on the basis of bipartite graph $\mathbf{S}$, i.e., $\mathbf{R} = \begin{bmatrix} \mathbf{0} & \mathbf{S} \\ \mathbf{S}^T & \mathbf{0} \end{bmatrix} \in \mathbb{R}^{(n+m)\times(n+m)}$, thus the label propagation based on the learned bipartite graph is formulated as follows:

$$\min_{\mathbf{Q}} \mathrm{Tr}(\mathbf{Q}^T\mathbf{L}_R\mathbf{Q}) + \mathrm{Tr}((\mathbf{Q}-\mathbf{Y})^T\mathbf{U}(\mathbf{Q}-\mathbf{Y})), \tag{6}$$

where $\mathbf{Q} = \begin{bmatrix} \mathbf{F} \\ \mathbf{G} \end{bmatrix} \in \mathbb{R}^{(n+m)\times c}$ denotes the predicted labels of unlabeled samples and anchors, $\mathbf{L}_R = \begin{bmatrix} \mathbf{D}_s & -\mathbf{S} \\ -\mathbf{S}^T & \boldsymbol{\Lambda} \end{bmatrix}$ is the Laplacian matrix of the augmented graph $\mathbf{R}$, in which $\mathbf{D}_s$ ($\mathbf{D}_s = \mathbf{I}_n$ due to the constraint $\mathbf{S}\mathbf{1} = \mathbf{1}$) and $\boldsymbol{\Lambda}$ are diagonal matrices with their diagonal elements being the row and column sums of $\mathbf{S}$, respectively. $\mathbf{Y} = \begin{bmatrix} \mathbf{Y}_n \\ \mathbf{0} \end{bmatrix} \in \mathbb{R}^{(n+m)\times c}$ contains the known labels of training samples and anchors, and $\mathbf{U} = \begin{bmatrix} \mathbf{U}_n & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$ is an $(n+m)$-order diagonal matrix, in which the diagonal element is a large const (e.g.,10^5) if its corresponding sample is labeled and 0 otherwise. To select discriminative features, the feature selection in Eq. (2), the similarity fusion and learning model in Eq. (5) and the label propagation in Eq. (6) are incorporated into a unified framework can mutually reinforce each other, formulating our final objective function:

$$\begin{aligned}\min_{\mathbf{Q},\mathbf{W},\mathbf{S},\boldsymbol{\delta},\mathbf{b}} &\sum_{i=1}^{n}\sum_{j=1}^{m} \|\mathbf{W}^T\mathbf{x}_i - \mathbf{W}^T\widetilde{\mathbf{z}}_j\|_2^2 s_{ij} + \theta \sum_{i=1}^{n} \|\mathbf{s}_i - \sum_{r=1}^{v} \delta_i^r \widetilde{\mathbf{a}}_i^r\|_2^2 \\ &+ \lambda\mathrm{Tr}(\mathbf{Q}^T\mathbf{L}_R\mathbf{Q}) + \mathrm{Tr}((\mathbf{Q}-\mathbf{Y})^T\mathbf{U}(\mathbf{Q}-\mathbf{Y})) + \mu\|\mathbf{X}^T\mathbf{W} + \mathbf{1}_n\mathbf{b}^T - \mathbf{F}\|_F^2 \\ \text{s.t.}\ \ &\mathbf{S}\mathbf{1} = \mathbf{1}, \mathbf{S} \geq 0, \boldsymbol{\delta}\mathbf{1} = \mathbf{1}, \boldsymbol{\delta} \geq 0, \|\mathbf{W}\|_{2,0} = h,\end{aligned} \tag{7}$$

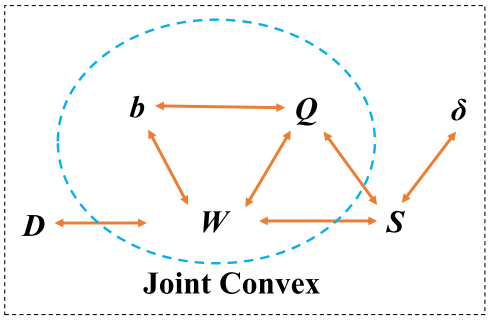


Fig. 3. The relationship between the variables in Eq. (8).

where θ, λ, and μ are regularization parameters to balance different terms. In Eq. (7), the first term utilizes the neighbor relations between samples and anchors in the projected subspace to update their similarity structures, reducing the impact of noisy features on the bipartite graph $\mathbf{S}$. The second term conducts the similarity fusion from the aspect of view-to-sample, balancing the contribution of different samples across multiple views, such that the adverse effects of outliers are largely shielded, enhancing the quality of graph $\mathbf{S}$ and effectively facilitating the ultimate feature selection. By adaptively learning a bipartite graph that conforms the consistent and complementary information among multiple views, the proposed ASFL in Eq. (7) not only reduces the cost of graph construction but also tactfully transforms the inversion of an n-order dense matrix into simple operations on low-order matrices. As a result, ASFL can address large-scale multi-view semi-supervised feature selection tasks efficiently due to its low computational complexity.

2.3. Optimization procedures

As mentioned above, Eq. (7) is difficult to solve immediately since it contains the non-convex $l_{2,0}$-norm constraint. To obtain the optimal solution, an auxiliary variable $\mathbf{D} = \mathbf{W}$ is introduced, so as to transform Eq. (7) into the following equivalent problem:

$$\begin{aligned}\min_{\mathbf{Q},\mathbf{W},\mathbf{S},\boldsymbol{\delta},\mathbf{b},\mathbf{D}} &\sum_{i=1}^{n}\sum_{j=1}^{m} \|\mathbf{W}^T\mathbf{x}_i - \mathbf{W}^T\widetilde{\mathbf{z}}_j\|_2^2 s_{ij} + \theta \sum_{i=1}^{n} \|\mathbf{s}_i - \sum_{r=1}^{v} \delta_i^r \widetilde{\mathbf{a}}_i^r\|_2^2 + \lambda\mathrm{Tr}(\mathbf{Q}^T\mathbf{L}_R\mathbf{Q}) \\ &+ \mathrm{Tr}((\mathbf{Q}-\mathbf{Y})^T\mathbf{U}(\mathbf{Q}-\mathbf{Y})) + \mu\|\mathbf{X}^T\mathbf{W} + \mathbf{1}_n\mathbf{b}^T - \mathbf{F}\|_F^2 + \frac{\eta}{2}\|\mathbf{D} - \mathbf{W} + \frac{\boldsymbol{\Pi}}{\eta}\|_F^2 \\ \text{s.t.}\ \ &\mathbf{S}\mathbf{1} = \mathbf{1}, \mathbf{S} \geq 0, \boldsymbol{\delta}\mathbf{1} = \mathbf{1}, \boldsymbol{\delta} \geq 0, \|\mathbf{D}\|_{2,0} = h,\end{aligned} \tag{8}$$

where η is the penalty parameter and $\boldsymbol{\Pi} \in \mathbb{R}^{n\times c}$ represents the Lagrangian multipliers. Although Eq. (8) is not jointly convex w.r.t. all variables, we can prove that the subproblem of $\mathbf{b}$, $\mathbf{Q}$ and $\mathbf{W}$ is convex. The detailed proof is given in Appendix. Fig. 3 depicts the logical relationships among the variables in Eq. (8), and we have developed an efficient strategy by dividing the objective function into a joint subproblem (w.r.t. $\mathbf{b}, \mathbf{W}$ and $\mathbf{Q}$) as well as another three subproblems, where the detailed steps are given as follows:

• **The subproblem w.r.t. b, Q** (i.e., $\begin{bmatrix} \mathbf{F} \\ \mathbf{G} \end{bmatrix}$) **and W**: When other variables are fixed, the problem of Eq. (8) becomes:

$$\begin{aligned}\min_{\mathbf{Q},\mathbf{W},\mathbf{b}} &\sum_{i=1}^{n}\sum_{j=1}^{m} \|\mathbf{W}^T\mathbf{x}_i - \mathbf{W}^T\widetilde{\mathbf{z}}_j\|_2^2 s_{ij} + \mu\|\mathbf{X}^T\mathbf{W} + \mathbf{1}\mathbf{b}^T - \mathbf{F}\|_F^2 \\ &+ \lambda\mathrm{Tr}(\mathbf{Q}^T\mathbf{L}_R\mathbf{Q}) + \mathrm{Tr}((\mathbf{Q}-\mathbf{Y})^T\mathbf{U}(\mathbf{Q}-\mathbf{Y})) + \frac{\eta}{2}\|\mathbf{D} - \mathbf{W} + \frac{\boldsymbol{\Pi}}{\eta}\|_F^2.\end{aligned} \tag{9}$$

Setting the derivative of Eq. (9) w.r.t. $\mathbf{b}$ to zero, the optimal solution of $\mathbf{b}$ is obtained as follows:

$$\mathbf{b} = (\mathbf{F}^T\mathbf{1} - \mathbf{W}^T\mathbf{X}\mathbf{1})/n. \tag{10}$$

Then, by substituting Eq. (10) into Eq. (9), we obtain the following subproblem:

$$\min_{\mathbf{Q},\mathbf{W}} \mathrm{Tr}(\begin{bmatrix}\mathbf{F}-\mathbf{Y}_n\\ \mathbf{G}\end{bmatrix}^T \begin{bmatrix}\mathbf{U}_n & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix} \begin{bmatrix}\mathbf{F}-\mathbf{Y}_n\\ \mathbf{G}\end{bmatrix}) + \mu\|\mathbf{X}_c^T\mathbf{W}-\mathbf{H}\mathbf{F}\|_F^2 + \lambda \mathrm{Tr}(\begin{bmatrix}\mathbf{F}\\ \mathbf{G}\end{bmatrix}^T \begin{bmatrix}\mathbf{I}_n & -\mathbf{S}\\ -\mathbf{S}^T & \boldsymbol{\Lambda}\end{bmatrix} \begin{bmatrix}\mathbf{F}\\ \mathbf{G}\end{bmatrix}), \quad (11)$$

where $\mathbf{X}_c = \mathbf{X}\mathbf{H}, \mathbf{H} = \mathbf{I}_n - \mathbf{1}\mathbf{1}^T/n$, and $\mathbf{H} = \mathbf{H}^2 = \mathbf{H}^T$. Taking the derivative of Eq. (11) w.r.t. $\mathbf{G}$ to zero, we have:

$$\boldsymbol{\Lambda}\mathbf{G} - \mathbf{S}^T\mathbf{F} = 0 \Rightarrow \mathbf{G} = \boldsymbol{\Lambda}^{-1}\mathbf{S}^T\mathbf{F}. \quad (12)$$

After that, substituting the solution of $\mathbf{G}$ into Eq. (11) and setting its derivative w.r.t. $\mathbf{F}$ to zero, the optimal $\mathbf{F}$ is calculated as:

$$\mathbf{F} = \left(\mathbf{V} - \mathbf{K}\mathbf{M}\mathbf{K}^T\right)^{-1}\mathbf{E}, \quad (13)$$

where $\mathbf{V} = (\lambda+\mu)\mathbf{I}_n + \mathbf{U}_n$ is an n-order diagonal matrix, $\mathbf{K} = [\mathbf{S}, \mathbf{1}] \in \mathbb{R}^{n\times(m+1)}$, $\mathbf{M} = \begin{bmatrix}\lambda\boldsymbol{\Lambda}^{-1} & 0\\ 0 & \mu/n\end{bmatrix} \in \mathbb{R}^{(m+1)\times(m+1)}$ and $\mathbf{E} = \mathbf{U}_n\mathbf{Y}_n + \mu\mathbf{X}_c^T\mathbf{W}$. We note that Eq. (13) still involves the inversion of an n-order dense matrix (i.e., $\mathbf{V} - \mathbf{K}\mathbf{M}\mathbf{K}^T$), requiring $\mathcal{O}(n^3)$ at least. We can reformulate the solution of $\mathbf{F}$ according to the Woodbury matrix identity[1]:

$$\mathbf{F} = \mathbf{V}^{-1}\mathbf{E} + \mathbf{V}^{-1}\mathbf{K}(\mathbf{M}^{-1} - \mathbf{K}^T\mathbf{V}^{-1}\mathbf{K})^{-1}\mathbf{K}^T\mathbf{V}^{-1}\mathbf{E}, \quad (14)$$

where the inversion of $(\mathbf{V}-\mathbf{K}\mathbf{M}\mathbf{K}^T)$ is replaced by the inverse operations on the diagonal matrix $\mathbf{V}$ and the $(m+1)$-order matrix $(\mathbf{M}^{-1}-\mathbf{K}^T\mathbf{V}^{-1}\mathbf{K})$, as well as several matrix multiplications. By calculating Eq. (14) from right to left, the computational complexity of solving $\mathbf{F}$ can be reduced from $\mathcal{O}\left(n^3\right)$ to $\mathcal{O}(n(m+1)^2 + (m+1)^3)$. Since $m \ll n$ in practical applications, the computational complexity of ASFL is linearly related to the number of training samples.

Then, substituting the solutions of $\mathbf{b}$, $\mathbf{G}$ and $\mathbf{F}$ into Eq. (11), the problem w.r.t. $\mathbf{W}$ becomes:

$$\begin{aligned}&\min_{\mathbf{W}} \mathrm{Tr}(\mathbf{W}^T(\mathbf{X}\mathbf{X}^T + \widetilde{\mathbf{Z}}\boldsymbol{\Lambda}\widetilde{\mathbf{Z}}^T - \mathbf{X}\mathbf{S}\widetilde{\mathbf{Z}}^T - \widetilde{\mathbf{Z}}\mathbf{S}^T\mathbf{X}^T + \mu\mathbf{X}_c\mathbf{X}_c^T)\mathbf{W})\\ &- \mathrm{Tr}\left(\mathbf{E}^T(\mathbf{V}-\mathbf{K}\mathbf{M}\mathbf{K}^T)^{-1}\mathbf{E}\right) + \frac{1}{2}\eta\|\mathbf{D} - \mathbf{W} + \frac{1}{\eta}\boldsymbol{\Pi}\|_F^2. \end{aligned} \quad (15)$$

Setting its derivative w.r.t. $\mathbf{W}$ to zero, we achieve the optimal solution of $\mathbf{W}$:

$$\mathbf{W} = (\mathbf{N} + \frac{\eta}{2}\mathbf{I})^{-1}\mathbf{B}, \quad (16)$$

where $\mathbf{N} = \mathbf{C} + \mu\mathbf{X}_c\mathbf{X}_c^T - \mu^2\mathbf{X}_c(\mathbf{V}-\mathbf{K}\mathbf{M}\mathbf{K}^T)^{-1}\mathbf{X}_c^T$, and $\mathbf{B} = \mu\mathbf{X}_c(\mathbf{V}-\mathbf{K}\mathbf{M}\mathbf{K}^T)^{-1}\mathbf{U}_n\mathbf{Y}_n + \frac{1}{2}\eta\mathbf{D} + \frac{1}{2}\boldsymbol{\Pi}$.

• **The subproblem w.r.t $\mathbf{D}$**: When other variables are fixed, the optimization problem for $\mathbf{D}$ becomes:

$$\min_{\|\mathbf{D}\|_{2,0}=h} \|\mathbf{D} - \mathbf{W} + \frac{\boldsymbol{\Pi}}{\eta}\|_F^2. \quad (17)$$

$\mathbf{D}$ can be directly solved by setting the $d-h$ rows of $\mathbf{W} - \frac{\boldsymbol{\Pi}}{\eta}$ with the smallest l_2-norm to zeros.

• **The subproblem w.r.t $\mathbf{S}$**: When other variables are fixed, the optimization problem for $\mathbf{S}$ is:

$$\begin{aligned}&\min_{\mathbf{S}\mathbf{1}=\mathbf{1},\mathbf{S}\geq 0} \sum_{i=1}^{n}\sum_{j=1}^{m}\|\mathbf{W}^T\mathbf{x}_i - \mathbf{W}^T\widetilde{\mathbf{z}}_j\|_2^2 s_{ij} + \theta\sum_{i=1}^{n}\|\mathbf{s}_i - \sum_{r=1}^{v}\delta_i^r\widetilde{\mathbf{a}}_i^r\|_2^2\\ &+ \lambda\sum_{i=1}^{n}\sum_{j=1}^{m}\|\mathbf{f}_i - \mathbf{g}_j\|_2^2 s_{ij}. \end{aligned} \quad (18)$$

Since the optimization of $\mathbf{S}$ is independent in rows (i.e., $\mathbf{s}_i$), thus we can solve $\mathbf{S}$ by rows:

$$\min_{\mathbf{s}_i\mathbf{1}=1,\mathbf{s}_i\geq 0} \|\mathbf{s}_i - \frac{\mathbf{p}_i}{2\theta}\|_2^2, \quad (19)$$

where $\mathbf{p}_i$ is a row vector with $p_{ij} = 2\theta\sum_{r=1}^{v}\delta_i^r a_{ij}^r - \|\mathbf{W}^T\mathbf{x}_i - \mathbf{W}^T\widetilde{\mathbf{z}}_j\|_2^2 - \lambda\|\mathbf{f}_i - \mathbf{g}_j\|_2^2$. Eq. (19) can be directly solved by an adaptive neighbor strategy, where the parameter θ can be determined according to the number of nearest anchors (i.e., k) [43].

[1] $(\hat{\mathbf{A}} - \hat{\mathbf{B}}\hat{\mathbf{C}}\hat{\mathbf{B}}^T)^{-1} = \hat{\mathbf{A}}^{-1} + \hat{\mathbf{A}}^{-1}\hat{\mathbf{B}}(\hat{\mathbf{C}}^{-1} - \hat{\mathbf{B}}^T\hat{\mathbf{A}}^{-1}\hat{\mathbf{B}})^{-1}\hat{\mathbf{B}}^T\hat{\mathbf{A}}^{-1}$

Algorithm 1 : The optimization algorithm for ASFL

Input: Multi-view data $\mathbf{X}$, labels $\mathbf{Y}_l$ of labeled samples $\mathbf{X}_l$, the number of anchors m, and the parameters λ, μ;

Initialize $\mathbf{W}$ by least squares regression on $\{\mathbf{X}_l, \mathbf{Y}_l\}$, $\boldsymbol{\delta} = 1/v$, $\mathbf{F} = [\mathbf{Y}_l; \mathbf{0}]$, $\mathbf{G} = \mathbf{0}$; $k = 10$, $\boldsymbol{\Pi} = \mathbf{0}$, $\eta = 0.01$, $\rho = 1.1$;
Generate the anchor sets $\{\mathbf{Z}^r\}_{r=1}^v$, and construct the bipartite graphs $\{\mathbf{A}^r\}_{r=1}^v$ for each view;
Learn the permutation matrices $\{\mathbf{O}^r\}_{r=1}^v$ to align graphs and anchors across different views by solving Eq. (3);
repeat
  Update $\mathbf{b}$ by Eq. (10);
  Update $\mathbf{G}$ by Eq. (12);
  Update $\mathbf{F}$ by Eq. (14);
  Update $\mathbf{W}$ by Eq. (16);
  Update $\mathbf{D}$ by solving Eq. (17);
  Update each row of $\mathbf{S}$ by solving Eq. (19);
  Update $\boldsymbol{\delta}$ by solving Eq. (23);
  Update ALM parameters $\boldsymbol{\Pi}$ and η by Eq. (24);
until the objective function of Eq. (8) converges;

Output: Feature selection matrix $\mathbf{W}$ with h non-zero rows.

• **The subproblem w.r.t $\boldsymbol{\delta}$**: When other variables are fixed, the optimization problem for $\boldsymbol{\delta}$ is:

$$\min_{\boldsymbol{\delta}\mathbf{1}=\mathbf{1}} \sum_{i=1}^{n}\|\mathbf{s}_i - \sum_{r=1}^{v}\delta_i^r\widetilde{\mathbf{a}}_i^r\|_2^2. \quad (20)$$

Considering that each row for $\boldsymbol{\delta}$ (i.e., $\boldsymbol{\delta}_i$) is independent, thus $\boldsymbol{\delta}$ can be solved by rows:

$$\min_{\boldsymbol{\delta}_i\mathbf{1}=1,\boldsymbol{\delta}_i\geq 0} \|\mathbf{s}_i - \sum_{r=1}^{v}\delta_i^r\widetilde{\mathbf{a}}_i^r\|_2^2. \quad (21)$$

Due to $\sum_{r=1}^{v}\delta_i^r = 1$, we can rewrite Eq. (21) as:

$$\|\mathbf{s}_i - \sum_{r=1}^{v}\delta_i^r\widetilde{\mathbf{a}}_i^r\|_2^2 = \|\sum_{r=1}^{v}\delta_i^r\mathbf{s}_i - \sum_{r=1}^{v}\delta_i^r\widetilde{\mathbf{a}}_i^r\|_2^2 = \|\sum_{r=1}^{v}\delta_i^r(\mathbf{s}_i - \widetilde{\mathbf{a}}_i^r)\|_2^2 = \|\boldsymbol{\delta}_i\mathbf{C}_i\|_2^2, \quad (22)$$

where $\mathbf{C}_i = [\mathbf{c}_i^1; \ldots; \mathbf{c}_i^v] \in \mathbb{R}^{v\times m}$ with $\mathbf{c}_i^r = \mathbf{s}_i - \widetilde{\mathbf{a}}_i^r$. Thus, the optimization problem for $\boldsymbol{\delta}_i$ becomes:

$$\min_{\boldsymbol{\delta}_i\mathbf{1}=1,\boldsymbol{\delta}\geq 0} \boldsymbol{\delta}_i\mathbf{C}_i\mathbf{C}_i^T\boldsymbol{\delta}_i^T. \quad (23)$$

Since $\mathbf{C}_i\mathbf{C}_i^T$ is positive semi-definite, Eq. (23) is a quadratic programming problem and can be solved efficiently by [44].

• **Update ALM parameters:** In each iteration, the penalty parameter η and the Lagrange multipliers $\boldsymbol{\Pi}$ are updated as:

$$\begin{aligned}&\boldsymbol{\Pi} = \boldsymbol{\Pi} + \eta(\mathbf{D} - \mathbf{W})\\ &\eta = \min(\rho\eta, \eta_{\max}), \end{aligned} \quad (24)$$

where $\rho = 1.1$ and $\eta_{\max} = 100$ are constants in this paper.

2.4. Complexity analysis

Algorithm 1 summarizes the optimization process of the proposed ASFL. By constructing bipartite graphs between samples and anchors, the computational complexity of ASFL is linearly related to the number of training samples. Specifically, in the initialization phase, ASFL uses k-means to generate m independent cluster centers in each view, with the computational complexity of $\mathcal{O}(ndm)$. Constructing bipartite

Table 2
The detailed information of multi-view datasets.

Dataset	Classes	Samples	Features (dimension of each view)
Reuters	6	1200	10 000(2000/2000/2000/2000/2000)
COIL20	20	1440	2801(512/420/1239/630)
Leaves	100	1600	192(64/64/64)
BDGP	5	2500	1750(1000/500/250)
Scene15	15	4485	4420(1800/1180/1240)
ALOI	100	10 800	279(77/13/64/125)
Mnist	10	70 000	459(256/144/59)
Youtube	31	95 000	832(113/322/110/287)

graphs and learning the permutation matrices cost $\mathcal{O}(nmdk)$ and $\mathcal{O}(m^3v)$, respectively. During the optimization process, updating $\mathbf{b}$ needs $O(nd)$; Calculating $\mathbf{G}$, $\mathbf{F}$ and $\mathbf{W}$ involves matrix inversions, requiring $\mathcal{O}(nmc)$, $\mathcal{O}(n(m+1)^2+(m+1)^3)$, and $\mathcal{O}(nd^2+d^3)$, respectively; Optimizing $\mathbf{S}$ costs $\mathcal{O}(nmdk)$. Besides, the optimization of $\mathbf{D}$ and δ take $\mathcal{O}(dh+dc)$ and $\mathcal{O}(nmv^2)$. Since v, c and m are small constants in practice, the overall computational complexity of the ASFL is $\mathcal{O}(n(m+1)^2+nd^2+d^3)$.

3. Experiments

In this section, comprehensive experiments are conducted to evaluate the effectiveness of the proposed ASFL in terms of multi-view semi-supervised feature selection. All experiments are implemented in Matlab R2018a and run on a Windows 10 computer with a 3.2 GHz CPU and 32 GB of RAM. To demonstrate the superiority of ASFL, we compare it with seven state-of-the-art feature selection methods, including the supervised Multi-view Sparse Feature Selection (MSFS) [45], the unsupervised Feature Selection with Bipartite Graph and l_{20}-norm Constraint (BGCFS) [41], the Semi-supervised Feature Selection with Binary Label Learning (SFS-BLL) [26], Multi-view Laplacian Sparse Feature Selection (MLSFS) [28], Multi-view Adaptive Semi-supervised Feature Selection (MASFS) [30], Efficient Multi-view Semi-supervised Feature Selection EMSFS [34] and Semi-supervised Multi-View Feature Selection (SMFS) [31].

3.1. Experimental setting

To fully validate the effectiveness of the proposed ASFL, experiments are conducted on eight real-world multi-view datasets, where Table 2 summarizes the details of the experimental datasets. In the experiments, Reuters, COIL20, Leaves, BDGP, Scene15 and ALOI are considered as small-scale datasets, while Mnist and Youtube are used as large-scale datasets. Specifically, 80% and 30% of samples are randomly selected as the training set for small-scale datasets and large-scale datasets. The number of anchors is varied based on the scale of datasets (i.e., $m = 5\% \times n$ and $m = 1\% \times n$ for small and large datasets, respectively). According to the conventions of feature selection, features selected by each method on the training set are utilized to train the linear SVM classifier (the regularization parameter is fixed as 0.1). Then, the effectiveness of selected features can be evaluated by calculating the classification accuracy of SVM on test samples represented by selected features. For a fair comparison, the parameters of other methods are tuned in the same ways as given in the respective research works. In the proposed ASFL, we tune the regularization parameters λ and μ from $\{10^{-3}, 10^{-2}, \ldots, 10^{3}\}$. Specifically, during the training stage, the supervised method (i.e., MSFS) will utilize labeled samples for training, while the unsupervised method (i.e., BGCFS) leverages the unlabeled samples. To reduce the statistical variability of experimental results, all methods are independently run 20 times on disparate training data partitions, and the average results on test samples under optimal parameter configurations are reported.

3.2. Experimental results

To compare the quality of selected features by different methods, Tables 3 and 4 record the mean and standard deviation of classification accuracy on test samples when using selected feature subsets, in which the numbers of selected features vary within the range of $\{5\%, 10\%, \ldots, 30\%\} \times d$, and the ratios of labeled samples in training sets are fixed at 30% for small-scale datasets and 3% for large-scale datasets. The best and second-best results are indicated in bold and underlined, respectively. "OM" (Out of Memory) denotes the out-of-memory error during the training process. From Tables 3 and 4, we can conclude that: (1) With the varied numbers of selected features, ASFL consistently attains competitive results, validating its effectiveness in multi-view feature selection tasks. (2) ASFL significantly outperforms the supervised method and the unsupervised method (i.e., MSFS and BGCFS), indicating that comprehensively exploiting the discriminative label information and the similarity structure of samples indeed facilitates the feature selection. (3) Compared to the single-view semi-supervised feature selection method (i.e., SFS-BLL), multi-view methods achieve superior performance, demonstrating that utilizing consistent and complementary information among multiple views can enhance the discriminative of selected features. Moreover, ASFL surpasses other methods in most cases, highlighting the significance of learning view-to-sample weights to integrate graphs as well as exploiting the similarity structures in the projected subspace.

Further, to evaluate the effectiveness of ASFL under different numbers of labeled samples, we fix the number of selected features at $20\% \times d$ and conduct experiments on small-scale datasets (i.e., Reuters, COIL20, Leaves, BDGP, Scene15 and ALOI) and large-scale datasets (i.e., Mnist and Youtube) by varying the ratio of labeled samples as $\{10\%, 20\%, 30\%, 40\%\}$ and $\{1\%, 2\%, 3\%, 4\%\}$, respectively. Fig. 4 records the experimental results of all methods with different ratios of labeled samples, and the following conclusion can be drawn: As the number of labeled samples increases, the performance of ASFL shows significant improvements, which highlights that utilizing the discriminative label information indeed facilitates the feature selection process. Meanwhile, in most situations, ASFL achieves superior results than other feature selection methods, validating the effectiveness of incorporating the view-to-sample similarity fusion, bipartite graph learning and the $l_{2,0}$-norm constraint into a unified framework. Additionally, when handling large-scale datasets (i.e., Mnist and Youtube), most semi-supervised feature selection methods (i.e., SFS-BLL, MLSFS, MASFS and SMFS) involve the inversion of n-order dense matrices, resulting in the OM issues during the optimization procedure. For further analysis, Table 5 displays the computational complexities of ASFL and its competitors, revealing that the computational complexity of most semi-supervised feature selection methods scales cubically with the number of training samples, while ASFL scales linearly with the number of training samples, making it more efficient to address large-scale tasks. To visually evaluate the computational efficiency of the proposed ASFL, Fig. 5 shows the running time of each feature selection method versus the proportion of training samples. We can find that as the number of the training samples increases, the running time of ASFL and EMSFS exhibits a different growth trend compared to other semi-supervised feature selection methods, indicating that constructing the bipartite graph to measure the similarity between training samples and anchors can effectively improve the computational efficiency. Moreover, although EMSFS can address larger-scale datasets, it overlooks the impact of outliers within multiple views, thus resulting in poor performance. These results fully validate the superiority of the proposed ASFL compared to other methods.

3.3. Quantitative analysis of ASFL

Visualization: To visually demonstrate the effectiveness of the proposed ASFL, t-SNE [46] is employed to project the selected features on

Table 3
Results (ACC ± STD %) on testing samples with various numbers of selected features.

Datasets	Methods	5%*d	10%*d	15%*d	20%*d	25%*d	30%*d
COIL20	MSFS [45]	95.00 ± 1.01	97.50 ± 0.72	98.19 ± 0.94	98.65 ± 0.97	98.68 ± 1.10	98.79 ± 0.91
	BGCFS [41]	94.20 ± 1.57	97.19 ± 1.05	97.64 ± 0.76	98.09 ± 0.98	98.51 ± 1.08	98.61 ± 0.88
	SFS-BLL [26]	95.28 ± 1.44	97.47 ± 1.15	98.37 ± 0.58	98.58 ± 0.57	98.78 ± 0.52	98.96 ± 0.52
	MLSFS [28]	97.33 ± 1.35	98.44 ± 1.06	98.72 ± 0.84	98.89 ± 0.62	98.92 ± 0.55	99.03 ± 0.71
	MASFS [30]	96.42 ± 1.51	97.97 ± 1.42	98.28 ± 1.31	98.65 ± 0.96	98.70 ± 1.06	98.68 ± 1.14
	EMSFS [34]	95.97 ± 1.68	97.69 ± 1.01	98.24 ± 0.83	98.66 ± 0.84	98.72 ± 0.88	98.82 ± 0.82
	SMFS [31]	**97.57** ± 1.11	98.65 ± 0.60	99.09 ± 0.71	**99.31** ± 0.45	99.26 ± 0.64	99.22 ± 0.74
	ASFL	97.50 ± 1.01	**98.96** ± 0.96	**99.17** ± 0.55	99.24 ± 0.61	**99.38** ± 0.57	**99.31** ± 0.60
Reuters	MSFS [45]	55.73 ± 2.18	61.46 ± 2.68	63.54 ± 3.33	64.64 ± 2.31	64.27 ± 1.48	64.95 ± 2.14
	BGCFS [41]	49.00 ± 2.73	55.42 ± 1.69	58.79 ± 3.02	59.00 ± 2.87	62.08 ± 2.85	63.25 ± 2.19
	SFS-BLL [26]	56.63 ± 2.16	62.17 ± 2.36	64.13 ± 2.93	64.92 ± 3.60	64.79 ± 1.81	65.83 ± 2.11
	MLSFS [28]	59.63 ± 2.80	**62.79** ± 1.84	62.33 ± 2.63	64.83 ± 3.03	65.63 ± 2.97	66.58 ± 2.83
	MASFS [30]	59.46 ± 2.82	60.67 ± 1.80	61.33 ± 2.53	63.96 ± 3.14	65.38 ± 2.60	65.96 ± 2.44
	EMSFS [34]	58.92 ± 2.92	62.08 ± 2.95	63.75 ± 2.51	64.17 ± 2.47	66.25 ± 2.19	67.50 ± 1.27
	SMFS [31]	60.88 ± 2.51	62.50 ± 1.96	64.29 ± 1.89	**65.71** ± 1.55	67.21 ± 1.28	67.54 ± 1.54
	ASFL	**61.17** ± 1.97	62.58 ± 2.99	**64.58** ± 2.36	65.50 ± 1.85	**67.25** ± 1.95	**67.83** ± 2.65
Leaves	MSFS [45]	17.95 ± 2.78	37.14 ± 3.96	49.69 ± 4.11	58.03 ± 3.24	63.43 ± 3.19	70.31 ± 3.37
	BGCFS [41]	16.56 ± 2.91	35.63 ± 4.13	50.81 ± 2.95	58.44 ± 3.64	65.00 ± 3.42	67.19 ± 2.50
	SFS-BLL [26]	18.21 ± 3.13	37.50 ± 2.91	51.63 ± 2.46	60.31 ± 2.50	67.18 ± 3.28	70.63 ± 3.15
	MLSFS [28]	18.47 ± 2.67	40.12 ± 2.10	55.12 ± 4.32	65.81 ± 3.15	73.09 ± 4.10	77.19 ± 3.03
	MASFS [30]	19.25 ± 3.49	40.66 ± 4.49	55.86 ± 3.64	67.37 ± 3.51	73.50 ± 3.62	78.00 ± 3.30
	EMSFS [34]	20.34 ± 2.19	48.06 ± 2.93	60.47 ± 2.58	68.59 ± 4.03	73.42 ± 4.59	**79.41** ± 4.22
	SMFS [31]	20.47 ± 1.39	42.34 ± 2.25	56.37 ± 3.69	66.58 ± 3.41	73.81 ± 2.10	78.72 ± 2.00
	ASFL	**22.13** ± 2.97	**49.34** ± 3.01	**61.81** ± 4.03	**69.97** ± 2.50	**74.56** ± 2.68	78.96 ± 2.56
BDGP	MSFS [45]	68.02 ± 1.98	71.38 ± 1.49	73.50 ± 0.61	75.86 ± 1.22	78.32 ± 1.16	79.12 ± 1.54
	BGCFS [41]	67.44 ± 2.28	72.98 ± 2.79	74.08 ± 1.16	75.80 ± 2.11	77.60 ± 3.54	78.40 ± 1.78
	SFS-BLL [26]	69.06 ± 1.42	73.88 ± 1.50	75.98 ± 2.12	76.82 ± 1.96	79.52 ± 1.97	80.54 ± 1.63
	MLSFS [28]	72.66 ± 2.61	77.24 ± 2.07	78.94 ± 1.88	80.72 ± 1.99	81.58 ± 1.43	82.54 ± 1.89
	MASFS [30]	73.44 ± 2.44	77.36 ± 2.32	79.26 ± 2.34	80.88 ± 2.43	82.22 ± 2.11	82.90 ± 2.01
	EMSFS [34]	76.06 ± 2.44	**78.84** ± 2.32	80.18 ± 2.34	81.36 ± 2.43	82.54 ± 2.11	83.38 ± 2.01
	SMFS [31]	74.78 ± 1.61	78.42 ± 1.37	**80.66** ± 1.48	**81.62** ± 1.72	82.80 ± 1.76	83.46 ± 1.04
	ASFL	**76.18** ± 1.84	78.72 ± 1.71	80.24 ± 1.69	81.46 ± 2.06	**83.22** ± 1.94	**83.65** ± 1.55

Table 4
Results (ACC ± STD %) on testing samples with various numbers of selected features.

Datasets	Methods	5%*d	10%*d	15%*d	20%*d	25%*d	30%*d
Scene15	MSFS [45]	54.40 ± 1.56	62.98 ± 1.09	65.89 ± 1.46	69.01 ± 1.06	69.34 ± 1.08	69.68 ± 1.13
	BGCFS [41]	46.26 ± 1.58	60.16 ± 1.71	63.88 ± 0.87	66.41 ± 0.61	68.04 ± 1.90	68.56 ± 1.93
	SFS-BLL [26]	63.61 ± 1.20	67.80 ± 1.87	69.90 ± 3.09	70.96 ± 2.63	71.79 ± 2.70	72.49 ± 2.64
	MLSFS [28]	67.29 ± 1.33	70.94 ± 1.11	71.58 ± 1.25	72.07 ± 1.27	72.56 ± 1.03	72.61 ± 1.07
	MASFS [30]	67.90 ± 2.45	70.96 ± 2.10	72.02 ± 1.91	72.47 ± 1.80	72.62 ± 1.86	72.60 ± 1.71
	EMSFS [34]	70.33 ± 2.27	71.29 ± 1.99	72.15 ± 1.25	72.17 ± 1.63	72.80 ± 1.60	73.65 ± 1.93
	SMFS [31]	69.23 ± 1.30	**72.60** ± 1.38	72.27 ± 1.16	72.74 ± 0.99	73.26 ± 1.10	73.43 ± 1.08
	ASFL	**71.91** ± 1.37	72.20 ± 1.89	**72.49** ± 2.33	**72.98** ± 1.93	**73.98** ± 1.99	**74.05** ± 1.86
ALOI	MSFS [45]	17.64 ± 3.82	30.69 ± 2.06	36.81 ± 2.27	46.55 ± 2.51	47.45 ± 1.65	49.68 ± 1.90
	BGCFS [41]	17.91 ± 5.17	29.02 ± 3.15	36.62 ± 5.51	45.92 ± 6.64	51.66 ± 4.90	53.51 ± 5.72
	SFS-BLL [26]	18.20 ± 1.32	31.56 ± 2.24	39.15 ± 1.55	46.85 ± 2.19	51.94 ± 2.14	55.13 ± 1.79
	MLSFS [28]	26.34 ± 2.02	35.94 ± 1.59	45.37 ± 1.76	50.46 ± 2.03	54.55 ± 3.91	59.53 ± 3.61
	MASFS [30]	26.68 ± 2.48	36.04 ± 2.64	45.68 ± 3.14	54.42 ± 3.01	60.61 ± 4.75	63.14 ± 4.93
	EMSFS [34]	20.90 ± 1.08	43.16 ± 0.93	55.89 ± 1.04	**62.15** ± 0.94	65.63 ± 0.80	67.05 ± 2.92
	SMFS [31]	22.53 ± 1.62	33.51 ± 1.86	44.47 ± 1.88	57.07 ± 4.04	**66.50** ± 2.51	67.57 ± 2.07
	ASFL	**34.03** ± 1.84	**53.43** ± 2.60	**57.98** ± 2.04	61.54 ± 1.67	65.95 ± 1.52	**67.75** ± 1.29
Mnist	MSFS [45]	61.79 ± 2.37	70.18 ± 0.83	73.80 ± 0.82	75.15 ± 0.87	76.21 ± 1.00	76.96 ± 1.13
	BGCFS [41]	58.04 ± 2.59	69.98 ± 1.11	72.91 ± 1.09	73.98 ± 0.72	75.40 ± 1.16	76.74 ± 0.64
	SFS-BLL [26]	OM	OM	OM	OM	OM	OM
	MLSFS [28]	OM	OM	OM	OM	OM	OM
	MASFS [30]	OM	OM	OM	OM	OM	OM
	EMSFS [34]	67.94 ± 2.26	74.91 ± 1.09	75.54 ± 1.22	77.16 ± 0.98	77.68 ± 0.86	79.03 ± 0.60
	SMFS [31]	OM	OM	OM	OM	OM	OM
	ASFL	**68.87** ± 0.98	**74.58** ± 1.03	**76.69** ± 0.74	**78.01** ± 0.51	**78.63** ± 0.52	**79.26** ± 0.44
Youtube	MSFS [45]	19.19 ± 1.72	25.02 ± 2.81	29.52 ± 3.32	32.97 ± 2.52	37.29 ± 2.20	40.06 ± 2.48
	BGCFS [41]	21.01 ± 4.05	27.31 ± 4.52	37.79 ± 3.22	47.00 ± 3.56	49.42 ± 3.89	53.29 ± 3.81
	SFS-BLL [26]	OM	OM	OM	OM	OM	OM
	MLSFS [28]	OM	OM	OM	OM	OM	OM
	MASFS [30]	OM	OM	OM	OM	OM	OM
	EMSFS [34]	42.45 ± 1.90	61.08 ± 1.99	70.67 ± 1.93	74.69 ± 1.98	**76.06** ± 2.06	76.42 ± 1.87
	SMFS [31]	OM	OM	OM	OM	OM	OM
	ASFL	**63.07** ± 0.89	**70.89** ± 1.28	**73.95** ± 1.81	**75.57** ± 1.40	75.94 ± 1.40	**76.87** ± 1.21

the subset of ALOI (750 samples from 15 classes) into two-dimensional space. Fig. 6 shows the results when each method selects $20\% \times d$ features. We can observe that the feature subset selected by ASFL can effectively separate samples from different classes (see Fig. 6(h)),

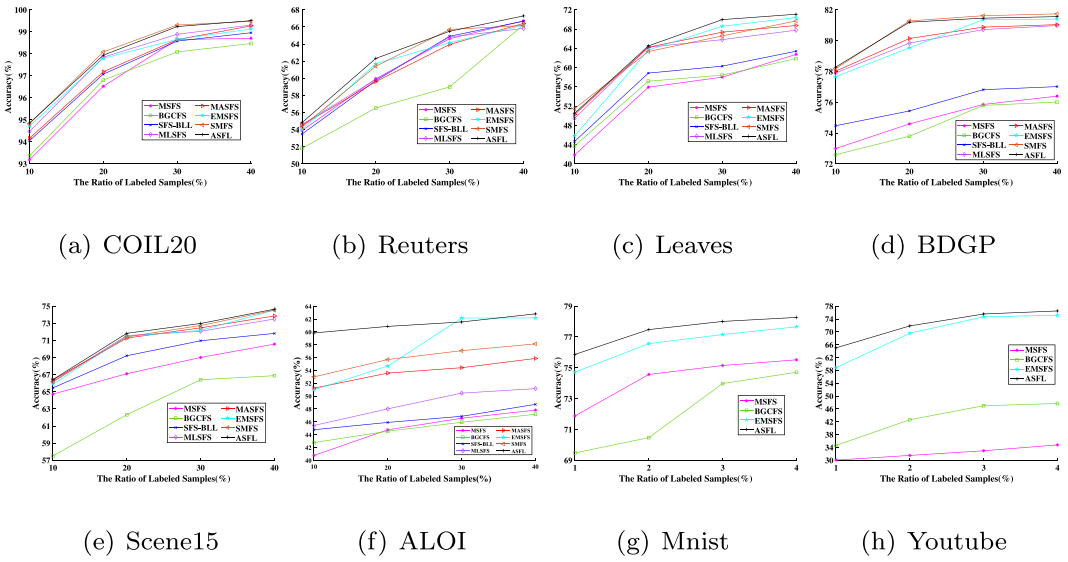


Fig. 4. Results of each method on the testing data with different numbers of labeled samples.

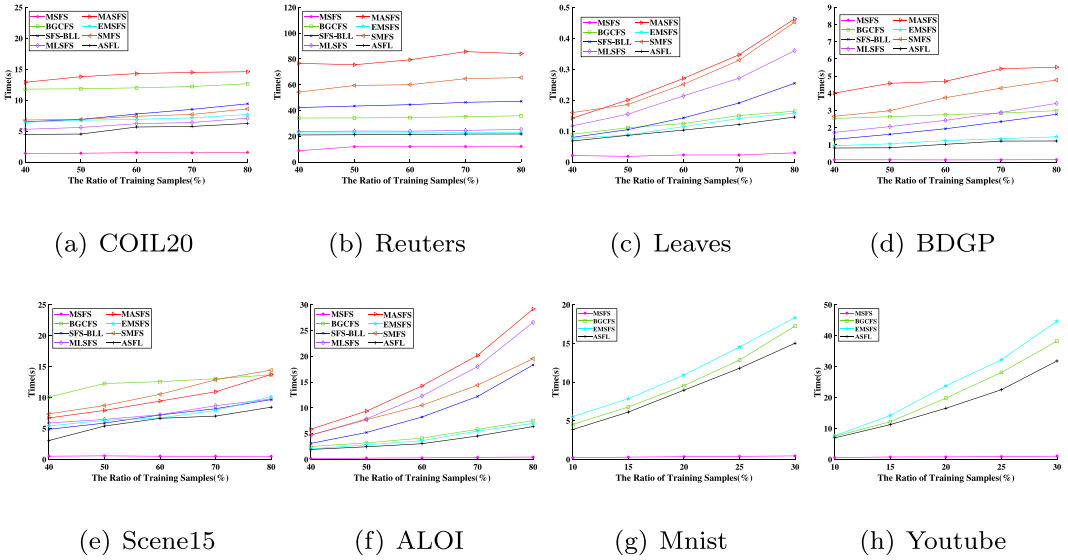


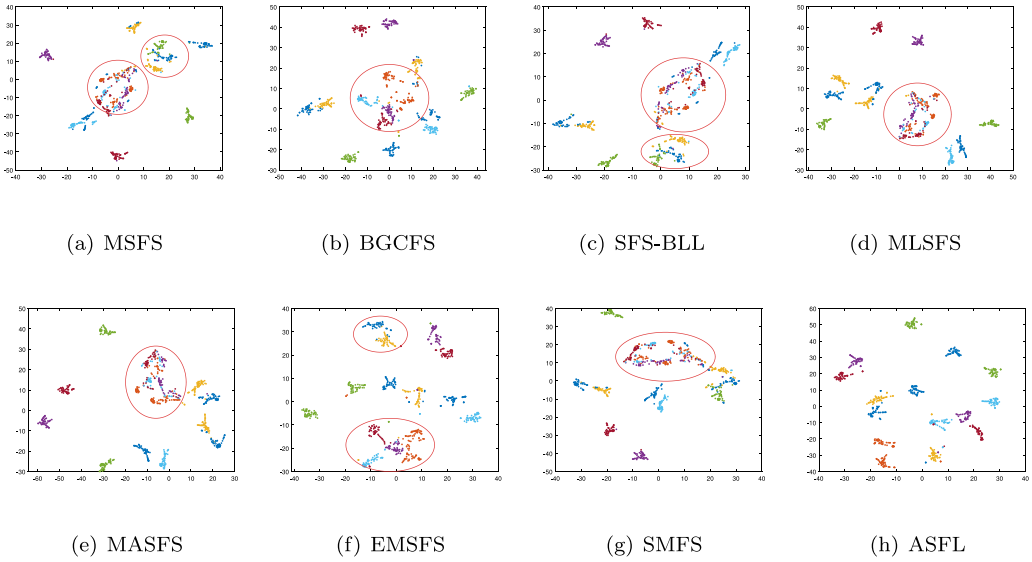
Fig. 5. Running time of ASFL and other methods versus the number of training samples.

Table 5
The computational complexity of semi-supervised feature selection methods.

Methods	Computational costs
SFS-BLL [26](2023)	$\mathcal{O}(n^3 + n^2d + d^3)$
MLSFS [28](2015)	$\mathcal{O}(n^3 + n^2d + d^3)$
MASFS [30](2020)	$\mathcal{O}(n^3 + n^2d + d^3)$
EMSFS [34](2023)	$\mathcal{O}(nm^2 + n^2d + d^3)$
SMFS [31](2024)	$\mathcal{O}(n^3 + n^2d + d^3)$
ASFL	$\mathcal{O}(n(m+1)^2 + nd^2 + d^3)$

whereas other methods show varying degrees of overlap among samples (overlaps are emphasized with red circles). This indicates that ASFL can select discriminative features by leveraging the similarity structure in the feature projection space as well as learning view-to-sample weights to coalesce aligned graphs, thereby improving the performance of subsequent classification tasks.

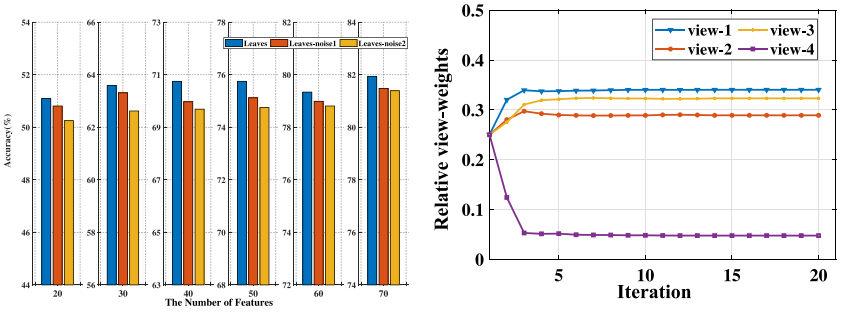
Robustness: To investigate the robustness of the proposed ASFL to noisy features, noisy views, and outlier samples within views, the robustness experiments are conducted by introducing noise into Leaves. Specifically, two variants of Leaves are designed: Leaves-noise-1, where 64 Gaussian noise features are added to view-3 of the Leaves; Leaves-noise-2 creates an additional view that contains 64 Gaussian noise features (i.e., one entirely noisy view named view-4). Fig. 7(a) shows the experimental results of ASFL across different numbers of selected features, in which the performance of ASFL on the noisy datasets (i.e., Leaves-noise-1 and Leaves-noise-2) remains comparable to the original Leaves dataset, indicating that ASFL can adaptively discriminate the noisy features and low-quality views during optimization to enhance the effectiveness of the final feature selection. Meanwhile, to demonstrate the robustness of ASFL for noisy views, we sum the columns of the view-to-sample weight matrix $\boldsymbol{\delta}$ to obtain the weight w_r for each view ($w_r = \sum_{i=1}^{n} \delta_i^r$), and then calculate the relative weight factor $\hat{w}_j = \frac{\hat{w}_j}{\sum_{r=1}^{v} \hat{w}_r}$. Fig. 7(b) displays the variation of $\hat{w}_j$ over iterations on the Leaves-noise-2, which verifies that the weight of the noisy view (view-4) decreases significantly in the initial iterations and then stabilizes at a small constant, highlighting that ASFL can reduce the impact of noisy views by adaptively learning view-to-sample weights. Subsequently, to analyze the influence of outliers, we create another



(a) MSFS (b) BGCFS (c) SFS-BLL (d) MLSFS

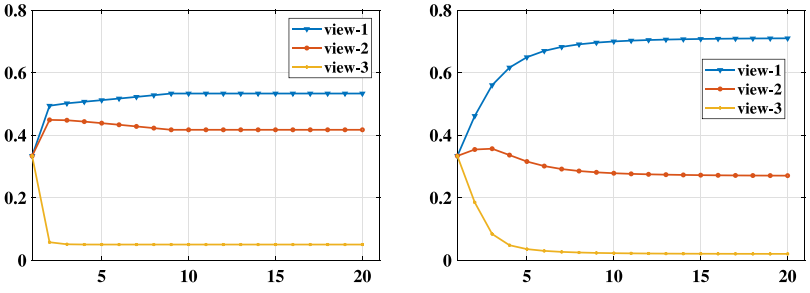
(e) MASFS (f) EMSFS (g) SMFS (h) ASFL

Fig. 6. The visualizations of selected features by different methods on ALOI.



(a) Accuracy (b) Relative view weight

Fig. 7. Robustness experiments of ASFL on the Leaves dataset.



(a) Weights of single sample (b) Average weights of all samples

Fig. 8. The sample weights versus the number of iterations on Leaves-noise-3.

variant named Leaves-noise-3, where samples in view-3 are outliers (all features set to −10). As shown in Fig. 8, Fig. 8(a) shows the weights on a single sample, while Fig. 8(b) exhibits the average weights of all samples, demonstrating that the proposed ASFL can effectively discriminate different views/samples and assign appropriate weights for them, thereby reducing the adverse effects of low-quality samples and facilitating the subsequent tasks.

Ablation Study: To validate the effectiveness of the proposed adaptive graph fusion and learning model, the ablation experiments are conducted on eight datasets. Specifically, three variants of ASFL are compared: ASFL-1 directly fuses the similarity graphs that are constructed from original space; ASFL-2 solely considers the structure information in the projected feature subspace and fuses the similarity graphs at the view level; ASFL-3, which excludes the graph fusion model and obtains the consistent graph solely from the projected

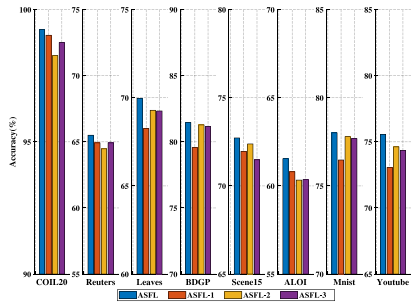


Fig. 9. Accuracy of ASFL and three variants when selecting 20%*d features.

subspace. Fig. 9 shows the experimental results of ASFL and its two variants when selecting $20\% \times d$ features (with 3% and 30% labeled ratio on training samples for large-scale and small-scale datasets, respectively). It can be observed that ASFL significantly outperforms its variants, confirming that simultaneously learning the adaptive view-to-sample weights to fuse similarity graphs across views, as well as exploiting similarity structures in the projected subspace, indeed enhances the quality of the similarity graphs, which is beneficial for selecting the effective feature subset.

Effect of the Anchor Number: To analyze the impact of the number of anchors on the performance and efficiency, we conduct experiments on ALOI and Mnist, where the number of anchors varied in the ranges $\{1\%, 2\%, \ldots, 9\%\} \times n$ and $\{0.2\%, 0.4\%, \ldots, 1.8\%\} \times n$, respectively. Fig. 10 illustrates the results, where Fig. 10(a) and (c) show the relationship between the performance and the number of anchors, and Fig. 10 (b) and (d) represent the time versus the number of anchors. We can observe that the performance of ASFL remains stable across varying numbers of anchors, whereas the runtime gradually increases with the number of anchors, demonstrating that generating an appropriate number of anchors not only ensures the effectiveness of the selected features but also enhances computational efficiency.

3.4. Parameter sensitivity and convergence analysis

The proposed ASFL encompasses two regularization parameters: λ and μ, in which λ controls the smoothness of the predicted label matrix, and μ balances the importance of the regression loss. To assess the impact of these parameters on the ASFL, we fix one parameter and record the accuracy as the other parameter varies within the range of $\{10^{-3}, 10^{-2}, \cdots, 10^{3}\}$. Fig. 11 shows the parameter sensitivity of ASFL on the Leaves and BDGP datasets. As depicted in Fig. 11, when fixing the number of selected features, the performance of ASFL first enhances and then decreases with the increase of λ. Additionally, ASFL shows superior performance when μ is less than 1, indicating that the integration of graph learning and label propagation can effectively enrich the label information and further guide the feature selection. Moreover, we experimentally analyze the convergence of ASFL, and Fig. 12 presents the variation of the objective function value versus the number of iterations. As seen in Fig. 12, the objective function converges quickly within a few iterations across all datasets, demonstrating the convergence and the effectiveness of the optimization strategy.

4. Conclusion

This paper proposes a novel multi-view semi-supervised feature selection with adaptive similarity fusion and learning (ASFL) to break the inherent limitations of existing methods. By learning the bipartite graphs between samples and anchors, ASFL not only avoids directly computing the inverse of the n-order dense matrix during the solution process but also reduces the computational cost of graph learning. Therefore, the computational complexity of ASFL is $\mathcal{O}(n(m+1)^2 + nd^2 + d^3)$, facilitating the processing of large-scale feature selection tasks. Moreover, unlike existing multi-view semi-supervised feature selection methods, ASFL devises an adaptive view-to-sample fusion manner to coalesce the aligned graphs while simultaneously exploiting the similarity structures in the projected subspace, effectively reducing the interference of noisy features and outliers in graph learning. Extensive experiments validate the effectiveness and superiority of ASFL in multi-view feature selection tasks.

Despite the advancements of the proposed ASFL method, some potential directions can be considered in the future: (1) extending ASFL to address unsupervised multi-view feature selection problems without the label information; (2) exploring how to select the optimal view from multi-view data to achieve efficient multi-view data fusion.

CRediT authorship contribution statement

Bingbing Jiang: Writing – review & editing, Writing – original draft, Methodology. **Jun Liu:** Validation. **Zidong Wang:** Writing – review & editing, Supervision. **Chenglong Zhang:** Writing – review & editing, Visualization, Software. **Jie Yang:** Writing – review & editing. **Yadi Wang:** Writing – review & editing. **Weiguo Sheng:** Writing – review & editing, Supervision. **Weiping Ding:** Writing – review & editing, Supervision.

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 is supported in part by the "Pioneer" and "Leading Goose" R&D Program of Zhejiang Province, China (Grant No. 2023C01022), the National Natural Science Foundation of China (Grant Nos. 62106066 and 62006065), the Henan Province Science Foundation of Excellent Young Scholars (Grant No. 242300421171), and the Key Research and Promotion Projects in Henan Province (Grant No. 232102110301).

Appendix. Proof of the joint convexity of b, Q and W

Proof. Fixing other variables, the problem of Eq. (8) w.r.t $\mathbf{b}$, $\mathbf{Q}$ and $\mathbf{W}$ is:

$$\min_{\mathbf{Q},\mathbf{W},\mathbf{b}} \sum_{i=1}^{n}\sum_{j=1}^{m} \|\mathbf{W}^T\mathbf{x}_i - \mathbf{W}^T\widetilde{\mathbf{z}}_j\|_2^2 s_{ij} + \mu\|\mathbf{X}^T\mathbf{W} + \mathbf{1}_n\mathbf{b}^T - \mathbf{F}\|_F^2 + \lambda \mathrm{Tr}(\mathbf{Q}^T\mathbf{L}_R\mathbf{Q}) + \mathrm{Tr}((\mathbf{Q}-\mathbf{Y})^T\mathbf{U}(\mathbf{Q}-\mathbf{Y})) + \frac{\eta}{2}\|\mathbf{D} - \mathbf{W} + \frac{\boldsymbol{\Pi}}{\eta}\|_F^2.$$

For convenience, we denote the above equation as T($\mathbf{b}$, $\mathbf{G}$, $\mathbf{F}$, $\mathbf{W}$). According to the definition of matrix trace, we have:

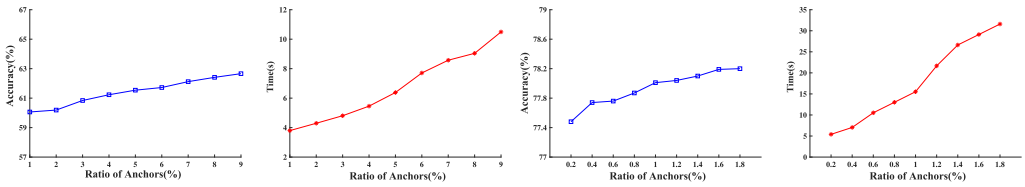
$$T(\mathbf{b},\mathbf{G},\mathbf{F},\mathbf{W}) = \mathrm{Tr}\left(\begin{bmatrix}\mathbf{b}^T\\ \mathbf{G}\\ \mathbf{F}\\ \mathbf{W}\end{bmatrix}^T \boldsymbol{\Gamma} \begin{bmatrix}\mathbf{b}^T\\ \mathbf{G}\\ \mathbf{F}\\ \mathbf{W}\end{bmatrix}\right) - \mathrm{Tr}\left(\begin{bmatrix}\mathbf{b}^T\\ \mathbf{G}\\ \mathbf{F}\\ \mathbf{W}\end{bmatrix}^T \begin{bmatrix}\mathbf{0}\\ \mathbf{0}\\ 2\mathbf{U}_n\mathbf{Y}_n\\ \eta\mathbf{D}+\boldsymbol{\Pi}\end{bmatrix}\right),$$

where

$$\boldsymbol{\Gamma} = \begin{bmatrix} \mu\mathbf{1}_n^T\mathbf{1}_n & \mathbf{0} & -\mu\mathbf{1}_n^T & \mu\mathbf{1}_n^T\mathbf{X}^T \\ \mathbf{0} & \lambda\boldsymbol{\Lambda} & -\lambda\mathbf{S}^T & \mathbf{0} \\ -\mu\mathbf{1}_n & -\lambda\mathbf{S} & (\lambda+\mu)\mathbf{I}_n + \mathbf{U}_n & -\mu\mathbf{X}^T \\ \mu\mathbf{X}\mathbf{1}_n & \mathbf{0} & -\mu\mathbf{X} & \mathbf{C} + \mu\mathbf{X}\mathbf{X}^T + \frac{\eta}{2}\mathbf{I}_d \end{bmatrix},$$

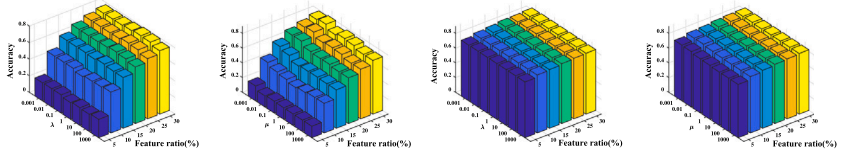
and $\mathbf{C} = \begin{bmatrix}\mathbf{X}\ \widetilde{\mathbf{Z}}\end{bmatrix}\begin{bmatrix}\mathbf{I}_n & -\mathbf{S}\\ -\mathbf{S}^T & \boldsymbol{\Lambda}\end{bmatrix}\begin{bmatrix}\mathbf{X}\ \widetilde{\mathbf{Z}}\end{bmatrix}^T$. For an arbitrary vector $\mathbf{v} = [v_1; \mathbf{v}_2; \mathbf{v}_3; \mathbf{v}_4] \in \mathbb{R}^{(1+m+n+d)\times 1}$, where $v_1 \in \mathbb{R}^{1\times 1}$, $\mathbf{v}_2 \in \mathbb{R}^{m\times 1}$, $\mathbf{v}_3 \in \mathbb{R}^{n\times 1}$, $\mathbf{v}_4 \in \mathbb{R}^{d\times 1}$, we have:

$$\mathbf{v}^T\boldsymbol{\Gamma}\mathbf{v} = \mu(v_1\mathbf{1}_n - \mathbf{v}_3 + \mathbf{X}^T\mathbf{v}_4)^T(v_1\mathbf{1}_n - \mathbf{v}_3 + \mathbf{X}^T\mathbf{v}_4)$$



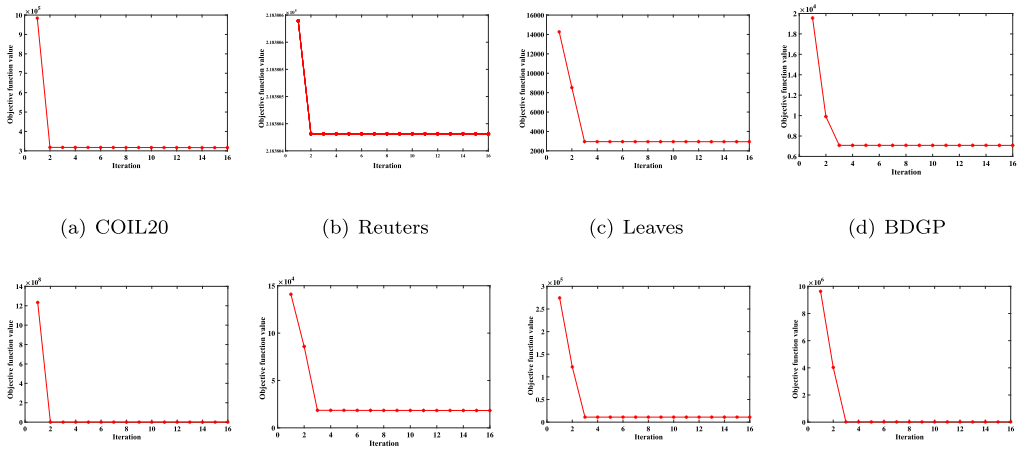
(a) Accuracy on ALOI (b) Time on ALOI (c) Accuracy on Mnist (d) Time on Mnist

Fig. 10. Average classification accuracy and running time of ASFL versus the numbers of anchors on the ALOI and Mnist datasets.



(a) Leaves-λ (b) Leaves-μ (c) BDGP-λ (d) BDGP-μ

Fig. 11. Accuracy of ASFL with different parameters on Leaves and BDGP.



(a) COIL20 (b) Reuters (c) Leaves (d) BDGP

(e) Scene15 (f) ALOI (g) Mnist (h) Youtube

Fig. 12. Curves of the objective function values of ASFL versus the number of iterations.

$$+ \mathbf{v}_3^T \mathbf{U}_n \mathbf{v}_3 + \mathbf{v}_4^T (\mathbf{C} + \frac{\eta}{2}\mathbf{I}_d)\mathbf{v}_4 + \lambda \begin{bmatrix} \mathbf{v}_3 \\ \mathbf{v}_2 \end{bmatrix}^T \mathbf{L}_R \begin{bmatrix} \mathbf{v}_3 \\ \mathbf{v}_2 \end{bmatrix}$$

Considering μ and η are positive parameters, $\mathbf{U}_n$, $\mathbf{C}$ and $\mathbf{L}_R$ are positive semi-definite, we have $\mathbf{v}^T \boldsymbol{\Gamma} \mathbf{v} \geq 0$ is true for any $\mathbf{v}$, proving that $\boldsymbol{\Gamma}$ is positive semi-definite. Thus, Eq. (8) is jointly convex w.r.t. $\mathbf{b}$, $\mathbf{Q}$ (including $\mathbf{G}$ and $\mathbf{F}$) and $\mathbf{W}$, which means that their solutions can be substituted into each other during optimization processes.

Data availability

Data will be made available on request.

References

[1] Lusi Li, Haibo He, Bipartite graph based multi-view clustering, IEEE Trans. Knowl. Data Eng. 34 (7) (2020) 3111–3125.

[2] Chenhang Cui, Yazhou Ren, Jingyu Pu, Jiawei Li, Xiaorong Pu, Tianyi Wu, Yutao Shi, Lifang He, A novel approach for effective multi-view clustering with information-theoretic perspective, Adv. Neural Inf. Process. Syst. 36 (2024).

[3] Han Zhang, Maoguo Gong, Yannian Gu, Feiping Nie, Xuelong Li, Side-constrained graph fusion for semi-supervised multi-view clustering, Neurocomputing 570 (2024) 127102.

[4] Yazhou Ren, Jingyu Pu, Zhimeng Yang, Jie Xu, Guofeng Li, Xiaorong Pu, S Yu Philip, Lifang He, Deep clustering: A comprehensive survey, IEEE Trans. Neural Netw. Learn. Syst. (2024).

[5] Rongyao Hu, Liang Peng, Jiangzhang Gan, Xiaoshuang Shi, Xiaofeng Zhu, Complementary graph representation learning for functional neuroimaging identification, in: Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 3385–3393.

[6] Xinyan Liang, Pinhan Fu, Qian Guo, Keyin Zheng, Yuhua Qian, DC-NAS: Divide-and-conquer neural architecture search for multi-modal classification, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2024, pp. 13754–13762.

[7] Zhihao Wu, Jie Wen, Yong Xu, Jian Yang, Xuelong Li, David Zhang, Enhanced spatial feature learning for weakly supervised object detection, IEEE Trans. Neural Netw. Learn. Syst. 35 (1) (2022) 961–972.

[8] Chang Tang, Xiao Zheng, Wei Zhang, Xinwang Liu, Xinzhong Zhu, En Zhu, Unsupervised feature selection via multiple graph fusion and feature weight learning, Sci. China Inf. Sci. 66 (5) (2023) 1–17.

[9] Cheng Liang, Lianzhi Wang, Li Liu, Huaxiang Zhang, Fei Guo, Multi-view unsupervised feature selection with tensor robust principal component analysis and consensus graph learning, Pattern Recognit. 141 (2023) 109632.

[10] Junqi Lu, Xijiong Xie, Yujie Xiong, Multi-view hypergraph regularized lp norm least squares twin support vector machines for semi-supervised learning, Pattern Recognit. (2024) 110753.

[11] Jia Zhang, Zhiming Luo, Candong Li, Changen Zhou, Shaozi Li, Manifold regularized discriminative feature selection for multi-label learning, Pattern Recognit. 95 (2019) 136–150.

[12] Xiaoping Li, Yadi Wang, Rubén Ruiz, A survey on sparse learning models for feature selection, IEEE Trans. Cybern. 52 (3) (2022) 1642–1660.
[13] Jie Xu, Yazhou Ren, Huayi Tang, Zhimeng Yang, Lili Pan, Yang Yang, Xiaorong Pu, S Yu Philip, Lifang He, Self-supervised discriminative feature learning for deep multi-view clustering, IEEE Trans. Knowl. Data Eng. 35 (7) (2022) 7470–7482.
[14] Zhiwen Cao, Xijiong Xie, Structure learning with consensus label information for multi-view unsupervised feature selection, Expert Syst. Appl. 238 (2024) 121893.
[15] Yadi Wang, Jun Wang, Dacheng Tao, Neurodynamics-driven supervised feature selection, Pattern Recognit. 136 (2023) 109254.
[16] Danyang Wu, Wei Chang, Jitao Lu, Feiping Nie, Rong Wang, Xuelong Li, Adaptive-order proximity learning for graph-based clustering, Pattern Recognit. 126 (2022) 108550.
[17] Kui Yu, Lin Liu, Jiuyong Li, Wei Ding, Thuc Duy Le, Multi-source causal feature selection, IEEE Trans. Pattern Anal. Mach. Intell. 42 (9) (2020) 2240–2256.
[18] Chang Tang, Jiajia Chen, Xinwang Liu, Miaomiao Li, Pichao Wang, Minhui Wang, Peng Lu, Consensus learning guided multi-view unsupervised feature selection, Knowl.-Based Syst. 160 (2018) 49–60.
[19] Shanhua Zhan, Jigang Wu, Na Han, Jie Wen, Xiaozhao Fang, Unsupervised feature extraction by low-rank and sparsity preserving embedding, Neural Netw. 109 (2019) 56–66.
[20] Chenglong Zhang, Yang Fang, Xinyan Liang, Xingyu Wu, Bingbing Jiang, et al., Efficient multi-view unsupervised feature selection with adaptive structure learning and inference, in: Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, 2024, pp. 5443–5452.
[21] Zhiwen Cao, Xijiong Xie, Multi-view unsupervised complementary feature selection with multi-order similarity learning, Knowl.-Based Syst. 283 (2024) 111172.
[22] Hebing Nie, Qun Wu, Haifeng Zhao, Weiping Ding, Muhammet Deveci, Flexible adaptive graph embedding for semi-supervised dimension reduction, Inf. Fusion (2023) 101872.
[23] Guodong Du, Jia Zhang, Ning Zhang, Hanrui Wu, Peiliang Wu, Shaozi Li, Semi-supervised imbalanced multi-label classification with label propagation, Pattern Recognit. 150 (2024) 110358.
[24] Yaojin Lin, Zhuoxin He, Lei Guo, Weiping Ding, Multi-label feature selection via positive or negative correlation, IEEE Trans. Emerg. Top. Comput. Intell. 8 (1) (2024) 401–414.
[25] Jingliu Lai, Hongmei Chen, Tianrui Li, Xiaoling Yang, Adaptive graph learning for semi-supervised feature selection with redundancy minimization, Inform. Sci. 609 (2022) 465–488.
[26] Dan Shi, Lei Zhu, Jingjing Li, Zhiyong Cheng, Zhenguang Liu, Binary label learning for semi-supervised feature selection, IEEE Trans. Knowl. Data Eng. 35 (3) (2023) 2299–2312.
[27] Shudong Huang, Wei Shi, Zenglin Xu, Ivor W Tsang, Jiancheng Lv, Efficient federated multi-view learning, Pattern Recognit. 131 (2022) 108817.
[28] Caijuan Shi, Qiuqi Ruan, Gaoyun An, Chao Ge, Semi-supervised sparse feature selection based on multi-view Laplacian regularization, Image Vis. Comput. 41 (2015) 1–10.
[29] Caijuan Shi, Gaoyun An, Ruizhen Zhao, Qiuiqi Ruan, Qi Tian, Multiview hessian semisupervised sparse feature selection for multimedia analysis, IEEE Trans. Circuits Syst. Video Technol. 27 (9) (2017) 1947–1961.
[30] Caijuan Shi, Zhibin Gu, Changyu Duan, Qi Tian, Multi-view adaptive semi-supervised feature selection with the self-paced learning, Signal Process. 168 (2020) 107332.
[31] Bingbing Jiang, Xingyu Wu, Xiren Zhou, Yi Liu, Anthony G Cohn, Weiguo Sheng, Huanhuan Chen, Semi-supervised multiview feature selection with adaptive graph learning, IEEE Trans. Neural Netw. Learn. Syst. 35 (3) (2024) 3615–3629.
[32] Jianglin Lu, Jie Zhou, Yudong Chen, Witold Pedrycz, Kwok-Wai Hung, Asymmetric transfer hashing with adaptive bipartite graph learning, IEEE Trans. Cybern. 54 (1) (2024) 533–545.
[33] Hong Chen, Feiping Nie, Rong Wang, Xuelong Li, Fast unsupervised feature selection with bipartite graph and $l_{2,0}$-norm constraint, IEEE Trans. Knowl. Data Eng. 35 (5) (2023) 4781–4793.
[34] Chenglong Zhang, Bingbing Jiang, Zidong Wang, Jie Yang, Yangfeng Lu, Xingyu Wu, Weiguo Sheng, Efficient multi-view semi-supervised feature selection, Inform. Sci. 649 (2023) 119675.
[35] Zhihui Lai, Foping Chen, Jiajun Wen, Multi-view robust regression for feature extraction, Pattern Recognit. 149 (2024) 110219.
[36] Jie Wen, Shijie Deng, Lunke Fei, Zheng Zhang, Bob Zhang, Zhao Zhang, Yong Xu, Discriminative regression with adaptive graph diffusion, IEEE Trans. Neural Netw. Learn. Syst. 35 (2) (2024) 1797–1809.
[37] Tianji Pang, Feiping Nie, Junwei Han, Xuelong Li, Efficient feature selection via L20-norm constrained sparse regression, IEEE Trans. Knowl. Data Eng. 31 (5) (2019) 880–893.
[38] Dimitri P. Bertsekas, Constrained Optimization and Lagrange Multiplier Methods, Academic Press, 2014.
[39] Chengrui Zhang, Lei Zhu, Dan Shi, Jiecai Zheng, Haibao Chen, Bo Yu, Semi-supervised feature selection with soft label learning, IEEE/CAA J. Autom. Sin. (2022).
[40] Jiangzhang Gan, Rongyao Hu, Yujie Mo, Zhao Kang, Liang Peng, Yonghua Zhu, Xiaofeng Zhu, Multigraph fusion for dynamic graph convolutional network, IEEE Trans. Neural Netw. Learn. Syst. 35 (1) (2024) 196–207.
[41] Bin Zhang, Qianyao Qiang, Fei Wang, Feiping Nie, Fast multi-view semi-supervised learning with learned graph, IEEE Trans. Knowl. Data Eng. 34 (1) (2022) 286–299.
[42] Siwei Wang, Xinwang Liu, Suyuan Liu, Jiaqi Jin, Wenxuan Tu, Xinzhong Zhu, En Zhu, Align then fusion: Generalized large-scale multi-view clustering with anchor matching correspondences, Adv. Neural Inf. Process. Syst. 35 (2022) 5882–5895.
[43] Feiping Nie, Xiaoqian Wang, Michael Jordan, Heng Huang, The constrained laplacian rank algorithm for graph-based clustering, in: Proceedings of the 30th AAAI Conference on Artificial Intelligence, Vol. 30, 2016, 1.
[44] Magnus R. Hestenes, Multiplier and gradient methods, J. Optim. Theory Appl. 4 (5) (1969) 303–320.
[45] Yongshan Zhang, Jia Wu, Zhihua Cai, S. Yu Philip, Multi-view multi-label learning with sparse feature selection for image annotation, IEEE Trans. Multimed. 22 (11) (2020) 2844–2857.
[46] Laurens Van der Maaten, Geoffrey Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (11) (2008) 2579–2605.

Bingbing Jiang received the Ph.D. degree from the University of Science and Technology of China, Hefei, China, in 2019. He is currently an associate Professor at Hangzhou Normal University, Hangzhou, China. His research interests include semi-supervised learning, multi-view learning, sparse learning, feature selection, and Bayesian inference. He has published more than 40 papers in prestigious journals and conferences.

Jun Liu is working toward an MS.c degree at Hangzhou Normal University, Hangzhou, China. His research interests include semi-supervised learning, and multi-view learning.

Zidong Wang (Fellow, IEEE) received the Ph.D. degree from the Nanjing University of Science and Technology, Nanjing, China, in 1994. He is currently a Professor at Brunel University London, Uxbridge, U.K. He has authored or co-authored more than 700 papers in international journals. His research interests include dynamical systems, bioinformatics, control theory, machine learning, signal processing, and applications. He is the Editor-in-Chief of Neurocomputing and the Editor-in-Chief of the International Journal of Systems Science. He served as the Associate Editor of several prestigious journals, like IEEE TAC, IEEE TCST, IEEE TNNLS, and IEEE TSP. He is a Member of the Academia Europaea, a Member of the European Academy of Sciences and Arts, an Academician of the International Academy for Systems and Cybernetic Sciences, and a Fellow of the Royal Statistical Society.

Chenglong Zhang is working toward an MS.c degree at Hangzhou Normal University, Hangzhou, China. His research interests include semisupervised learning, feature selection, and multi-view clustering.

Jie Yang received the Ph.D. degree from the University of Technology Sydney, Sydney, Australia, in 2022. He is currently a Postdoctoral Fellow with the Computational Intelligence and Brain–Computer Interface Lab of the University of Technology Sydney. His research interests include unsupervised learning, clustering, and EEG data processing.

Yadi Wang received the Ph.D. degree from the School of Computer Science and Engineering, Southeast University, Nanjing, China, in 2020. She is currently an associate Professor at Henan University, Kaifeng, China. Her current research interests include data mining, machine learning, and neurodynamic optimization.

Weiguo Sheng received the Ph.D. degree from Brunel University London, Uxbridge, U.K., in 2005. He is currently a Professor at Hangzhou Normal University, Hangzhou, China. His research interests include evolutionary computation, data mining, and machine learning. He serves as the Associate Editor of Neurocomputing.

Weiping Ding received the Ph.D. degree from Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2013. He is a Full Professor at Nantong University, Nantong, China, and also the supervisor of Ph.D. postgraduate by the Faculty of Data Science at City University of Macau, China. His main research directions involve deep neural networks, multimodal machine learning, and medical image analysis. He has published over 250 articles, including over 90 IEEE Transactions papers. He serves as an Associate Editor/Editorial Board member of IEEE TNNLS, IEEE TFS, IEEE TITS, IEEE TIV, IEEE TETCI, IEEE TAI, Information Fusion, Information Sciences, Neurocomputing, and Applied Soft Computing. He is the Co-Editor-in-Chief of both the Journal of Artificial Intelligence and Systems and the Journal of Artificial Intelligence Advances.