Neurocomputing 657 (2025) 131573

Contents lists available at ScienceDirect

Neurocomputing

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

Multi-view semi-supervised feature selection with multi-order similarity and tensor learning

Hangyu Chen[a, b], Xijiong Xie[a, b,*], Yujie Xiong[c]

[a] Faculty of Electrical Engineering and Computer Science, Ningbo University, Ningbo, 315211, China
[b] Key Laboratory of Mobile Network Application Technology of Zhejiang Province, Ningbo, 315211, China
[c] School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai, 201620, China

ARTICLE INFO

Communicated by J. Zhao

Keywords:
Multi-view learning
Semi-supervised feature selection
Multi-order similarity graph
Tensor learning
High-order information

ABSTRACT

Multi-view data has attracted extensive attention because it can better characterize samples, and multi-view semi-supervised feature selection can not only effectively improve multi-view performance, but also maintain the original real structure of the data. To this end, many scholars have proposed various models to achieve this goal. However, most of the existing methods rely on the graph structure constructed from the original data and use the constructed graph as a guide for feature selection. This not only ignores multi-order domain knowledge, but also ignores the high-order relations between views. Therefore, this study effectively integrates multi-order domain information with graph learning, and performs tensor low-rank learning on the graph structure between multiple views. A multi-view semi-supervised feature selection method based on multi-order similarity and tensor learning is proposed, which not only integrates multi-order domain information, but also takes into account the relationship between views. Based on this, we propose an iterative method to solve the objective function and prove the superiority of our method on multiple basic datasets.

1. Introduction

In real life, a multitude of data needs to be gathered from various heterogeneous aspects or sources. For example, an image can be described from numerous views such as text and pixels. The surface of the planet can be represented by multiple heterogeneous aspects such as spectral data and spatial information [1,2]. With the development of machine learning, this type of data with multiple domain characteristics is defined as multi-view data [3]. Because it describes objects from different aspects, it can represent objects more comprehensively semantically [4]. Nevertheless, the “curse of dimensionality” [5] will inevitably affect multi-view data because of the presence of noisy and redundant features, which will impair the learning task performance. Therefore, how to better reduce dimensionality has become a hot topic of research.

An important strategy for reducing dimensionality is feature selection [6]. This method does not change the original features of the sample, it only obtains a subset of the original sample. As a result, it has strong interpretability in addition to being able to eliminate elements that are unnecessary and noisy. Numerous scholars have put forth numerous feature selection models, and classified them into unsupervised [7,8], semi-supervised [9,10], and supervised [11,12] categories based on the proportion of samples that include labels. Among these methods, the semi-supervised approach chooses feature subsets from the original features by utilizing the label information of a limited number of data samples. Because it can make up for the problem of overfitting or underfitting caused by using only a small number of labels, it can also make up for the problem that only unlabeled data may lose the real structure of the data. For that reason, multi-view semi-supervised feature selection is the main topic of this paper.

Many multi-view semi-supervised feature selection models have been suggested recently, and they may be broadly classified into two groups according to the way they utilize labeled samples. One of them is to extend the label propagation technique, which can reduce noisy features while maintaining critical information, and use a tiny bit of known label information to influence feature selection [13,14]. The other paradigm is to combine label information with the feature selection process, and select features that can better reflect label relationships by constructing a hybrid graph structure or joint optimization [15,16]. With the label propagation technology, the former may reduce noise and effectively use only a limited amount of labels for feature selection while capturing

* Corresponding author at: Faculty of Electrical Engineering and Computer Science, Ningbo University, Ningbo, 315211, China.
Email addresses: 2211100084@nbu.edu.cn (H. Chen), xjxie11@gmail.com (X. Xie), xiong@sues.edu.com (Y. Xiong).

https://doi.org/10.1016/j.neucom.2025.131573
Received 13 January 2025; Received in revised form 21 July 2025; Accepted 13 September 2025
Available online 18 September 2025

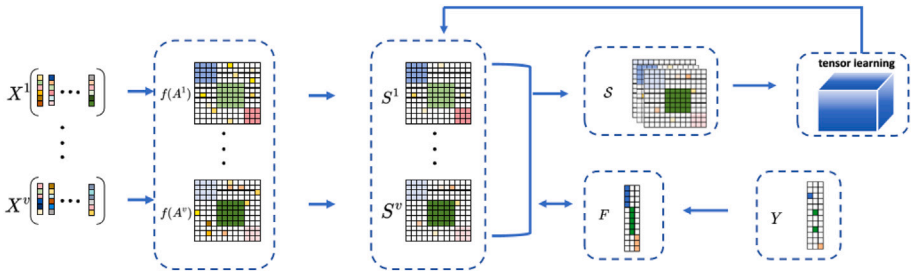


Fig. 1. This is the framework of MSFSMT we proposed, which mainly consists of three parts: multi-order similarity learning, tensor learning, and label information learning. Given a multi-view data $X = \{X^{(1)}, \ldots, X^{(V)}\}$ and labeled data Y_l, we first perform multi-order similarity learning on the original data to obtain the initial fixed similarity graph matrix for each view, and then obtain the adaptive graph matrix S^v by minimizing the difference, and stack S^v into a third-order tensor $\mathcal{S}$. At the same time, we investigate high-order linkages between views using tensor robust principal component analysis. Then, we integrate the adaptive graph matrices of multiple views into a spectral graph to learn a consensus indicator matrix F, and finally fuse the existing label information into the indicator matrix.

the global structure of the data. By utilizing both the intrinsic structure of unlabeled data and the label information simultaneously, the latter can improve the accuracy and consistency of feature selection. But they are both graph-based, semi-supervised feature selection techniques use a similarity graph matrix to preserve the geometric structure among data points [17,18]. In this case, their performance depends largely on the quality of the graph structure. Especially for multi-view data, obtaining a consensus graph that better reflects the true structure of the data has become a decisive factor in multi-view semi-supervised feature selection.

According to the different ways that numerous views are processed, the multi-view semi-supervised feature selection techniques that are now in use can be broadly classified into two categories. One is to simply splice the views and then select features from the original data using the single-view method [19,20]. However, this method easily ignores the connection and differences between views and has poor effectiveness and adaptability in multi-view applications. The other is to assign a weight to the similarity graph matrix of each view to obtain a fused consensus graph, and then use the consensus graph to guide feature selection [21,22]. However, the existing methods are limited by the noise in the original data, and the obtained similarity graph matrix deviates from the real data structure, resulting in a decrease in the final performance.

In order to obtain a more reliable similarity graph structure, we propose a multi-view semi-supervised feature selection model with multi-order similarity and tensor learning (MSFSMT). First, we build each view's initial similarity matrix using multi-order similarity learning as a graph filter. This technique guarantees that the neighborhood data of various orders is integrated into the learning graph. The various views' graph structures are combined into a third-order tensor, and the tensor low-rank learning effectively reduces the effect of noise. Finally, multiple graph structures are integrated into a consensus indicator matrix after spectral clustering. Furthermore, the indicator matrix is guided by a small quantity of label information and integrated into a unified learning framework model. Fig. 1 simply shows the process and framework of our model. The following are the main contributions of this paper:

- By leveraging the neighborhood knowledge of multi-order to construct a more dependable graph structure that will better support subsequent feature selection, the MSFSMT algorithm efficiently blends multi-order similarity learning with semi-supervised feature selection.
- The MSFSMT algorithm uses the tensor low-rank learning to improve the robustness of the model, which can better utilize the high-order relationships between views to keep the consensus graph structure consistent as much as possible.
- We integrate multi-order similarity learning, tensor low-rank learning, and semi-supervised learning into a unified framework and design an effective iterative update algorithm to solve the objective function.
- To prove the superiority of our suggested strategy, we do tests on several datasets and compare the experimental findings.

This is how the remainder of the paper is structured. We provide a brief overview of the relevant semi-supervised feature selection work in Section 2. We provide the MSFSMT method's formulation and details in Section 3. We present the MSFSAT optimization algorithms in Section 4. To illustrate the efficacy of the approach, we perform extensive experiments on multiple benchmark datasets in Section 5. In Section 6, we provide a summary of the MSFSMT methodology.

2. Related work

In this section, we mainly introduce the main models of semi-supervised feature selection.

2.1. Single-view semi-supervised feature selection

Most semi-supervised feature selection techniques were initially based on filtering techniques to select features by evaluating the relationship between features and the target variable. Zhao and Liu [23] proposed the semi-supervised feature selection via spectral analysis (SSFSSA) model, which addresses the problem of a small number of labeled samples and incorporates unlabeled samples into a regularized framework. Doquire and Verleysen [24] proposed a semi-supervised feature selection algorithm for regression problems, which primarily uses the Laplace score to evaluate the importance of features. A semi-supervised feature selection algorithm based on the maximum relevance and minimum redundancy criterion of Pearson's correlation coefficient (RRPC) was proposed [25], which selects features through incremental search technology. Although these methods can effectively utilize the information of the features themselves, they do not fully utilize the relationship between the features. Later, the rise of embedded methods, which can be integrated with feature selection, has certain advantages over other methods, and scholars have also proposed many embedded semi-supervised feature selection methods. The semi-supervised feature selection via manifold regularization (SFSVMR) model was presented by Xu et al. [26]. It picks features by increasing the degree of classification between various categories. The semi-supervised approach of feature selection via sparse rescaled linear square regression (SRLSR) was introduced by Chen et al. [27]. This method selects features using the regression coefficients from least squares regression. The local preserving logistic I-Relief for semi-supervised feature selection (LPLIR) approach was proposed by Tang et al. [28]. It maintains the consistency of the local sample structure while simultaneously maximizing the labeled data boundary. Because these methods are greatly affected by outliers, the similarity matrix constructed based on k-nearest neighbors may be locally optimal. Zeng et al. [29] suggested a semi-supervised feature selection technique with a global sparsity constraint and a local adaptive loss function that considers the sparsity of both the local and

global sample structures in order to reduce its impact. Zhong et al. [30] proposed an adaptive semi-supervised discriminative feature selection analysis (SADA), which can reduce the impact of noise on the similarity matrix by learning the adaptive similarity matrix and projection matrix in an iterative process. In order to enhance the quality of feature selection, Lai et al. [31] presented adaptive graph learning for semi-supervised feature selection with redundancy minimization (AGLRM), which combines adaptive graph learning and redundancy minimization regularization. Kang et al. [15] proposed the structured graph learning framework with single kernel (SGSK), which uses the self-expression of samples to obtain the global structure and adapts to the field to maintain the local structure. In order to make better use of label information, Wang et al. [32] proposed sparse discriminative semi-supervised feature selection (SDSSFS), which iteratively combines learning regression coefficients and predicting unknown labels. Sheikhpour et al. [33] proposed a Hessian-based semi-supervised feature selection framework using the generalized uncorrelated constraint (HSFSGU), which uses topological structure and generalized uncorrelated constraints to make the projection matrix suitable for feature selection.

2.2. Multi-view semi-supervised feature selection

In recent years, multi-view data has become a hot topic of research due to its advantages. On the basis of single-view semi-supervised feature selection, multi-view semi-supervised feature selection has also been developed and innovated, and various useful models have been put forth. At the beginning, researchers simply spliced the features of multiple views together [34,35], but this method was just a simple migration of single-view semi-supervised feature selection, and did not consider the differences and complementarities between views. Later, Shi et al. [36] proposed a multi-view semi-supervised feature selection using the Laplace regularization method, which effectively utilized the connection between views. Shi et al. [21] proposed a multi-view Hessian semi-supervised sparse feature selection (MHSFS) model, which uses the Hessian regularization method to encode the local geometric structure of unlabeled samples to keep the local structure of data samples consistent. Nie et al. [37] proposed multi-view learning with adaptive neighbors (MLAN), which can be used in semi-supervised classification tasks. This method can adaptively weight each view. Although the performance of these methods is higher than that of the corresponding single-view methods, they all construct similarity matrices through original features and are easily affected by noise and redundancy in data. In addition, the similarity matrix remains unchanged during the iterative solution process, resulting in a certain deviation in the final solved projection matrix. In order to enable the Laplacian graph to adjust to the prediction data throughout the iteration process, Shi et al. [38] introduced self-paced learning into the multi-view adaptive semi-supervised feature selection (MASFS) approach. Ziraki et al. [13] proposed a multi-view consistent graph construction and label propagation algorithm (MVCGL), which combines multi-view graph structure information and label information. However, these methods cannot handle large-scale data. Zhang et al. [39] proposed a multi-view semi-supervised feature selection for bipartite graphs based on adaptive learning, which greatly reduced the complexity of the calculation. In order to better explore the different structures between multiple views, Guo et al. [40] proposed a robust semi-supervised multi-view graph learning framework based on the shared and individual structure (RSSMvSI). This model obtains clean data by sparsely denoising the original data, thereby improving the robustness of feature selection.

There are always opportunities for improvement even though earlier approaches have had some successes with the multi-view semi-supervised feature selection. To be more precise, they all create the graph structure from raw data without utilizing multi-order neighbor information, which leads to a less-than-ideal network structure. As a result, we investigate how multi-view semi-supervised feature selection incorporates multi-order similarity learning. Furthermore, the aforementioned techniques all immediately acquire a consensus graph structure from every view, which gives scant consideration to the diversity and consistency of information across views. Therefore, we study how to make up for this defect through the tensor learning.

3. Proposed method

3.1. Notations and definitions

To introduce the proposed model more clearly, we first provide a brief explanation of the symbols and definitions in this paper. We use bold capital letters to represent a matrix, for example, $X^v \in \mathbb{R}^{n\times d^v}$ represents the data matrix of the v th view, $S^v \in \mathbb{R}^{n\times n}$ represents the similarity matrix of the v th view, $F \in \mathbb{R}^{n\times k}$ represents the indicator matrix, $W^v \in \mathbb{R}^{d^v\times k}$ represents the projection matrix of the v th view, $Y_l \in \mathbb{R}^{l\times k}$ represents the label matrix of the labeled samples, and $Y = [Y_l; 0] \in \mathbb{R}^{n\times k}$ represents the label matrix. $\|X\|_F$, $\|X\|_{2,1}$ denote the Frobenius norm and $l_{2,1}$-norm of X, respectively. We also use uppercase calligraphic letters to denote third-order tensors, for example, $\mathcal{S} \in \mathbb{R}^{n_1\times n_2\times n_3}$ denotes a third-order tensor consisting of the similarity matrices of V views. In addition, we use $\|\mathcal{S}\|_{\circledast}$ to denote the t-SVD based tensor nuclear norm of $\mathcal{S}$ [41,42].

3.2. Framework of MSFSMT

In this paper, in order to allow the graph structure guiding feature selection to incorporate more multi-order neighbor information, we need to make the similarity matrix S as close as possible to the neighborhood information $f(A)$ of different orders, which can be defined by mathematical formula.

$$\min_S \|S - f(A)\|_F^2 \tag{1}$$

where $f(A)$ contains p different orders of domain information of $A, A^2, \dots A^p$, and its mathematical formula [43,44] is defined as

$$f(A) = A + A^2 + \cdots + A^p. \tag{2}$$

Among them, A describes the first-order domain information, that is, the probability of a node reaching another node through one step of random walk. For example, A_{ij} represents the probability of node x_i transitioning to x_j through one step. Similarly, A^p represents the probability of a node reaching another node through p steps of random walk, that is, $A^P = \underbrace{A \cdot A \cdots A}_{P}$. However, because $f(A)$ simply adds domain information of different orders, its value may have different degrees of amplitude deviation, so we restrict its projection to $[0, 1]$ and set the value of the diagonal elements to 0, and extend formula 1 to multiple views and obtain

$$\begin{aligned} &\min_{S^v} \sum_{v=1}^{V} \|S^v - f(A^v)\|_F^2 \\ &\text{s.t. } (S^v)^\top \mathbf{1} = \mathbf{1}, 0 \le S^v \le 1. \end{aligned} \tag{3}$$

In order to obtain high-order correlations between views, we stack the obtained similarity matrix S^v into a third-order tensor $\mathcal{S}$ and use t-SVD to ensure the low rank of $\mathcal{S}$ [45]. The formula (3) can be expanded to

$$\begin{aligned} &\min_{S} \sum_{v=1}^{V} \|S^v - f(A^v)\|_F^2 + \|\mathcal{S}\|_{\circledast} \\ &\text{s.t. } (S^v)^\top \mathbf{1} = \mathbf{1}, 0 \le S^v \le 1. \end{aligned} \tag{4}$$

In addition, we reasonably assume that the cluster labels corresponding to each view have relatively large similarities, so we use a common indicator matrix to utilize the consensus information between views. According to spectral graph theory, the higher the similarity of two samples in the similarity graph matrix, the greater the probability that they

have the same label [46]. Next, we may include a regularization term to ensure that the local graph structure of every view remains consistent.

$$\begin{aligned}&\min_{F}\sum_{v=1}^{v}\mathrm{Tr}(F^{\top}L^{v}F)\\&\text{s.t. } F\geq 0, F^{\top}F=I,\end{aligned}\tag{5}$$

where $L^v = D^v - \frac{S^v+(S^v)^\top}{2}$, D^v is the degree matrix of S^v, and its diagonal elements are defined as $D^v_{ii} = \sum_{j=1}^{n} S^v_{ij}$.

The samples that were labeled in the indicator matrix need to match the true labels in order to fully utilize the labeled data. This will yield the following penalty terms.

$$\min_{F}\sum_{i=1}^{l}\left[\sum_{j=1}^{k}U_{ii}\left(F_{ij}-Y_{ij}\right)^{2}\right]\tag{6}$$

Among them, U is the decision rule matrix, which is a diagonal matrix. For unlabeled samples x_i, the diagonal elements of the decision rule matrix $U_{ii} = 1$; if the sample is labeled, U_{ii} is a large value to ensure that the predicted label is consistent with the true label. For convenience, the above formula can be converted to

$$\min_{F}\quad \mathrm{Tr}((F-Y)^{T}U(F-Y))\tag{7}$$

Finally, the projection matrix between the eigenvalue and the indicator matrix can be used as the discriminant matrix for feature selection [47]. The definition is as follows.

$$\min_{W}\sum_{v=1}^{V}(\|X^{v}W^{v}-F\|_{F}^{2}+\alpha\|W^{v}\|_{2,1})\tag{8}$$

We apply the $l_{2,1}$-norm to the projection matrix W^v. When α is large enough, most of the values in the projection matrix are almost zero, that is, only a small number of values are retained, thereby increasing the difference between features and making the feature subset selected by the feature more representative.

Combining with the formulas (4), (5), (7) and (8), we can derive the overall objective function of MSFSMT.

$$\begin{aligned}\min_{W^{v},S^{v},F}&\sum_{v=1}^{v}(\|X^{v}W^{v}-F\|_{F}^{2}+\alpha\|W^{v}\|_{2,1}+\beta\|S^{v}-f(A^{v})\|_{F}^{2}\\&+\gamma\,\mathrm{Tr}(F^{\top}L^{v}F))+\lambda\|S\|_{\circledast}+\mathrm{Tr}((F-Y)^{\top}U(F-Y))\\&\text{s.t. } S^{v\top}\mathbf{1}=\mathbf{1}, 0\leq S^{v}\leq 1, F\geq 0, F^{\top}F=I,\end{aligned}\tag{9}$$

Among them, α, β, λ, and γ are balanced hyperparameters for balancing various constraints.

From the formula (9), we can see three advantages of the MSFSMT model: First, unlike other methods that guide the similarity matrix by constructing a Laplacian graph or a self-representing graph, it constructs a multi-order graph matrix with multi-order domain information, and then lets the similarity matrix approximate, which can better capture the global structural information related to the graph. Second, using tensor low-rank constraints can better utilize high-order connections between views and increase the robustness of the model. Third, a consensus similarity graph matrix is obtained in all views by employing the complementary information and consistency information between various views.

4. Optimization algorithm

In this section, we design an efficient iterative optimization method to find the optimal solution. As can be seen from formula (9), the objective function involves three variables: F, W^v, and S^v. It is quite challenging to solve these variables at the same time. To deal with this problem, we decompose the objective function into multiple subproblems. When optimizing one of the variables, the other variables remain unchanged and are optimized alternately.

4.1. Update W^v

In formula (9), W^v for each view can be solved separately, and other variables are fixed to obtain the optimization subproblem.

$$\min_{W^{v}}\|X^{v}W^{v}-F\|_{F}^{2}+\beta\|W^{v}\|_{2,1}\tag{10}$$

We can transform the above formula (10) into a differentiable formula.

$$\min_{W^{v}}\mathrm{Tr}((X^{v}W^{v}-F)^{\top}(X^{v}W^{v}-F))+\alpha\,\mathrm{Tr}(W^{v\top}G^{v}W^{v})\tag{11}$$

where G^v is a diagonal matrix, and its diagonal elements are: $G^v_{ii} = \frac{1}{2\|W^v(i,:)\|^2}$, then calculate the derivative of the formula (11) with respect to W^v and let its value equal to zero, we can get the following update rule.

$$W^{v}=\left(X^{v\top}X^{v}+\alpha G^{v}\right)^{-1}X^{v\top}F.\tag{12}$$

4.2. Update F

When other variables are fixed, the update to formula (9) becomes

$$\begin{aligned}\min_{F}&\sum_{v=1}^{v}(\|X^{v}W^{v}-F\|_{F}^{2}+\gamma\,\mathrm{Tr}(F^{\top}L^{v}F))\\&+\mathrm{Tr}((F-Y)^{\top}U(F-Y))\\&\text{s.t. } F\geq 0, F^{\top}F=I.\end{aligned}\tag{13}$$

There are multiple constraints on F, and we can use the Lagrange multiplier method to solve the update rule for F [48]. The Lagrange function of the above formula (13) is

$$\begin{aligned}\min_{F}&\sum_{v=1}^{v}(\mathrm{Tr}((X^{v}W^{v}-F)^{\top}(X^{v}W^{v}-F))+\gamma\,\mathrm{Tr}(F^{\top}L^{v}F))\\&+\mathrm{Tr}((F-Y)^{\top}U(F-Y))+\frac{\eta}{2}(F^{\top}F-I)+\mathrm{Tr}(\Gamma F^{\top})\end{aligned}\tag{14}$$

where η is the Lagrange multiplier of the orthogonal constraint. The larger the value, the more F can maintain orthogonality, that is, $F^\top F = I$. In addition, Γ is the Lagrange multiplier of the constraint $F \geq 0$. By deriving the formula (14), when F_{ij} is equal to zero, according to the Karush–Kuhn–Tucker (KKT) condition, we can get $\Gamma_{ij}F_{ij} = 0$, so we can get the update rule for F_{ij}.

$$F_{ij}\longleftarrow F_{ij}\frac{\left(\sum_{v=1}^{V}X^{v}W^{v}+\eta F+UY\right)_{ij}}{\left(VF+\gamma\sum_{v=1}^{V}L^{v}F+\eta FF^{\top}F+UF\right)_{ij}}.\tag{15}$$

4.3. Update S^v

By fixing the F and W^v variables, we can derive the update formula for S^v.

$$\begin{aligned}\min_{S^{v}}&\sum_{v=1}^{V}(\|S^{v}-f(A^{v})\|_{F}^{2}+\gamma\,\mathrm{Tr}(F^{\top}L^{v}F))+\lambda\|S\|_{\circledast}\\&\text{s.t. } (S^{v})^{\top}\mathbf{1}=\mathbf{1}, 0\leq S^{v}\leq 1.\end{aligned}\tag{16}$$

In order to make the above formula separable, we introduce a new tensor $\mathcal{J}$, and we obtain

$$\begin{aligned}\min_{S^{v}}&\sum_{v=1}^{V}(\|S^{v}-f(A^{v})\|_{F}^{2}+\gamma\,\mathrm{Tr}(F^{\top}L^{v}F))+\lambda\|\mathcal{J}\|_{\circledast}\\&+\frac{\mu}{2}\|S-\mathcal{J}+\frac{\mathcal{M}}{\mu}\|_{F}^{2}\\&\text{s.t. } (S^{v})^{\top}\mathbf{1}=\mathbf{1}, 0\leq S^{v}\leq 1.\end{aligned}\tag{17}$$

When the variable $\mathcal{J}$ is fixed, the update of S^v for each view is consistent, and its formula can be written as

$$\min_{S^v} \|S^v - f(A^v)\|_F^2 + \gamma \operatorname{Tr}(F^\top L^v F) + \frac{\mu}{2}\|S^v - J^v + \frac{M^v}{\mu}\|_F^2 \quad \text{s.t. } (S^v)^\top \mathbf{1} = \mathbf{1}, 0 \le S^v \le 1. \tag{18}$$

For a more convenient and clear expression, we introduce $B^v = J^v - \frac{M^v}{\mu}$, and because $\operatorname{Tr}(F^\top L^v F) = \sum_{i,j=1}^n \frac{1}{2}\|\mathrm{f}_i - \mathrm{f}_j\| S_{ij}^v$, we can rewrite the above formula as follows.

$$\min_{S_i^{v\top}\mathbf{1}=1, 0\le S_{ij}^v \le 1} \sum_{i,j=1}^n \left(\beta (S_{ij}^v - f(A^v)_{ij})^2 + \frac{\mu}{2}(S_{ij}^v - B_{ij}^v)^2 + \frac{\gamma}{2}\|\mathrm{f}_i - \mathrm{f}_j\|_2^2 S_{ij}^v\right) \tag{19}$$

Let $\mathrm{v}_{ij} = \|\mathrm{f}_i - \mathrm{f}_j\|_2^2$, and let v_i represent the j th element in the i th row. Then, it is comparable to maximizing the following minimization problem in order to find the optimal S^v.

$$\min_{S_i^{v\top}\mathbf{1}=1, 0\le S_{ij}^v \le 1} \|S_i^v - \frac{2\beta f(A^v)_i + \mu B_i^v - \frac{\gamma}{2}\mathrm{v}_i}{2\beta + \frac{\mu}{4}}\|_2^2 \tag{20}$$

We can solve the optimal S^v value according to the method proposed by Huang et al. [49]. For the variable $\mathcal{J}$, which we introduced additionally, we also need to iteratively update it, and we can get

$$\min_{\mathcal{J}} \|\mathcal{J}\|_{\circledast} + \frac{\mu}{2}\|S - \mathcal{J} + \frac{\mathcal{M}}{\mu}\|_F^2 \tag{21}$$

According to the t-SVD tensor nuclear norm minimization theory [50], it can be solved by the following method.

$$\mathcal{J}^* = \mathcal{U} * \mathcal{C} * \mathcal{V}^\top, \tag{22}$$

where $\mathcal{S} + \frac{\mathcal{M}}{\mu} = \mathcal{U} * \mathcal{O} * \mathcal{V}^\top, \mathcal{C} = \mathcal{O} * \mathcal{D}$. $\mathcal{D}$ represents a tensor composed of diagonal matrices, whose diagonal items are defined as $\mathcal{D}(i,i,j) = \left(1 - \frac{n/\mu}{\mathcal{O}(i,i,j)}\right)_+$. Combining the above update rules for each variable, we summarize the basic algorithm flow of MSFSMT, as shown in Algorithm 1.

4.4. Complexity and convergence analysis

An analysis of the complexity of various proposed approaches may be conducted by approximately dividing the optimization process of the MSFSMT algorithm into four stages. The first part is the update of W^v, and its cost for each view is $O_1 = O(d^{v3})$. The second part is the updating of F. Since it updates every element in the matrix, its complexity is $O_2 = O(knc + cn^2)$. Furthermore, updating S^v and the tensor $\mathcal{J}$ has computational complexity $O_3 = O(n^2)$ and $O_4 = O(2V^2 + 2n^2 V log(n))$, respectively. Consequently, $O = O_1 + O_2 + O_3 + O_4$ is the computational complexity of one iteration of the MSFSMT algorithm. This algorithm employs an alternating optimization framework where each sub-problem is solved optimally, typically yielding a closed-form solution. This guarantees that the objective function value decreases monotonically with each iteration until convergence.

Algorithm 1 MSFSMT.

1: **Input:** Muti-view data: $X = [X^1, \ldots, X^v]$, labeled matrix Y, regularization parameters $\alpha, \beta, \lambda, \gamma$, initialize diagonal matrix U and matrix W^v, where U is a diagonal matrix defined as $U_{ii} = 9$ for $i = 1, 2, \ldots, l$ and $U_{ii} = 1$ otherwise. The elements W_{ij} in W^v are random numbers between 0 and 1. $\rho = 1.1$, $\mu = 10^{-5}$, $\mu_{max} = 10^5$.
2: **Repeat**
3: Update W^v by Eq. (12).
4: Update F by Eq. (15).
5: Update S^v by Eq. (20).
6: Update $\mathcal{J}$ by Eq. (22).
7: Update $\mathcal{M}$ by $\mathcal{M} = \mathcal{M} + \mu(S - \mathcal{J})$.
8: Update μ by $\min(\rho\mu, \mu_{max})$.
9: **Until** Convergence
10: **Output:** Calculate $\|W_i\|_2 (i = 1, 2, \ldots, d)$ and sort its values. Take the largest h as the discriminant matrix, and the corresponding original dataset is the final feature subset.

5. Experiment

In this section, we designed a series of experiments on multiple base datasets to demonstrate the superiority and effectiveness of the MSFSMT algorithm. All experiments can be roughly divided into two parts, one is to compare the MSFSMT algorithm with other representative feature selection models, and the other is to evaluate the MSFSMT algorithm from different aspects.

5.1. Dataset and experimental settings

To evaluate the outstanding MSFSMT models, we used 6 benchmark multi-view datasets, including MSRCV1, 3Sources, Handwritten (HW), Caltech101-7 (Cal-7), WebKB, and ORL. Table 1 briefly introduces the basic information of these datasets, including the feature size, number of samples, and number of categories of each view.

Next, we will further introduce the experimental settings. We randomly select 70 % of the samples from each dataset as the training set and the remaining 30 % as the validation set. Because it belongs to semi-supervised learning, the training set needs to be divided. We set 10 %, 20 %, and 30 % as the labeled ratio and randomly divide the training set into labeled sample sets and unlabeled sample sets. The MSFSMT model has multiple hyperparameters, among which the step size of the random walk is set in {2, 3, 4, 5, 6, 7, 8, 9}, that is, the value of p is set in {2, 3, 4, 5, 6, 7, 8, 9}. For other parameters $\alpha, \beta, \lambda, \gamma$, they are all set in {0.01, 0.1, 1, 10, 100}. In addition to the setting of hyperparameters, the feature subset size also needs to be set. We set the feature subset size to a fixed value of {50, 100, …, 500}. The best linear SVM classifier for the feature subset is chosen using five-fold cross validation on the labeled samples. It is then tested on the validation set, and the accuracy is recorded. We perform the above experiment for five times, recording only the best accuracy for each time, to minimize the contingency induced by random sample selection.

5.2. Comparison methods

We compared the MSFSMT algorithm with seven feature selection techniques, comprising two multi-view unsupervised feature selection techniques, two single-view semi-supervised feature selection techniques, and four multi-view semi-supervised feature selection techniques, in order to show the algorithm's superiority and progress. The particular techniques are as follows.

- TRCA-CGL: This technique combines adaptive learning with tensor resilient principal component analysis to produce a trustworthy pseudo-label that directs feature selection [51].
- CFSMO: The model applies multi-order similarity learning to learn the graph structure of each view and maintains the complementary information of multiple views through a consensus latent representation [52].
- SFSS: It is a single-view semi-supervised feature selection model using sparse regression and manifold regularization [35].
- SFS-SLL: A single-view semi-supervised feature selection model that effectively combines soft label learning and sparse regression feature selection methods [53].
- MLSFS: A multi-view semi-supervised feature selection method using multi-view Laplacian regularization unifies the graph structure information of multiple views into a consensus indicator matrix [36].

Table 1
Details of the multi-view dataset.

Feature	MSRCV1	3Sources	HW	Cal-7	WebKB	ORL
1	CMT(24)	BBC(3560)	FCCS(76)	WM(40)	View 1(1703)	GIST(512)
2	HOG(576)	REUTERS(3631)	FAC(216)	CENTRIST(254)	View 2(230)	LBP(59)
3	GIST(512)	GUARDIAN(3068)	KAR(64)	HOG(1984)	View 3(230)	HOG(864)
4	CENTRIST(254)	–	PA(240)	GIST(512)	–	CENTRIST(254)
5	LBP(256)	–	ZER(47)	LBP(256)	–	–
6	–	–	MOR(6)	GABOR(48)	–	–
Instance	210	169	2000	1474	203	400
Class	7	6	10	7	4	40

Table 2
Comparison of the classification results from several datasets using feature selection strategies (%). The best-performing row is bolded.

Datasets	labeled ratios	TRCA-CGL	CCSFS	SFSS	SFS-SLL	MLSFS	MASFS	SMFS	EMSFS	MSFSAT
3Sources	10 %	62.97 ± 3.78	64.58 ± 4.04	60.57 ± 5.86	64.75 ± 4.62	66.75 ± 3.87	61.78 ± 3.92	63.82 ± 3.59	68.62 ± 3.91	**75.43 ± 2.11**
	20 %			63.37 ± 5.25	68.51 ± 3.47	71.23 ± 3.27	65.62 ± 3.01	73.59 ± 2.88	71.67 ± 2.56	**80.39 ± 1.96**
	30 %			65.72 ± 2.91	72.52 ± 3.95	75.92 ± 2.24	70.51 ± 2.04	77.73 ± 2.66	78.43 ± 2.97	**84.31 ± 2.24**
HW	10 %	95.06 ± 0.18	95.11 ± 0.24	90.54 ± 0.42	95.67 ± 0.32	95.74 ± 0.26	94.17 ± 0.49	95.12 ± 0.52	93.49 ± 0.50	**96.51 ± 0.26**
	20 %			94.62 ± 0.52	**97.27 ± 0.06**	96.53 ± 0.13	96.02 ± 0.31	96.41 ± 0.36	96.83 ± 0.17	96.23 ± 0.43
	30 %			95.26 ± 0.21	**97.67 ± 0.12**	97.33 ± 0.21	96.83 ± 0.38	97.51 ± 0.29	97.23 ± 0.27	97.13 ± 0.20
Cal-7	10 %	92.43 ± 1.21	93.66 ± 2.06	89.59 ± 2.52	92.35 ± 2.07	93.79 ± 1.15	88.21 ± 1.18	92.71 ± 2.92	95.25 ± 1.81	**97.51 ± 0.45**
	20 %			92.96 ± 0.49	96.21 ± 0.86	96.72 ± 0.34	89.69 ± 1.61	95.84 ± 0.32	95.71 ± 1.35	**98.29 ± 0.13**
	30 %			96.32 ± 0.28	97.25 ± 0.36	97.92 ± 0.28	88.46 ± 1.83	97.62 ± 0.13	97.07 ± 0.90	**98.79 ± 0.18**
MSRCV1	10 %	77.46 ± 1.33	77.14 ± 1.42	54.60 ± 4.84	64.76 ± 7.81	61.90 ± 6.05	59.05 ± 9.76	80.32 ± 2.88	79.36 ± 2.13	**90.16 ± 1.80**
	20 %			75.23 ± 5.85	80.01 ± 5.23	79.04 ± 4.14	67.94 ± 4.94	84.44 ± 3.68	82.53 ± 1.58	**92.38 ± 1.31**
	30 %			85.40 ± 3.29	92.06 ± 1.12	86.67 ± 3.29	77.78 ± 3.58	86.98 ± 2.61	85.71 ± 1.87	**91.75 ± 1.86**
ORL	10 %	42.50 ± 3.83	45.16 ± 3.24	49.50 ± 3.07	38.63 ± 6.78	45.33 ± 4.73	48.67 ± 3.67	50.85 ± 3.78	53.98 ± 0.65	**60.33 ± 1.12**
	20 %			68.72 ± 3.26	61.18 ± 4.69	70.26 ± 3.13	69.28 ± 2.51	71.67 ± 3.95	68.79 ± 0.46	**80.73 ± 0.90**
	30 %			76.21 ± 1.58	72.36 ± 5.17	81.52 ± 2.13	82.89 ± 2.73	84.29 ± 1.86	82.14 ± 0.58	**85.67 ± 0.64**
WebKB	10 %	80.66 ± 1.37	77.04 ± 2.07	75.08 ± 2.93	83.93 ± 1.76	87.21 ± 2.14	86.56 ± 2.75	81.72 ± 3.28	85.12 ± 1.76	**92.79 ± 1.87**
	20 %			80.23 ± 2.26	86.85 ± 0.92	89.26 ± 1.54	87.13 ± 1.85	81.98 ± 2.95	88.96 ± 0.44	**93.11 ± 1.37**
	30 %			81.83 ± 1.92	85.97 ± 1.26	91.34 ± 0.97	84.25 ± 2.53	82.63 ± 2.79	93.19 ± 0.25	**94.09 ± 0.83**

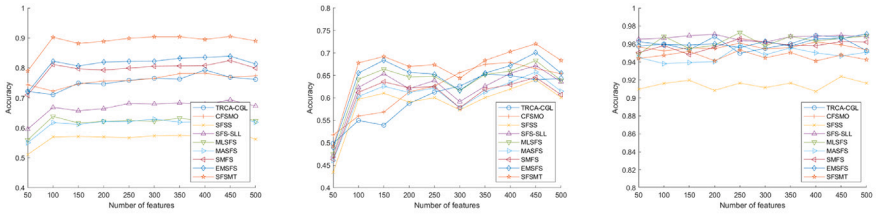
- MASFS: This method is an extension of MLSFS. It adds an adaptive step size to MLSFS so that the graph structure can also be updated during the iteration process [38].
- SMFS: This method unifies the feature learning and graph learning and can adaptively weight the projection matrix [54].
- EMSFS: The model integrates graph learning, label propagation and multi-view feature selection into a unified framework, adaptively constructs a bipartite graph between training samples and anchor points, and significantly reduces the computational complexity [39].

These comparison methods also have hyperparameters. In order to control the variables, we set these hyperparameters to {0.01, 0.1, 1, 10, 100}. The settings of the feature subset size, classification method, and verification method are also consistent with the above paragraph. The unsupervised feature selection method does not use labeled information. The results of this method are consistent under different label ratios, so we only record the results once.

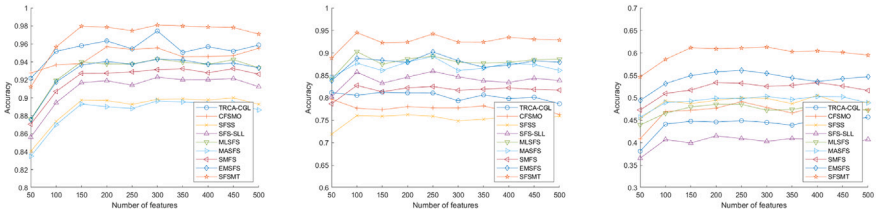
5.3. Comparative analysis of accuracy

We use accuracy as the evaluation indicator, and its value ranges from 0 to 100. The higher the accuracy value, the better the classification effect of the model. Table 2 shows the experimental results of all models, with the best results marked in bold. At the same time, Fig. 2 shows the highest classification accuracy of each method under different numbers of features, that is, when different numbers of features are selected, the hyperparameters of the optimal model are different, and we select the highest accuracy for recording. Through the experimental results, we can draw the following conclusions.

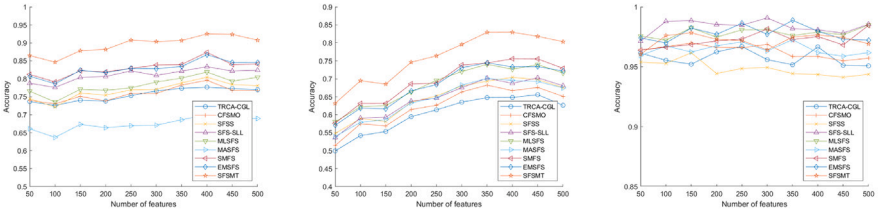
(1) In most experiments, the MSFSMT method outperforms the comparison methods. For example, in the 3Sources dataset with 10 % labeled samples, the accuracy is nearly 10 % higher than the second-best method, and it is 7 % higher in the 30 % labeled sample set. Although our method does not perform best on the HW dataset with 20 % and 30 % labeled samples, it still achieves the highest accuracy for the 10 % labeled sample set, with a 1 % improvement based on the original baseline.
(2) The classification accuracy of the MSFSMT method essentially increases with the increase in the proportion of the number of labels. For example, in the ORL dataset, the accuracy of the 20 % labeled sample set is 20 % higher than that of the 10 % labeled sample set, and the accuracy of the 30 % labeled sample set is 5 % higher than that of the 20 % labeled sample set, which proves that the MSFSMT method can effectively utilize labeled samples.
(3) The accuracy of 10 % labeled samples is improved by the SFSMT approach more than that of 20 % and 30 % labeled samples in most datasets. In the Cal-7 dataset, for instance, the accuracy of the 10 % labeled sample set is 5 % greater than the second best; nevertheless, the 20 % and 30 % labeled sample sets have 3 % and 1 % higher accuracy, respectively, than the second best, respectively. This proves that the MSFSMT method performs better when there are fewer labels, which further proves that our proposed method can make full use of the structural information of unlabeled data.
(4) Selecting different numbers of feature subsets will have a certain impact on the final accuracy, and the more features selected does not necessarily mean higher accuracy. For example, in the Cal-7 dataset with 10 % labeled samples, the accuracy fluctuates as the number of features increases. In the 3Sources dataset with 20 % labeled samples, the accuracy first increases and then decreases as the number of features increases. The MSFSMT method performs best under almost all different numbers of feature subsets, and the fluctuation range is smaller than that of other methods, which proves that the features selected by this method are more representative and have stronger noise resistance.



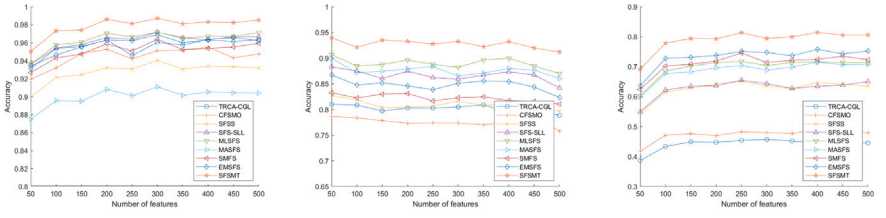
(a) MSRCV1, 10% (b) 3Sources,10% (c) HW,10%



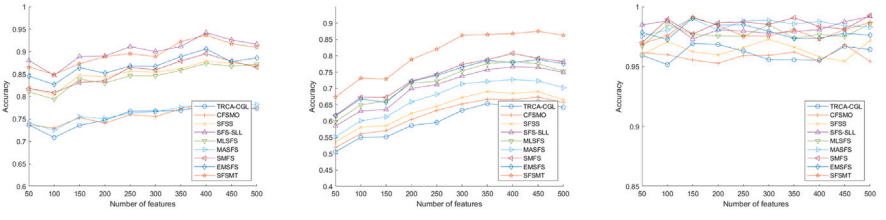
(d) Cal-7,10% (e) WebKB, 10% (f) ORL,10%



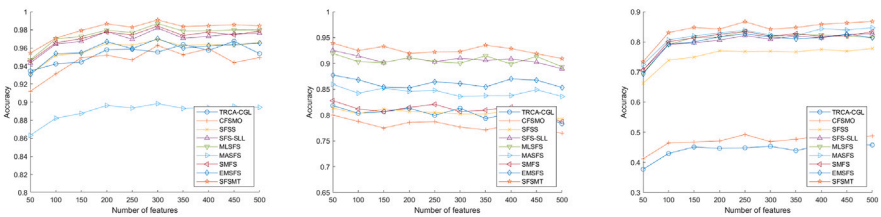
(g) MSRCV1, 20% (h) 3Sources,20% (i) HW,20%



(j) Cal-7,20% (k) WebKB, 20% (l) ORL,20%

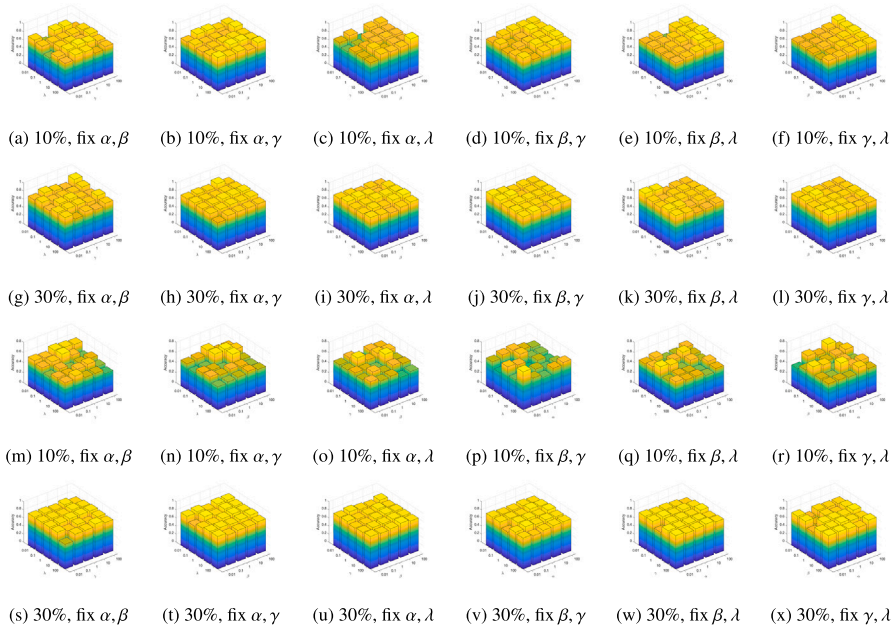


(m) MSRCV1, 30% (n) 3Sources,30% (o) HW,30%



(p) Cal-7,30% (q) WebKB, 30% (r) ORL,30%

Fig. 2. The accuracy of all models at different numbers of features.



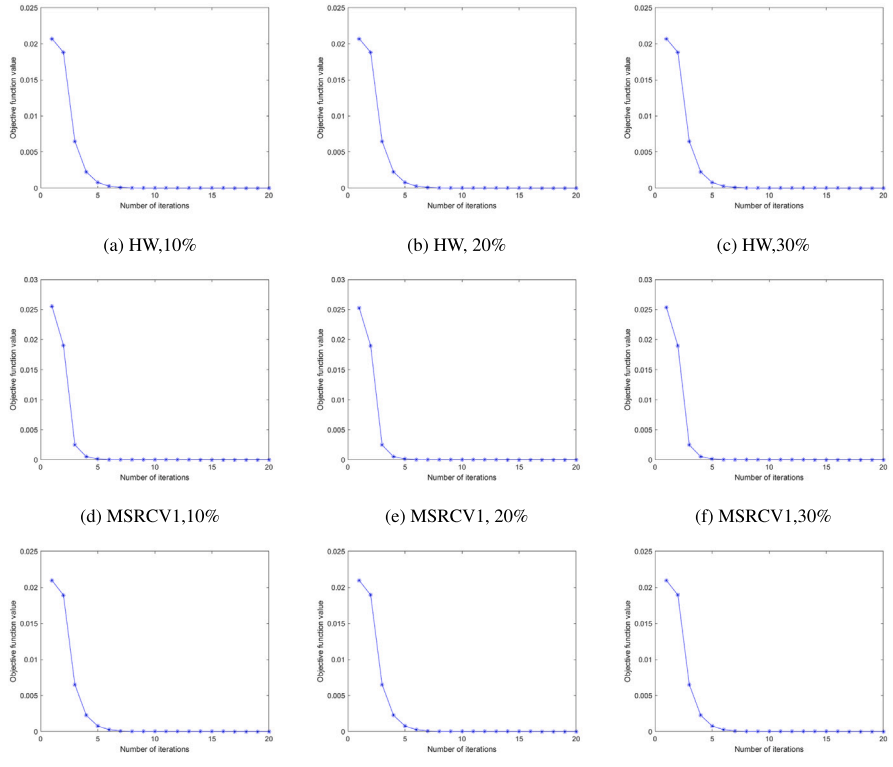
(a) 10%, fix α, β (b) 10%, fix α, γ (c) 10%, fix α, λ (d) 10%, fix β, γ (e) 10%, fix β, λ (f) 10%, fix γ, λ

(g) 30%, fix α, β (h) 30%, fix α, γ (i) 30%, fix α, λ (j) 30%, fix β, γ (k) 30%, fix β, λ (l) 30%, fix γ, λ

(m) 10%, fix α, β (n) 10%, fix α, γ (o) 10%, fix α, λ (p) 10%, fix β, γ (q) 10%, fix β, λ (r) 10%, fix γ, λ

(s) 30%, fix α, β (t) 30%, fix α, γ (u) 30%, fix α, λ (v) 30%, fix β, γ (w) 30%, fix β, λ (x) 30%, fix γ, λ

Fig. 3. Experimental results of parameter sensitivity on the datasets MSRCV1 and 3Sources. (The first two rows (experiments (a)–(l)) are the experimental results of MSRCV1, and the last two rows (experiments (o)–(x)) are the experimental results of 3Sources).



(a) HW,10% (b) HW, 20% (c) HW,30%

(d) MSRCV1,10% (e) MSRCV1, 20% (f) MSRCV1,30%

(g) Cal-7,10% (h) Cal-7, 20% (i) Cal-7,30%

Fig. 4. Convergence curves of the MSFSMT algorithm on HW, MSRACV1, and Cal-7 datasets with different label ratios.

5.4. Impact of parameter sensitivity

In this subsection, we use the grid search method to explore the impact of the balance parameters $\alpha, \beta, \lambda, \gamma$ in the MSFSMT model on the classification accuracy. We fix two of them and set their values to 1, set the other two parameters in {0.01, 0.1, 1, 10, 100}, and set the random walk step size p to 6. The feature subset size is still a fixed value of {50, 100, …, 500}, and we record the highest accuracy under different numbers of features. The experimental results of the MSRCV1 and 3Sources datasets are shown in Fig. 3. It is not difficult to find that the hyperparameters will have a certain degree of fluctuation in the accuracy, but remain roughly stable. For different situations, the optimal parameters are often different, so the grid search method is also needed when selecting hyperparameters.

5.5. Convergence analysis

In this subsection, we verify the convergence of the MSFSMT method through experiments. Fig. 4 shows the figure of the objective function value as the number of iterations changes. From the figure, we can see that the MSFSMT method has good convergence under different label ratios and in different datasets, and the convergence speed is very fast, which proves the effectiveness of the iterative method we proposed.

5.6. Ablation experiment

In order to further verify the effectiveness of each component, we conducted an ablation study on the proposed model. MSFSMT can be roughly divided into three parts: the basic model of multi-view semi-supervised feature selection, multi-order similarity graph, and tensor low-rank learning. We denote the basic model of semi-supervised feature selection as MSFS, the basic model of multi-view semi-supervised feature selection plus the model of multi-order similarity graph as MSFSM, and finally add tensor low-rank learning to the MSFSM model to obtain the MSFSMT model proposed in this paper. Because tensor low rank learning is based on multi-order similarity graphs, it is impossible to give a model of the basic model of multi-view semi-supervised feature selection plus tensor low-rank learning. Below we give the formulas of each ablation model.

- MSFS:

$$\begin{aligned} &\min_{W^v,F} \sum_{v=1}^{v} (\|X^v W^v - F\|_F^2 + \alpha \|W^v\|_{2,1}) + \mathrm{Tr}((F-Y)^\top U (F-Y)) \\ &\text{s.t. } F \geq 0, F^\top F = I, \end{aligned} \tag{23}$$

- MSFSM:

$$\begin{aligned} &\min_{W^v,S^v,F} \sum_{v=1}^{v} (\|X^v W^v - F\|_F^2 + \alpha \|W^v\|_{2,1} + \beta \|S^v - f(A^v)\|_F^2 \\ &\quad + \gamma \,\mathrm{Tr}(F^\top L^v F)) + \mathrm{Tr}((F-Y)^\top U (F-Y)) \\ &\text{s.t. } {S^v}^\top \mathbf{1} = \mathbf{1}, 0 \leq S^v \leq 1, F \geq 0, F^\top F = I, \end{aligned} \tag{24}$$

- MSFSMT:

$$\begin{aligned} &\min_{W^v,S^v,F} \sum_{v=1}^{v} (\|X^v W^v - F\|_F^2 + \alpha \|W^v\|_{2,1} + \beta \|S^v - f(A^v)\|_F^2 \\ &\quad + \gamma \,\mathrm{Tr}(F^\top L^v F)) + \lambda \|S\|_{\circledast} + \mathrm{Tr}((F-Y)^\top U (F-Y)) \\ &\text{s.t. } {S^v}^\top \mathbf{1} = \mathbf{1}, 0 \leq S^v \leq 1, F \geq 0, F^\top F = I, \end{aligned} \tag{25}$$

From the classification accuracy of the above three models in different proportions of labeled data sets shown in Figs. 5–7, we can see that no matter what the ratio of labeled numbers is, the classification accuracy of MSFSM is higher than that of MSFS in most data sets, and the classification accuracy of MSFSMT is higher than that of MSFSM in most data sets, which proves the effectiveness and irreplaceability of each part of our model.

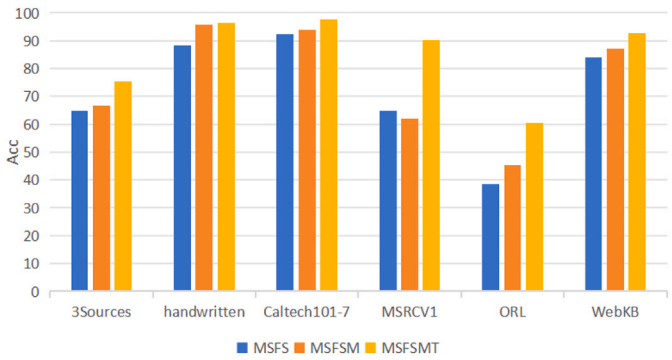


Fig. 5. The classification accuracy of different ablation models on a 10 % labeled dataset.

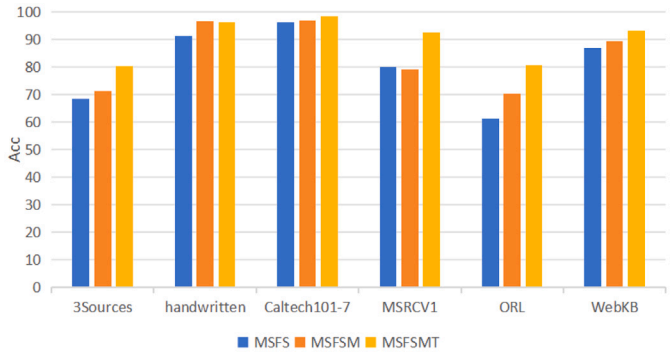


Fig. 6. The classification accuracy of different ablation models on a 20 % labeled dataset.

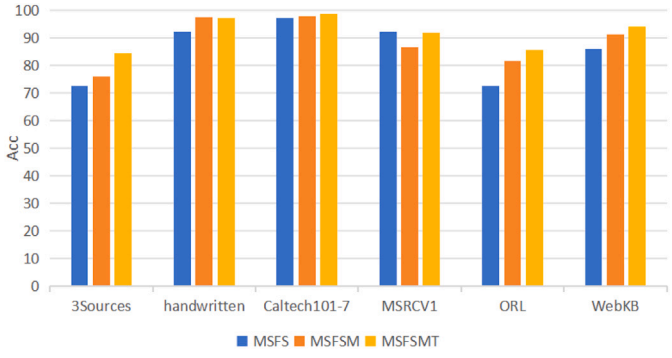


Fig. 7. The classification accuracy of different ablation models on a 30 % labeled dataset.

5.7. Running time analysis

In Section 4.4, we not only provide a complexity analysis of the model, but we also record the actual running time of the model to prove the practical feasibility of the model. Table 3 shows the time it takes to run all feature selection models once. From Table 3, we can see that the running time of our model is not much different from the fastest multi-view semi-supervised feature selection model, and is within a reasonable range, which proves the practical feasibility of the model proposed in this paper.

Table 3
The running time (in seconds) of one iteration for each comparison method.

	TRCA-CGL	CFSMO	SFSS	SFS-SLL	MLSFS	MASFS	SMFS	EMSFS	MSFSMT
3Sources	89.9595	26.8606	12.5363	30.2559	36.5288	128.2355	102.5842	0.0411	10.9015
handwritten	52.6884	30.2337	26.3050	0.8648	77.5661	304.7852	206.2384	0.2596	20.8065
Caltech101-7	53.5734	27.7353	16.5431	19.0421	39.5523	194.2334	146.5327	1.2248	13.6631
MSRCV1	3.0132	29.1155	3.2604	2.0087	9.2487	50.8465	20.2086	0.1772	1.1097
ORL	3.8105	5.9161	3.5446	1.645	21.9067	74.5211	49.7991	0.5575	1.9482
WebKB	3.4570	6.0994	5.0367	3.6153	7.2665	33.9982	13.8652	0.2849	1.1416

6. Conclusion

In this paper, we propose a new MSFSMT method, which is not only an effective extension of the few existing multi-view semi-supervised feature selection methods, but also studies the application of multi-order similarity learning and tensor learning in multi-view semi-supervised feature selection. This method can not only use the neighborhood information of different orders to build a more reliable graph structure, but also use tensor low rank to explore high-order connections between different views. In addition, the model can adaptively obtain a consensus indicator matrix and make full use of complementary information between views. We not only designed an iterative method for the objective function, but also proved its convergence experimentally. Experiments on multiple basic datasets show the superiority of the MSFSMT method.

Although this method has achieved good results, in the future, we still need to reduce hyperparameters to enhance the applicability of the model, and we can also extend this method to handle incomplete data.

CRediT authorship contribution statement

Hangyu Chen: Writing – original draft. **Xijiong Xie:** Writing – review & editing. **Yujie Xiong:** Writing – review & editing.

Funding head

This work is supported by National Natural Science Foundation of China (No. 61906101). This work is also supported by the Ningbo Municipal Natural Science Foundation of China (No. 2023J115).

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.

Data availability

Data will be made available on request.

References

[1] D. Hong, L. Gao, N. Yokoya, J. Yao, J. Chanussot, Q. Du, B. Zhang, More diverse means better: multimodal deep learning meets remote-sensing imagery classification, IEEE Trans. Geosci. Remote Sens. 59 (2020) 4340–4354.
[2] S. Shi, F. Nie, R. Wang, X. Li, Self-weighting multi-view spectral clustering based on nuclear norm, Pattern Recognit. 124 (2022) 108429.
[3] J. Lu, X. Xie, Y. Xiong, Multi-view hypergraph regularized Lp norm least squares twin support vector machines for semi-supervised learning, Pattern Recognit. 156 (2024) 110753.
[4] Y. Wang, X. Lin, L. Wu, W. Zhang, Q. Zhang, X. Huang, Robust subspace clustering for multi-view data by exploiting correlation consensus, IEEE Trans. Image Process. 24 (2015) 3939–3949.
[5] E.J. Keogh, A. Mueen, Curse of dimensionality, Encycl. Mach. Learn. Data Min. 2017 (2017) 314–315.
[6] R. Shang, Y. Meng, W. Wang, F. Shang, L. Jiao, Local discriminative based sparse subspace learning for feature selection, Pattern Recognit. 92 (2019) 219–230.
[7] X. He, D. Cai, P. Niyogi, Laplacian score for feature selection, Adv. Neural Inf. Process. Syst. 18 (2005) 507–514.
[8] F. Nie, W. Zhu, X. Li, Unsupervised feature selection with structured graph optimization, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 30, 2016.
[9] K. Benabdeslem, M. Hindawi, Efficient semi-supervised feature selection: constraint, relevance, and redundancy, IEEE Trans. Knowl. Data Eng. 26 (2013) 1131–1143.
[10] R. Sheikhpour, M.A. Sarram, S. Gharaghani, M.A.Z. Chahooki, A survey on semi-supervised feature selection methods, Pattern Recognit. 64 (2017) 141–158.
[11] F. Nie, H. Huang, X. Cai, C. Ding, Efficient and robust feature selection via joint l-2, 1-norms minimization, Adv. Neural Inf. Process. Syst. 23 (2010) 1813–1821.
[12] X. Wu, X. Xu, J. Liu, H. Wang, B. Hu, F. Nie, Supervised feature selection with orthogonal regression and feature weighting, IEEE Trans. Neural Netw. Learn. Syst. 32 (2020) 1831–1838.
[13] N. Ziraki, F. Dornaika, A. Bosaghzadeh, Multiple-view flexible semi-supervised classification through consistent graph construction and label propagation, Neural Networks 146 (2022) 174–180.
[14] F. Wang, L. Zhu, L. Xie, Z. Zhang, M. Zhong, Label propagation with structured graph learning for semi-supervised dimension reduction, Knowl.-Based Syst. 225 (2021) 107130.
[15] Z. Kang, C. Peng, Q. Cheng, X. Liu, X. Peng, Z. Xu, L. Tian, Structured graph learning for clustering and semi-supervised classification, Pattern Recognit. 110 (2021) 107627.
[16] F. Dornaika, Y.E. Traboulsi, Joint sparse graph and flexible embedding for graph-based semi-supervised learning, Neural Networks 114 (2019) 91–95.
[17] X. Dong, L. Zhu, X. Song, J. Li, Z. Cheng, Adaptive collaborative similarity learning for unsupervised multi-view feature selection, arXiv preprint arXiv:1904.11228, 2019.
[18] X. Bai, L. Zhu, C. Liang, J. Li, X. Nie, X. Chang, Multi-view feature selection via nonnegative structured graph learning, Neurocomputing 387 (2020) 110–122.
[19] X. Chen, R. Chen, Q. Wu, F. Nie, M. Yang, R. Mao, Semisupervised feature selection via structured manifold learning, IEEE Trans. Cybern. 52 (2021) 5756–5766.
[20] D. Shi, L. Zhu, J. Li, Z. Cheng, Z. Liu, Binary label learning for semi-supervised feature selection, IEEE Trans. Knowl. Data Eng. 35 (2021) 2299–2312.
[21] C. Shi, G. An, R. Zhao, Q. Ruan, Q. Tian, Multiview hessian semisupervised sparse feature selection for multimedia analysis, IEEE Trans. Circuits Syst. Video Technol. 27 (2016) 1947–1961.
[22] Y. Li, X. Shi, C. Du, Y. Liu, Y. Wen, Manifold regularized multi-view feature selection for social image annotation, Neurocomputing 204 (2016) 135–141.
[23] Z. Zhao, H. Liu, Semi-supervised feature selection via spectral analysis, in: Proceedings of the 2007 SIAM International Conference on Data Mining, SIAM, 2007, pp. 641–646.
[24] G. Doquire, M. Verleysen, A graph laplacian based approach to semi-supervised feature selection for regression problems, Neurocomputing 121 (2013) 5–13.
[25] J. Xu, B. Tang, H. He, H. Man, Semisupervised feature selection based on relevance and redundancy criteria, IEEE Trans. Neural Netw. Learn. Syst. 28 (2016) 1974–1984.
[26] Z. Xu, I. King, M.R.-T. Lyu, R. Jin, Discriminative semi-supervised feature selection via manifold regularization, IEEE Trans. Neural Netw. 21 (2010) 1033–1047.
[27] X. Chen, G. Yuan, F. Nie, Z. Ming, Semi-supervised feature selection via sparse rescaled linear square regression, IEEE Trans. Knowl. Data Eng. 32 (2018) 165–176.
[28] B. Tang, L. Zhang, Local preserving logistic I-relief for semi-supervised feature selection, Neurocomputing 399 (2020) 48–64.
[29] Z. Zeng, X. Wang, F. Yan, Y. Chen, Local adaptive learning for semi-supervised feature selection with group sparsity, Knowl.-Based Syst. 181 (2019) 104787.
[30] W. Zhong, X. Chen, F. Nie, J.Z. Huang, Adaptive discriminant analysis for semi-supervised feature selection, Inf. Sci. 566 (2021) 178–194.
[31] J. Lai, H. Chen, T. Li, X. Yang, Adaptive graph learning for semi-supervised feature selection with redundancy minimization, Inf. Sci. 609 (2022) 465–488.
[32] C. Wang, X. Chen, G. Yuan, F. Nie, M. Yang, Semisupervised feature selection with sparse discriminative least squares regression, IEEE Trans. Cybern. 52 (2021) 8413–8424.
[33] R. Sheikhpour, K. Berahmand, S. Forouzandeh, Hessian-based semi-supervised feature selection using generalized uncorrelated constraint, Knowl.-Based Syst. 269 (2023) 110521.
[34] J. Zhao, K. Lu, X. He, Locality sensitive semi-supervised feature selection, Neurocomputing 71 (2008) 1842–1849.
[35] Z. Ma, F. Nie, Y. Yang, J.R. Uijlings, N. Sebe, A.G. Hauptmann, Discriminating joint feature analysis for multimedia data understanding, IEEE Trans. Multimed. 14 (2012) 1662–1672.
[36] C. Shi, Q. Ruan, G. An, C. Ge, Semi-supervised sparse feature selection based on multi-view laplacian regularization, Image Vis. Comput. 41 (2015) 1–10.
[37] F. Nie, G. Cai, J. Li, X. Li, Auto-weighted multi-view learning for image clustering and semi-supervised classification, IEEE Trans. Image Process. 27 (2017) 1501–1511.
[38] C. Shi, Z. Gu, C. Duan, Q. Tian, Multi-view adaptive semi-supervised feature selection with the self-paced learning, Signal Process. 168 (2020) 107332.

[39] C. Zhang, B. Jiang, Z. Wang, J. Yang, Y. Lu, X. Wu, W. Sheng, Efficient multi-view semi-supervised feature selection, Inf. Sci. 649 (2023) 119675.
[40] W. Guo, Z. Wang, W. Du, Robust semi-supervised multi-view graph learning with sharable and individual structure, Pattern Recognit. 140 (2023) 109565.
[41] Z. Zhang, G. Ely, S. Aeron, N. Hao, M. Kilmer, Novel methods for multilinear data completion and de-noising based on tensor-SVD, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 3842–3849.
[42] J. Wu, Z. Lin, H. Zha, Essential tensor learning for multi-view spectral clustering, IEEE Trans. Image Process. 28 (2019) 5910–5922.
[43] Z. Lin, Z. Kang, Graph filter-based multi-view attributed graph clustering., in: IJCAI, 2021, pp. 2723–2729.
[44] S. Cao, W. Lu, Q. Xu, GraRep: learning graph representations with global structural information, in: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, 2015, pp. 891–900.
[45] E.J. Candès, X. Li, Y. Ma, J. Wright, Robust principal component analysis?, J. ACM 58 (2011) 1–37.
[46] P. Zhu, W. Zhu, Q. Hu, C. Zhang, W. Zuo, Subspace clustering guided unsupervised feature selection, Pattern Recognit. 66 (2017) 364–374.
[47] S.T. Roweis, L.K. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (2000) 2323–2326.
[48] Y. Liu, R. Jin, L. Yang, Semi-supervised multi-label learning by constrained non-negative matrix factorization, in: AAAi, vol. 6, 2006, pp. 421–426.
[49] J. Huang, F. Nie, H. Huang, A new simplex sparse learning model to measure data similarity for clustering, in: Twenty-Fourth International Joint Conference on Artificial Intelligence, 2015.
[50] W. Hu, D. Tao, W. Zhang, Y. Xie, Y. Yang, The twist tensor nuclear norm for video completion, IEEE Trans. Neural Netw. Learn. Syst. 28 (2016) 2961–2973.
[51] C. Liang, L. Wang, L. Liu, H. Zhang, F. Guo, Multi-view unsupervised feature selection with tensor robust principal component analysis and consensus graph learning, Pattern Recognit. 141 (2023) 109632.
[52] Z. Cao, X. Xie, Multi-view unsupervised complementary feature selection with multi-order similarity learning, Knowl.-Based Syst. 283 (2024) 111172.
[53] C. Zhang, L. Zhu, D. Shi, J. Zheng, H. Chen, B. Yu, Semi-supervised feature selection with soft label learning, IEEE/CAA J. Autom. Sin. (2022) 1–13.
[54] B. Jiang, X. Wu, X. Zhou, Y. Liu, A.G. Cohn, W. Sheng, H. Chen, Semi-supervised multiview feature selection with adaptive graph learning, IEEE Trans. Neural Netw. Learn. Syst. 35 (2024) 3615–3629.

Author biography

Hangyu Chen received a bachelor's degree in engineering from Lanzhou University of Finance and Economics, China, in 2020. He is currently pursuing a master's degree at Ningbo University in Ningbo, China. His research interests include support vector machines, multi-view learning, feature selection, semi-supervised learning etc.

Xijiong Xie received the Ph.D. degree from the Pattern Recognition and Machine Learning Research Group, Department of Computer Science and Technology, East China Normal University, in 2016. He is currently an Associate Professor with the Faculty of Electrical Engineering and Computer Science, Ningbo University, China. He has over 30 publications at peer-reviewedjournals and conferences in his research areas, such as IEEE TRANSACTIONS ON KNOWLEDGE AND DATAENGINEERING(TKDE), IEEE TRANSACTIONS ON NEURALNETWORKS AND LEARNING SYSTEMS(TNNLS), IEEE TRANSACTIONS ON CYBERNETICS, Expert systems with application, Pattern Recognition, Neurocomputing, and Information Fusion. His research interests include kernel methods, support vector machines, multi-view learning, and deep learning.

Yujie Xiong received his Ph.D. in Computer Science from the East China Normal University in June 2018. He is currently an Associate Professor in the School of Electronic and Electrical Engineering at Shanghai University of Engineering Science, China. His research interests include biometric, document image analysis, and knowledge graph based application, where he has published more than 30 academic publications.