Neurocomputing 413 (2020) 72–84

Contents lists available at ScienceDirect

Neurocomputing

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

Subspace learning for unsupervised feature selection via adaptive structure learning and rank approximation

Ronghua Shang *, Kaiming Xu, Licheng Jiao

Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, Xidian University, Xi'an, Shaanxi Province 710071, China

ARTICLE INFO

Article history:
Received 21 January 2020
Revised 21 May 2020
Accepted 26 June 2020
Available online 7 July 2020
Communicated by Xin Luo

Keywords:
Subspace learning
Adaptive structure learning
Rank constraint
Projection matrix
Feature selection

ABSTRACT

Traditional unsupervised feature selection methods usually construct a fixed similarity matrix. This matrix is sensitive to noise and becomes unreliable, which affects the performance of feature selection. The researches have shown that both the global reconstruction information and local structure information are important for feature selection. To solve the above problem effectively and make use of the global and local information of data simultaneously, a novel algorithm is proposed in this paper, called subspace learning for unsupervised feature selection via adaptive structure learning and rank approximation (SLASR). Specifically, SLASR learns the manifold structure adaptively, thus the preserved local geometric structure can be more accurate and more robust to noise. As a result, the learning of the similarity matrix and the low-dimensional embedding is completed in one step, which improves the effect of feature selection. Meanwhile, SLASR adopts the matrix factorization subspace learning framework. By minimizing the reconstruction error of subspace learning residual matrix, the global reconstruction information of data is preserved. Then, to guarantee more accurate manifold structure of the similarity matrix, a rank constraint is used to constrain the Laplacian matrix. Additionally, the $l_{2,1/2}$ regularization term is used to constrain the projection matrix to select the most sparse and robust features. Experimental results on twelve benchmark datasets show that SLASR is superior to the six comparison algorithms from the literature.

1. Introduction

In machine learning, pattern recognition, data mining, and other areas, data often has high dimensionality [1,2]. How to deal with high-dimensional data efficiently and improve the efficiency of learning algorithms is a difficult task. As noise and redundant features are inevitable, the efficiency of data processing is reduced. And only a small number of features are discriminative and important, so it is necessary to reduce the dimension of data [3,4]. The commonly used dimensionality reduction techniques include feature extraction and feature selection [5,6]. Feature extraction finds a low-dimensional subspace based on a projection so that the original data can be well represented [7,8]. Feature selection selects a representative feature subset from the feature set to represent the original data [9]. Compared with feature extraction, feature selection preserves the semantic information of the original features, thus improving the interpretability of the corresponding models [10,11].

Generally, feature selection algorithms can be divided into three categories: supervised, semi-supervised and unsupervised feature selection algorithms [12,13]. Supervised feature selection methods require the class labels of samples, but marking a large amount of data requires high labor costs [14]. The semi-supervised feature selection methods use both labeled and unlabeled samples to select the discriminative features [15,16]. In unsupervised feature selection, the intrinsic information of data is used to select features without any class information [17]. In real-world applications, the class labels of dattrima are often unavailable, so the unsupervised feature selection algorithms show great advantages. According to different search strategies, feature selection methods include three categories, namely filter, wrapper and embedded methods [18]. The filtering methods use the intrinsic properties of data to evaluate the importance of features [19]. The wrap methods rely on specific learning algorithms to select features [20,21]. For embedded methods, a model should be built and feature selection is completed during the learning process of this model [22]. The filter methods have low time cost. The wrapped methods can achieve great effectiveness, but the time cost is relatively high. The embedded methods can achieve great performance with low time cost. Therefore, we propose a novel embedded unsupervised feature selection algorithm in this paper.

Subspace learning can obtain the low-dimensional representation of the original data, and can be applied to feature selection

* Corresponding author.
E-mail address: rhshang@mail.xidian.edu.cn (R. Shang).

https://doi.org/10.1016/j.neucom.2020.06.111
0925-2312/

methods by using the matrix decomposition strategy. In [23], the matrix factorization feature selection (MFFS) method is proposed. This method applies the matrix decomposition strategy to the subspace learning framework, so the subspace spanned by feature subset can well characterize the subspace spanned by feature set. Then, unsupervised feature selection via maximum projection and minimum redundancy (MPMR) is proposed in [24]. In this approach, feature subset is well evaluated and the low redundancy of the selected features is guaranteed. The two methods mentioned above use the matrix factorization technique to select a representative feature subset to obtain good feature selection performance, but they both neglect the local manifold structure of data. On the contrary, the local information of data is fully considered in the SLASR proposed in this paper. And the manifold structure can be learned adaptively, so the performance of feature selection is improved.

Researches have shown that high-dimensional data often contains important manifold structures. And making use of the local geometric structures can improve the nonlinear learning performance of algorithms. Many manifold learning algorithms, such as Local Linear Embedding (LLE) [25], Laplacian Eigenmap (LE) [26], and Locality Preserving Projections (LPP) [27] have been proposed to preserve the local structures of data into low-dimensional embedding. Many classic manifold structure preservation feature selection methods have been proposed, such as Laplacian Score (LapScor) [21], spectral feature selection (SPEC) [28], unsupervised discriminant feature selection (UDFS) [29], multi-clustering feature selection (MCFS) [14], joint embedding learning and sparse regression (JELSR) [30] and Non-negative spectral learning and sparse regression-based dual-graph regularized feature selection (NSSRD) [4]. For MCFS, JELSR and NSSRD, they first construct the similarity matrix to preserve the manifold structures, and then learn the low-dimensional embedding to select features. As this process is completed in two steps, so it is difficult to obtain the optimal results. In order to obtain more accurate manifold information, unsupervised feature selection with structured graph optimization is proposed in [31]. In this method, the similarity matrix can be learned adaptively, so manifold structures can be well preserved. Then $l_{2,1}$-*norm* is used to constrain the transformation matrix to select the representative features, so SOGFS can obtain good feature selection effectiveness. However, this algorithm still has some inadequacies. Although the manifold structures can be well preserved, the global reconstruction information of data is not considered. In other words, the selected features cannot reconstruct the feature set of the original data well, so the effect of feature selection needs to be improved.

In this paper, we propose a novel algorithm called subspace learning for unsupervised feature selection via adaptive structure learning and rank approximation (SLASR). Inspired by the method in unsupervised feature selection with structured graph optimization (SOGFS), SLASR constructs an adaptive similarity matrix, so the manifold structure can be learned adaptively during the iteration process, making the manifold structure more accurate and the similarity matrix more robust to noise. Additionally, SLASR not only preserves the local structure information, but also retains the global reconstruction information of data by using the subspace learning framework. By minimizing the reconstruction error of subspace learning residual matrix, the subspace spanned by the selected feature subset can well characterize the subspace spanned by the original feature set, resulting in more accurate reconstruction information of data. Since SOGFS ignores the global information of data, the feature selection performance is reduced. In contrast, SLASR makes full use of the advantages of SOGFS on the one hand, and makes up for its shortcomings effectively by using the subspace learning framework on the other hand, so SLASR can obtain better feature selection effect. Then, since the original data has class c samples, they can be clustered into c categories. To make the manifold structure more accurate, the rank constraint is used to guarantee that the similarity matrix contains c connected components. Compared with $l_{2,1}$-*norm*, $l_{2,1/2}$ regularization term can select the most sparse and robust features, so we apply this term to SLASR to improve the effect of feature selection. We present efficient iterative update rules and evaluate the importance of different features according to the matrix $\boldsymbol{W}$. Then we perform clustering for the selected features. By comparing SLASR with six comparison algorithms, the experimental results on twelve datasets show that SLASR achieves better performance.

The novelties and contributions are highlighted as follows:

1) An adaptive similarity matrix is constructed under the framework of subspace learning. Thus, the manifold structure can be updated adaptively in the learning process of low-dimensional embedding, and the local structure information can be well preserved. By using the subspace learning framework, the global reconstruction information of data can also be retained.
2) The rank constraint is used to constrain the Laplacian matrix. Since this constraint guarantees that the similarity matrix contains c connection components, the learned manifold information will be more accurate.
3) The projection matrix is constrained by $l_{2,1/2}$ regularization term. By using this term, the most sparse and robust features are selected.

The rest of this paper is organized as follows. In Section 2, we introduce the related work of the proposed algorithm. And in Section 3, we present the proposed SLASR, give the iterative updating formulas and computational complexity analysis. Then, we demonstrate the convergence of SLASR. In Section 4, we present the experimental results of SLASR and compare SLASR with other algorithms. Then in Section 5, we summarize the whole paper.

2. Related work

The related notations and the related researches of the proposed SLASR are introduced in this section. First, some related notations used in SLASR are presented, then a brief introduction to the framework of subspace learning and the probabilistic neighbors-based manifold structure preservation strategy are given.

2.1. Related notations

Some related notations are listed in this section. Here, scalars, vectors, and matrices are represented as italic letters, italic bold lowercase letters, and italic bold uppercase letters, respectively. Suppose $X = [x_1{}^T; x_2{}^T; \ldots; x_n{}^T] \in \Re^{n\times d}$ is the original data matrix, where $x_i \in \Re^{d\times 1}$ represents a data sample, n is the total number of samples, and d is the number of features contained in each sample. For the given square matrix $\boldsymbol{M}$, $\boldsymbol{Tr}(\boldsymbol{M})$ represents the trace of $\boldsymbol{M}$. The f_r-*norm* of vector $x \in \Re^d$ is defined as:

$$\| x \|_r = \left(\sum_{i=1}^{d} |x_i|^r \right)^{1/r} \tag{1}$$

where $\boldsymbol{x}_i$ is the ith element of the vector $\boldsymbol{x}$. For the matrix $X \in \Re^{n\times d}$, its l_p-*norm* is defined as follows:

$$\| X \|_p = \left(\sum_{i=1}^{n} \sum_{j=1}^{d} |X_{ij}|^p \right)^{1/p} \tag{2}$$

Its $l_{p,q}$-*norm* has the following form:

$$\| X \|_{p,q} = \left(\sum_{i=1}^{n} \left(\sum_{j=1}^{d} |X_{ij}|^p \right)^{q/p} \right)^{1/q} \tag{3}$$

where X_{ij} represents the element of the ith row and the jth column of the matrix $\boldsymbol{X}$. In this paper, the importance of different features can be evaluated by calculating f_2-*norm* for each row of the projection matrix. When $p = 2$, $q = 1/2$, the l_2-*norm* of the subspace learning residual matrix and the $l_{2,1/2}$ regularization term for constraining the projection matrix can be obtained.

2.2. The framework of subspace learning

In order to characterize the subspace spanned by the original feature set using the subspace spanned by the feature subset, Wang et al. proposed MFFS [23] by applying the matrix factorization strategy to the measurement of subspace distance. And in [32], Zhou et al. gave a more complete description of the MFFS framework, which can be expressed as follows:

$$\begin{aligned} &\arg \min_{W,H} \| X - XWH \|_2^2 \\ &s.t. \quad W \in \{0,1\}^{d\times l},\ W^T 1_{d\times 1} = 1_{l\times 1}, \| W 1_{l\times 1} \|_0 = l. \end{aligned} \tag{4}$$

where $X \in \Re^{n\times d}$ is the original data matrix, of which each row represents a data sample, and each column represents a feature. l is the number of the selected features. $H \in \Re^{l\times d}$ denotes a coefficient matrix and can be used to reconstruct the original feature set. W is the feature selection matrix, indicating that l representative features can be selected from d features. For the matrix $\boldsymbol{W}$, only 0–1 elements can be taken. And the constraints above guarantee $\boldsymbol{W}$ can play a role. Specifically, expression $W^T 1_{d\times 1} = 1_{l\times 1}$ indicates that each column of the matrix $\boldsymbol{W}$ has only one 1 element, and the other elements are all zero. Since 1 element of different columns may arise in the same row, at most l features can be selected. The expression $\| W 1_{l\times 1} \|_0 = l$ indicates that the matrix $\boldsymbol{W}$ has l non-zero rows, so each feature can only be selected once. Finally, exact l features can be selected.

2.3. The probabilistic neighbors-based manifold structure preservation strategy

High-dimensional data often contains important local structure information, which can improve the nonlinear learning ability of algorithms. In recent years, local structure preservation strategy has been applied to feature selection. To preserve the manifold structure, a fixed similarity matrix can be constructed. Since noise is unavoidable in practical scenarios, to make the preserved manifold structure more accurate and guarantee the robustness to noise, Nie et al. introduced a probabilistic neighbors-based manifold structure preservation method [31], which can be described as follows:

Denote $X \in \Re^{n\times d}$ as a data matrix, and $x \in \Re^{d\times 1}$ as a data sample. According to the method in [31], define z_{ij} as the probability that the sample $\boldsymbol{x}_i$ is connected to the jth sample. Since similar samples have a greater probability of being interconnected, z_{ij} is inversely proportional to the distance between the samples. The square of the Euclidean distance can be used as the distance metric. To get the value of z_{ij}, the following expression should be solved:

$$\min_{z_i^T 1 = 1, 0 \leqslant z_{ij} \leqslant 1} \sum_{ij} \left(\| x_i - x_j \|_2^2 z_{ij} + \alpha z_{ij}^2 \right) \tag{5}$$

Since the connection probability of two samples can measure the similarity between them, z_{ij} can be regarded as the element of the ith row and the jth column of the similarity matrix $\boldsymbol{Z}$. In Eq. (5), $\boldsymbol{z}_i$ represents the ith column of the similarity matrix. α is a regularization term parameter, which is necessary and prevents the trivial solution from the Eq. (5).

3. The proposed method

In this section, we introduce the proposed SLASR. SLASR mainly consists of three parts: sparse and robust subspace learning, adaptive manifold structure preservation and feature selection. Then we provide the updating formulas, computational complexity analysis, and convergence proof of SLASR.

3.1. Sparse and robust subspace learning

Since the issue of Eq. (4) is a combinatorial optimization problem, it is difficult to solve. To deal with this issue efficiently, Eq. (4) can be relaxed to a continuous optimization problem [32]. In general, since the number of selected features l is much smaller than the feature dimension d, the feature selection matrix $\boldsymbol{W}$ is sparse and has many zero values. So the constraint $W \in \{0,1\}^{d\times l}$ can be relaxed to a non-negative constraint. And the hard constraints $W^T 1_{d\times 1} = 1_{l\times 1}, \| W 1_{l\times 1} \|_0 = l$ can be relaxed to $h(W) \leqslant l$, which can be used to measure the sparsity of rows. Based on the above analyses, Eq. (4) can be relaxed as the following form:

$$\begin{aligned} &\arg \min_{W,H} \| X - XWH \|_2^2 + \gamma_1 h(W) \\ &\qquad s.t. \quad W \in \Re_+^{d\times K} \end{aligned} \tag{6}$$

For $h(W)$, traditional methods usually select the group lasso, thus $h(W) = \| W \|_{2,1}$. Recently, researches have shown that $l_{2,1/2}$ regularization term can obtain great performance. For $l_{2,1/2}$ regularization term, Wang et al. [33] proposed $l_{2,p}$ matrix pseudo norm based least square regression feature selection framework. To verify the effectiveness of the proposed algorithm, the variable p is adjusted in the range of [0, 1]. And the experimental results show that when p takes 1/2, the algorithm achieves the lowest classification error rate on all test datasets when the number of selected features is different, which fully demonstrates that the $l_{2,1/2}$ regularization term can select the discriminative and robust features.

Additionally, to observe the sparsity of $l_{2,1/2}$ regularization term intuitively, here we presented the contour maps [34] of $l_{2,2}$-norm, $l_{2,1}$-norm, and $l_{2,1/2}$ regularization term, as shown in Fig. 1.

It can be seen from Fig. 1 that $l_{2,1/2}$ regularization term tends to obtain sparser results than $l_{2,2}$-norm and $l_{2,1}$-norm during the optimization process, so the sparsity can be improved by using this regularization term.

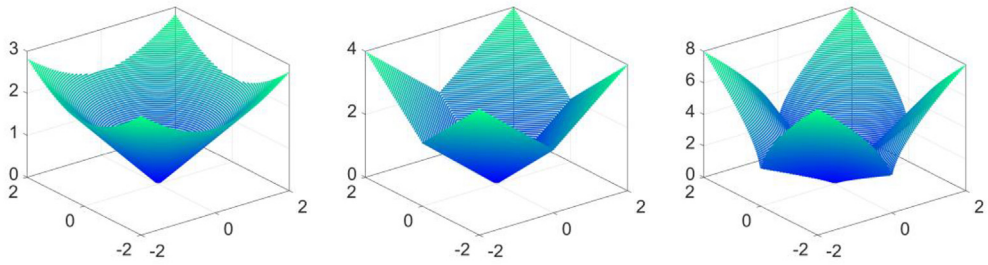
Benefiting from the robustness and sparsity of $l_{2,1/2}$ term, the matrix $\boldsymbol{W}$ is constrained by this term and the following formula is obtained:

$$h(W) = \| W \|_{2,1/2}^{1/2} \tag{7}$$

Substituting Eq. (7) into Eq. (6), we get:

$$\begin{aligned} &\arg \min_{W,H} \| X - XWH \|_2^2 + \gamma_1 \| W \|_{2,1/2}^{1/2} \\ &\qquad s.t. \quad W \in \Re_+^{d\times K} \end{aligned} \tag{8}$$

where γ_1 is a balance parameter, and $\Re_+^{d\times K}$ is the set of matrices with dimension $d \times K$. Now $\boldsymbol{W}$ is regard as a projection matrix, and K is the dimension of the subspace. The subspace dimension K is not always equal to the number of selected features l [32]. In order to complete feature selection more effectively under the subspace learning framework, it is usually set $K \geqslant l$.



(a) $l_{2,2}$-norm (b) $l_{2,1}$-norm (c) $l_{2,1/2}$ regularization term

Fig. 1. The contour maps of $l_{2,2}$-norm, $l_{2,1}$-norm, and $l_{2,1/2}$ regularization term.

3.2. Adaptive manifold structure preservation

Since the fixed similarity matrix is sensitive to noise, an adaptive manifold structure preservation strategy is adopted in this section to preserve the manifold structure which is robust to noise. Through the Eq. (5), the similarity matrix $\boldsymbol{Z}$ can be defined. Denote c as the number of sample categories. For the matrix $\boldsymbol{Z}$, it is expected to contain c connected components to retain a more accurate manifold structure. However, it is difficult to get a similarity matrix with c connection components directly. Researches have shown that when the following rank constraint is applied to the Laplacian matrix $\boldsymbol{LZ}$ [35], the similarity matrix will satisfy the above condition:

$$\text{rank}(\boldsymbol{L_Z}) = n\text{-}c \tag{9}$$

For the obtained similarity matrix $\boldsymbol{Z}$, the Laplacian matrix $\boldsymbol{LZ}$ is defined as follows:

$$L_Z = D - \frac{Z^T + Z}{2} \tag{10}$$

where $\boldsymbol{D}$ is a diagonal matrix and the ith element of the diagonal is defined as $\sum_j \frac{(z_{ij}+z_{ji})}{2}$.

Then, the ith smallest eigenvalue of the Laplacian matrix $\boldsymbol{L_Z}$ is defined as $\sigma_i(L_Z)$. Since the matrix $\boldsymbol{L_Z}$ is semi-definite, all its eigenvalues are greater than or equal to zero. It can be proved that the Eq. (9) is equivalent to the following equation:

$$\sum_{i=1}^{c} \sigma_i(L_Z) = 0 \tag{11}$$

where c is the number of categories of samples. Since Eq. (11) is difficult to solve, the following formulas can be obtained based on Ky Fan's theory [36]:

$$\begin{aligned} &\sum_{i=1}^{c} \sigma_i(L_Z) = \min \ Tr\left(G^T L_Z G\right) \\ &s.t. \quad G \in \Re^{n\times c}, G^T G = I \end{aligned} \tag{12}$$

To guarantee that the constructed similarity matrix has c connected components, the Eqs. (5) and (12) are combined to obtain the following expression:

$$\begin{aligned} &\min \ \sum_{i,j} \left(\| x_i - x_j \|_2^2 z_{ij} + \alpha z_{ij}^2 \right) + 2\lambda Tr(G^T L_Z G) \\ &s.t. \quad \forall i, z_i^T 1 = 1, 0 \leqslant z_{ij} \leqslant 1, G \in \Re^{n\times c}, G^T G = I \end{aligned} \tag{13}$$

Since the item $Tr(\boldsymbol{G^T L_Z G})$ should be close to zero, the parameter λ should be set large enough. In the updating process of the similarity matrix $\boldsymbol{Z}$, if the number of connected components is less than c, the parameter λ should be increased, otherwise it should be decreased. In this paper, the parameter λ is adjusted adaptively and is set to 2 times and 1/2 times of the original value when λ needs to be increased and decreased, respectively. By solving Eq. (13), the similarity matrix $\boldsymbol{Z}$ which has c connected components can be obtained, preserving the manifold structure information more accurately.

$S \in \Re^{d\times m}$ is defined as the projection matrix, where d is the total number of features and m is the dimension of the low-dimensional embedding. It is clear that $\boldsymbol{XS}$ is the obtained low-dimensional embedded matrix. In order to preserve the manifold structure of high-dimensional data into low-dimensional embedding, the original data in Eq. (13) is replaced with low-dimensional embedding. And the obtained expression is as follows:

$$\begin{aligned} &\min \ \sum_{i,j} \left(\| S^T x_i - S^T x_j \|_2^2 z_{ij} + \alpha z_{ij}^2 \right) + 2\lambda Tr(G^T L_Z G) \\ &s.t. \quad \forall i, z_i^T 1 = 1, 0 \leqslant z_{ij} \leqslant 1, G \in \Re^{n\times c}, G^T G = I, S^T S = I \end{aligned} \tag{14}$$

Comparing Eqs. (13) and (14), it can be seen that the neighbor information of $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ of the original samples has been preserved into low-dimensional embeddings $\boldsymbol{S^T x_i}$ and $\boldsymbol{S^T x_j}$. Meanwhile, the orthogonal constraint $\boldsymbol{S^T S = I}$ is applied to the projection matrix in the above formula, which can make the feature space more discriminative after dimension reduction [31]. In the formula above, the similarity matrix $\boldsymbol{Z}$ and the projection matrix $\boldsymbol{S}$ can be learned in a single step, which means that manifold learning and feature selection can be performed simultaneously. Thus, better performance can be obtained.

3.3. The framework of SLASR

To preserve the local structure while retaining the global reconstruction information of data, we set $m = K$. Where m is the dimension of low-dimensional embedding and K is the dimension of subspace. So the projection matrix $\boldsymbol{W}$ and the projection matrix $\boldsymbol{S}$ are unified. Then Eqs. (8) and (14) are combined and the obtained objective function of SLASR is as follows:

$$\begin{aligned} &\arg \ \min_{W,H,Z,G} \beta \| X - XWH \|_2^2 + \tfrac{1}{2} \sum_{i,j} \left(\| W^T x_i - W^T x_j \|_2^2 z_{ij} + \alpha z_{ij}^2 \right) \\ &+ \lambda Tr(G^T L_Z G) + \gamma \| W \|_{2,1/2}^{1/2} + \tfrac{\mu}{2} \| W^T W - I \|_2^2 \\ &s.t. \quad \forall i, z_i^T 1 = 1, 0 \leqslant z_{ij} \leqslant 1, G \in \Re^{n\times c}, G^T G = I, W \geqslant 0, H \geqslant 0 \end{aligned} \tag{15}$$

where $\beta > 0$, $\gamma > 0$, and $\mu > 0$ are the balance parameters.

3.4. Feature selection

After solving the objective function of SLASR, the projection matrix $\boldsymbol{W}$ can be obtained, where $W = [w_1; w_2; \ldots; w_d]$, and w_i is the ith row of $\boldsymbol{W}$. Then the value of $||w_i||_2$ is calculated, which can be used to evaluate the importance of the ith feature. And the larger the value of $||w_i||_2$, the more important the ith feature

is. Then all the values of $||w_i||_2$ are sorted in descending order and the corresponding first l features are selected to construct a new data matrix $X_{new} \in \Re^{n\times l}$ to complete the feature selection.

3.5. Connection with SOGFS

As can be seen from Eq. (15), when removing the term of the subspace learning residual matrix, using $l_{2,1}$-*norm* to replace $l_{2,1/2}$ regularization term to constrain the projection matrix $\boldsymbol{W}$, and removing the non-negative constraints imposed on the matrices $\boldsymbol{W}$ and $\boldsymbol{H}$, SLASR degenerates into SOGFS. The objective function of SOGFS is as follows:

$$\begin{aligned} &\arg \min_{W,Z,G} \sum_{i,j} \left(\| W^T x_i - W^T x_j \|_2^2 z_{ij} + \alpha z_{ij}^2 \right) + 2\lambda Tr(G^T L_Z G) \\ &+\gamma \| W \|_{2,1} + \tfrac{\mu}{2} \| W^T W - I \|_2^2 \\ &s.t. \quad \forall i, z_i^T 1 = 1, 0 \leqslant z_{ij} \leqslant 1, G \in \Re^{n\times c}, G^T G = I \end{aligned} \tag{16}$$

3.6. Update rules for SLASR

The update rules of the objective function in Eq. (15) is given in this section. Since it contains four variables $\boldsymbol{W}$, $\boldsymbol{H}$, $\boldsymbol{Z}$, and $\boldsymbol{G}$, it is difficult to solve this objective function directly. Therefore, the alternating iterative update method is used to solve this problem [37]. Two Lagrange multipliers ψ_{ij} and ϕ_{ij} are introduced to guarantee $W \geqslant 0$ and $H \geqslant 0$. First, only the terms related to $\boldsymbol{W}$ and $\boldsymbol{H}$ in Eq. (15) are preserved. By applying the Lagrange multipliers, the following equation can be obtained:

$$\begin{aligned} L(W,H) = \beta \| X - XWH \|_2^2 + \frac{1}{2} \sum_{i,j} \| W^T x_i - W^T x_j \|_2^2 z_{ij} + \gamma \| W \|_{2,1/2}^{1/2} \\ + \frac{\mu}{2} \| W^T W - I \|_2^2 + Tr\left(\psi W^T\right) + Tr\left(\phi H^T\right) \end{aligned} \tag{17}$$

A diagonal matrix is first defined and its ith element is as follows:

$$U_{ii} = \frac{1}{max\left(\| w_i \|_2^{3/2} + \varepsilon \right)} \tag{18}$$

where ε is a small constant and can avoid the denominator being zero.

By using the matrix $\boldsymbol{U}$, $\| W \|_{2,1/2}^{1/2}$ can be rewritten as $Tr\left(W^T U W\right)$. So Eq. (17) can obtain the following form:

$$\begin{aligned} L(W,H) = \beta Tr\left((X - XWH)(X - XWH)^T\right) + Tr\left(W^T X^T L_Z X W\right) \\ +\gamma Tr\left(W^T U W\right) + \tfrac{\mu}{2} Tr\left(\left(W^T W - I_K\right)\left(W^T W - I_K\right)^T\right) \\ +Tr\left(\psi W^T\right) + Tr\left(\phi H^T\right) \end{aligned} \tag{19}$$

The matrices $\boldsymbol{H}$ and $\boldsymbol{U}$ in Eq. (19) are first fixed to update $\boldsymbol{W}$. By taking the partial derivative of Eq. (19) with respect to $\boldsymbol{W}$, the following formula can be obtained:

$$\begin{aligned} \frac{\partial \mathrm{L}}{\partial W} = 2\beta\left(X^T X W H H^T - X^T X H^T\right) + 2X^T L_Z X W + 2\gamma U W \\ + 2\mu\left(W W^T W - W\right) + \psi \end{aligned} \tag{20}$$

By using the Karush–Kuhn–Tucker (KKT) conditions [38] $\psi_{ij} W_{ij} = 0$, the obtained formula is as follows:

$$\begin{aligned} \left[\beta\left(X^T X W H H^T - X^T X H^T\right) + X^T (D - Z) X W \right. \\ \left. + \gamma U W + \mu\left(W W^T W - W\right)\right]_{ij} W_{ij} = 0 \end{aligned} \tag{21}$$

Thus, the update rule for $\boldsymbol{W}$ is as follows:

$$W_{ij} \leftarrow W_{ij} \frac{\left[\beta X^T X H^T + X^T Z X W + \mu W\right]_{ij}}{\left[\beta X^T X W H H^T + X^T D X W + \gamma U W + \mu W W^T W\right]_{ij}} \tag{22}$$

Then, the matrices $\boldsymbol{W}$ and $\boldsymbol{U}$ in Eq. (19) are fixed to update $\boldsymbol{H}$. By taking the partial derivative of Eq. (19) with respect to $\boldsymbol{H}$, the following formula can be obtained:

$$\frac{\partial \mathrm{L}}{\partial H} = 2\beta\left(W^T X^T X W H - W^T X^T X\right) + \phi \tag{23}$$

By using the Karush–Kuhn–Tucker (KKT) conditions $\phi_{ij} H_{ij} = 0$, the obtained formula is as follows:

$$\left[\beta\left(W^T X^T X W H - W^T X^T X\right)\right]_{ij} H_{ij} = 0 \tag{24}$$

So the update rule for the matrix $\boldsymbol{H}$ is as follows:

$$H_{ij} \leftarrow H_{ij} \frac{\left[W^T X^T X\right]_{ij}}{\left[W^T X^T X W H\right]_{ij}} \tag{25}$$

Then, the items related to $\boldsymbol{G}$ in Eq. (15) are preserved and the following formula can be obtained:

$$\begin{aligned} &\arg \min_G Tr(G^T L_Z G) \\ &s.t. \quad G \in \Re^{n\times c}, G^T G = I \end{aligned} \tag{26}$$

It is obvious that the optimal solution $\boldsymbol{G}$ in above formula consists of the feature vectors corresponding to the c smallest eigenvalues of the Laplacian matrix $\boldsymbol{L_Z}$.

In order to calculate the matrix $\boldsymbol{Z}$, the terms related to $\boldsymbol{Z}$ in Eq. (15) is preserved. The obtained formula is as follows:

$$\begin{aligned} &\arg \min_Z \sum_{i,j} \left(\| W^T x_i - W^T x_j \|_2^2 z_{ij} + \alpha z_{ij}^2 \right) + 2\lambda Tr(G^T L_Z G) \\ &s.t. \quad \forall i, z_i^T 1 = 1, 0 \leqslant z_{ij} \leqslant 1, G \in \Re^{n\times c}, G^T G = I \end{aligned} \tag{27}$$

By using the spectrum theory, the following formula can be obtained:

$$\sum_{i,j} \| g_i - g_j \|_2^2 z_{ij} = 2Tr\left(G^T L_Z G\right) \tag{28}$$

where $\boldsymbol{g}_i$ is the ith row of the matrix $\boldsymbol{G}$. By substituting Eq. (28) into Eq. (27), the following expression can be obtained:

$$\begin{aligned} &\arg \min_Z \sum_{i,j} \left(\| W^T x_i - W^T x_j \|_2^2 z_{ij} + \alpha z_{ij}^2 \right) + \lambda \sum_{i,j} \| g_i - g_j \|_2^2 z_{ij} \\ &s.t. \quad \forall i, z_i^T 1 = 1, 0 \leqslant z_{ij} \leqslant 1, G \in \Re^{n\times c}, G^T G = I \end{aligned} \tag{29}$$

For each sample, we need to find its similarity vector. And the similarity vectors of different samples are independent to each other. For a single sample, the following problem should be solved:

$$\begin{aligned} &\arg \min_Z \sum_{j} \left(\| W^T x_i - W^T x_j \|_2^2 z_{ij} + \alpha z_{ij}^2 \right) + \lambda \sum_{j} \| g_i - g_j \|_2^2 z_{ij} \\ &s.t. \quad z_i^T 1 = 1, 0 \leqslant z_{ij} \leqslant 1, G \in \Re^{n\times c}, G^T G = I \end{aligned} \tag{30}$$

A matrix $R \in \Re^{n\times n}$ is defined and r_{ij} represents the element of the ith row and the jth column of the matrix $\boldsymbol{R}$. And the equation $r_{ij} = \| W^T x_i - W^T x_j \|_2^2$ holds. Then, a matrix $T \in \Re^{n\times n}$ is defined and t_{ij} represents the element of the ith row and the jth column of the matrix $\boldsymbol{T}$. And the equation $t_{ij} = \| g_i - g_j \|_2^2$ holds. We write $e_i \in \Re^{n\times 1}$ and there is $e_{ij} = r_{ij} + \lambda t_{ij}$, then Eq. (30) can be written as follows:

$$\begin{aligned} &\min \ \| z_i + \tfrac{1}{2\alpha} e_i \|_2^2 \\ &s.t. \ \ z_i^T 1 = 1, 0 \leqslant z_{ij} \leqslant 1 \end{aligned} \tag{31}$$

It can be seen that Eq. (31) is easy to solve. According to [31], the parameter α controls the number of neighborhood samples. The vector z_i will contain only one non-zero element when $\alpha \to 0$, and the values of all elements in vector z_i will be $1/n$ when $\alpha \to \infty$. To make the vector z_i contain exactly k non-zero elements when the subscript i takes different values, according to the conclusion in [31], the value of α is set as follows:

$$\alpha = \frac{1}{n}\sum_{i=1}^{n}\alpha_i = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{k}{2}e_{i,k+1} - \frac{1}{2}\sum_{j=1}^{k}e_{ij}\right) \tag{32}$$

Based on the above analyses, the optimization process of SLASR is summarize in Table 1.

3.7. Computational complexity analysis

In this section, we analyze the time and space complexity of SLASR. First, we give the time complexity of SLASR. Where n and d represent the number of samples and features, respectively, c represents the number of sample categories, and *NIter* is the maximum number of iterations. In each iteration, to update the similarity matrix $\boldsymbol{Z}$ and the matrix $\boldsymbol{G}$, the computational complexity is $O((c+n)n^2)$. To update the matrices $\boldsymbol{U}$, $\boldsymbol{W}$, and $\boldsymbol{H}$, the time cost is $O(dn(d+n))$. Since SLASR iterates *NIter* times, the total complexity is $O\left(NIter\left(cn + n^2 + dn + d^2\right)n\right)$. For the proposed SLASR, the space complexity of allocating storage space for parameter variables in the parameter list is $O(nd)$, and the space complexity for defining local variables in sub functions is $O\left(n^2 + d^2 + n(d+c)\right)$. So the total space complexity is $O\left(n^2 + d^2 + n(d+c)\right)$.

3.8. Convergence analysis

Here, we demonstrate the convergence of SLASR. Since Eqs. (26) and (31) have closed form solutions, the convergence of SLASR under the update rules of variables $\boldsymbol{W}$ and $\boldsymbol{H}$ is demonstrated. Similar to the method in [4,9], the convergence of SLASR under the variable $\boldsymbol{H}$ is first proved.

Table 1
The procedure of SLASR.

Input: Data matrix $X \in \Re^{n\times d}$; maximum number of iterations *NIter*; balance parameters β, γ, μ; a large enough number λ; number of sample categories c; dimension of the subspace K; number of selected features l.
Output: Index set of the selected features I; new data matrix $X_{new} \in \Re^{n\times l}$.
Initialize the matrix $\boldsymbol{W}$ with the matrix of all 1 elements. Initialize the matrix $\boldsymbol{H}$ into a cluster indicator matrix by using the *k-means* clustering algorithm.
Initialize the similarity matrix $\boldsymbol{Z}$ by solving Eq. (5), and obtain the Laplace matrix $\boldsymbol{L_Z}$.
Initialize the matrix $\boldsymbol{G}$ by solving Eq. (26).
Calculate the vector $\boldsymbol{z}_i$ and update the similarity matrix $\boldsymbol{Z}$ by solving the Eq. (31).
Update the matrix $\boldsymbol{G}$ by solving Eq. (26), and the matrix $\boldsymbol{G}$ consists of the eigenvectors corresponding to the c smallest eigenvalues of $\boldsymbol{L_Z}$.
Update the matrices $\boldsymbol{U}$, $\boldsymbol{W}$, and $\boldsymbol{H}$ according to the update rules in Eqs. (18), (22), and (25).
Repeat the steps 4, 5, and 6 until the maximum number of iterations is satisfied.
Calculate all the values of $||w_i||_2$ when the subscript i varies and sort these values in descending order. Then select the features corresponding to the first l evaluation values. Then the index set I of the selected features and a new data matrix $X_{new} \in \Re^{n\times l}$ can be obtained to complete feature selection.

Definition 1.. Given a function J(h, h′), if C(h) satisfies the following condition:

$$J(h,h') \geqslant C(h), J(h,h) = C(h) \tag{33}$$

Then C is non-increasing under the update rule of Eq. (34):

$$h^{(t+1)} = \arg\ \min_h J\left(h, h^{(t)}\right) \tag{34}$$

where $J(h,h')$ is an auxiliary function of $C(h)$.

Proof. By retaining the terms of the variable $\mathbf{H}$ in the objective function (15), the following function is obtained:

$$C(H) = \beta Tr\left((X - XWH)(X - XWH)^T\right) \tag{35}$$

The following formulas can be obtained by taking the first-order and the second-order partial derivatives of $C(\boldsymbol{H})$ with respect to $\boldsymbol{H}$:

$$C_{ij}' = \left[\frac{\partial C}{\partial H}\right]_{ij} = \left[2\beta\left(W^T X^T XWH - W^T X^T X\right)\right]_{ij} \tag{36}$$

$$C_{ij}'' = \left[2\beta W^T X^T XW\right]_{ii} \tag{37}$$

Lemma 1: Giving the auxiliary functions of C_{ij}, which is as follows:

$$\begin{aligned} J\left(H_{ij}, H_{ij}^{(t)}\right) = {} & C_{ij}\left(H_{ij}^{(t)}\right) + C_{ij}'\left(H_{ij}^{(t)}\right)\left(H_{ij} - H_{ij}^{(t)}\right) \\ & + \frac{\left[\beta W^T X^T XWH^{(t)}\right]_{ij}}{H_{ij}^{(t)}}\left(H_{ij} - H_{ij}^{(t)}\right)^2 \end{aligned} \tag{38}$$

Denoting the Taylor expansion of $C_{ij}(H_{ij})$ as follows:

$$\begin{aligned} C_{ij}(H_{ij}) = {} & C_{ij}\left(H_{ij}^{(t)}\right) + C_{ij}'\left(H_{ij}^{(t)}\right)\left(H_{ij} - H_{ij}^{(t)}\right) \\ & + \left[\beta W^T X^T XW\right]\left(H_{ij} - H_{ij}^{(t)}\right)^2 \end{aligned} \tag{39}$$

It can be seen from Eqs. (38) and (39) that $J\left(H_{ij}, H_{ij}^{(t)}\right) \geqslant C_{ij}(H_{ij})$ is equivalent to:

$$\frac{\left[\beta W^T X^T XWH^{(t)}\right]_{ij}}{H_{ij}^{(t)}} \geqslant \beta W^T X^T XW \tag{40}$$

And it is obvious that:

$$\begin{aligned} \left[\beta W^T X^T XWH^{(t)}\right]_{ij} &= \sum_{b=1}^{K}\left[\beta W^T X^T XW\right]_{ib} H_{bj}^{(t)} \\ &\geqslant \left[\beta W^T X^T XW\right]_{ii} H_{ij}^{(t)} \end{aligned} \tag{41}$$

So inequality (40) holds, that is, $J\left(H_{ij}, H_{ij}^{(t)}\right) \geqslant C_{ij}(H_{ij})$ holds. And it can be seen that equation $J(H_{ij}, H_{ij}) = C_{ij}(H_{ij})$ holds.

Then, we prove that the update rule of variable $\boldsymbol{H}$ satisfies the update formula (34) that makes C_{ij} non-increasing.

By substituting $J\left(H_{ij}, H_{ij}^{(t)}\right)$ in Eq. (38) into Eq. (34), the following formula can be obtained:

$$H_{ij}^{(t+1)} = H_{ij}^{(t)} - H_{ij}^{(t)}\frac{C_{ij}'\left(H_{ij}^{(t)}\right)}{2\left[\beta W^T X^T XWH^{(t)}\right]_{ij}} \tag{42}$$

Substituting Eq. (36) into Eq. (42) presents the following expression:

Table 2
The information of thirteen datasets.

Dataset	Size	Dim	Classes	Type
Ionosphere	351	34	2	Text image
JAFFE	213	676	10	Face image
Umist	575	644	20	Face image
Lung_dis	73	325	7	Biological
YaleB	2414	1024	38	Face image
COIL20	1440	1024	20	Digital image
ORL	400	1024	40	Face image
TOX_171	171	5748	4	Biological
Isolet	1560	617	26	Letter image
AR10P	130	2400	10	Face image
Usps	9298	256	10	Digital image
Orlraws	100	10,304	10	Face image
AT&T	400	10,304	40	Face image

$$H_{ij}^{(t+1)} = H_{ij}^{(t)} \frac{\left[W^T X^T X\right]_{ij}}{\left[W^T X^T X W H^{(t)}\right]_{ij}} \tag{43}$$

It can be seen that Eq. (43) is the update rule of variable $\boldsymbol{H}$. Therefore, it can be concluded that the objective function is non-increasing under the update rule (25). It can also be proved that the objective function is non-increasing under the update rule of variable $\boldsymbol{W}$. So the objective function in Eq. (15) is non-increasing under the update rules (22) and (25).

4. Experiments

In this section, we test the performance of SLASR and six comparison algorithms. The *k-means* [39,40] method is used to measure the performance of feature selection algorithms. The effectiveness of SLASR is verified, and the experimental results are given and analyzed. Then, the parameter sensitivity and the convergence of SLASR are presented.

4.1. Datasets

Thirteen test datasets are used in this experiment. And these datasets can be divided into four categories, including face image, text image, digital image, and biological image [41,42]. The detailed information of datasets is shown in Table 2.

4.2. Comparison algorithms

To test the performance of the proposed SLASR, six comparison algorithms are used in this paper:

1) Baseline: Performing clustering operation on all features directly without feature selection.
2) LapScor: Laplacian Score [21] makes use of the local manifold structure of data to calculate the feature scores to select the discriminative features.
3) UDFS: unsupervised discriminant feature selection [29]. It uses the local discriminant information and local structure information of data simultaneously to guide feature selection.
4) MFFS: matrix factorization feature selection [23], which selects the most representative feature subset under the subspace learning framework of matrix decomposition to complete feature selection.
5) MCFS: multi-cluster data feature selection [14], for which spectral analysis is first performed, then sparse regression is executed to select features.
6) SOGFS: unsupervised feature selection with structured graph optimization [31]. In this method, the adaptive manifold structure is used to guarantee the robustness to noise and the $l_{2,1}$-*norm* regularization term is used to select the discriminative features.

4.3. Evaluation metrics

In this paper, the metrics Normalized Mutual Information (NMI) [4,43] and the Clustering Accuracy (ACC) [4,9] are used to evaluate the performance of feature selection algorithms. And the greater the values of the two metrics, the better the performance of the corresponding algorithms.

The formula of Normalized Mutual Information (NMI) can be described as follows:

$$NMI(P,Q) = \frac{I(P,Q)}{\sqrt{H(P)H(Q)}} \tag{44}$$

where P and Q are two random variables, H(P) and H(Q) represent the entropy values of P and Q, respectively. I(P,Q) represents the mutual information of random variables P and Q. In the clustering tasks, P and Q represent the clustering labels and the real labels of samples, respectively.

The formula of Clustering Accuracy (ACC) is as follows:

$$\mathrm{ACC} = \frac{1}{n}\sum_{j=1}^{n} \delta(d_j, map(e_j)) \tag{45}$$

where $\delta(x,y)$ is an indicator function, and$\delta(x,y) = 1$ when x is equal to y, otherwise$\delta(x,y) = 0$. In Eq. (45), d_j represents the real label of the sample $\mathbf{x}_j$, and e_j represents the label obtained by performing the clustering task. To match the clustering labels with the real labels, Hungarian algorithm [44] is used in this paper. map(.) represents a mapping function and implements the Hungarian algorithm.

4.4. Experimental results and analyses

4.4.1. Experimental settings

In this section, the parameter settings of SLASR and six comparison algorithms are given. For SOGFS and SLASR, to construct the adaptive similarity matrix, the parameter of nearest neighbors k is set to 5. For Lapscor and MCFS, the parameter of nearest neighbors k is set to 5, and the Gaussian scale parameter is searched in the range of $\{10^{+1}, 10^{+2}, 10^{+3}, 10^{+4}, 10^{+5}\}$. For SLASR, the parameters β and γ are adjusted in the range of $\{10^{-8}, 10^{-7}, \ldots, 10^{+7}, 10^{+8}\}$ and the parameter μ is searched in the range of $\{10^{+0}, 10^{+1}, \ldots, 10^{+7}, 10^{+8}\}$. For the subspace dimension K, the range is set to {d/3, d/2, (2*d)/3} to obtain the best ACC and NMI accuracy. For SOGFS, the

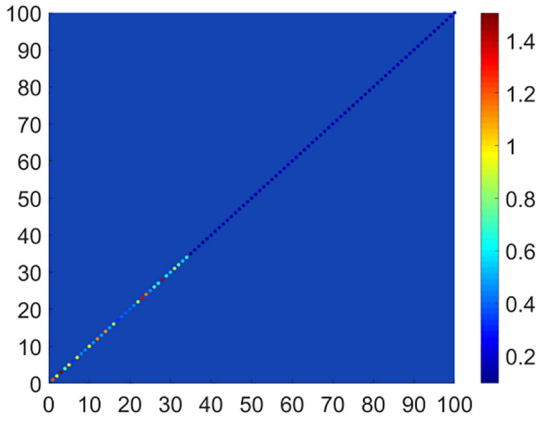


Fig. 2. The diagonal matrix of evaluation values of 100 features.

setting of parameter γ is the same as that of SLASR, and the dimension of low-dimensional embedding is set to {d/3, d/2, (2*d)/3}. For MFFS, UDFS, MCFS, SOGFS, and SLASR, the maximum number of iterations is set to 30. And the number of selected features l is set to {20, 30, 40, 50, 60, 70, 80, 90, 100}. Since this algorithm is sensitive to initial values, *k-means* clustering is performed 20 times to obtain an average value to represent the final clustering result. The balance parameters β, γ, and μ are adjusted to achieve the best ACC and NMI values.

It can be seen from Eq. (15) that the objective function contains six hyper parameters α, β, λ, γ, μ, K. As can be seen from the description of Eq. (13), the parameter can be updated adaptively during the algorithm optimization process and is set to 2 times and 1/2 times of the original value when needs to be increased and decreased, respectively. Meanwhile, the parameter α can be calculated according to the Eq. (32). So we do not need to adjust parameters λ and α. For the subspace dimension K, the search range is set to {d/3, d/2, (2*d)/3} to obtain the best ACC and NMI accuracy, which is easy to adjust. For the remaining parameters β, γ, μ to adjust them more efficiently in real applications, the order of magnitude interval between adjacent parameters during grid search can be increased. For instance, the searching range of parameters β and γ can be set to {10^{-8}, 10^{-6}, 10^{-4}, … , 10^{+4}, 10^{+6}, 10^{+8}}, and the parameter μ can be searched in the range of {10^{+0}, 10^{+2}, …, 10^{+6}, 10^{+8}}. By using this method, the number of hyper parameters to be adjusted can be reduced on a large scale, making the algorithm more suitable for practical scenarios.

4.4.2. The effectiveness evaluation of SLASR

To verify that SLASR can select the representative features, the Ionosphere dataset is used to test the performance of SLASR. For the given Ionosphere dataset, its dimension is 351×34, where 351 is the number of samples and 34 represents the number of features. In this experiment, 66 new features are generated by using the linear combination of 34 original features. For different new features, the corresponding linear combination coefficients are different. And these coefficients are normalized and randomly generated. All features are put together, resulting in a total of 100 features. And the first 34 features are original features. Then the projection matrix $\boldsymbol{W}$ can be obtained by solving the Eq. (15). For each row vector $\boldsymbol{w}_i$ of $\boldsymbol{W}$, its f_2-*norm* can be calculated and used

Table 3
Clustering accuracy of seven algorithms on twelve datasets (ACC ± STD%).

Dataset	Baseline	LapScor	UDFS	MFFS	MCFS	SOGFS	SLASR
JAFFE	79.98 ± 5.32	69.27 ± 5.19	67.82 ± 5.29	69.06 ± 7.37	74.86 ± 6.56	83.92 ± 4.35	**90.05 ± 5.47**
Umist	43.61 ± 2.16	42.09 ± 1.94	44.64 ± 2.67	44.41 ± 3.67	47.35 ± 2.62	38.65 ± 2.17	**52.65 ± 3.18**
Lung_dis	74.25 ± 4.94	70.41 ± 7.34	71.99 ± 5.51	70.75 ± 4.87	73.15 ± 6.26	77.60 ± 5.25	**81.58 ± 5.25**
YaleB	9.69 ± 0.49	8.69 ± 0.25	9.66 ± 0.25	15.88 ± 0.65	14.47 ± 0.79	9.20 ± 0.29	**18.34 ± 0.64**
COIL20	65.64 ± 3.94	60.41 ± 2.11	62.44 ± 2.57	66.35 ± 3.13	64.06 ± 2.36	60.13 ± 2.48	**69.24 ± 3.08**
ORL	51.96 ± 2.57	44.44 ± 1.88	45.95 ± 2.17	46.24 ± 2.11	48.58 ± 2.34	52.61 ± 3.08	**56.09 ± 2.33**
TOX_171	43.36 ± 1.90	41.35 ± 2.96	41.52 ± 1.80	37.84 ± 2.20	41.52 ± 2.60	46.43 ± 1.31	**49.27 ± 0.53**
Isolet	61.41 ± 2.38	55.83 ± 2.14	40.36 ± 1.56	57.08 ± 2.00	61.60 ± 3.60	50.36 ± 2.23	**66.36 ± 2.99**
AR10P	21.96 ± 2.92	33.92 ± 2.57	33.54 ± 2.05	41.08 ± 2.41	31.54 ± 2.51	**45.65 ± 2.41**	44.73 ± 4.42
Usps	65.38 ± 2.39	61.84 ± 0.13	45.16 ± 0.53	59.70 ± 3.15	**66.51 ± 4.04**	56.17 ± 0.36	65.76 ± 3.50
Orlraws	76.60 ± 6.17	66.65 ± 5.32	53.50 ± 3.35	66.35 ± 5.29	59.90 ± 4.02	73.35 ± 4.07	**79.55 ± 4.26**
AT&T	61.45 ± 4.56	49.30 ± 1.92	51.67 ± 2.37	49.36 ± 1.44	53.39 ± 3.60	55.18 ± 2.31	**62.61 ± 2.73**

Table 4
Normalized Mutual Information of seven algorithms on twelve datasets (NMI ± STD%).

Dataset	Baseline	LapScor	UDFS	MFFS	MCFS	SOGFS	SLASR
JAFFE	84.07 ± 3.25	76.64 ± 3.83	70.82 ± 5.05	74.23 ± 5.44	80.76 ± 6.21	85.87 ± 2.21	**90.73 ± 3.38**
Umist	64.47 ± 1.46	62.33 ± 1.91	60.63 ± 1.86	65.65 ± 2.04	67.07 ± 1.72	57.92 ± 1.51	**69.69 ± 1.36**
Lung_dis	69.97 ± 3.38	64.86 ± 5.71	68.52 ± 3.52	64.42 ± 3.42	67.22 ± 4.34	73.36 ± 3.16	**75.57 ± 4.49**
YaleB	12.97 ± 0.58	13.24 ± 0.30	15.49 ± 0.29	26.77 ± 0.61	23.25 ± 0.68	14.67 ± 0.20	**30.80 ± 0.42**
COIL20	76.62 ± 1.92	69.67 ± 1.18	73.18 ± 1.27	75.64 ± 1.67	73.94 ± 1.29	70.35 ± 1.30	**77.22 ± 1.42**
ORL	72.36 ± 1.71	66.62 ± 1.33	69.25 ± 1.10	67.61 ± 1.63	68.86 ± 1.39	72.61 ± 1.40	**75.09 ± 0.99**
TOX_171	14.32 ± 1.55	9.90 ± 1.80	10.98 ± 1.03	9.26 ± 4.42	11.36 ± 3.13	20.17 ± 1.31	**26.07 ± 0.64**
Isolet	75.66 ± 1.00	69.45 ± 0.91	55.57 ± 1.18	70.19 ± 1.00	74.81 ± 1.28	64.30 ± 0.88	**76.43 ± 1.79**
AR10P	19.17 ± 2.77	35.08 ± 1.40	28.37 ± 1.57	43.46 ± 1.71	30.67 ± 3.07	**50.36 ± 2.88**	47.83 ± 3.27
Usps	**60.98 ± 0.60**	57.35 ± 0.08	41.00 ± 0.63	58.21 ± 0.49	59.09 ± 1.52	51.00 ± 0.28	60.40 ± 0.99
Orlraws	81.67 ± 4.70	74.54 ± 2.79	62.12 ± 2.20	73.24 ± 4.31	62.99 ± 2.45	81.50 ± 4.52	**82.92 ± 3.37**
AT&T	80.26 ± 2.25	72.54 ± 1.02	73.66 ± 1.20	70.27 ± 1.04	73.06 ± 2.10	74.29 ± 1.40	**80.60 ± 1.28**

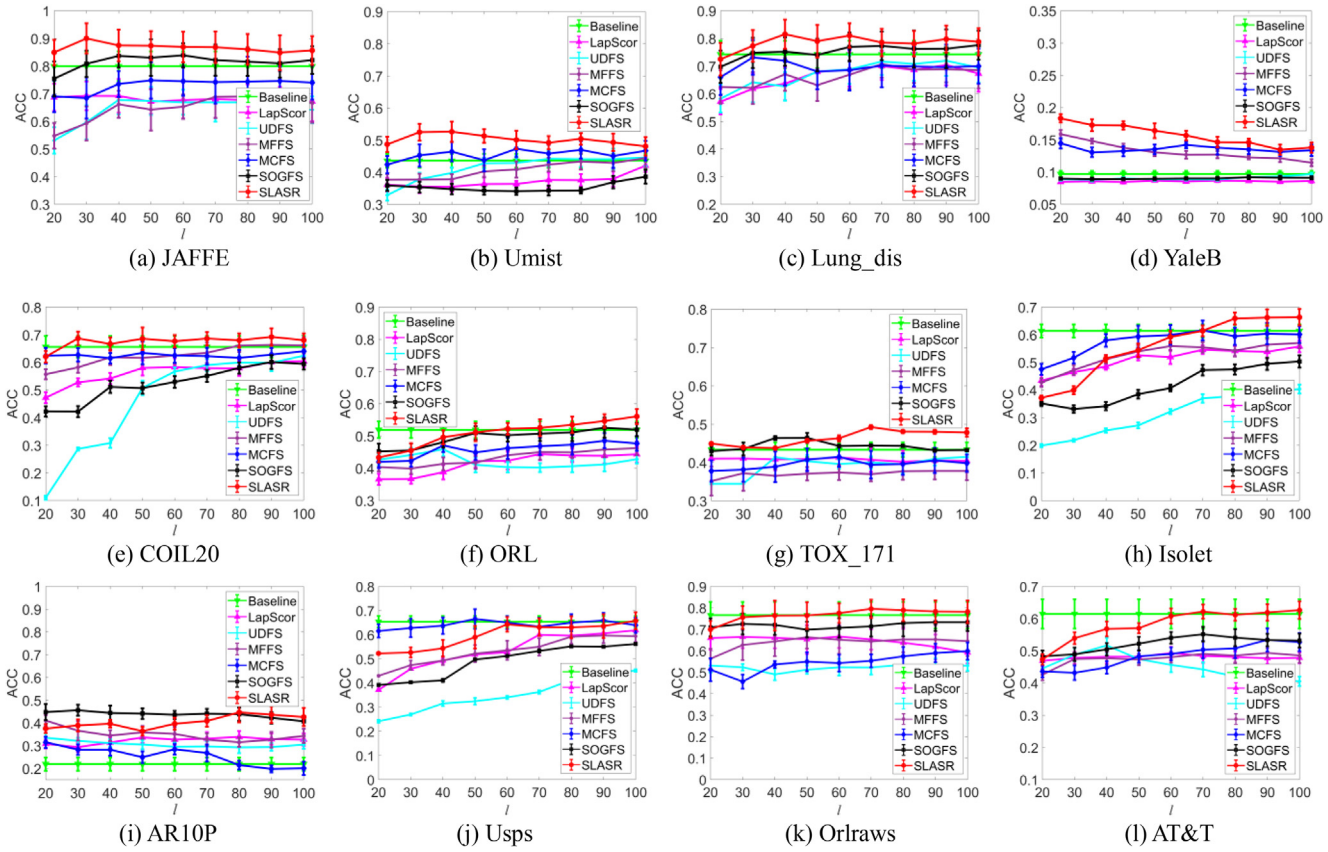


Fig. 3. Clustering accuracy of seven algorithms on twelve datasets with different number of selected features.

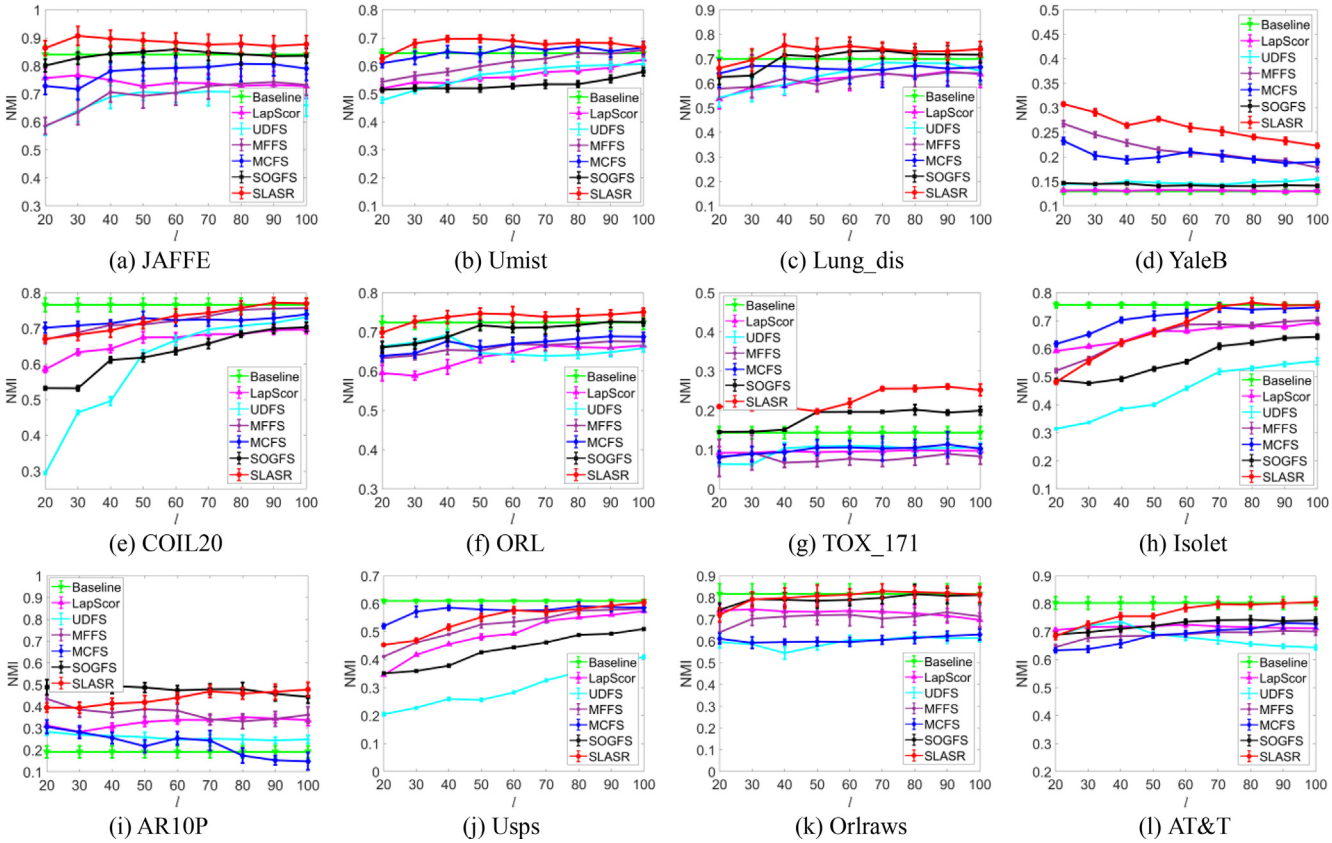


Fig. 4. Normalized Mutual Information of seven algorithms on twelve datasets with different number of selected features.

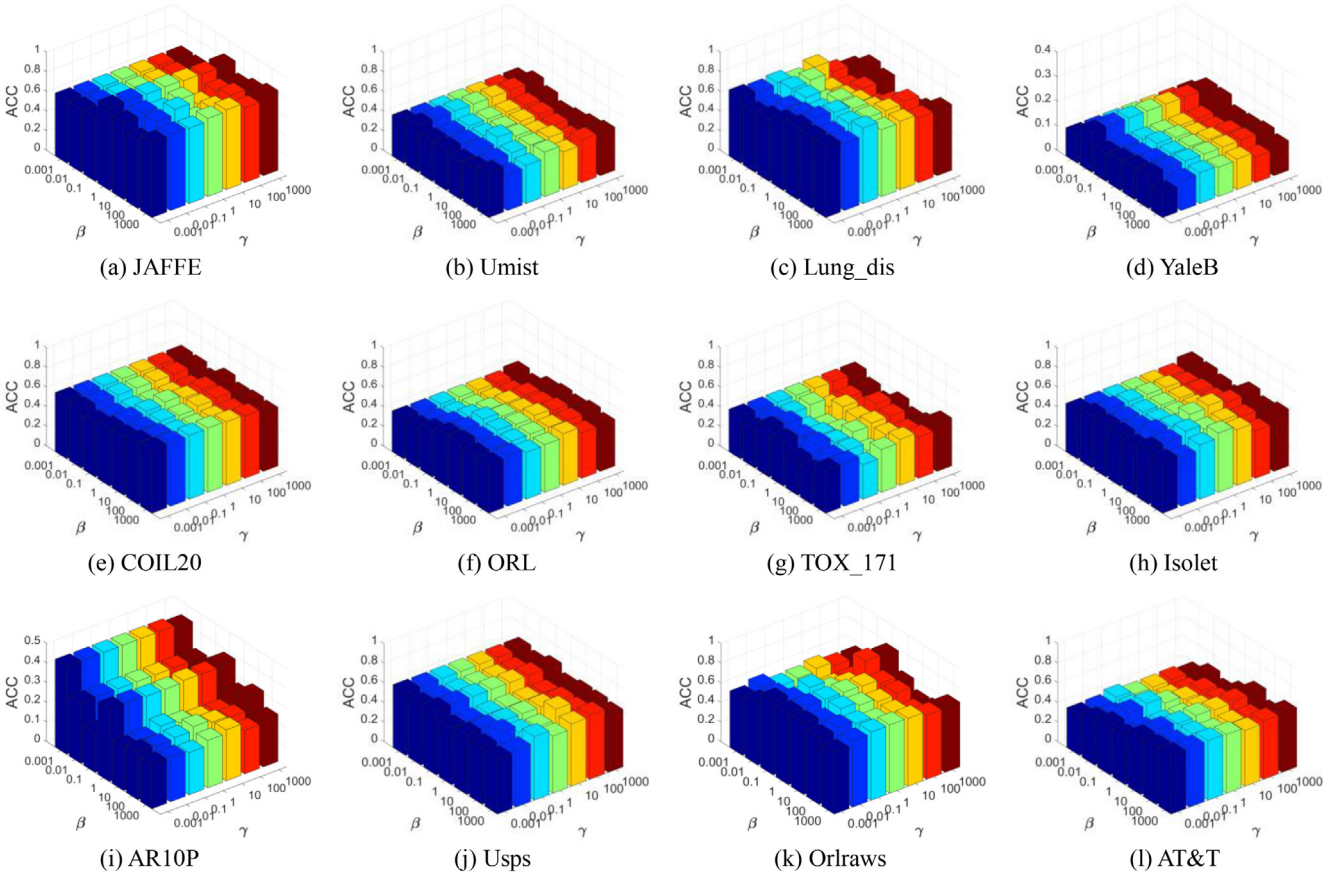


Fig. 5. Clustering accuracy of SLASR on twelve datasets using different β and γ.

as the evaluation value of the ith feature. Then all the evaluation values can be used to generate a diagonal matrix, which is as follows.

It can be seen clearly from Fig. 2 that the evaluation values of the first 34 features are significantly greater than those of the last 66 features, indicating that SLASR is effective, and the original representative features can be selected.

4.4.3. Experimental results and analyses

The clustering results of seven algorithms on twelve datasets are shown in Tables 3 and 4. And ACC and NMI values are shown in Tables 3 and 4, respectively. The bold marked and the underline marked values are the best and the second best results, respectively.

It can be seen from Tables 3 and 4 that the proposed SLASR can obtain the optimal ACC and NMI values on most datasets. Compared with SOGFS, SLASR not only learns the manifold structure adaptively, but also preserves the global reconstruction information of data through the matrix factorization subspace learning framework. So SLASR can obtain better performance. Meanwhile, SLASR can obtain better clustering results than those of baseline method on most datasets, indicating that SLASR can select the representative features. So the proposed SLASR has great effectiveness.

Figs. 3 and 4 show the variation of the clustering results of the seven algorithms on twelve datasets with the number of selected features.

For each figure, the horizontal axis indicates the number of selected features l. And the vertical axis indicates the clustering accuracy (ACC) and standard deviation (STD) in Fig. 3 and the normalized mutual information (NMI) and standard deviation (STD) in Fig. 4. It can be seen from Figs. 3 and 4 that the proposed SLASR outperforms other comparison algorithms on most datasets. And on the datasets JAFFE, Umist, Lung_dis, YaleB, TOX_171, and AR10P, SLASR is significantly better than Baseline method, which fully demonstrates the advantages of SLASR. For the datasets Isolet, AR10P and Usps, the performance of SLASR is not so good when the number of selected features is small. However, as the number of selected feature increases, the performance of SLASR improves and eventually exceeds that of other comparison algorithms. So the proposed SLASR has great effectiveness.

4.4.4. Parameter sensitivity analysis

For the proposed SLASR, the parameters that need to be adjusted include: the coefficient of subspace learning term β, the coefficient of sparse constraint term γ, the coefficient of orthogonal constraint term μ, and the dimension of the subspace K. Since β and γ have a great influence on accuracy, the sensitivity of SLASR to the balance parameters β and γ is tested in this paper. The parameters β and γ are searched in the range of $\{10^{-3}, 10^{-2}, 10^{-1}, 10^{+0}, 10^{+1}, 10^{+2}, 10^{+3}\}$, and Figs. 5 and 6 show the three-dimensional histograms of ACC and NMI values on twelve datasets, respectively.

It can be seen from Figs. 5 and 6 that when parameters β and γ vary, the ACC and NMI values can keep relatively stable on most datasets, especially for Umist, COIL20, ORL, and Usps. Thus, SLASR has better parameter sensitivity result and is not sensitive to parameters β and γ.

4.4.5. Convergence test

The convergence of SLASR on twelve datasets is tested in this section. The convergence curves are shown in Fig. 7. In each subfig-

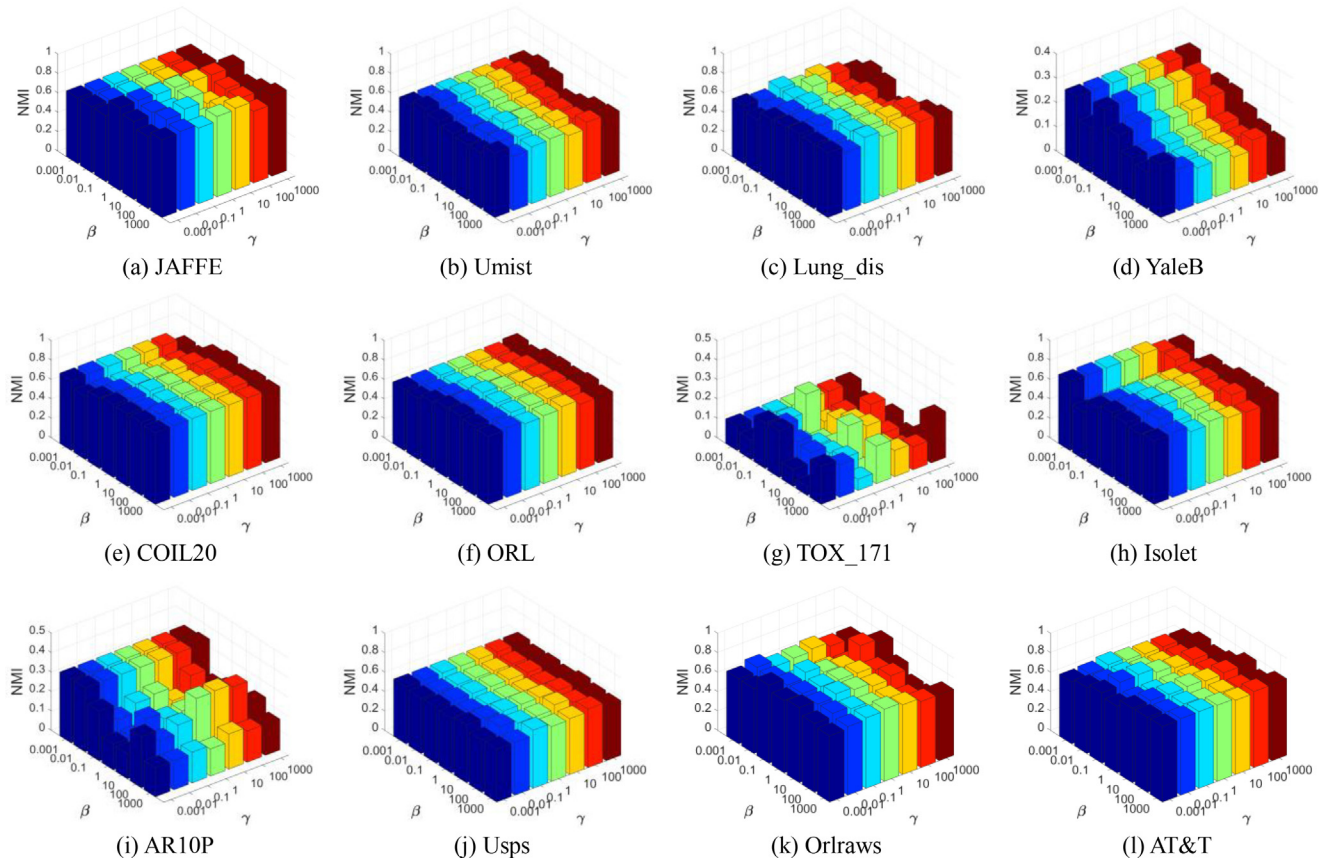


Fig. 6. Normalized Mutual Information of SLASR on twelve datasets using different β and γ.

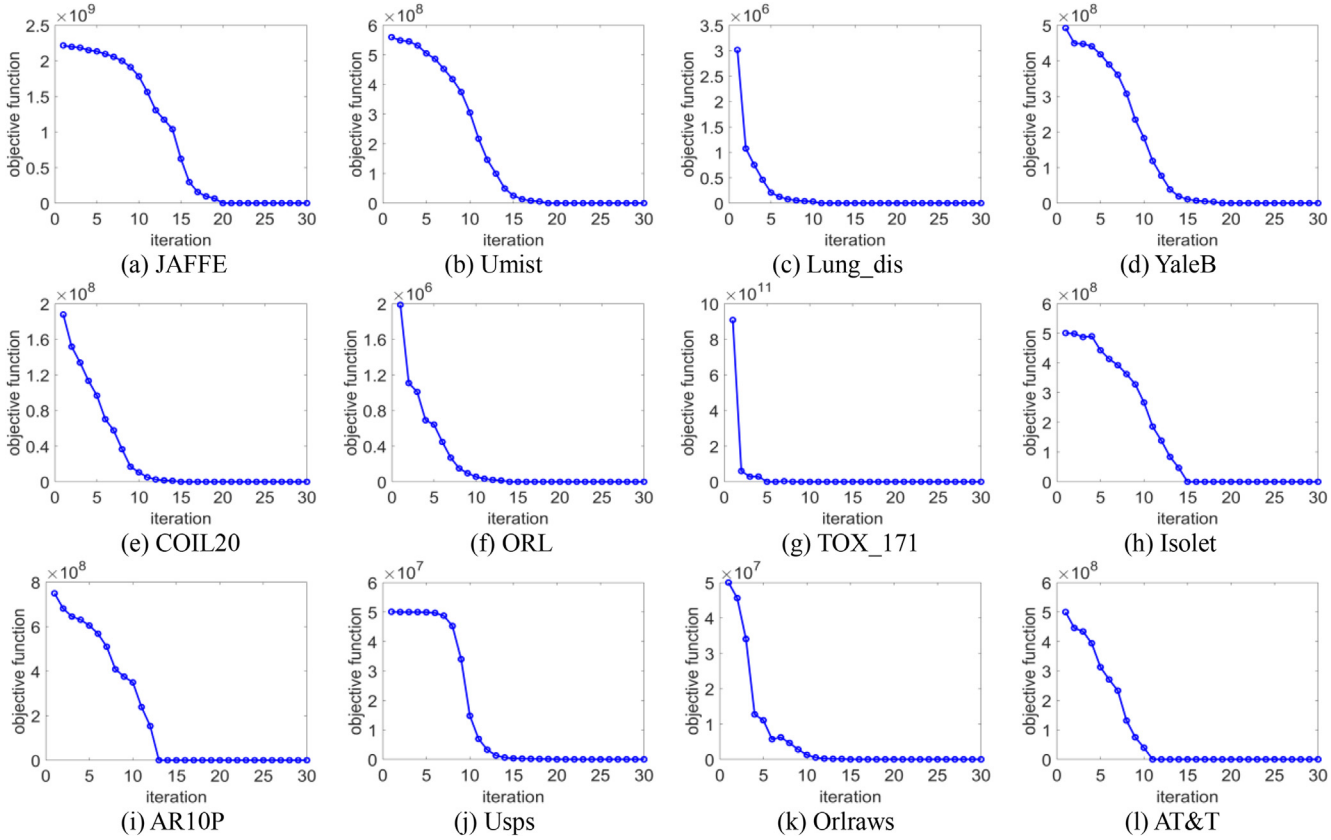


Fig. 7. The convergence curves of the objective function on twelve datasets.

ure, the horizontal axis and the vertical axis indicate the number of iterations and the value of objective function, respectively.

It can be seen from Fig. 7 that the value of objective function decreases as the number of iterations increases. And on all datasets, the objective function converges within 20 iterations, especially for the datasets Lung_dis, COIL20, ORL, TOX_171, Orlraws, and AT&T. So it is reasonable to set the maximum number of iterations to 30. The convergence of SLASR is proved.

5. Conclusions

In this paper, a novel algorithm is proposed, called subspace learning for unsupervised feature selection via adaptive structure learning and rank approximation (SLASR). By introducing the adaptive manifold learning strategy into the framework of subspace learning, not only the local manifold structure of data is well preserved, but also the global reconstruction information is retained. It can be seen from the experimental results that SLASR achieves the best accuracy in most cases, which fully demonstrates the effectiveness of the proposed SLASR. Additionally, the rank constraint is imposed on the Laplacian matrix, which ensures the similarity matrix containing c connected components. So the manifold information becomes more accurate. And the $l_{2,1/2}$ regularization term applied to the projection matrix improves the sparsity and robustness of the selected features. It can also be seen from the experiments of parameter sensitivity analysis and the convergence test that SLASR is not sensitive to the key parameters and can achieve great convergence performance on all datasets. All experimental results show that SLASR can obtain better performance than other comparison algorithms. Since the stochastic optimization methods can achieve better optimization results than the traditional alternating iterative multiplier algorithms, novel stochastic optimization mechanisms are hoped to be developed to make the algorithms obtain better optimization results.

CRediT authorship contribution statement

Ronghua Shang: Conceptualization, Methodology, Writing - review & editing. **Kaiming Xu:** Methodology, Data curation, Writing - original draft, Software. **Licheng Jiao:** Conceptualization, 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.

Acknowledgement

This work was partially supported by the National Natural Science Foundation of China under Grants 61773304, 61836009, 61871306, 61772399 and U1701267, the Fund for Foreign Scholars in University Research and Teaching Programs (the 111 Project) under Grants No. B07048, the Major Research Plan of the National Natural Science Foundation of China under Grants 91438201 and 91438103, and the Program for Cheung Kong Scholars and Innovative Research Team in University under Grant IRT1170.

References

[1] C. Tang, X. Liu, M. Li, et al., Robust unsupervised feature selection via dual self-representation and manifold regularization, Knowl.-Based Syst. 145 (2018) 109–120.
[2] S. Wang, H. Wang, Unsupervised feature selection via low-rank approximation and structure learning, Knowl.-Based Syst. 124 (2017) 70–79.
[3] Y. Wan, X. Chen, J. Zhang, Global and intrinsic geometric structure embedding for unsupervised feature selection, Expert Syst. Appl. 93 (2018) 134–142.
[4] R. Shang, W. Wang, R. Stolkin, et al., Non-negative spectral learning and sparse regression-based dual-graph regularized feature selection, IEEE Trans. Cybern. 48 (2) (2017) 793–806.
[5] W. Zheng, H. Yan, J. Yang, Robust unsupervised feature selection by nonnegative sparse subspace learning, Neurocomputing 334 (2019) 156–171.
[6] F. Nie, S. Xiang, Y. Liu, et al., Orthogonal vs. uncorrelated least squares discriminant analysis for feature extraction, Pattern Recogn. Lett. 33 (5) (2012) 485–491.
[7] M. Qi, T. Wang, F. Liu, et al., Unsupervised feature selection by regularized matrix factorization, Neurocomputing 273 (2018) 593–610.
[8] P. Huang, C. Chen, Z. Tang, et al., Feature extraction using local structure preserving discriminant analysis, Neurocomputing 140 (2014) 104–113.
[9] R. Shang, W. Wang, R. Stolkin, et al., Subspace learning-based graph regularized feature selection, Knowl.-Based Syst. 112 (2016) 152–165.
[10] J. Wang, L. Wu, J. Kong, et al., Maximum weight and minimum redundancy: a novel framework for feature subset selection, Pattern Recogn. 46 (6) (2013) 1616–1627.
[11] Y. Saeys, I. Inza, P. Larrañaga, A review of feature selection techniques in bioinformatics, Bioinformatics 23 (19) (2007) 2507–2517.
[12] X. Wang, X. Zhang, Z. Zeng, et al., Unsupervised spectral feature selection with l1-norm graph, Neurocomputing 200 (2016) 47–54.
[13] P. Zhu, W. Zhu, Q. Hu, et al., Subspace clustering guided unsupervised feature selection, Pattern Recogn. 66 (2017) 364–374.
[14] D. Cai, C. Zhang, X. He, Unsupervised feature selection for multi-cluster data, in: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, 2010, pp. 333–342.
[15] Z. Xu, I. King, M.R.T. Lyu, et al., Discriminative semi-supervised feature selection via manifold regularization, IEEE Trans. Neural Networks 21 (7) (2010) 1033–1047.
[16] J. Xu, B. Tang, H. He, et al., Semisupervised feature selection based on relevance and redundancy criteria, IEEE Trans. Neural Networks Learn. Syst. 28 (9) (2016) 1974–1984.
[17] X. Du, F. Nie, W. Wang, et al., Exploiting combination effect for unsupervised feature selection by ℓ2, 0 norm, IEEE Trans. Neural Networks Learn. Syst. 99 (2018) 1–14.
[18] P. Zhu, Q. Xu, Q. Hu, et al., Co-regularized unsupervised feature selection, Neurocomputing 275 (2018) 2855–2863.
[19] P. Mitra, C.A. Murthy, S.K. Pal, Unsupervised feature selection using feature similarity, IEEE Trans. Pattern Anal. Mach. Intell. 24 (3) (2002) 301–312.
[20] M. Breaban, H. Luchian, A unifying criterion for unsupervised clustering and feature selection, Pattern Recogn. 44 (4) (2011) 854–865.
[21] X. He, D. Cai, P. Niyogi, Laplacian score for feature selection, Adv. Neural Inf. Process. Syst. (2006) 507–514.
[22] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, J. Mach. Learn. Res. 3 (Mar) (2003) 1157–1182.
[23] S. Wang, W. Pedrycz, Q. Zhu, et al., Subspace learning for unsupervised feature selection via matrix factorization, Pattern Recogn. 48 (1) (2015) 10–19.
[24] S. Wang, W. Pedrycz, Q. Zhu, et al., Unsupervised feature selection via maximum projection and minimum redundancy, Knowl.-Based Syst. 75 (2015) 19–29.
[25] S.T. Roweis, L.K. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (5500) (2000) 2323–2326.
[26] M. Belkin, P. Niyogi, Laplacian eigenmaps and spectral techniques for embedding and clustering, Adv. Neural Inf. Process. Syst. (2002) 585–591.
[27] X. He, P. Niyogi, Locality preserving projections, Adv. Neural Inf. Process. Syst. (2004) 153–160.
[28] Z. Zhao, H. Liu, Spectral feature selection for supervised and unsupervised learning, Proceedings of the 24th International Conference on Machine Learning, ACM, 2007, pp. 1151–1157..
[29] Y. Yang, H.T. Shen, Z. Ma, et al., l2, 1-norm regularized discriminative feature selection for unsupervised learning, IJCAI Proceedings-International Joint Conference on Artificial Intelligence. 2011, 22(1): 1589..
[30] C. Hou, F. Nie, D. Yi, et al., Feature selection via joint embedding learning and sparse regression, IJCAI Proc. Int. Joint Conf. Artif. Intell.. 22 (1) (2011) 1324.
[31] F. Nie, W. Zhu, X. Li, Unsupervised feature selection with structured graph optimization, Thirtieth AAAI Conference on Artificial Intelligence, 2016.
[32] N. Zhou, Y. Xu, H. Cheng, et al., Global and local structure preserving sparse subspace learning: an iterative approach to unsupervised feature selection, Pattern Recogn. 53 (2016) 87–101.
[33] W. Wang, S. Chen, l2, p Matrix Norm and Its Application in Feature Selection. arXiv preprint arXiv:1303.3987, 2013..
[34] Y. Shi, J. Miao, Z. Wang, et al., Feature Selection With $l_{2,\,1-2}$ Regularization, IEEE Trans. Neural Networks Learn. Syst. 29 (10) (2018) 4967–4982.
[35] B. Mohar, Y. Alavi, G. Chartrand, O. Oellermann, The laplacian spectrum of graphs, Graph Theory Combinat. Appl. 2 (1991) 871–898.
[36] K. Fan, On a theorem of weyl concerning eigenvalues of linear transformations i, Proc. Nat. Acad. Sci. USA 35 (11) (1949) 652.
[37] F. Shang, L.C. Jiao, F. Wang, Graph dual regularization non-negative matrix factorization for co-clustering, Pattern Recogn. 45 (6) (2012) 2237–2250.
[38] Y. Li, C. Lei, Y. Fang, et al., Unsupervised feature selection by combining subspace learning with feature self-representation, Pattern Recogn. Lett. 109 (2018) 35–43.
[39] Y. Yang, D. Xu, F. Nie, et al., Image clustering using local discriminant models and global integration, IEEE Trans. Image Process. 19 (10) (2010) 2761–2773.

[40] A. Rakhlin, A. Caponnetto, Stability of K-Means clustering, Adv. Neural Inf. Process. Syst. (2007) 1121–1128.
[41] H. Lu, Z. Fu, X. Shu, Non-negative and sparse spectral clustering, Pattern Recogn. 47 (1) (2014) 418–426.
[42] http://featureselection.asu.edu/datasets.php..
[43] J. Han, Z. Sun, H. Hao, Selecting feature subset with sparsity and low redundancy for unsupervised learning, Knowl.-Based Syst. 86 (2015) 210–223.
[44] C.H. Papadimitriou, K. Steiglitz, Combinatorial Optimization: Algorithms and Complexity, Courier Corporation, 1998.

Ronghua Shang (M'09) received the B.S. degree in information and computation science and the Ph.D. degree in pattern recognition and intelligent systems from Xidian University in 2003 and 2008, respectively. She is currently a professor with Xidian University. Her current research interests include optimization problems, machine learning, image processing, and data mining.

Kaiming Xu received the B.S. degree in electronic information engineering from Hebei University of Engineering, Handan, China. He is currently working toward the master's degree in electronic and communication engineering from Xidian University, Xi'an, China. His current research interests include machine learning and data mining.

Licheng Jiao (SM'89-F'17) received the B.S. degree from Shanghai Jiaotong University, Shanghai, China, in 1982, the M.S. and Ph.D. degrees from Xi'an Jiaotong University, Xi'an, China, in 1984 and 1990, respectively, all in electronic engineering. From 1990 to 1991, he was a Postdoctoral Fellow with the National Key Laboratory for Radar Signal Processing, Xidian University, Xi'an. Since 1992, he has been a Professor with the School of Electronic Engineering, Xidian University. He is currently the Director of the Key Lab of Intelligent Perception and Image Understanding of Ministry of Education of China, Xidian University. He is in charge of about 40 important scientific research projects, and published more than 20 monographs and 100 papers in international journals and conferences. His research interests include image processing, natural computation, machine learning, and intelligent information processing. **Dr. Jiao is a member of the IEEE Xi'an Section Execution Committee and the Chairman of awards and recognition committee, the Vice Board Chairperson of the Chinese Association of Artificial Intelligence, the Councilor of the Chinese Institute of Electronics, the Committee Member of the Chinese Committee of Neural Networks, and an expert of academic degrees committee of the state council.**