Pattern Recognition 132 (2022) 108966

Contents lists available at ScienceDirect

Pattern Recognition

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

Uncorrelated feature selection via sparse latent representation and extended OLSDA

Ronghua Shang[a], Jiarui Kong[a], Weitong Zhang[a,*], Jie Feng[a], Licheng Jiao[a], Rustam Stolkin[b]

[a] Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, School of Artificial Intelligence, Xidian University, Xi'an, Shanxi Province 710071, China

[b] The Extreme Robotics Lab, University of Birmingham, UK

ARTICLE INFO

Article history:
Received 22 April 2022
Revised 30 June 2022
Accepted 9 August 2022
Available online 10 August 2022

Keywords:
Unsupervised feature selection
Sparse latent representation
OLSDA
Pseudo-labels
Uncorrelated constraints

ABSTRACT

Modern unsupervised feature selection methods predominantly obtain the cluster structure and pseudo-labels information through spectral clustering. However, the pseudo-labels obtained by spectral clustering are usually mixed between positive and negative. Moreover, the Laplacian matrix in spectral clustering typically affects feature selection. Additionally, spectral clustering does not consider the interconnection information between data. To address these problems, this paper proposes uncorrelated feature selection via sparse latent representation and extended orthogonal least square discriminant analysis (OLSDA), which we term SLREO). Firstly, SLREO retains the interconnection between data by latent representation learning, and preserves the internal information between the data. In order to remove redundant interconnection information, an $l_{2,1}$-norm constraint is applied to the residual matrix of potential representation learning. Secondly, SLREO obtains non-negative pseudo-labels through orthogonal least square discriminant analysis (OLSDA) of embedded non-negative manifold structure. It not only avoids the appearance of negative pseudo-labels, but also eliminates the effect of the Laplacian matrix on feature selection. The manifold information of the data is also preserved. Furthermore, the matrix of the learned latent representation and OLSDA is used as pseudo-labels information. It not only ensures that the generated pseudo-labels are non-negative, but also makes the pseudo-labels closer to the true class labels. Finally, in order to avoid trivial solutions, an uncorrelated constraint and $l_{2,1}$-norm constraint are imposed on the feature transformation matrix. These constraints ensure row sparsity of the feature transformation matrix, select low-redundant and discriminative features, and improve the effect of feature selection. Experimental results show that the Clustering Accuracy (ACC) and Normalized Mutual Information (NMI) of SLREO are significantly improved, as compared with six other published algorithms, tested on 11 benchmark datasets.

1. Introduction

With the rapid development of information technology, many different fields of research, and industrial applications, are handling increasingly large amounts of high-dimensional data. This is increasingly apparent in fields such as computer vision and pattern recognition [1]. However, the learning performance of classification and clustering algorithms is easily affected by redundant features and noise in high-dimensional data [2]. Moreover, demands on both computer memory, and also computational complexity, will also be affected by high-dimensional data. Therefore, before directly processing high-dimensional data, dimensionality reduction is generally performed [3]. Dimensionality reduction methods can remove redundant features and noise, reduce time complexity [4] and improve algorithm performance [5]. Commonly used dimensionality reduction methods include feature extraction and feature selection [6]. Feature extraction forms new features from the original data through a series of conversion methods, thereby obtaining a new feature space [7]. Feature selection attempts to select the most representative features from the original feature set as the optimal feature subset, which can retain the information content, and physical meaning, of the original features [8], while minimizing the required number of dimensions. Moreover, feature selection can remove redundant features and noise, so it can also improve the performance of the algorithm. Therefore, feature selection is widely used in image processing, text processing and other fields [6].

Feature selection methods can be broadly divided into supervised feature selection [2], semi-supervised feature selection

[*] Corresponding author.
E-mail address: wtzhang_1@xidian.edu.cn (W. Zhang).

https://doi.org/10.1016/j.patcog.2022.108966
0031-3203/© 2022 Elsevier Ltd. All rights reserved.

[7] and unsupervised feature selection [3], according to the degree to which class labelled training data is used. Supervised feature selection provides label information, and uses the correlation between samples and labels to select discriminative features [9]. Semi-supervised feature selection provides partial labels, and combines some labeled samples with unlabeled samples to perform feature selection. Unsupervised feature selection is based on information within the data, and does not require labels. In real-world, industrial applications, or new applications which lack public benchmark datasets, labeling enough data to enable fully supervised learning methods may be prohibitively time and resource consuming. Therefore, unsupervised feature selection algorithms have are attracting increasing attention from the research community [10].

An alternative way to view feature selection methods, is to divide them according to different strategies such as filter [11], wrapper [12] and embedded [13]. Filter feature selection methods work by filtering the initial feature set through setting methods, and then use the filter features for feature selection [11]. For example, the LapScor algorithm [11] assesses the importance of features by calculating their Laplacian score through the local geometric structure of the data. Wrapper feature selection evaluates the feature subset according to the performance of a learning method such as SVM [14], and finally selects the optimal feature subset [12]. Embedded feature selection is a combination of filter and wrapper, which performs feature selection in the process of classifier learning. Compared with wrapper feature selection, embedded feature selection can reduce the computational cost. Compared with filtered feature selection, the classifier in embedded feature selection can select more accurate feature subsets. Due to these advantages, these methods are attracting increasing attention [15].

The question of how to obtain the label information of data samples in unsupervised feature selection remains an open research challenge. Most unsupervised feature selection algorithms generate pseudo-label information through spectral clustering [4]. For example, Zhao and Liu used the frequency spectrum of the graph to measure feature correlation, and proposed a feature selection framework named spectral feature selection algorithm (SPEC) [16]. Cai et al. proposed multi-cluster feature selection (MCFS) by selecting the features of multi cluster structure of data can be reserved through spectral analysis [17]. MCFS adopts a two-step strategy. The first step preserves the manifold structure, and the second step performs sparse regression with l_1-norm regularization. Hou et al. adopted a single-step strategy to unify embedded learning and sparse regression for feature selection, and proposed joint embedding learning and sparse regression (JELSR) [13]. The difference between JELSR and MCFS is that JELSR uses a single-step strategy, while MCFS uses a two-step strategy. Based on the JELSR feature selection framework, non-negative spectral learning and sparse regression-based dual-graph regularized (NSSRD) is proposed by Shang et al. [18]. NSSRD introduced the idea of the feature graph to provide discriminative information for feature selection, and imposed non-negative and $l_{2,1}$-norm constraint on the feature selection matrix. In Ref. [19], the concept of graph regularization is introduced into matrix factorization, and [19] proposes subspace learning-based graph regularized feature selection (SGFS). SGFS used the spectral embedding method to preserve the local geometric structure and imposed $l_{2,1}$-norm constraint on the feature selection matrix to select sparse discriminative features. Li et al. used the generalized uncorrelated regression model to select uncorrelated but discriminative features, and proposed generalized uncorrelated regression with adaptive graph unsupervised feature selection (URAFS) [20]. Tang et al. proposed robust unsupervised feature selection via dual self-representation and manifold regularization (DSRMR), which used feature self-representation and data self-representation to learn a sparse feature matrix and similarity matrix respectively to retain the manifold information more accurately [21]. Shang et al. proposed sparse and low-redundant subspace learning-based dual-graph regularized (SLSDR) [22]. SLSDR introduced data graph into SGFS framework, using the local geometric information of the data manifold and feature manifold to guide subspace learning. Additionally, SLSDR imposed $l_{2,1}$-norm sparsity constraint on the subspace learning residual matrix to ensure robustness to outlier samples. In order to retain the local structure information more accurately [23], embedded the adaptive similarity matrix into the subspace learning framework and proposed subspace learning for unsupervised feature selection via adaptive structure learning and rank approximation (SLASR). Compared with SLSDR, SLASR can adaptively learn manifold structure. The pseudo-labels obtained by spectral clustering are used to guide feature selection, and good results have been obtained.

Generally speaking, spectral clustering uses the method of constructing an adjacency matrix to provide pseudo-label information, which improves the result of feature selection algorithms to a certain extent. However, for large datasets, constructing an adjacency matrix is very time-consuming. Furthermore, the quality of the constructed Laplacian matrix also affects the performance of the algorithm. Zhang et al. proposed unsupervised feature selection with extended OLSDA via embedding nonnegative manifold structure (NMSFS) to solve this problem [24]. This method applied orthogonal least square discriminant analysis (OLSDA) [25] to unsupervised feature selection method, and embedded the manifold structure retained by the discrete clustering indicator matrix, which did not involve the Laplace matrix. While preserving the manifold structure, although NMSFS eliminated the impact of the Laplacian matrix on performance, it also ignored the internal connections between data.

More recently, methods are emerging that also consider the interconnection of data, yielding good feature selection results. Tang et al. proposed unsupervised feature selection via latent representation learning and manifold regularization (LRLMR) [26]. LRLMR embedded latent representation learning into the feature selection matrix to preserve the interconnection between data. And LRLMR preserved the manifold structure of the data space to select features in the latent representation space. Shang et al. proposed dual space latent representation learning for unsupervised feature selection (DSLRL) [27]. DSLRL mined the correlation between data from data space and feature space, respectively. The non-negative and orthogonal constraints are applied to the sparse transformation matrix to avoid trivial solutions, and the correlation between data space and feature space is used to guide feature selection.

Yuan et al. proposed convex non-negative matrix factorization with adaptive graph for unsupervised feature selection (CNAFS) [28]. CNAFS used the convex matrix factorization with adaptive graph constraints to mine the correlation between the data, and integrated the pseudo-label matrix learning into the self-expression module for feature selection. Among the three methods, LRLMR and DSLRL both preserved the interconnection between data by embedding latent representation learning into the feature selection framework, and CNAFS used self-adaptation to mine more internal information between data. Although the correlation between data is preserved to a certain extent, multiple Laplacian matrices need to be calculated. The Laplacian matrices have a certain impact on the efficiency of the algorithm, and it affects the performance of feature selection.

In feature selection, sparse constraints are usually used, and many scholars have proposed different constraint methods. Ye et al. proposed an NPDA algorithm that adopts the cut l_1-norm as a distance metric, which can handle even a small value well and can better eliminate outliers [29]. Ye et al. Proposed Lp- and Ls-Norm Distance Based Robust Linear Discriminant Analysis (FLDA-Lsp), which achieved robustness by using the Lp and Ls norms

[30]. Li et al. proposed the GZA-PNMCC algorithm and can be used to estimate sparse recognition problems [31]. Albu et al. proposed MFx-LPPNLMS for ANC systems and adapted to more sparse cases [32]. Li et al. proposed discriminative and uncorrelated feature selection with constrained spectral analysis in unsupervised learning (DUCFS) [33], using a regularized regression model with generalized uncorrelated constraint to explore low-redundancy and discriminative features.

In summary, we consider the influence of the Laplacian matrix on feature selection, the relationship between some features of data and hidden attributes, and the internal relationship between the data. In order to preserve the manifold structure of the data, aiming at the deficiencies of the two types of algorithms, this paper proposes uncorrelated feature selection via sparse latent representation and extended OLSDA in unsupervised learning (SLREO). SLREO embeds the non-negative manifold structure into OLSDA, which not only relaxes discrete cluster labels into continuous cluster labels, but also ensures the non-negativity of pseudo-labels. In contrast to the conventional manifold structure preservation approaches, SLREO does not use the Laplacian matrix. Hence, it can reduce the influence of the Laplacian matrix on feature selection. In addition, SLREO guarantees the non-negativity of the pseudo-labels, and the generated pseudo-labels are more accurate than other methods without restrictions. Some features of data are related to hidden attributes. In order to preserve the internal connection between data, SLREO introduces a latent representation matrix, and uses latent representation to preserve the interconnection between data. In order to make the pseudo-labels closer to the real class labels, in this paper we unify the latent representation matrix and the non-negative indication matrix with manifold structure. Furthermore, SLREO constrains the non-negativity of the pseudo-label and the interconnection of data through OLSDA and latent representation methods. To avoid the appearance of trivial solutions and excessive suppression of non-zero rows, SLREO imposes an uncorrelated constraint and $l_{2,1}$-norm constraint on the feature selection matrix. These constraints guarantee the row sparsity of the feature transformation matrix, select low-redundant features, and improve the feature selection performance.

The contributions of this paper are as follows:

(1) SLREO performs feature selection in the latent representation space, uses latent representation learning to mine the hidden information between data, and retains the interconnection between data. The latent representation matrix is used as a pseudo-label matrix. By considering the interconnection between data, the pseudo-labels more closely match physically meaningful labels in the real application.

(2) SLREO generates pseudo-label information through the OLSDA method embedded in a non-negative manifold structure. Compared with the spectral clustering methods commonly used, this algorithm can not only preserve the manifold structure, but also ensure the non-negativity of pseudo-labels. Additionally, compared with spectral clustering, SLREO also reduces the influence of the Laplacian matrix on the feature selection due to the use of the scaled discrete clustering indicator matrix.

(3) The $l_{2,1}$-norm constraint is imposed on the residual matrix of latent representation learning to ensure the effectiveness and robustness of the clustering indicator. The unified latent representation matrix and the clustering indicator matrix obtained by scaling the discrete clustering indicator matrix are compared with only latent representation learning or only OLSDA, the pseudo-labels are not only non-negative, but also contains the dependence between data.

(4) Applying uncorrelated constraint and $l_{2,1}$-norm constraint on the feature transformation matrix can avoid excessive suppression of non-zero rows and the appearance of redundant solutions. Therefore, the importance of each feature can be better reflected, so that more discriminative features can be selected.

The remainder of this paper is organized as follows. Section 2 presents the SLREO algorithm, optimization method, convergence analysis and computational complexity analysis. In Section 3, the experimental results and analysis of SLREO and compared algorithms are provided. Section 4 summarizes the paper and provides concluding remarks.

2. The proposed SLREO algorithm

This section introduces details of our proposed algorithm, SLREO, for uncorrelated feature selection via sparse latent representation and extended OLSDA in unsupervised learning. SLREO is designed to address two key problems: (i) spectral clustering cannot ensure the non-negativity of pseudo-labels; and (ii) spectral clustering ignores certain kinds of internal connections between data. The proposed SLREO methods, as well as the update rules of SLREO, computational complexity analysis and convergence analysis are provided in this section.

2.1. Uncorrelated feature selection via sparse latent representation

2.1.1. Sparse latent representation

A common assumption in unsupervised feature selection is that each data sample is independent of each other and has the same distribution as a prerequisite. However, in practical applications, samples may not be fully statistically independent of each other and may thus provide link information which can play a useful role in feature selection. In order to consider the dependency information between samples, the adjacency matrix $\boldsymbol{A}$ is constructed by Gaussian function [21] to represent the interconnection information between data. The adjacency matrix $\boldsymbol{A}$ is defined as Eq. (1).

$$\boldsymbol{A} = \exp\left(\frac{-\left\|\boldsymbol{x}^i - \boldsymbol{x}^j\right\|_2^2}{\sigma^2}\right) \tag{1}$$

where, $i, j = 1, 2, \ldots, n$. $\boldsymbol{x}^i$ represents the ith instance.$\sigma$is the Gaussian scale parameter.

The latent representation can be obtained through the adjacency matrix of the samples, because the more similar two samples are, the more likely the two samples affect each other. Generally, the information of latent representation is generated through non-negative factorization [34]. The adjacency matrix $\boldsymbol{A}$ is decomposed into the product of non-negative matrix $\boldsymbol{Q}$ and its transposed $\boldsymbol{Q}^T$. The specific form of the model is as Eq. (2).

$$\begin{aligned}&\min_{\boldsymbol{Q}} \left\|\boldsymbol{A} - \boldsymbol{Q}\boldsymbol{Q}^T\right\|_F^2\\&s.t. \boldsymbol{Q} \geq 0\end{aligned} \tag{2}$$

where$\boldsymbol{Q} \in \Re^{n\times c}$is the latent representation of n data instances, that is, the latent representation matrix. c is the number of categories, so $\boldsymbol{Q}$ is also the clustering structure of the data, which can guide feature selection.

However, there will also be repeated interconnection information between the data. This repeated interconnection information between data will increase the time complexity of the algorithm. In order to remove this redundant link information, we impose $l_{2,1}$-norm constraint on the residual matrix of the latent representation learning, to ensure the sparsity of the latent representation. Therefore, the sparse latent representation can be obtained as Eq. (3).

$$\begin{aligned}&\min_{\boldsymbol{Q}} \left\|\boldsymbol{A} - \boldsymbol{Q}\boldsymbol{Q}^T\right\|_{2,1}\\&s.t. \boldsymbol{Q} \geq 0\end{aligned} \tag{3}$$

2.1.2. Uncorrelated feature selection

According to the above content, in order to select discriminative features and give full play to the role of latent representation learning, a regression model can be obtained as Eq. (4).

$$\min_{\boldsymbol{W}} \left\|\boldsymbol{X}^T\boldsymbol{W} - \boldsymbol{Q}\right\|_F^2 \quad s.t. \boldsymbol{W} \geq 0 \tag{4}$$

where$\boldsymbol{X} \in \Re^{d\times n}$is the data matrix and$\boldsymbol{W} \in \Re^{d\times c}$is the projection transformation matrix. The importance of the ith feature is reflected by the 2-norm$\|\boldsymbol{W}(i,:)\|_2$which is regarded as the feature weight. And$\|\boldsymbol{W}(i,:)\|_2$is the ith row vector of $\boldsymbol{W}$. In general, in order to ensure the sparsity of $\boldsymbol{W}$, $l_{2,1}$-norm is used to constrain $\boldsymbol{W}$ to select the sparse feature. The form is as Eq. (5).

$$\min_{\boldsymbol{W}} \left\|\boldsymbol{X}^T\boldsymbol{W} - \boldsymbol{Q}\right\|_F^2 + \beta\|\boldsymbol{W}\|_{2,1} \quad s.t. \boldsymbol{W} \geq 0 \tag{5}$$

where, $\|\boldsymbol{W}\|_{2,1} = \sum_{i=1}^{n}(\sum_{j=1}^{d}|\boldsymbol{W}_{ij}|^2)^{1/2}$.$\beta > 0$is the balance parameter to control the sparsity of the model.

In order to improve the effect of feature selection, this paper imposes a simple uncorrelated constraint on the matrix $\boldsymbol{W}$ to select more discriminative and non-redundant features. Specifically, the matrix $\boldsymbol{W}$ is constrained by a data scatter matrix $\boldsymbol{S}$. Shang et al. proved that this constraint performs better than the orthogonal constraint in terms of uncorrelated and discriminative feature selection [23].

$$\min_{\boldsymbol{W}} \left\|\boldsymbol{X}^T\boldsymbol{W} - \boldsymbol{Q}\right\|_F^2 + \beta\|\boldsymbol{W}\|_{2,1} \quad s.t. \boldsymbol{W} \geq 0,\ \boldsymbol{W}^T\boldsymbol{S}\boldsymbol{W} = \boldsymbol{I} \tag{6}$$

where, $\boldsymbol{S} \in \Re^{d\times d}$is the common scatter matrix of data, and $\boldsymbol{S} = \boldsymbol{X}\boldsymbol{H}\boldsymbol{X}^T$. $\boldsymbol{H} \in \Re^{n\times n}$is the centering matrix and $\boldsymbol{H} = \boldsymbol{I} - (1/n)\boldsymbol{1}\boldsymbol{1}^T$. According to the value of $\boldsymbol{S}$, highly uncorrelated features can be selected.

Combining latent representation learning and uncorrelated constraint, we can get Eq. (7).

$$\min_{\boldsymbol{W},\boldsymbol{Q}} \left\|\boldsymbol{X}^T\boldsymbol{W} - \boldsymbol{Q}\right\|_F^2 + \alpha\left\|\boldsymbol{A} - \boldsymbol{Q}\boldsymbol{Q}^T\right\|_{2,1} + \beta\|\boldsymbol{W}\|_{2,1} \quad s.t. \boldsymbol{W} \geq 0,\ \boldsymbol{W}^T\boldsymbol{S}\boldsymbol{W} = \boldsymbol{I} \tag{7}$$

where α, $\beta$are the balanced parameters which are used to control the weight of latent representation learning in the model, and $\alpha > 0$, $\beta > 0$.

2.1.3. Extended OLSDA extended OLSDA

Through sparse latent representation learning, the data space containing interconnection information between data can be obtained. The pseudo-labels generated in data space are closer to the true class labels, but these pseudo-labels are not necessarily non-negative. There are methods of constraint through the Laplacian matrix, which have a certain impact on the clustering indicator matrix. Therefore, our approach combines latent representation with extended OLSDA. Extended OLSDA refers to the method of least square discriminant analysis (OLSDA) embedded in a non-negative manifold structure. A non-negative constraint can be directly imposed on the clustering indicator through this method, and the influence of the Laplacian matrix on the clustering index can be reduced. According to [25], the form of extended OLSDA application in unsupervised feature selection is as Eq. (8).

$$\min_{\boldsymbol{W}} Tr(\boldsymbol{W}^T\boldsymbol{X}\boldsymbol{H}\boldsymbol{X}^T\boldsymbol{W}) - Tr(\boldsymbol{L}^T\boldsymbol{H}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{W}^T\boldsymbol{X}\boldsymbol{H}\boldsymbol{L}) \quad s.t. \boldsymbol{W} \geq 0 \tag{8}$$

where$\boldsymbol{X} \in \Re^{d\times n}$is the data matrix which contains n samples, and each sample has d features.$\boldsymbol{H} \in \Re^{n\times n}$is the centering matrix and $\boldsymbol{H} = \boldsymbol{I} - (1/n)\boldsymbol{1}\boldsymbol{1}^T$. $\boldsymbol{L} \in \Re^{n\times c}$is a scaled discrete clustering indicator matrix.

Considering that the applicable scenario is unsupervised feature selection, only OLSDA is not enough. Moreover, the efficiency of unsupervised feature selection by OLSDA is low, so we use latent representation learning to make supplementary feature selection. $\boldsymbol{L}$ can be relaxed into a continuous clustering indicator matrix $\boldsymbol{Q}$. Since the true class labels are non-negative, non-negative constraint is imposed on the clustering indicator matrix $\boldsymbol{Q}$ to interact. At this time, a non-negative clustering indicator matrix with manifold structure can be obtained as Eq. (9).

$$\min_{\boldsymbol{W},\boldsymbol{Q}} Tr(\boldsymbol{W}^T\boldsymbol{X}\boldsymbol{H}\boldsymbol{X}^T\boldsymbol{W}) - Tr(\boldsymbol{Q}^T\boldsymbol{H}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{W}^T\boldsymbol{X}\boldsymbol{H}\boldsymbol{Q}) \quad s.t.\ \boldsymbol{W} \geq 0,\ \boldsymbol{Q} \geq 0 \tag{9}$$

where$\boldsymbol{Q} \in \Re^{n\times c}$is clustering indicator matrix. The $\boldsymbol{Q}$ in Eq. (9) is the latent representation matrix, that is, the clustering indicator matrix obtained through latent representation learning. Then, non-negative and more accurate pseudo-labels can be obtained by using extended OLSDA for non-negative constraint.

2.2. The objective function of SLREO

SLREO combines sparse latent representation learning and OLSDA, and the feature selection of SLREO is performed in the latent representation space. The clustering indicator matrix is generated by latent representation learning, and the feature selection is carried out under the constraint of OLSDA. This can ensure that the dependency relationship between the data is included, and it also ensures the non-negativity of the clustering indicators. Therefore, a balance parameter needs to be introduced to balance the OLSDA and the latent representation. Combining Eqs. (7) and (9), the objective function of SLREO can be obtained as Eq. (10).

$$\begin{aligned} \min_{\boldsymbol{W},\boldsymbol{Q}} & Tr(\boldsymbol{W}^T\boldsymbol{X}\boldsymbol{H}\boldsymbol{X}^T\boldsymbol{W}) - Tr(\boldsymbol{Q}^T\boldsymbol{H}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{W}^T\boldsymbol{X}\boldsymbol{H}\boldsymbol{Q}) \\ & + \lambda\left[\left\|\boldsymbol{X}^T\boldsymbol{W} - \boldsymbol{Q}\right\|_F^2 + \alpha\left\|\boldsymbol{A} - \boldsymbol{Q}\boldsymbol{Q}^T\right\|_{2,1} + \beta\|\boldsymbol{W}\|_{2,1}\right] \\ s.t.\ & \boldsymbol{W} \geq 0,\ \boldsymbol{Q} \geq 0,\ \boldsymbol{W}^T\boldsymbol{S}\boldsymbol{W} = \boldsymbol{I} \end{aligned} \tag{10}$$

where$\lambda$is the balance parameter, and $\lambda > 0$.

In feature selection, matrix $\boldsymbol{W}$ and matrix $\boldsymbol{Q}$ can be obtained by optimizing the objective function of SLREO. As mentioned earlier, the importance of each feature is measured by$\|\boldsymbol{w}_i\|_2$. The larger the value of $\|\boldsymbol{w}_i\|_2$, the more important the ith feature. We sort all$\|\boldsymbol{w}_i\|_2$in descending order, select the first l features, form a new data matrix$\boldsymbol{X}_{new} \in \Re^{n\times l}$, and complete feature selection.

2.3. Optimization procedure of SLREO

This section explains the process of updates and optimization of the objective function (10). To incorporate the uncorrelated constraints fully in the update process, we add this constraint to the Lagrangian function and solve it using the KKT condition [35]. For a single matrix $\boldsymbol{W}$ or matrix $\boldsymbol{Q}$, when the other variable is fixed, the problem is convex. Therefore, this paper adopts the alternate iterative update method [36] to optimize the objective function. There are two variables in the objective function, matrix $\boldsymbol{W}$ and matrix $\boldsymbol{Q}$. By fixing one of the variables and updating the other variable, the objective function can be decomposed into two sub-problems to update and optimize, respectively.

By introducing Lagrangian multipliers$\boldsymbol{\Psi}$and$\boldsymbol{\Phi}$to constrain$\boldsymbol{W} > 0$and$\boldsymbol{Q} > 0$, respectively, the Lagrangian function of the objective function (10) is as Eq. (11).

$$\begin{aligned} L(\boldsymbol{W},\boldsymbol{Q}) = & \lambda\left[\left\|\boldsymbol{X}^T\boldsymbol{W} - \boldsymbol{Q}\right\|_F^2 + \alpha\left\|\boldsymbol{A} - \boldsymbol{Q}\boldsymbol{Q}^T\right\|_{2,1} + \beta\|\boldsymbol{W}\|_{2,1}\right] \\ & + Tr(\boldsymbol{W}^T\boldsymbol{X}\boldsymbol{H}\boldsymbol{X}^T\boldsymbol{W}) - Tr(\boldsymbol{Q}^T\boldsymbol{H}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{W}^T\boldsymbol{X}\boldsymbol{H}\boldsymbol{Q}) \\ & + \gamma\left\|\boldsymbol{W}^T\boldsymbol{S}\boldsymbol{W} - \boldsymbol{I}\right\|_2^2 + Tr(\boldsymbol{\Psi}\boldsymbol{W}^T) + Tr(\boldsymbol{\Phi}\boldsymbol{Q}^T) \end{aligned} \tag{11}$$

Algorithm 1
The procedure of SLRE0.

```
Input: Data matrix$\boldsymbol{X} \in \Re^{n\times d}$, the balance parameter $\alpha$, $\beta$, $\lambda$ and $\gamma$, the maximum number of iterations $N_{iter}$, the number of feature selection $l$.
Initialize the matrix $\boldsymbol{W}$=$ones(d, c)$, $\boldsymbol{Q}$=$rand(n, c)$;
Initialize the matrix$\boldsymbol{H} = \boldsymbol{I} - (1/n)\mathbf{1}\mathbf{1}^T$;
Initialize the matrix$\boldsymbol{S} = \boldsymbol{XHX}^T$;
Compute the adjacency matrix$\boldsymbol{A} \in \Re^{n\times n}$;
Update $\boldsymbol{W}$ and $\boldsymbol{Q}$ according to Eqs. (18) and (21) until convergence;
The$||\boldsymbol{W}_i||_2$of the $i$th feature is arranged in descending order, and the first $l$ corresponding features are selected to construct a new data matrix$\boldsymbol{X}_{new}$.
Output: Select the feature index set $index$, the new data matrix$\boldsymbol{X}_{new} \in \Re^{n\times l}$.
```

Table 1
Computational complexity analysis of SLREO compared with five other well-known algorithms.

Algorithms	Computational complexity
MCFS	$O(dn^2)$
JELSR	$O(dn^2+N_{iter}(n^3+mdn))$
URAFS	$O(N_{iter}(d^3+dn^2))$
LRLMR	$O(dn^2+N_{iter}(d^3))$
CNAFS	$O(d^2n+n^2d+n^3)$
SLREO	$O(N_{iter}(n^2d+d^2n))$

where $\gamma > 0$ and γ is used to control the effect of unrelated constraints on the update formula.

The objective function of SLREO is solved by an alternating iterative method, that is, the objective function is decomposed into two sub-problems: fixed $\boldsymbol{Q}$ optimization $\boldsymbol{W}$ and fixed $\boldsymbol{W}$ optimization $\boldsymbol{Q}$.

The update rule of $\boldsymbol{W}$ can be obtained as Eq. (12)

$$\boldsymbol{W}_{ij} \leftarrow \boldsymbol{W}_{ij} \frac{\left[\boldsymbol{XHQQ}^T\boldsymbol{HX}^T\boldsymbol{W}+\lambda\boldsymbol{XQ}+2\gamma\boldsymbol{SWI}\right]_{ij}}{\left[\boldsymbol{XHX}^T\boldsymbol{W}+\lambda\left(\boldsymbol{XX}^T\boldsymbol{W}+\beta\boldsymbol{DW}\right)+2\gamma\boldsymbol{SWW}^T\boldsymbol{S}^T\boldsymbol{W}\right]_{ij}} \tag{12}$$

Therefore, the update rule of $\boldsymbol{Q}$ can be obtained as Eq. (13).

$$\boldsymbol{Q}_{ij} \leftarrow \boldsymbol{Q}_{ij} \frac{\left[\lambda\left(\boldsymbol{X}^T\boldsymbol{W}+\alpha\boldsymbol{A}^T\boldsymbol{UQ}\right)+\boldsymbol{HX}^T\boldsymbol{WW}^T\boldsymbol{XHQ}\right]_{ij}}{\left[\lambda\left(\boldsymbol{Q}+2\alpha\boldsymbol{QQ}^T\boldsymbol{UQ}\right)\right]_{ij}} \tag{13}$$

The specific update process is described in detail in Supplementary materials. The convergence of the objective function is proved in Supplementary materials. The optimization process of SLREO is summarized in Algorithm 1.

2.4. Computational and space complexity analysis

First, notation is introduced. n is the total number of samples and d is the number of features for each sample. c is the number of categories, l is the number of features selected, m represents the dimension of the low-dimensional space and N_{iter} is the number of iterations. Next, computational complexity analysis is performed. When the adjacency matrix A and the data scatter matrix S are constructed, the computational complexity is $O(dn^2+d^2n)$. In each iteration, the computational complexity of updating the matrix W is $O(d^2+n^2d+d^2n+d^2c+ndc)$, and the computational complexity of updating the matrix Q is $O(n^2+n^2c+n^2c+nc)$. Therefore, the computational complexity of SLREO is $O(N_{iter}(n^2+d^2+n^2d+d^2n+d^2c+n^2c+c^2n))$. Since in practical applications, $c<<d$ and $c<<n$, and $n>d$ or $n<d$, the total complexity of SLREO is $O(N_{iter}(n^2d+d^2n))$.

It can be found from Table 1 that, except for MCFS, the highest order of computational complexity among the other four comparison algorithms is 3. And the highest order of SLREO is 2. It can be seen from the order of magnitude that SLREO has certain advantages in computational complexity.

Table 2
Details of the datasets.

Dataset	Instance	Feature	Class	Type
COIL20	1440	1024	20	Digital image
Isolet1	1560	617	26	Speech signal
warpPIE10P	210	2420	10	Face image
AR10P	130	2400	10	Face image
Mnist	5000	784	10	Digital image
JAFFE	213	676	10	Face image
AT&T	400	10304	40	Face image
orlraws10P	100	10304	40	Face image
Optdigit	3823	64	10	Digital image
CLL-SUB-111	111	11340	10	Biological microarray
PIE32	11554	1024	32	Face image

When updating the objective function of SLREO, the required space complexity for matrices is $O(n^2+d^2+nd+dc+nc)$, and the space complexity for defining variables is $O(n^2+c^2)$. And the total space complexity is $O(n^2+\mathrm{d}^2)$.

3. Experiments

In order to evaluate the effectiveness of SLREO is compared against six other unsupervised feature selection algorithms, including well known classical methods, and recent state-of-the-art methods. Eleven datasets are used in the comparative experiment. This section first introduces the datasets and algorithms. Next, a parameter sensitivity experiment is presented, to verify that changes of key parameters do not substantially degrade the performance of SLREO. Finally, a convergence experiment is used to verify the convergence of SLREO.

3.1. Datasets and the compared algorithms

The SLREO algorithm is tested on eleven datasets, including face image datasets (warpPIE10P, AR10P, JAFFE [33], AT&T, orlraws10P and PIE32 [37]), digital image datasets (COIL20, Mnist and Optdigit) [28], a voice signal dataset (Isolet1) and a Biological microarray dataset (CLL-SUB-111) [28]. Table 2 shows these datasets.

In order to verify the effectiveness of the SLREO algorithm in feature selection, this experiment selects a baseline algorithm and six unsupervised feature selection algorithms for comparison. These algorithms are described as follows.

(1) Baseline: feature selection is not performed during clustering, and all original features are retained.
(2) MCFS [18]: multi-cluster feature selection selects features through spectral analysis and l_1-norm regularization regression model, and retains the manifold structure.
(3) JELSR [13]: JELSR is an unsupervised feature selection framework that adopts a one-step strategy. And JELSR unifies embedded learning and sparse regression for feature selection.
(4) URAFS [20]: URAFS uses preserved manifold structure and generalized irrelevant regression model to select irrelevant but discriminant features.
(5) LRLMR [27]: use the data space which retains the local geometric structure to learn latent representations, and perform feature selection in the latent representation space.
(6) CNAFS [28]: convex matrix factorization with adaptive graph constraint is used to mine the correlation between data, and the pseudo-label matrix learning is integrated into the self-expression module for feature selection.

3.2. Evaluation metrics

To evaluate the effectiveness of an unsupervised feature selection algorithm, one of the most direct methods is clustering. The

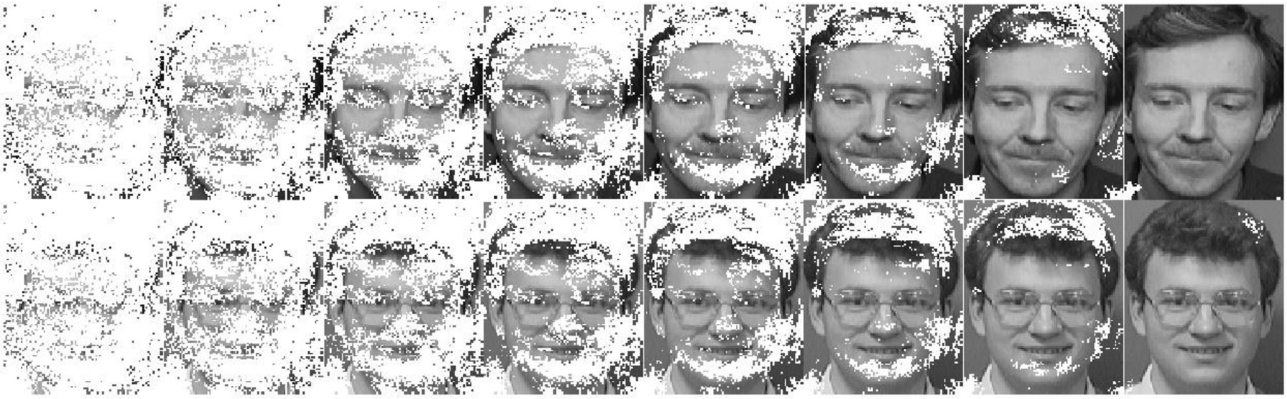


Fig. 1. The results of selecting different numbers of features for two test samples in the AT&T dataset.

selected features are clustered, and the effectiveness of the algorithm is judged by the clustering effect. The Clustering Accuracy (ACC) [37] and Normalized Mutual Information (NMI) [38] are the measurement standards of clustering effect. If the values of ACC and NMI are larger, the clustering effect is better, and the effect of feature selection algorithm is better. Given that p_i and c_i are the cluster label and true label of sample x_i, respectively, the definition of ACC is as Eq. (14).

$$ACC = \frac{1}{n}\sum_{i=1}^{n}\delta(c_i, map(p_i)) \tag{14}$$

where, n is the total number of samples. $map(\cdot)$ is the optimal mapping function. Kuhn-Munkres [38] algorithm is used to arrange the clustering labels to match the real labels.$\delta(p_i, c_i)$is an indicator function with a value of 0-1. When p_i=c_i $\delta(p_i, c_i)$= 1 otherwise it is 0. The ACC reflects the same proportion of the cluster label and the true label. The larger the ACC value, the more accurate the cluster labels are

Given two variables P and C, the definition of NMI is as Eq. (15).

$$\mathrm{NMI} = \frac{I(P,C)}{\sqrt{\mathrm{H}(P)H(C)}} \tag{15}$$

where H(P) and H(C) are the entropy and I(P, C) is the mutual information of P and C. Corresponding to the clustering problem, P and C represent the cluster label and the true label of the sample, respectively. NMI can reflect the consistency between the clustering results and the true labels, and the larger the NMI value is, the higher the consistency.

3.3. Experimental settings

This experiment performs feature selection on 7 algorithms on 11 datasets. *k-means* [39] performs cluster analysis on selected features. In the clustering process, the *k-means* algorithm is sensitive to the initial values, so this experiment performs clustering for each algorithm 20 times. Finally, the average of 20 results is selected as the final clustering result. For MCFS, JELSR, URAFS, LRLMR algorithms, the nearest neighbor parameter k is set to 5. The search range of the balance parametersα, β, λandγin this experiment is set to {10^{-8}, 10^{-7}, ..., 10^{+7}, 10^{+8}}. For all datasets, the adjustment range of the feature number l is {20, 30, 40, 50, 60, 70, 80, 90, 100}. The parameter maximum number of iterations N_{iter} is set to 50. On each dataset, the values of ACC and NMI are maximized by adjusting the value of the balance parameterα, β, λ, $\gamma$and the number of feature selection l. For the compared algorithms of MCFS, JELSR, URAFS, LRLMR, CNAFS, the parameters are adjusted according to the method proposed in the paper to select the best clustering result.

3.4. Experimental results and analysis

3.4.1. Feature selection on face images

In this experiment, two samples are randomly selected from the AT&T face dataset, which are the sixth sample of the third class and the second sample of the sixth class. Features are selected from the range of feature numbers {1280, 2560, 3840, 5120, 6400, 7680, 8960, 10240} by SLREO, and experiments are performed on test samples. The experimental results are shown in Fig. 1.

It can be found from Fig. 1 that with the increase of the number of features selected by the SLREO algorithm, the facial features in the two test samples become progressively clearer. This shows that SLREO can effectively select the more important features such as eyes, nose, mouth, ears and chin in the face. This also illustrates that SLREO can select representative features.

3.4.2. Comparison of experimental results and analysis

Tables 3 and 4 show the ACC, NMI and standard deviation (STD) of eight algorithms for feature selection on different datasets. Among them, the black boldface indicates the best value, and the underline indicates the second best value.

Table 3 mainly reflects the ACC values of SLREO and the compared algorithms on different datasets. From the perspective of ACC value, SLREO achieves the best value on most datasets. Especially on the dataset JAFFE, SLREO performs 12% better than the next best algorithms. On the dataset AR10P, SLREO is also 8% higher than the next best. This reflects the advantages of SLREO compared with the compared algorithms. On other datasets, SLREO is also more than 2% higher than the compared algorithms on average. On the dataset AT&T, due to the large amount of data, the compared algorithms do not exceed Baseline. Only the value of SLREO exceeds the compared algorithms, and it is much higher than the compared algorithms. It can be seen from Table 3 that SLREO performs strongly in comparison to other methods.

Table 4 shows the NMI values of SLREO and the compared algorithms on different datasets. In Table 4, SLREO consistently achieves the best NMI scores. Especially on the very large dataset AT&T, SLREO can still achieve the best score. Because the dataset AT&T is large and has many features, the NMI values of the comparison algorithms are not as good as that of Baseline. Only SLREO outperforms baseline on this large dataset Compared with MCFS, JELSR, URAFS, LRLMR and CNAFS, SLREO is 4% higher on average. On the dataset Mnist, SLREO is nearly 12% higher than the next best value. On the dataset JAFFE, SLREO is about 6% higher than the next best value. On other datasets, SLREO is significantly higher than the compared algorithms, demonstrating its significant advantages and benefits.

Table 3
The best clustering results of different algorithms on 11 datasets (ACC±STD%).

Datasets	Baseline	MCFS	JELSR	URAFS	LRLMR	CNAFS	SLREO
COIL20	65.75 ± 4.16	64.26 ± 3.91	63.16 ± 2.84	62.37 ± 3.10	64.44 ± 1.89	66.82 ± 2.92	68.14 ± 0.33
Isolet1	61.73 ± 2.77	56.15 ± 1.55	62.93 ± 1.91	58.13 ± 3.13	63.26 ± 1.65	65.71 ± 2.76	67.23 ± 2.32
warpPIE10P	26.60 ± 0.97	36.57 ± 1.71	28.95 ± 1.69	30.17 ± 2.11	47.15 ± 1.93	57.21 ± 3.89	57.74 ± 4.58
AR10P	21.50 ± 2.93	23.69 ± 1.96	44.58 ± 3.15	29.12 ± 1.80	27.73 ± 2.59	43.23 ± 2.64	52.81 ± 2.78
Mnist	53.84 ± 1.51	48.08 ± 1.65	50.01 ± 1.05	53.21 ± 0.51	51.55 ± 2.02	52.45 ± 1.59	55.47 ± 1.75
JAFFE	86.13 ± 5.66	87.23 ± 5.24	74.92 ± 2.69	79.79 ± 5.17	80.09 ± 4.20	80.84 ± 4.64	95.65 ± 5.75
AT&T	60.96 ± 3.30	56.30 ± 2.78	61.25 ± 2.04	57.21 ± 2.24	58.10 ± 2.48	55.15 ± 2.21	63.05 ± 2.51
orlraws10P	76.60 ± 6.17	78.40 ± 5.47	74.90 ± 4.24	75.95 ± 4.45	77.25 ± 4.19	80.30 ± 4.27	81.85 ± 4.36
Optdigit	80.35 ± 0.15	78.47 ± 0.60	81.53 ± 0.64	79.89 ± 0.27	80.35 ± 0.78	80.09 ± 1.31	84.62 ± 0.08
CLL-SUB-111	53.15 ± 0.00	53.66 ± 1.67	53.15 ± 0.00	53.15 ± 0.00	52.25 ± 2.51	53.15 ± 0.00	60.18 ± 0.37
PIE32	7.54 ± 0.23	7.74 ± 0.22	7.40 ± 0.22	8.08 ± 0.23	8.65 ± 0.33	8.56 ± 0.20	10.05 ± 0.31

Table 4
The best clustering results of different algorithms on 11 datasets (NMI ± STD%).

Datasets	Baseline	MCFS	JELSR	URAFS	LRLMR	CNAFS	SLREO
COIL20	76.69 ± 1.99	74.50 ± 1.50	73.85 ± 1.48	73.12±1.73	75.80 ± 1.09	76.93 ± 1.62	77.10 ± 1.79
Isolet1	76.06 ± 1.26	69.47 ± 0.85	73.74 ± 0.67	73.16 ± 1.08	75.61 ± 0.64	75.60 ± 1.09	76.61 ± 0.98
warpPIE10P	26.38 ± 2.69	40.80 ± 2.13	28.02 ± 2.33	27.44 ± 2.15	56.23 ± 2.2	58.19 ± 2.91	59.44 ± 2.99
AR10P	18.62 ± 2.98	22.24 ± 2.80	41.06 ± 3.49	25.87 ± 1.87	24.76 ± 2.78	41.31 ± 2.71	56.00 ± 2.11
Mnist	46.72 ± 0.71	43.05 ± 0.62	42.76 ± 0.39	42.71 ± 0.44	42.79 ± 0.77	43.05 ± 0.78	58.55 ± 2.88
JAFFE	87.57 ± 3.93	89.42 ± 2.99	75.86 ± 1.48	80.42 ± 2.96	82.30 ± 3.00	84.64 ± 2.95	95.30 ± 3.15
AT&T	78.96 ± 1.37	74.08 ± 1.35	79.22 ± 1.41	76.84 ± 1.87	77.85 ± 1.37	72.5 ± 1.05	80.57 ± 1.27
orlraws10P	81.67 ± 4.70	82.76 ± 3.65	80.24 ± 2.65	80.97 ± 3.61	86.40 ± 2.60	80.42 ± 2.10	87.10 ± 2.81
Optdigit	75.84 ± 0.21	73.78 ± 0.33	74.49 ± 0.34	73.69 ± 0.09	74.91 ± 0.41	70.73 ± 0.47	75.90 ± 1.06
CLL-SUB-111	18.07 ± 0.00	15.38 ± 1.59	17.87 ± 0.00	18.06 ± 0.00	15.37 ± 0.00	18.07 ± 0.35	34.74 ± 0.09
PIE32	18.90 ± 0.29	20.67 ± 0.20	20.51 ± 0.24	20.42 ± 0.28	21.07 ± 0.27	18.97 ± 0.17	22.90 ± 0.31

The superior performance of SLREO is because SLREO conducts latent representation learning initially, when generating the clustering indicator matrix and the interconnection information between the data is no longer ignored. Moreover, SLREO imposes $l_{2,1}$-norm constraint on the residual matrix of latent representation learning to ensure that the interconnection information is low-redundant. Then, SLREO uses extended OLSDA to ensure the non-negativity of clustering labels. Finally, when selecting features, the features selected by uncorrelated constraint and non-negative constraint are more discriminative and non-redundant.

Fig. 2 gives the ACC results of SLREO and the compared algorithms using different numbers of features.

It can be seen from Fig. 2 that SLREO can maintain the best scores, out of all compared algorithms, in most cases under different number of features over each dataset. Especially prominently on the dataset AR10P, for example, SLREO clearly maintains the highest performance value over all numbers of features. With the increase of the number of features, the feature selection effect of SLREO gradually improves. Fig. 2 can also reflect that SLREO has a strong feature selection performance.

Fig. 3 gives the NMI results of SLREO and the compared algorithms using different numbers of features. The abscissa indicates the selected feature number l, and the ordinate indicates NMI.

In Fig. 3, especially on the datasets Mnist and AR10P, SLREO algorithm clearly demonstrates superior feature selection performance. On the dataset AT&T, the compared algorithms are under the Baseline, but SLREO outperforms the Baseline. From the overall point of view of other datasets, SLREO has a small number of points, in a few cases, where performance dips below that of some of the compared algorithms; however, SLREO clearly demonstrates superior performance in most cases.

From Figs. 2 and 3, the feature clustering effect selected by SLREO is clearly superior to the compared algorithms. The main reason is that SLREO not only considers the interconnection information, but also filters the interconnection information, and the retained data information is more complete and less redundant. When providing clustering indicators, extended OLSDA can ensure the non-negativity of pseudo-labels. Therefore, the pseudo-labels of SLREO are closer to the real class labels, and the feature selection has better performance.

3.4.3. Parameter sensitivity analysis

This section fixes other parameters and adjusts the values of balance parametersβandγwithin the range of {0, 10^{-3}, 10^{-2}, 10^{-1}, 10^{+0}, 10^{+1}, 10^{+2}, 10^{+3}} to show the sensitivity of SLREO to balance parametersβandγ.From the experiment, the ACC and NMI values of each groupβandγwere plotted with a three-dimensional histogram. Figs. 4 and 5 are the 3D histograms of ACC and NMI, respectively.

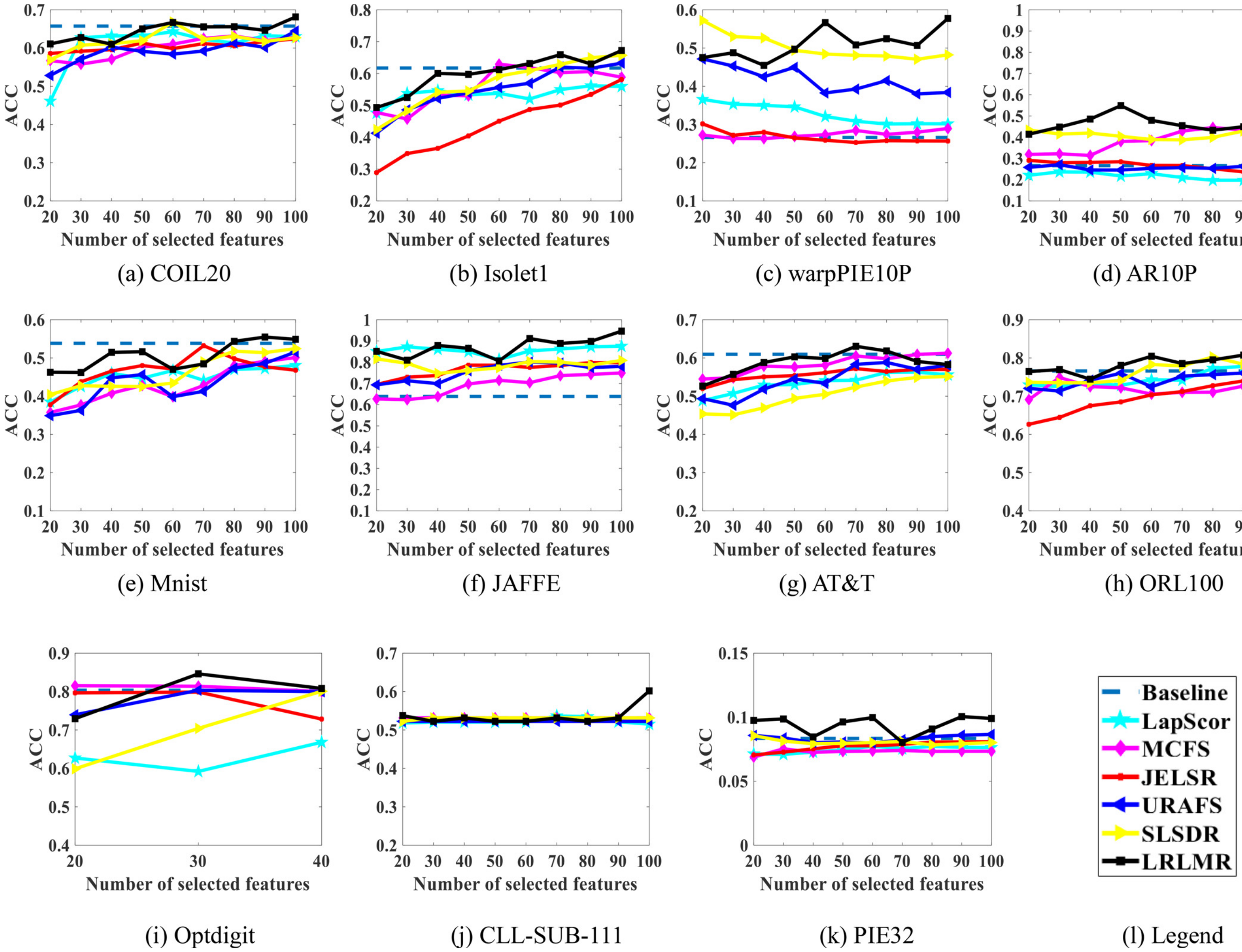
In Fig. 4, the value of ACC changes very little on the datasets COIL20, AR10P and JAFFE. With the change ofβandγ, the value of ACC of SLREO fluctuates, however, this fluctuation is not large. SLREO also tends to be stable under each parameter value in Fig. 4. This shows that ACC is not sensitive to the changes of parameters βandγ, and the changes do not significantly affect the ACC value.

In Fig. 5, on the dataset Mnist, asβandγincrease, the NMI value also increases. Due to the large amount of data in Mnist, the selected features appear redundant. On the datasets Isolet1 and JAFFE, the changes are very small. Especially on the dataset JAFFE, the change tends to be stable. On other datasets, the values of NMI do not change greatly with the change of parametersβandγ.

3.4.4. Ablation experiment

To verify the contribution of extended OLSDA and uncorrelated constraint to the algorithm, an ablation experiment is performed. This experiment is conducted on five datasets to obtain ACC and NMI values. The results are in Tables 5 and 6, where ’ × ’ indicate that no uncorrelated constraint is used or no extended OLSDA is used, and ’√’ indicates that uncorrelated constraint is used or extended OLSDA is used.

From Tables 5 and 6, it can be found that the effect when using uncorrelated constraint and extended OLSDA is superior to the performance without these components. When only one of them was



(a) COIL20 (b) Isolet1 (c) warpPIE10P (d) AR10P

(e) Mnist (f) JAFFE (g) AT&T (h) ORL100

(i) Optdigit (j) CLL-SUB-111 (k) PIE32 (l) Legend

Fig. 2. The ACC of eight algorithms for selecting different number of features on eleven datasets.

Table 5
The effect of individual constraints on ACC.

Extended OLSDA	Uncorrelated constraint	COIL20	Isolet1	warpPIE10P	AR10P	JAFFE	Optdigit
×	×	47.66 ± 2.07	64.95 ± 1.38	43.55 ± 2.08	50.88 ± 2.19	78.31 ± 4.22	82.78 ± 0.11
√	×	4 7.73 ± 1.13	66.91 ± 1.89	55.38 ± 2.09	50.88 ± 2.11	78.31 ± 4.22	82.78 ± 0.11
×	√	47.66 ± 2.07	66.89 ± 1.35	47.19 ± 2.55	44.42 ± 2.45	90.33 ± 5.98	80.80±0.10
√	√	68.14 ± 0.33	67.23 ± 2.32	57.74 ± 4.58	52.81 ± 2.78	95.65 ± 5.75	82.80 ± 0.09

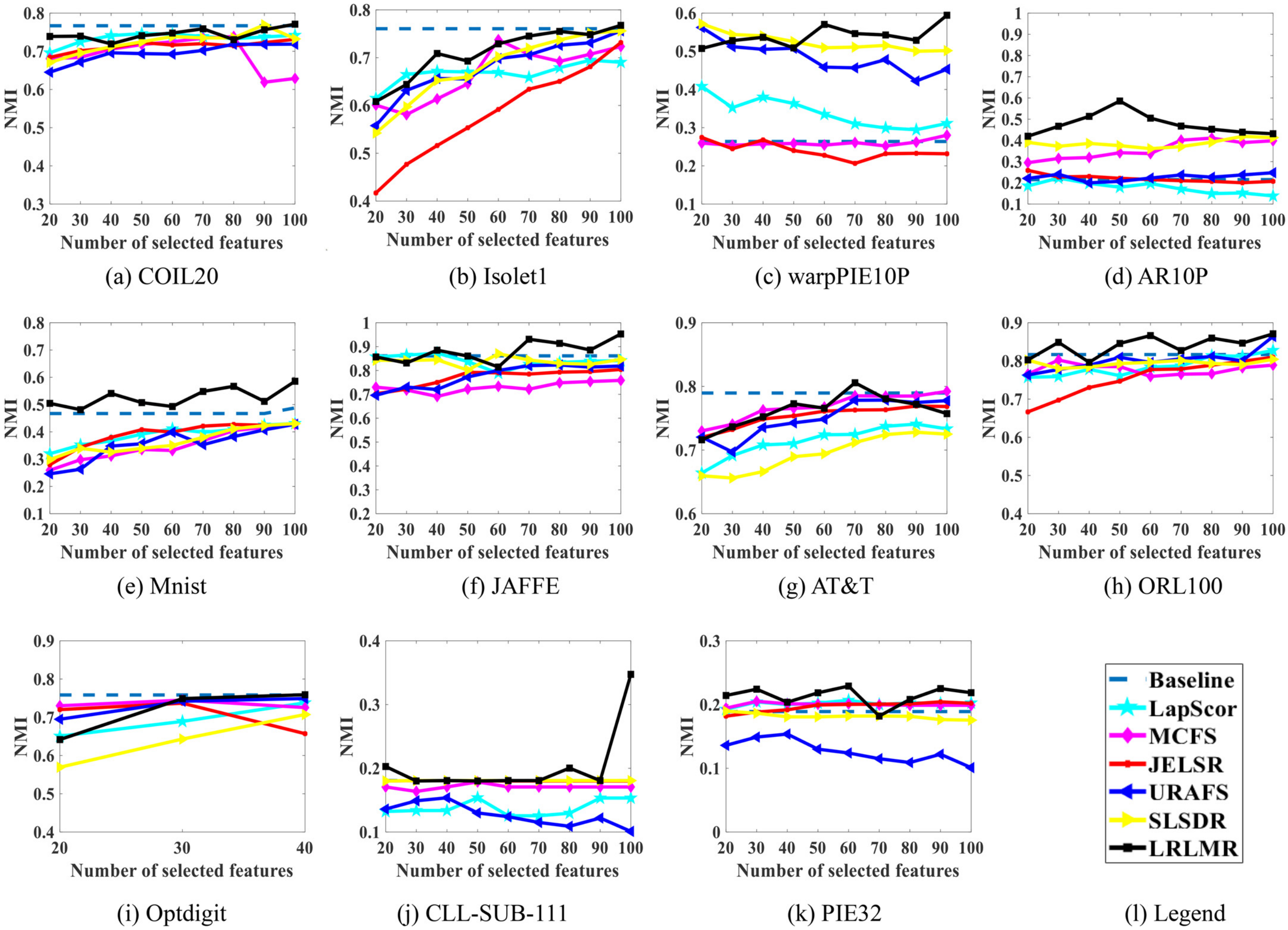
Table 6
The effect of individual constraints on NMI.

Extended OLSDA	Uncorrelated constraint	COIL20	Isolet1	warpPIE10P	AR10P	JAFFE	Optdigit
×	×	57.24 ± 0.94	73.99 ± 0.99	43.21 ± 2.94	53.27 ± 2.71	78.40 ± 2.55	74.84 ± 0.11
√	×	5 8.62 ± 0.93	76.04 ± 1.15	56.87 ± 2.18	53.48 ± 2.00	78.40 ± 2.55	74.84 ± 0.11
×	√	57.24 ± 0.94	75.12 ± 0.74	46.35 ± 3.83	43.21 ± 2.42	93.02 ± 3.62	72.89 ± 0.12
√	√	77.10 ± 1.79	76.61 ± 0.98	57.74 ± 4.58	56.00 ± 2.11	95.30 ± 3.15	75.90 ± 1.06

used, the clustering results do not change much. On the datasets Isolet and Optdigit, the results are not much different. But on the datasets COIL20 and JAFFE, it is clear that using uncorrelated constraint greatly improves performance. This shows that extended OLSDA and uncorrelated constraint in SLREO is effective and can improve the accuracy of the algorithm.

3.4.5. Convergence test

In the previous section, the theoretical convergence of SLREO was analyzed. Fig. 6 plots the change of the SLREO objective function value during the experiment to show the convergence of the SLREO algorithm as empirically measured. The objective function value and the number of iterations of SLREO are represented by

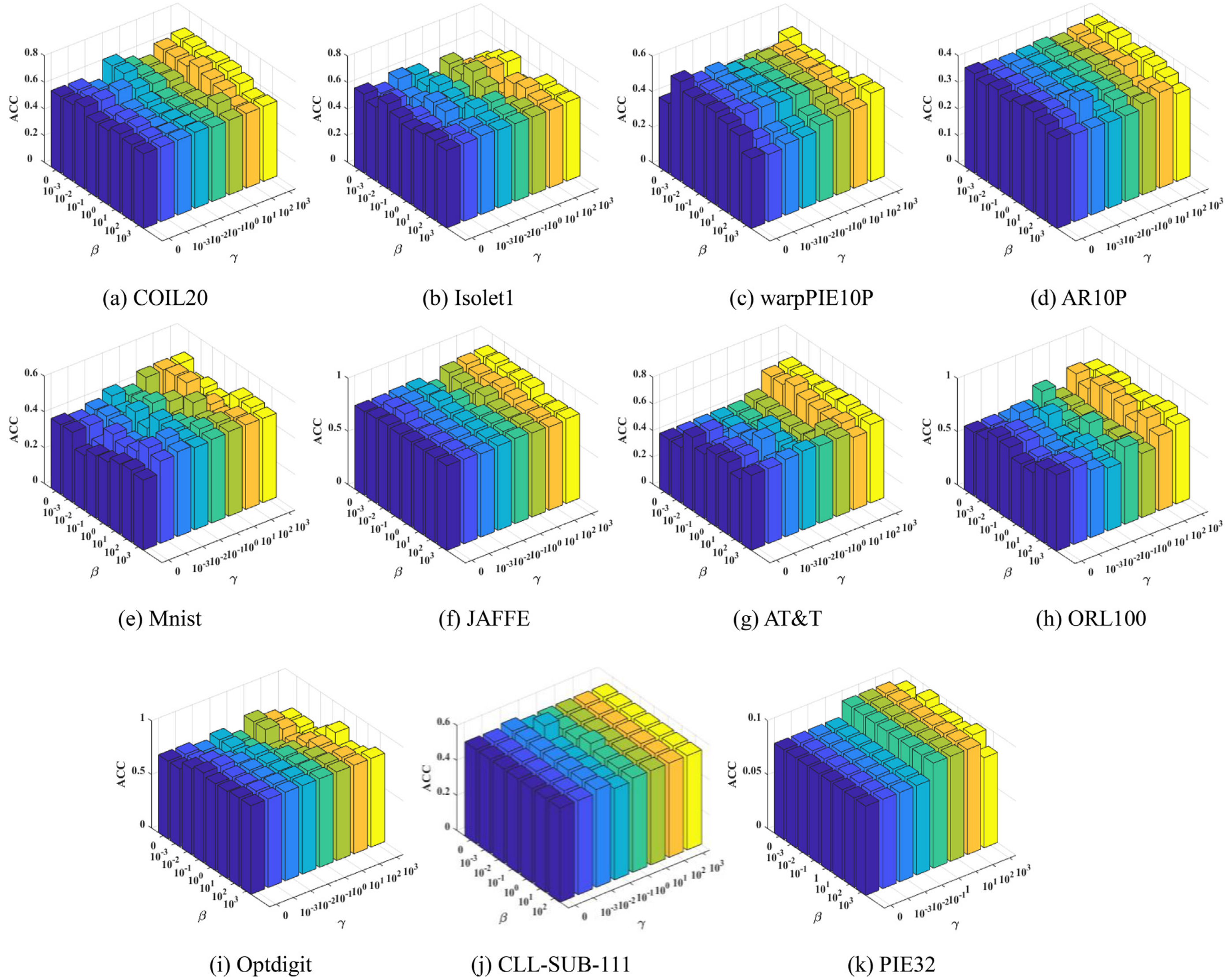


(a) COIL20 (b) Isolet1 (c) warpPIE10P (d) AR10P

(e) Mnist (f) JAFFE (g) AT&T (h) ORL100

(i) Optdigit (j) CLL-SUB-111 (k) PIE32 (l) Legend

Fig. 3. The NMI of eight algorithms for selecting different number of features on eleven datasets.



(a) COIL20

(b) Isolet1

(c) warpPIE10P

(d) AR10P

(e) Mnist

(f) JAFFE

(g) AT&T

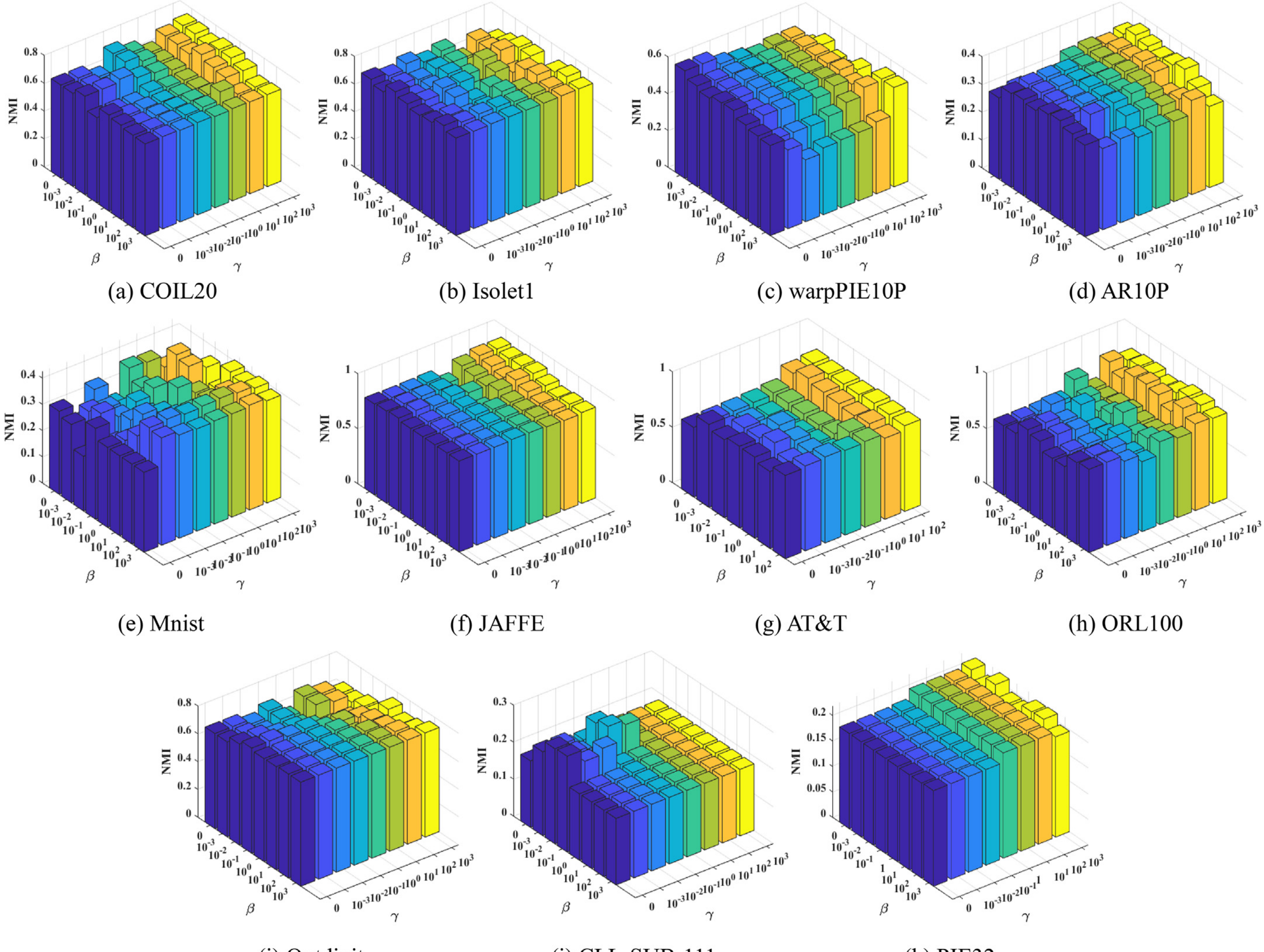
(h) ORL100

(i) Optdigit

(j) CLL-SUB-111

(k) PIE32

Fig. 4. The ACC of SLREO with differentβandγon eleven datasets.



(a) COIL20

(b) Isolet1

(c) warpPIE10P

(d) AR10P

(e) Mnist

(f) JAFFE

(g) AT&T

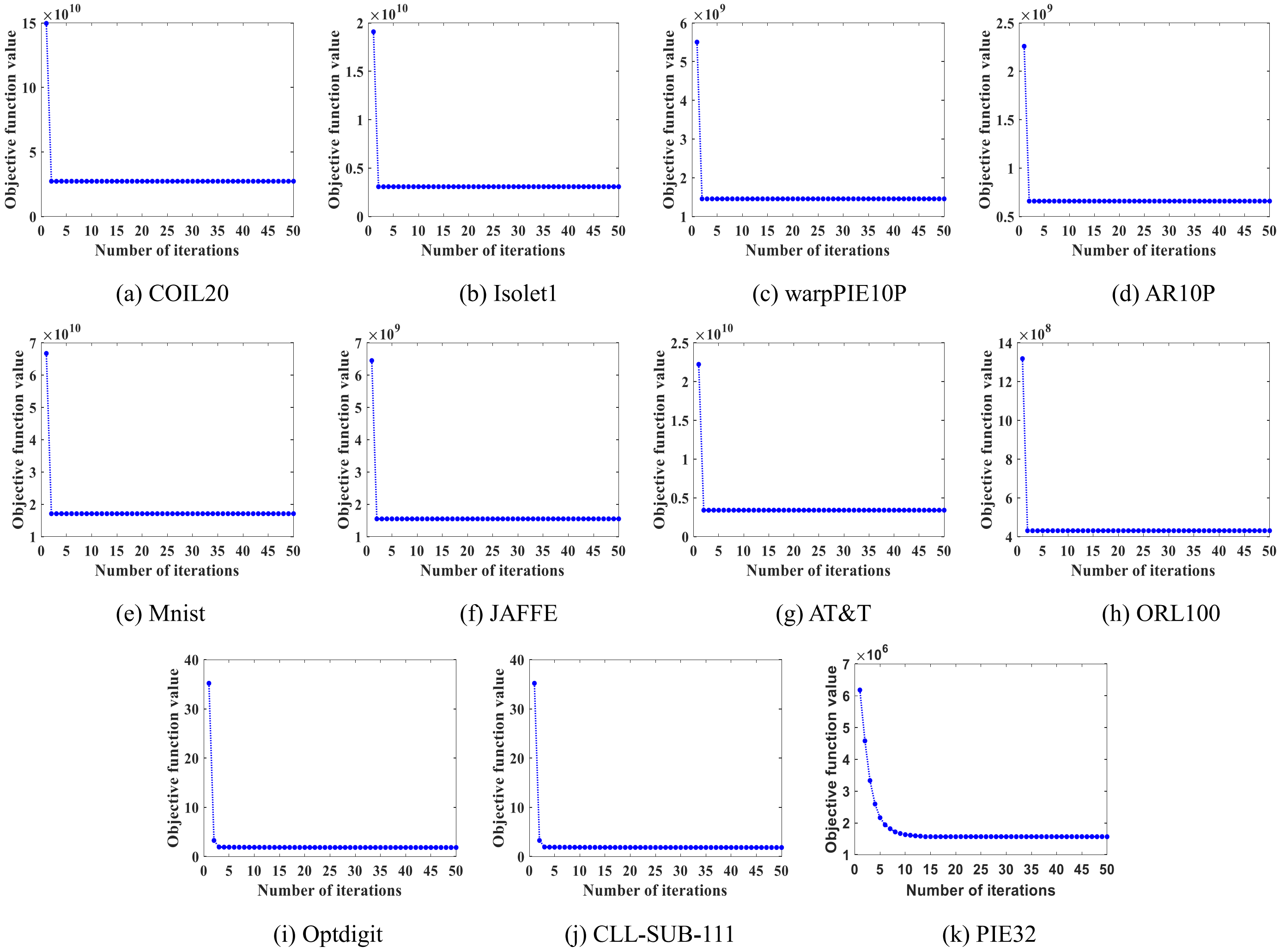
(h) ORL100

(i) Optdigit

(j) CLL-SUB-111

(k) PIE32

Fig. 5. The NMI of SLREO with differentβandγon eleven datasets.



(a) COIL20 (b) Isolet1 (c) warpPIE10P (d) AR10P

(e) Mnist (f) JAFFE (g) AT&T (h) ORL100

(i) Optdigit (j) CLL-SUB-111 (k) PIE32

Fig. 6. Convergence curve of SLREO on eleven datasets.

abscissa and ordinate respectively. The maximum number of iterations is uniformly specified as 50.

In Fig. 6, on the 11 datasets, the value of the objective function reduces rapidly and converges as the number of iterations increases. In other words, SLREO is demonstrably convergent on 10 datasets. Fig. 6 also verifies that a maximum number of iterations of 50 is appropriate to ensure sufficient convergence.

4. Conclusions

This paper proposes uncorrelated feature selection via sparse latent representation and extended OLSDA in unsupervised learning (SLREO). Since there are connections between data samples in practical applications, SLREO preserves the interconnection information between data through latent representation learning. In order to ensure the low redundancy of this interconnection information, SLREO imposes an $l_{2,1}$-norm constraint on the residual matrix latent representation learning. Considering that the real labels are non-negative, SLREO introduces the extended OLSDA to ensure the non-negativeness of the pseudo-labels. The combination of latent representation learning and extended OLSDA makes the generated pseudo-labels closer to the real class labels. In order to select low-redundancy and discriminative features, uncorrelated constraint and non-negative constraint are imposed on the feature transformation matrix. Therefore, the performance of SLREO on feature selection has been significantly improved. Extensive experiments, on multiple benchmark datasets, demonstrate that SLREO outperforms seven compared feature selection algorithms. SLREO exploits the interaction between extended OLSDA and latent representation learning, which makes the cluster labels closer to the real class labels. SLREO can show better performance in feature selection, but the final clustering labels are still continuous. Since the real class labels are all discrete, we hope to constrain the pseudo-labels and generate discrete cluster labels in the future work.

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.

Acknowledgment

We would like to express our sincere appreciation to the editors and the anonymous reviewers for their insightful comments, which have greatly helped us in improving the quality of the paper. This work was partially the National Natural Science Foundation of China under Grants Nos. 62176200, 61773304, and 61871306, the

Natural Science Basic Research Program of Shaanxi under Grant No.2022JC-45, 2022JQ-616 and the Open Research Projects of Zhejiang Lab under Grant 2021KG0AB03, the Guangdong Provincial Key Laboratory under Grant No. 2020B121201001 and the GuangDong Basic and Applied Basic Research Foundation under Grant No. 2021A1515110686.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.patcog.2022.108966.

References

[1] S. Wang, J. Chen, W. Guo, G. Liu, Structured learning for unsupervised feature selection with high-order matrix factorization, Expert Syst. Appl. 140 (2020) 112878.

[2] Z. Li, F. Nie, D. Wu, Z. Hu, X. Li, Unsupervised feature selection with weighted and projected adaptive neighbors, IEEE Trans. Cybern. (2021), doi:10.1109/TCYB.2021.3087632.

[3] S. Yi, Z. He, X. Jing, Y. Li, Yiu. Cheung, F. Nie, Adaptive weighted sparse principal component analysis for robust unsupervised feature selection, IEEE Trans. Neural Netw. Learn. Syst. 31 (6) (2020) 2153–2163.

[4] F. Nie, H. Huang, X. Cai, C. Ding, Efficient and robust feature selection via joint ℓ2, 1-norms minimization, in: Advances in Neural Information Processing Systems, 23, MIT Press, 2010, pp. 1813–1821.

[5] 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.

[6] S. Woo, C. Lee, Incremental feature extraction based on decision boundaries, Pattern Recognit. 77 (2018) 65–74.

[7] Y. Zhang, Q. Wang, D. Gong, X. Song, Nonnegative laplacian embedding guided subspace learning for unsupervised feature selection, Pattern Recognit. 93 (2019) 337–352.

[8] M. Zhao, Z. Zhang, T. Chow, B. Li, A general soft label based linear discriminant analysis for semi-supervised dimensionality reduction, Neural Netw. 55 (2014) 83–97.

[9] F. Nie, X. Dong, X. Li, Unsupervised and semisupervised projection with graph optimization, IEEE Trans. Neural Netw. Learn. Syst. 32 (4) (2020) 1547–1559.

[10] R. Zhang, F. Nie, X. Li, Self-weighted supervised discriminative feature selection, IEEE Trans. Neural Netw. Learn. Syst. 29 (8) (2018) 3913–3918.

[11] P. Mitra, C. Murthy, S. Pal, Unsupervised feature selection using feature similarity, IEEE Trans. Pattern Anal. Mach. Intell. 24 (3) (2002) 301–312.

[12] S. Solorio-Fernández, J. Martínez-Trinidad, J. Carrasco-Ochoa, A new unsupervised spectral feature selection method for mixed data: a filter approach, Pattern Recognit. 72 (2017) 314–326.

[13] C. Hou, F. Nie, X. Li, D. Yi, Y. Wu, Joint embedding learning and sparse regression: a framework for unsupervised feature selection, IEEE Trans. Cybern. 44 (6) (2014) 793–804.

[14] M.A. Hearst, S.T. Dumais, E. Osman, J. Platt, B. Scholkop, Support vector machines, IEEE Intell. Syst. Appl. 13 (4) (1998) 18–28.

[15] J. Dy, C. Brodley, Feature selection for unsupervised learning, J. Mach. Learn. Res. 5 (2004) 845–889.

[16] Z. Zhao, H. Liu, Spectral feature selection for supervised and unsupervised learning, in: Proceedings of the International Conference on Machine Learning, 2007, pp. 1151–1157.

[17] 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.

[18] R. Shang, W. Wang, R. Stolkin, L. Jiao, Non-negative spectral learning and sparse regression-based dual-graph regularized feature selection, IEEE Trans. Cybern. 48 (2) (2017) 793–806.

[19] R. Shang, W. Wang, R. Stolkin, L. Jiao, Subspace learning-based graph regularized feature selection, Knowl. Based Syst. 112 (2016) 152–165.

[20] X. Li, H. Zhang, R. Zhang, Y. Liu, F. Nie, Generalized uncorrelated regression with adaptive graph for unsupervised feature selection, IEEE Trans. Neural Netw. Learn. Syst. 30 (5) (2019) 1587–1595.

[21] C. Tang, X. Liu, M. Li, P. Wang, J. Chen, L. Wang, W. Li, Robust unsupervised feature selection via dual self-representation and manifold regularization, Knowl. Based Syst. 145 (2018) 109–120.

[22] R. Shang, K. Xu, F. Shang, L. Jiao, Sparse and low-redundant subspace learning-based dual-graph regularized robust feature selection, Knowl. Based Syst. 187 (2020) 104830.

[23] R. Shang, K. Xu, L. Jiao, Subspace learning for unsupervised feature selection via adaptive structure learning and rank approximation, Neurocomputing 413 (6) (2020) 72–84.

[24] R. Zhang, H. Zhang, X. Li, S. Yang, Unsupervised feature selection with extended OLSDA via embedding nonnegative manifold structure, IEEE Trans. Neural Netw. Learn. Syst. (2020) 3045053, doi:10.1109/TNNLS.2020.

[25] F. Nie, S. Xiang, Y. Liu, C. Hou, C. Zhang, Orthogonal vs uncorrelated least squares discriminant analysis for feature extraction, Pattern Recognit. Lett. 33 (5) (2012) 485–491.

[26] C. Tang, M. Bian, X. Liu, M. Li, H. Zhou, P. Wang, H. Yin, Unsupervised feature selection via latent representation learning and manifold regularization, Neural Netw. 117 (2019) 163–178.

[27] R. Shang, L. Wang, F. Shang, L. Jiao, Y. Li, Dual space latent representation learning for unsupervised feature selection, Pattern Recognit. 114 (2021) 107873.

[28] A. Yuan, M. You, D. He, X. Li, Convex non-negative matrix factorization with adaptive graph for unsupervised feature selection, IEEE Trans. Cybern. (2020), doi:10.1109/TCYB.2020.3034462.

[29] Q. Ye, Z. Li, L. Fu, Z. Zhang, W. Yang, G. Yang, Non-peaked discriminant analysis for data representation, IEEE Trans. Neural Netw. Learn. Syst. 30 (12) (2019) 3818–3832.

[30] Q. Ye, L. Fu, Z. Zhang, H. Zhao, M. Naiem, Lp-and Ls-norm distance based robust linear discriminant analysis, Neural Netw. 105 (2018) 393–404.

[31] Y. Li, Y. Wang, F. Albu, J. Jiang, A general zero attraction proportionate normalized maximum correntropy criterion algorithm for sparse system identification, Symmetry 9 (10) (2017) 229.

[32] F. Albu, I. Caciula, Y. Li, Y. Wang, The ℓp-norm proportionate normalized least mean square algorithm for active noise control, in: Proceedings of the 21st International Conference on System Theory, Control and Computing, 2017, pp. 396–400.

[33] X. Li, H. Zhang, R. Zhang, F. Nie, Discriminative and uncorrelated feature selection with constrained spectral analysis in unsupervised learning, IEEE Trans. Image Process. 29 (2020) 2139–2149.

[34] Z. He, S. Xie, R. Zdunek, G. Zhou, Symmetric nonnegative matrix factorization: algorithms and applications to probabilistic clustering, IEEE Trans. Neural Netw. 22 (12) (2011) 2117–2131.

[35] W. Xu, Y. Gong, Document clustering by concept factorization, in: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2004, pp. 202–209.

[36] D. Lee, H.S. Seung, Algorithms for non-negative matrix factorization, in: Advances in Neural Information Processing Systems, MIT Press, 2001, pp. 556–562.

[37] H. Lu, Z. Fu, X. Shu, Non-negative and sparse spectral clustering, Pattern Recognit. 47 (1) (2014) 418–426.

[38] H. Papadimitriou, K. Steiglitz, Combinatorial Optimization: Algorithms and Complexity, Courier Corporation, 1998.

[39] S. Ghosh, S. Dubey, Comparative analysis of k-means and fuzzy c-means algorithms, Int. J. Adv. Comput. Sci. Appl. 4 (4) (2013) 45–47.

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 machine learning, pattern recognition evolutionary computation, image processing, and data mining.

Jiarui Kong received the B.S. degree in college of computer science & engineering from Northwest Normal University, Lanzhou, China. She is currently working toward the master's degree in school of artificial intelligence from Xidian University, Xi'an, China. Her current research interests include machine learning and data mining.

Weitong Zhang (M' 21) received the B.E. degree in Electronic and Information Engineering from Changchun University of Science and Technology, Changchun, China, in 2013, the M.S. degree in Electronics and Communication Engineering, and the Ph.D. degree in Electronic science and technology from Xidian University, Xi'an, China, in 2017 and in 2021. She is currently a lecturer with Xidian University. Her current research interests include complex networks and machine learning.

Jie Feng (SM'15) received the B.S. degree from Chang'an University, Xi'an, China, in 2008, and the Ph.D. degree from Xidian University, Xi'an, China, in 2014. She is currently an Associate Professor in the Laboratory of Intelligent Perception and Image Understanding, Xidian University, Xi'an, China. Her current interests include remote sensing image processing, deep learning, and machine learning.

Licheng Jiao (SM'89) 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. From 1990 to 1991, he was a postdoctoral Fellow in the National Key Laboratory for Radar Signal Processing, Xidian University, Xi'an, China. Since 1992, Dr. Jiao has been a Professor in the School of Electronic Engineering at Xidian University. Currently, he is the Director of the Key Lab of Intelligent Perception and Image Understanding of Ministry of Education of China at Xidian University, Xi'an, China. Dr. Jiao is a Senior Member of IEEE, member of IEEE Xi'an Section Execution Committee and the Chairman of Awards and Recognition Committee, vice board chairperson of Chinese Association of Artificial Intelligence, councilor of Chinese Institute of Electronics, committee member of Chinese Committee of Neural Networks, and expert of Academic Degrees Committee of the State Council. His research interests include image processing, natural computation, machine learning, and intelligent information processing. He has charged of about 40 important scientific research projects, and published more than 20 monographs and a hundred papers in international journals and conferences.

Rustam Stolkin received the M.Eng. degree in Engineering Science from the University of Oxford, Oxford, UK in 1998, and the Ph.D. degree in computer vision from University College London, London, UK, in 2004. He is: Chair of Robotics at University of Birmingham UK; Royal Society Industry Fellow; Chair of the Expert Group on Robotic and Remote Systems for the OECD's Nuclear Energy Agency; and founded the UK National Centre for Nuclear Robotics in 2017. His research interests include computer vision and image processing, machine learning and AI, robotic grasping and manipulation, and human–robot interaction.