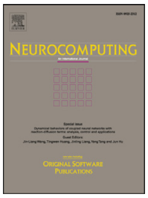
Neurocomputing 301 (2018) 36–45

Contents lists available at ScienceDirect

Neurocomputing

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



Structure preserving unsupervised feature selection

Quanmao Lu[a,b], Xuelong Li[a], Yongsheng Dong[a,c,*]

[a] *Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, Shaanxi, P.R. China*
[b] *University of Chinese Academy of Sciences, 19A Yuquanlu, Beijing 100049, P.R. China*
[c] *School of Information Engineering, Henan University of Science and Technology, Luoyang 471023, Henan, P.R. China*

ARTICLE INFO

Article history:
Received 6 December 2017
Revised 5 February 2018
Accepted 2 April 2018
Available online 5 April 2018

Communicated by Prof. Zidong Wang

Keywords:
Unsupervised feature selection
Self-expression model
Structure preserving

ABSTRACT

Spectral analysis was usually used to guide unsupervised feature selection. However, the performances of these methods are not always satisfactory due to that they may generate continuous pseudo labels to approximate the discrete real labels. In this paper, a novel unsupervised feature selection method is proposed based on self-expression model. Unlike existing spectral analysis based methods, we utilize self-expression model to capture the relationships between the features without learning the cluster labels. Specifically, each feature can be reconstructed by using a linear combination of all the features in the original feature space, and a representative feature should give a large weight to reconstruct other features. Besides, a structure preserved constraint is incorporated into our model for keeping the local manifold structure of the data. Then an efficient alternative iterative algorithm is utilized to solve our proposed model with the theoretical analysis on its convergence. The experimental results on different datasets show the effectiveness of our method.

1. Introduction

High-dimensional data are commonly used in data mining, machine learning and computer vision. To directly deal with high-dimensional data can significantly increase the time and space requirements for data processing. In practice, not all features are important and relevant, because part of features are redundant and contain noise that can degenerate the performance of algorithms. Therefore, it is important and necessary to reduce the dimensionality of the data [1–3]. As one of the effective methods to solve this problem, feature selection [4–9] has drawn much attention in recent years.

Feature selection tries to choose a subset of features from the original feature space by eliminating the irrelevant and redundant features based on certain criteria. In terms of the availability of the label information, feature selection methods can be classified into three types: supervised methods [10,11], semi-supervised methods [12,13] and unsupervised methods [14–18]. Supervised methods utilize training data that contain the intrinsic discriminative information to evaluate the importance of features. So they are able to select more representative features. Recently, sparsity-based methods have shown good performance in supervised feature selection. Semi-supervised methods are designed for solving small labeled-sample problems. Considering that the small labeled data do not provide enough information to choose discriminative features, semi-supervised methods exploit both labeled and unlabeled data to perform feature selection. However, high-dimensional data in real applications are often unlabeled, and it is very expensive and time consuming to annotate the data [19]. Therefore, unsupervised feature selection is an indispensable but challenging task for many real applications [20–24].

In the past decades, many unsupervised feature selection methods have been developed. They can be roughly divided into three categories, including filter methods [13,25], wrapper methods [2,26], and embedded methods [11,27–29]. The filter methods try to select representative features by using statistical properties of features without involving any learning algorithm. Variance and Laplacian score [25] can be regarded as two of the simplest filter methods. The wrapper methods usually “wrap” the feature selection procedures around a given learning algorithm and evaluate the performance of features based on predetermined learning algorithm. The embedded methods incorporate feature selection as a part of the model construction. Compared with the filter and wrapped methods, embedded based methods are able to capture various properties of the data, such as local manifold structure and data similarity preserving. In recent years, employing spectral analysis in unsupervised feature selection has become a common technique and shown its superiority in many tasks [30–32]. Most of these methods involve two stages in feature selection process. The first stage is to construct a similarity matrix by exploring the

[*] Corresponding author.
E-mail address: dongyongsheng98@163.com (Y. Dong).

https://doi.org/10.1016/j.neucom.2018.04.001
0925-2312/© 2018 Elsevier B.V. All rights reserved.

manifold structure of the data for learning the cluster indicator matrix. The second stage tries to embed feature selection in a sparsity regularization model, *i.e.* $l_{2,1}$-norm regularized regression, and uses the learned cluster indicator matrix to guide feature selection. Obviously, how to learn a good cluster indicator matrix is the key problem for spectral analysis based methods. However, previous work commonly utilize continuous pseudo labels to approximate the real cluster labels that are discrete in nature. As a result, it is inevitable to bring noise into the learned cluster indicator matrix that degenerates the performance of feature selection algorithms.

In this paper, we propose a novel unsupervised feature selection method based on self-expression model. In nature, self-expression model is popular used in subspace clustering problem [33,34]. Many subspace clustering methods [35,36] utilize self-expression model to explore the correlation between the samples and achieve better performance than previous subspace clustering algorithms. Besides, the extension version of self-expression model is capable of dealing with the outliers and noise in the raw data. Considering that feature selection problem aims to find the most representative feature subset, it is important to obtain the relationships between the features. Therefore, we utilize self-expression model to capture the correlation between the features and handle the noise in the data. Specifically, each feature can be reconstructed by using a linear combination of all the features in the original feature space, and a representative feature should give a large weight to reconstruct other features. Note that our proposed model performs feature selection without learning a pseudo cluster indicator matrix that avoids bringing the noise in feature selection process. Besides, considering that local manifold structure is usually better than global structure [31], a structure preserved constraint is constructed in our model for keeping the local structure of the data. Furthermore, aiming at feature selection, $l_{2,1}$-norm regularization is adopted on the feature section matrix for selecting valuable features. Then an efficient alternative iterative algorithm can be utilized to optimize our objective function with the theoretical analysis on its convergence. So our proposed model can perform local structure learning and feature selection simultaneously. Comprehensive experiments on six benchmark data sets show that our proposed method outperforms other state-of-the-art methods in different tasks.

The main contributions of this paper can be described as follows:

1. We propose a novel unsupervised feature selection method based on self-expression model. Then each feature can be reconstructed by using a linear combination of all the features in the original feature space, and a representative feature should give a large weight to reconstruct other features.
2. A structure preserved constraint is incorporated into our objective function for maintaining the local manifold structure of the data. So our method performs local structure learning and feature selection simultaneously.
3. An efficient alternative iterative algorithm is exploited to solve the proposed problem. The corresponding convergence is theoretically proved in the paper.
4. Finally, we verify the effectiveness of our proposed method on six benchmark data sets. The experimental results show that our method outperforms the stat-of-the-art methods.

The reminder of this paper is arranged as follows. The related work of unsupervised feature selection is briefly introduced in Section 2. In Section 3, we first present the proposed unsupervised feature selection model, and then employ an efficient alternative iterative algorithm to solve the objective function, followed by theoretical analysis on its convergence. Section 4 reports the performance of the proposed method on six benchmark data sets, and finally, Section 5 gives the conclusion with future work.

Notations. Throughout this paper, matrices are presented as boldface uppercase letters while vectors are written as boldface lowercase letters. Without specific explanation, for an arbitrary matrix $\mathbf{M}$, M_{ij} is the (i, j)th entry of $\mathbf{M}$, $\mathbf{m}_i$ denotes the transpose of the ith row of matrix $\mathbf{M}$, $\|\mathbf{M}\|_F$ is the Frobenius norm of $\mathbf{M}$ and $Tr(\mathbf{M})$ is the trace of $\mathbf{M}$ if $\mathbf{M}$ is square. For any $\mathbf{M} \in \mathbb{R}^{r\times t}$, its $l_{2,1}$-norm is defined as

$$\|\mathbf{M}\|_{2,1} = \sum_{i=1}^{r}\sqrt{\sum_{j=1}^{t} M_{ij}^2}. \tag{1}$$

2. Related work

Here, we briefly present the related work of unsupervised feature selection and self-expression based methods. As mentioned before, we introduce unsupervised feature selection methods in three types:filter, wrapper and embedded approaches.

The filter methods utilize a proxy measure to score a feature subset instead of the commonly used error rate. They try to distinguish the importance of all the features by using statistical properties without involving any learning algorithm. For example, Variance score assumes that larger variance means better representation ability, and chooses the feature with large variance. However, the noise in the features may have great influence on the variance of the features. Laplacain Score (LS) [25] evaluates the importance of a feature with its power of locality preserving. It first uses the heat kernel to obtain an adjacency matrix with k nearest neighbors for each data, and then constructs the corresponding graph Laplacian matrix based on adjacency matrix and degree matrix. At last, the Laplacian Score is computed for each feature. Although these methods provide a simple way to solve feature selection problem, they select features without taking the redundancy in the features into consideration that may result in selecting some redundant features and degenerating their performance. In order to solve this problem, Peng et al. [37] proposed a novel feature selection framework based on mutual information. This method employs a series of intuitive measures of redundancy and relevance to select appropriate features. Additionally, Masaeli et al. [38] converted transformation-based methods, including Hilbert–Schmidt Independence Criterion (HSIC) and Linear Discriminant Analysis (LDA), to two new feature selection algorithms by using l_0/l_∞ regularization. Note that the filter methods are usually easy to implement but might fail to select the most important feature subset for a particular task.

The wrapper approaches are often associated with a predictive model or a learning algorithm, and evaluate the performance of features based on the predetermined learning algorithm. Considering that clustering is the important problem in unsupervised learning, some learning algorithms and criteria are learned or checked for accomplishing feature selection and clustering. Dy and Brodley [39] wrapped the search for the best feature subset around a clustering algorithm and employed two performance criteria, *i.e.* scatter separability and maximum likelihood, to evaluate candidate feature subsets. Wolf and Shashua [40] performed feature selection based on least-squares optimization process and employed spectral properties of the candidate feature subsets to guide the search. Besides, a novel wrapper algorithm [41] was proposed based on Support Vector Machines (SVM) with kernel function. The basic idea of this method is to measure the importance of features by using the number of errors in a validation subset. Compared with the filter method, the wrapper method is able to achieve a better performance. However, the wrapper method has expensive time cost in feature selection process.

The embedded method incorporates feature selection into a part of model construction that is superior to others in many aspects and has drawn much attention in recent years. Inspired by the development on spectral analysis and l_1-regularized model for subsect selection, Multi-Cluster Feature Selection (MCFS) [27] was proposed for unsupervised feature selection that aims to preserve the multi-cluster structure of the raw data. However, MCFS fails to exploit the discriminative information in feature selection process. Unsupervised Discriminative Feature Selection (UDFS) [42] combines discriminative analysis with $l_{2,1}$-norm regularization into a joint framework for selecting the most discriminative feature subset for data representation. However, its orthogonal constraint on the feature selection matrix is unreasonable since feature weight vectors are not necessarily orthogonal with each other in nature. Nonnegative Discriminative Feature Selection (NDFS) [30] employs nonnegative spectral analysis to reflect the discriminative information of the data, and then obtains the corresponding cluster labels to guide feature selection. Both UDFS and NDFS ignore the outliers or noise in the raw data. In order to handle the outliers or noise in the data, Robust Unsupervised Feature Selection (RUFS) [31] utilizes nonnegative matrix factorization and $l_{2,1}$-norm minimization to perform cluster indicator matrix learning and feature learning simultaneously. Besides, Embedded Unsupervised Feature Selection (EUFS) [43] directly embeds feature selection into a clustering algorithm without transferring unsupervised feature selection problem into a sparse learning based supervised feature selection with the learned pseudo labels.

Another related area is self-expression based methods that are commonly designed for solving subspace clustering problem. Specifically, self-expression based methods regard the raw data as the dictionary to learn the representation matrix. The general model of self-expression based methods can be formulated as

$$\begin{aligned}&\min_{\mathbf{Z},\mathbf{E}} R(\mathbf{Z})+L(\mathbf{E})\\ &s.t.\ \mathbf{X}=\mathbf{XZ}+\mathbf{E},\end{aligned} \tag{2}$$

where $\mathbf{Z}$ denotes the representation matrix, $\mathbf{E}$ is the noise term, $R(\mathbf{Z})$ is the regularization term on $\mathbf{Z}$, and $L(\mathbf{E})$ represents the loss function of the noise term. Sparse Subspace Clustering (SSC) [44], as a first proposed subspace clustering method based on self-expression model, was designed for capturing the sparsest representation for each point. Although SSC can improve the performance of subspace clustering, its representation matrix may be too sparse to capture the relationships between the data points in the same subspace. Then Low Rank Representation (LRR) [45] was proposed to find the lowest rank representation of the samples jointly. Based on SCC and LRR, many subspace clustering methods have been developed by using self-expression model [46,47]. Considering that self-expression model is capable of exploring the correlation between the samples, we employ self-expression model to perform feature selection task in the paper.

3. Our method

Previous work usually learn a cluster indicator matrix to guide feature selection process. However, these methods commonly utilize continuous pseudo labels to approximate the real cluster labels that are discrete in nature. As a result, it is inevitable to introduce noise into the learned cluster labels and degenerate the performance of feature selection. In this paper, a novel unsupervised feature selection method is proposed without learning the cluster labels. Specifically, self-expression model is utilized to analyse the relationships between the features and handle the noise in the raw data. Besides, a structure preserved constraint is incorporated into our objective function for keeping the local manifold structure of the data. Furthermore, an efficient alternative iterative algorithm is designed for solving the above problem, and the theoretical analysis on its convergence is then given.

3.1. Problem formulation

Recall that unsupervised feature selection aims to find the representative features from all the features in the raw data. We employ self-expression model to capture the relationships between the features and find the representative features. Based on self-expression model, each feature can be reconstructed by using a linear combination of all the features in the original feature space, and a representative feature should give a large weight to represent other features. Let $\mathbf{X}\in\mathbb{R}^{n\times d}$ denote the raw data, in which n is the number of points and d is the number of features. $\mathbf{x}_i\in\mathbb{R}^{d\times 1}$ represents the transpose of the ith row of $\mathbf{X}$. For each feature vector $\mathbf{f}_i$, the corresponding linear representation optimization problem can be written as

$$\min \sum_{j=1}^{d}\left|w_{ji}\right|_p,\quad s.t.\ \ \mathbf{f}_i=\sum_{j=1}^{d} w_{ji}\mathbf{f}_j, \tag{3}$$

where w_{ji} denotes the representation coefficient and $|\bullet|_p$ represents the p-norm. The corresponding matrix form can be formulated as

$$\min_{\mathbf{W}} \|\mathbf{W}\|_p,\quad s.t.\ \ \mathbf{X}=\mathbf{XW}, \tag{4}$$

where $\mathbf{W}\in\mathbb{R}^{d\times d}$ is the representation matrix. Considering that the data from real applications are always contaminated by the noise, the Frobenius norm is utilized to deal with the noise in the data that is a popular way in many tasks. Then the problem (4) can be rewritten as

$$\min_{\mathbf{W},\mathbf{E}} \|\mathbf{E}\|_F^2+\alpha\|\mathbf{W}\|_p,\quad s.t.\ \ \mathbf{X}=\mathbf{XW}+\mathbf{E}, \tag{5}$$

which is equivalent to

$$\min_{\mathbf{W}} \|\mathbf{X}-\mathbf{XW}\|_F^2+\alpha\|\mathbf{W}\|_p, \tag{6}$$

where the first term represents the residual term and α is a weight factor to balance the two terms. Let $\mathbf{w}_i\in\mathbb{R}^{d\times 1}$ denote the transpose of the i-th row of $\mathbf{W}$. Note that w_{ij} is the representation coefficient of the ith feature to reconstruct the jth feature. Therefore, for the ith feature, the larger value of $\|\mathbf{w}_i\|_2$ means better representation ability. Then $l_{2,1}$ minimization regularization is reasonably utilized to constrain the representation matrix $\mathbf{W}$ for guaranteeing that $\mathbf{W}$ is sparse in rows. The corresponding feature selection problem becomes

$$\min_{\mathbf{W}} \|\mathbf{X}-\mathbf{XW}\|_F^2+\alpha\|\mathbf{W}\|_{2,1}. \tag{7}$$

Furthermore, a structure preserved constraint is added into our objective function for keeping the local manifold structure of the data. Specifically speaking, the studies in manifold learning theory and spectral graph theory have shown that the local manifold structure can be effectively modeled by constructing a nearest neighbor graph based on the data points [31,48]. Then we employ the popular Gaussian kernel function to calculate the corresponding weight matrix $\mathbf{S}$. Given two points $\mathbf{x}_i$ and $\mathbf{x}_j$, the weight between them can be calculated by

$$S_{ij}=e^{-\frac{\|\mathbf{x}_i-\mathbf{x}_j\|^2}{\sigma}}. \tag{8}$$

In order to preserve the local manifold structure, the reconstructed data points should keep their nearest neighbor relationships in the original data space. It means that if two points $\mathbf{x}_i$ and $\mathbf{x}_j$ are similar in the original data space, their reconstructed samples $\mathbf{W}^T\mathbf{x}_i$ and $\mathbf{W}^T\mathbf{x}_j$ should have small distance. Then the local structure preserving problem can be formulated as

$$\min_{\mathbf{W}} \frac{1}{2}\sum_{i,j}\left\|\mathbf{W}^T\mathbf{x}_i-\mathbf{W}^T\mathbf{x}_j\right\|_2^2 S_{ij}. \tag{9}$$

It can be seen that the value of $\left\|\mathbf{W}^T\mathbf{x}_i-\mathbf{W}^T\mathbf{x}_j\right\|_2^2$ is likely to be small when S_{ij} is large. Therefore, by minimizing the function in Eq. (9), the reconstructed data samples are able to keep the nearest neighbor relationships in the original data space. We apply Eq. (9) to Eq. (7), and finally we get

$$\min_{\mathbf{W}} \|\mathbf{X}-\mathbf{XW}\|_F^2+\alpha\|\mathbf{W}\|_{2,1}+\frac{\beta}{2}\sum_{i,j}\left\|\mathbf{W}^T\mathbf{x}_i-\mathbf{W}^T\mathbf{x}_j\right\|_2^2 S_{ij}, \tag{10}$$

where $\beta > 0$ is a weight factor. Note that our method can perform local structure learning and feature selection simultaneously.

By solving the problem (10), we can learn the representation matrix $\mathbf{W}$ which can be regarded as the feature selection matrix.

3.2. Optimization

Considering that the problem (10) contains $l_{2,\,1}$ regularization, it is hard to find its optimal solution directly. Therefore, we utilize an alternative iterative algorithm to solve this problem.

It is easy to verify that

$$\begin{aligned}
&\frac{1}{2}\sum_{i,j}\left\|\mathbf{W}^T\mathbf{x}_i-\mathbf{W}^T\mathbf{x}_j\right\|_2^2 S_{ij}\\
&=\sum_{i=1}^{n}\left(\mathbf{W}^T\mathbf{x}_i\right)^T\mathbf{W}^T\mathbf{x}_i D_{ii}-\sum_{i,j=1}^{n}\left(\mathbf{W}^T\mathbf{x}_i\right)^T\mathbf{W}^T\mathbf{x}_j S_{ij}\\
&=Tr(\mathbf{W}^T\mathbf{X}^T\mathbf{DXW})-Tr(\mathbf{W}^T\mathbf{X}^T\mathbf{SXW})\\
&=Tr(\mathbf{W}^T\mathbf{X}^T\mathbf{L_S}\mathbf{XW}),
\end{aligned} \tag{11}$$

where $\mathbf{D}$ is a diagonal matrix whose each entry is row (or column, because $\mathbf{S}$ is symmetric matrix) sum of $\mathbf{S}$, $D_{ii}=\sum_j S_{ij}$. $\mathbf{L_S}$ is the Laplacian matrix [49,50] which is calculated by formula $\mathbf{L_S}=\mathbf{D}-\mathbf{S}$. Based on Eq. (11), and replace $\|\mathbf{W}\|_{2,\,1}$ with $\sum_i\|\mathbf{w}_i\|_2$, the problem (10) can be reformulated as

$$\min_{\mathbf{W}} \|\mathbf{X}-\mathbf{XW}\|_F^2+\alpha\sum_i\|\mathbf{w}_i\|_2+\beta Tr(\mathbf{W}^T\mathbf{X}^T\mathbf{L_S}\mathbf{XW}). \tag{12}$$

Since the problem (10) is designed for obtaining a matrix $\mathbf{W}$ that is sparse in rows, $\|\mathbf{w}_i\|_2$ is likely to be zero in theory that can make Eq. (12) non-differentiable. In order to avoid this situation, we rewrite $\|\mathbf{w}_i\|_2$ as $\sqrt{\mathbf{w}_i^T\mathbf{w}_i}$, and regularize $\sqrt{\mathbf{w}_i^T\mathbf{w}_i}$ as $\sqrt{\mathbf{w}_i^T\mathbf{w}_i+\varepsilon}$, where ε is a small enough constant. Then we have

$$\min_{\mathbf{W}} \|\mathbf{X}-\mathbf{XW}\|_F^2+\alpha\sum_i\sqrt{\mathbf{w}_i^T\mathbf{w}_i+\varepsilon}+\beta Tr(\mathbf{W}^T\mathbf{X}^T\mathbf{L_S}\mathbf{XW}). \tag{13}$$

Note that when ε is infinitely close to zero, Eq. (13) is equal to the problem (10). Let

$$\mathcal{J}=\|\mathbf{X}-\mathbf{XW}\|_F^2+\alpha\sum_i\sqrt{\mathbf{w}_i^T\mathbf{w}_i+\varepsilon}+\beta Tr(\mathbf{W}^T\mathbf{X}^T\mathbf{L_S}\mathbf{XW}). \tag{14}$$

Setting the derivative of $\mathcal{J}$ respect to $\mathbf{W}$ to be zero, we have

$$\begin{aligned}
\frac{\partial\mathcal{J}}{\partial\mathbf{W}}&=-2(\mathbf{X}^T\mathbf{X}-\mathbf{X}^T\mathbf{XW})+2\alpha\mathbf{QW}+2\beta\mathbf{X}^T\mathbf{L_S}\mathbf{XW}\\
&=2(\beta\mathbf{X}^T\mathbf{L_S}\mathbf{XW}+\mathbf{X}^T\mathbf{XW}+\alpha\mathbf{QW}-\mathbf{X}^T\mathbf{X})=0,
\end{aligned} \tag{15}$$

where $\mathbf{Q}\in\mathbb{R}^{d\times d}$ is a diagonal matrix, and Q_{ii} is denoted as

$$Q_{ii}=\frac{1}{2\sqrt{\mathbf{w}_i^T\mathbf{w}_i+\varepsilon}}. \tag{16}$$

Considering that $\mathbf{Q}$ depends on $\mathbf{W}$, it is unable to obtain $\mathbf{W}$ directly. Thus we utilize an alternative iterative algorithm to find the optimal solution. When $\mathbf{W}$ is fixed, we can obtain $\mathbf{Q}$ by Eq. (16). When $\mathbf{Q}$ is fixed, $\mathbf{W}$ can be easily obtained by solving Eq. (15) and shown as follows:

$$\mathbf{W}=(\beta\mathbf{X}^T\mathbf{L_S}\mathbf{X}+\mathbf{X}^T\mathbf{X}+\alpha\mathbf{Q})^{-1}\mathbf{X}^T\mathbf{X}. \tag{17}$$

After obtaining the final solution $\mathbf{W}$, we can sort all features according to $\|\mathbf{w}_i\|_2$ in descending order and select the top h ranked features. The whole procedures of our optimization algorithm is given in Algorithm 1.

Algorithm 1 Alternative iterative algorithm to solve problem (10).

```
Input: Data matrix $\mathbf{X}\in\mathbb{R}^{n\times d}$, weight matrix $\mathbf{S}$, parameters $\alpha$ and $\beta$, number of selected features $h$.
Output: $h$ selected features
Initialize $\mathbf{Q}=\mathbf{I}, \mathbf{Q}\in\mathbb{R}^{d\times d}$
Calculate $\mathbf{L_S}=\mathbf{D}-\mathbf{S}$
repeat
    Update $\mathbf{W}=(\beta\mathbf{X}^T\mathbf{L_S}\mathbf{X}+\mathbf{X}^T\mathbf{X}+\alpha\mathbf{Q})^{-1}\mathbf{X}^T\mathbf{X}$
    Calculate the diagonal matrix $\mathbf{Q}$ by
        $Q_{ii}=\frac{1}{2\sqrt{\mathbf{w}_i^T\mathbf{w}_i+\varepsilon}}$
until convergence
sort all features according to $\|\mathbf{w}_i\|_2$ in descending order and select the top $h$ ranked features.
```

3.3. Convergence analysis

In this subsection, we prove the convergence of our optimization algorithm in Algorithm 1. We begin with the following lemma that is easily to prove.

Lemma 1. *The following inequality holds for any positive real number u and v.*

$$\sqrt{u}-\frac{u}{2\sqrt{v}}\le\sqrt{v}-\frac{v}{2\sqrt{v}}. \tag{18}$$

Then we have the following theorem.

Theorem 1. *The objective function in problem (13) is decreasing by following the procedures in Algorithm 1.*

Proof. *Note that solving Eq. (15) is equivalent to solve the problem*

$$\min_{\mathbf{W}} \|\mathbf{X}-\mathbf{XW}\|_F^2+\alpha Tr(\mathbf{W}^T\mathbf{QW})+\beta Tr(\mathbf{W}^T\mathbf{X}^T\mathbf{L_S}\mathbf{XW}). \tag{19}$$

Given a update $\widehat{\mathbf{W}}$ of $\mathbf{W}$, we have

$$\begin{aligned}
&\left\|\mathbf{X}-\mathbf{X}\widehat{\mathbf{W}}\right\|_F^2+\alpha Tr(\widehat{\mathbf{W}}^T\mathbf{Q}\widehat{\mathbf{W}})+\beta Tr(\widehat{\mathbf{W}}^T\mathbf{X}^T\mathbf{L_S}\mathbf{X}\widehat{\mathbf{W}})\\
&\le\|\mathbf{X}-\mathbf{XW}\|_F^2+\alpha Tr(\mathbf{W}^T\mathbf{QW})+\beta Tr(\mathbf{W}^T\mathbf{X}^T\mathbf{L_S}\mathbf{XW}).
\end{aligned} \tag{20}$$

We add $\alpha\sum_i\frac{\varepsilon}{2\sqrt{\mathbf{w}_i^T\mathbf{w}_i+\varepsilon}}$ into both sides of the inequality (20), and substitute $\mathbf{Q}$ in Eq. (16), then inequality (20) can be formulated as

$$\begin{aligned}
&\left\|\mathbf{X}-\mathbf{X}\widehat{\mathbf{W}}\right\|_F^2+\alpha\sum_i\frac{\widehat{\mathbf{w}}_i^T\widehat{\mathbf{w}}_i+\varepsilon}{2\sqrt{\mathbf{w}_i^T\mathbf{w}_i+\varepsilon}}+\beta Tr(\widehat{\mathbf{W}}^T\mathbf{X}^T\mathbf{L_S}\mathbf{X}\widehat{\mathbf{W}})\\
&\le\|\mathbf{X}-\mathbf{XW}\|_F^2+\alpha\sum_i\frac{\mathbf{w}_i^T\mathbf{w}_i+\varepsilon}{2\sqrt{\mathbf{w}_i^T\mathbf{w}_i+\varepsilon}}+\beta Tr(\mathbf{W}^T\mathbf{X}^T\mathbf{L_S}\mathbf{XW}).
\end{aligned} \tag{21}$$

Based on Lemma 1, we have

$$\begin{aligned}
&\alpha\sum_i\sqrt{\widehat{\mathbf{w}}_i^T\widehat{\mathbf{w}}_i+\varepsilon}-\alpha\sum_i\frac{\widehat{\mathbf{w}}_i^T\widehat{\mathbf{w}}_i+\varepsilon}{2\sqrt{\mathbf{w}_i^T\mathbf{w}_i+\varepsilon}}\\
&\le\alpha\sum_i\sqrt{\mathbf{w}_i^T\mathbf{w}_i+\varepsilon}-\alpha\sum_i\frac{\mathbf{w}_i^T\mathbf{w}_i+\varepsilon}{2\sqrt{\mathbf{w}_i^T\mathbf{w}_i+\varepsilon}}.
\end{aligned} \tag{22}$$

Sum over the inequality (21) and inequality (22), we have

Table 1
Statistics of six data sets.

Dataset	# of samples	# of features	# of classes
COIL20	1440	1024	20
JAFFE	213	676	10
Yale	165	1024	15
MNIST	3000	784	10
lung	203	3312	5
TOX_171	171	5748	4

$$\begin{aligned} &\|\mathbf{X}-\mathbf{X}\widehat{\mathbf{W}}\|_F^2+\alpha\sum_i\sqrt{\widehat{\mathbf{w}}_i^T\widehat{\mathbf{w}}_i+\varepsilon}+\beta Tr(\widehat{\mathbf{W}}^T\mathbf{X}^T\mathbf{L_S}\mathbf{X}\widehat{\mathbf{W}}) \\ &\le\|\mathbf{X}-\mathbf{X}\mathbf{W}\|_F^2+\alpha\sum_i\sqrt{\mathbf{w}_i^T\mathbf{w}_i+\varepsilon}+\beta Tr(\mathbf{W}^T\mathbf{X}^T\mathbf{L_S}\mathbf{X}\mathbf{W}). \end{aligned} \tag{23}$$

□

4. Experiments

In this section, we verify the effectiveness of our proposed method on six benchmark data sets, regarding to image data and biological data. Following previous work on unsupervised feature selection, clustering accuracy (AC) [51] and normalized mutual information (NMI) [52] are employed to evaluate the performance of our method for clustering.

4.1. Datasets

The evaluation is performed on six publicly available benchmark datasets, including four image datasets (COIL20, JAFFE, Yale and MNIST) and two biological datasets (lung and TOX_171). Table 1 gives a summary of these datasets.

4.2. Experiment setup

Following the common way to measure the performance of unsupervised feature selection methods, we employ two widely used evaluation metrics, namely AC and NMI, to evaluate different methods in terms of clustering performance.

Suppose the predicted label vector of the raw data $\mathbf{X}$ is $\mathbf{p}$ which is obtained by clustering method, and $\mathbf{g}$ denotes the corresponding ground truth label vector. Let $\hat{\mathbf{p}}=map(\mathbf{p})$, where $map(\mathbf{p})$ is the best permutation mapping function. Then AC can be defined as

$$AC=\frac{\sum_{i=1}^n\delta(g_i,\hat{p}_i)}{n}, \tag{24}$$

where

$$\delta(x,y)=\begin{cases}1, & if\ x=y\\ 0, & else\end{cases}, \tag{25}$$

n is the number of data points. It can be seen that the more predicted label vector matches the ground truth, the more accuracy clustering result can be obtained.

Suppose the clusters in the predicted label vector $\mathbf{p}$ and the ground truth label vector $\mathbf{g}$ are C_p and C_g respectively. Let n_p be the number of points in cluster C_p, n_g be the number of points in cluster C_g and $n_{p,\,g}$ denote the number of points that are in cluster C_p as well as in cluster C_g. Then the Mutual Information (MI) is defined as

$$MI(\mathbf{p},\mathbf{g})=\sum_{p,g=1}^k\frac{n_{p,g}}{n}\log\left(\frac{\frac{n_{p,g}}{n}}{\frac{n_p}{n}\cdot\frac{n_g}{n}}\right). \tag{26}$$

In order to normalize MI, the normalized mutual information (NMI) is defined as follows:

$$NMI(\mathbf{p},\mathbf{g})=\frac{2\cdot MI(\mathbf{p},\mathbf{g})}{H(\mathbf{p})+H(\mathbf{g})}, \tag{27}$$

where $H(\bullet)$ is the entropy function. Based on (26) and (27), we have

$$NMI(\mathbf{p},\mathbf{g})=\frac{-2\sum_{p,g=1}^k n_{p,g}\log\left(\frac{n_{p,g}\cdot n}{n_p\cdot n_g}\right)}{\sum_{p=1}^k n_p\log(\frac{n_p}{n})+\sum_{g=1}^k n_g\log(\frac{n_g}{n})}. \tag{28}$$

To validate the effectiveness of our proposed method in feature selection, we compare it with one baseline and six representative unsupervised feature selection methods that are presented as follows:

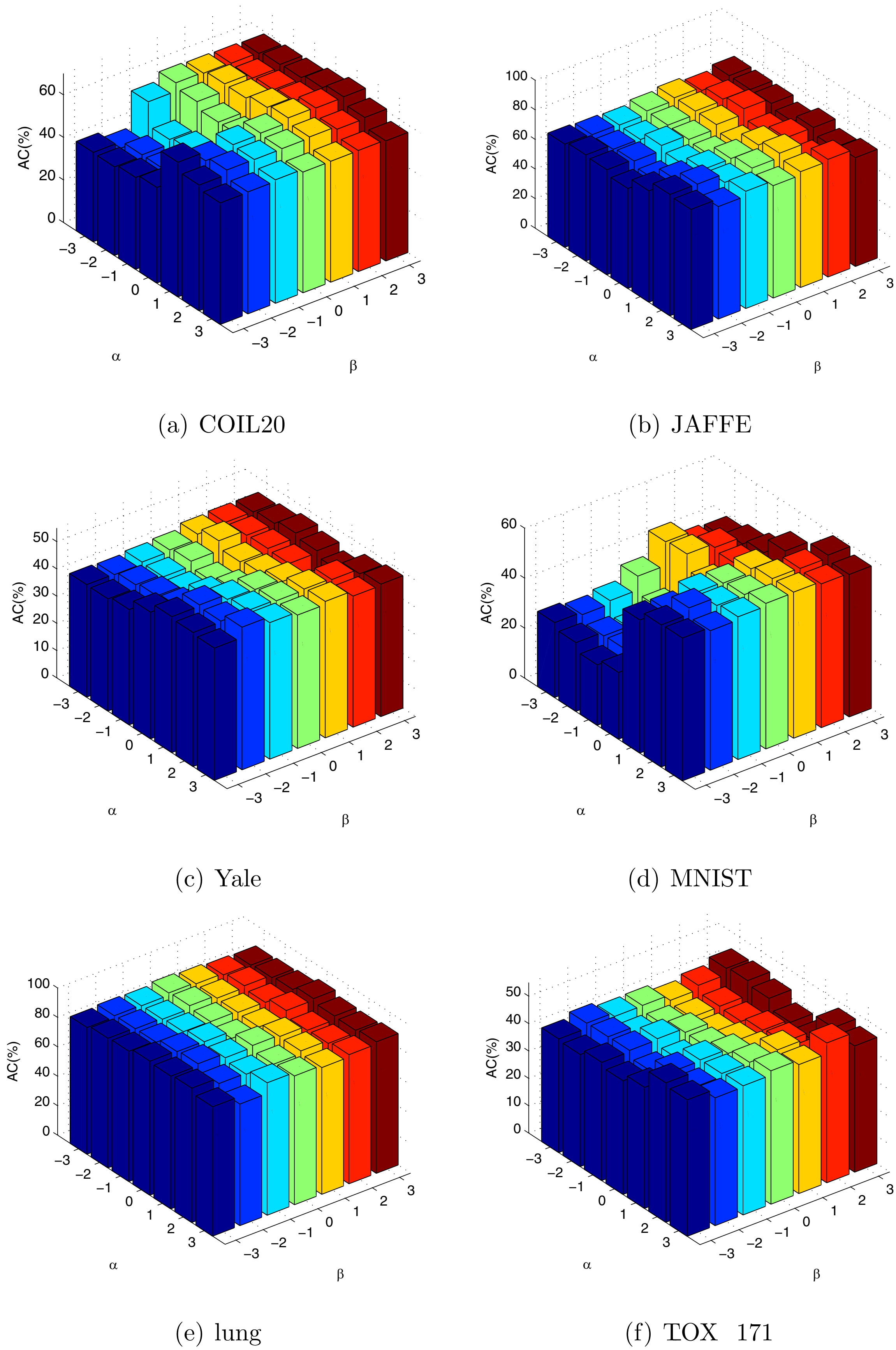
1. All-Fea: All original features are adopted. It is used as the baseline.
2. Laplacian Score (LS) [25]: It evaluates the importance of each feature based on its locality preserving power.
3. Multi-Cluster Feature Selection (MCFS) [27]: This method tries to select the features that can preserve the multi-cluster structure of the raw data.
4. Unsupervised Discriminative Feature Selection (UDFS) [42]: UDFS combines discriminative analysis with $l_{2,1}$ minimization regularization into a joint model to formulate unsupervised feature selection problem.
5. Nonnegative Discriminative Feature Selection (NDFS) [30]: In order to exploit discriminative information in the data, NDFS performs the cluster label learning and feature selection simultaneously. Besides, a nonnegative constraint is utilized for obtaining a more accurate cluster label.
6. Robust Unsupervised Feature Selection (RUFS) [31]: It utilizes nonnegative matrix factorization and $l_{2,\,1}$-norm minimization to perform label learning and feature learning simultaneously.
7. Embedded Unsupervised Feature Selection (EUFS) [43]: Instead of transforming unsupervised feature selection into sparse learning problem with the learned cluster label, EUFS directly embeds feature selection into a clustering method via sparse learning.

In order to fairly evaluate the performance of different unsupervised methods, we choose the parameters for them following the previous work. Specifically, for LS, MCFS, UDFS, NDFS, RUFS, EUFS and our method, we set the size of nearest neighbors to be 5 for all the datasets. For the weight factors of different methods, we utilize a "grid-search" strategy from $\{10^{-3},10^{-2},\ \ldots\ ,10^2,10^3\}$ to find the best parameters. The number of selected features are set as {50, 100, 150, 200, 250, 300} for all the datasets. To evaluate different methods, K-means algorithm is employed to cluster the data points based on the selected features. Considering that K-means algorithm is sensitive to initialization, following previous work, we repeat K-means algorithm for 20 times and report the average clustering results with standard deviation.

4.3. Results and analysis

The experimental results of different unsupervised feature selection methods on the whole datasets are presented in Tables 2 and 3. From the experimental results, we have the following observations:

1. Unsupervised feature selection methods achieve better performance than the baseline method (All-Fea) for most situations. Therefore, feature selection is necessary and effective to deal with the noise or redundant information in the raw data. It can not only make the subsequent processing algorithms more efficient, but also improve the performance.
2. The discriminative information is useful for feature selection. More specifically, MCFS, UDFS, NDFS, RUFS, and our method take discriminative information into account, which commonly achieve more accurate clustering results than All-Fea and LS.



(a) COIL20

(b) JAFFE

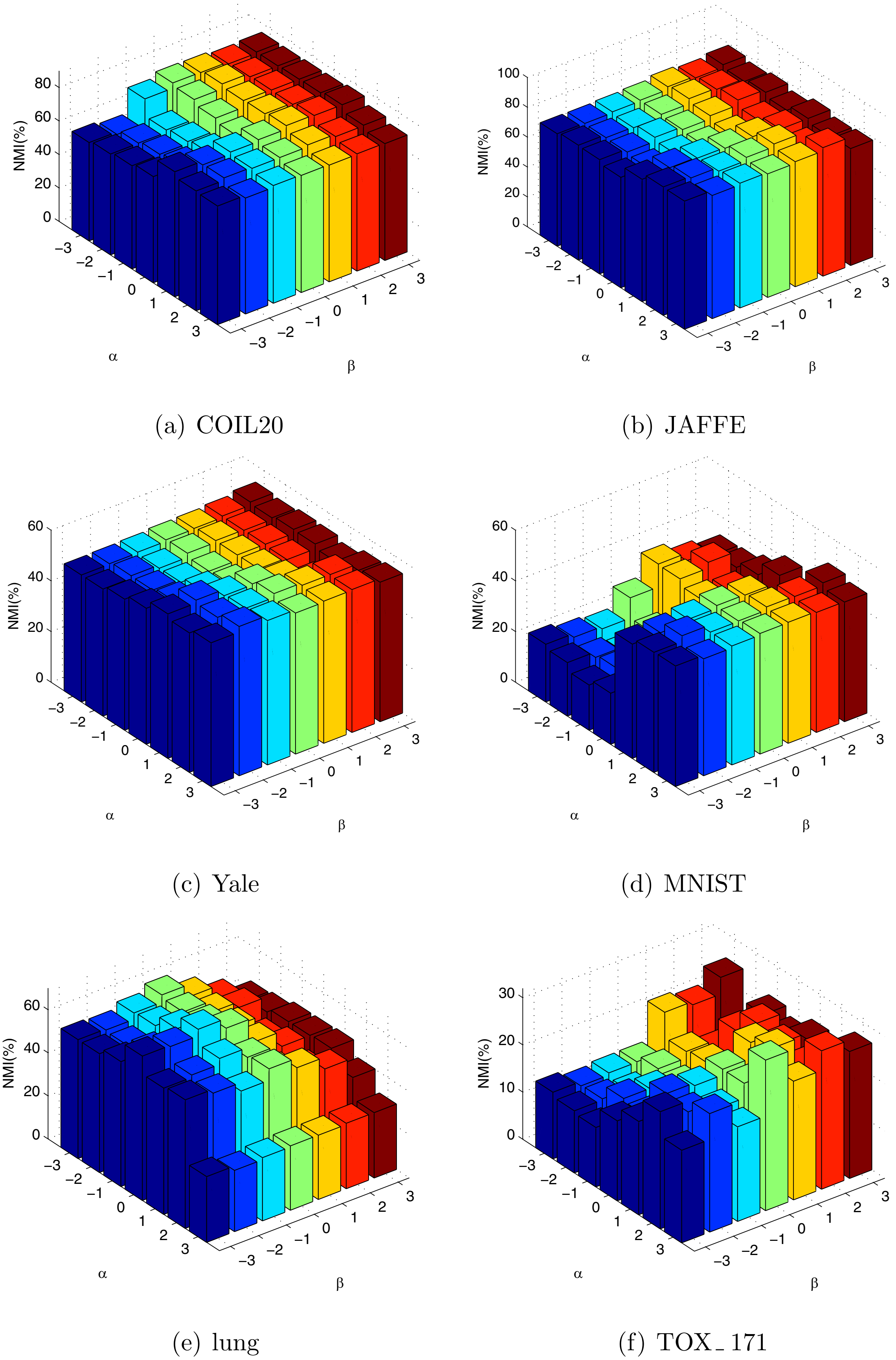
(c) Yale

(d) MNIST

(e) lung

(f) TOX 171

Fig. 1. AC versus the parameters α and β of the proposed method on different datasets.



(a) COIL20

(b) JAFFE

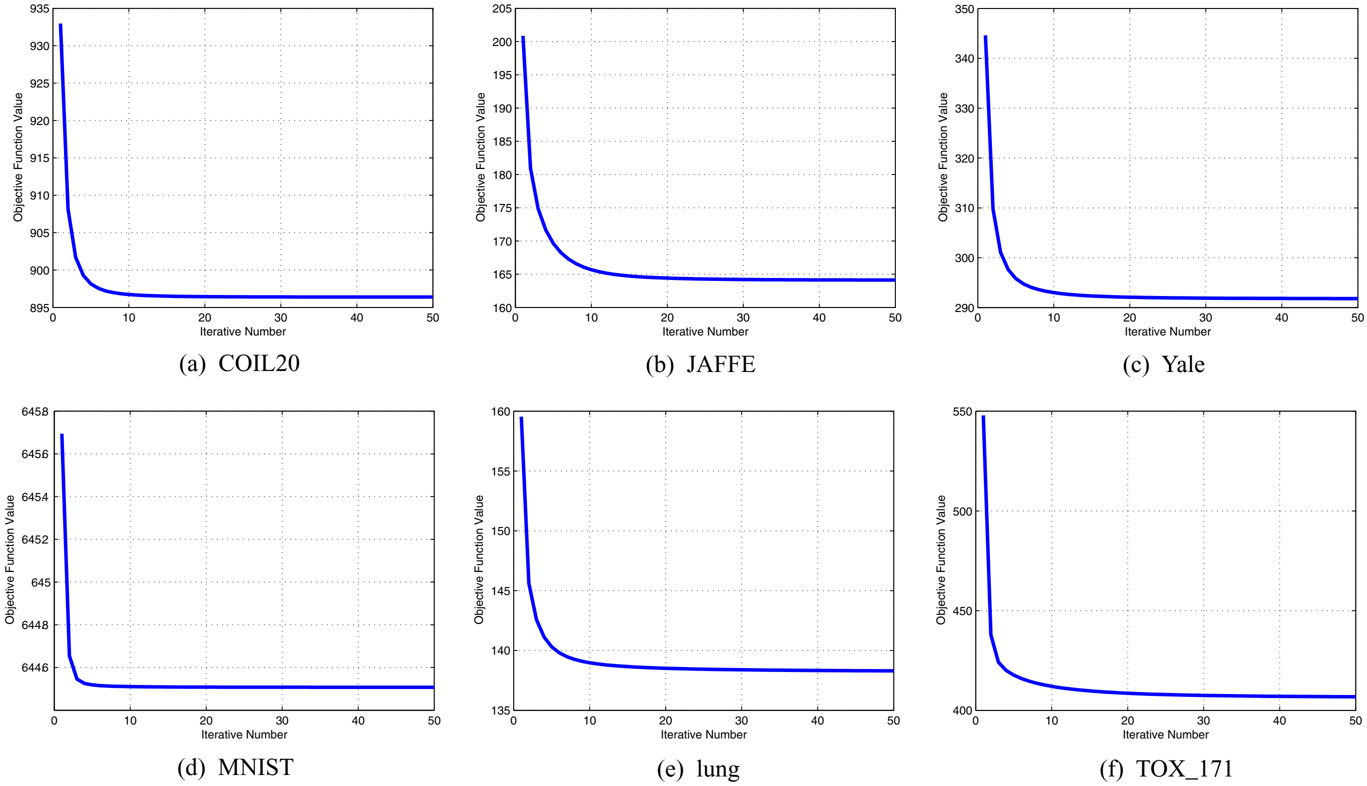
(c) Yale

(d) MNIST

(e) lung

(f) TOX_171

Fig. 2. NMI versus the parameters α and β of the proposed method on different datasets.



(a) COIL20

(b) JAFFE

(c) Yale

(d) MNIST

(e) lung

(f) TOX_171

Fig. 3. Convergence curves of our optimization algorithm on different datasets.

Table 2
Clustering results (AC% ± std) of different feature selection algorithms on six benchmark datasets. The best results are highlighted in bold.

Dataset	COIL20	JAFFE	Yale
All-Fea	54.33 ± 3.67	75.12 ± 5.01	47.70 ± 3.90
LS	61.00 ± 2.38	80.08 ± 4.51	48.18 ± 2.81
MCFS	61.70 ± 4.05	80.64 ± 6.73	48.52 ± 2.23
UDFS	61.65 ± 3.74	80.56 ± 5.33	50.55 ± 2.73
NDFS	61.60 ± 3.62	81.08 ± 6.00	46.55 ± 2.68
RUFS	62.30 ± 4.03	81.03 ± 5.78	48.00 ± 3.28
EUFS	62.96 ± 3.83	80.32 ± 5.64	48.97 ± 2.63
Ours	**64.27** ± **2.32**	**82.02** ± **4.27**	**52.36** ± **3.53**
Dataset	MNIST	lung	TOX_171
All-Fea	40.78 ± 2.18	90.44 ± 3.74	46.26 ± 3.71
LS	41.86 ± 2.30	86.35 ± 2.65	46.18 ± 1.52
MCFS	56.60 ± 2.86	91.18 ± 2.17	47.39 ± 6.58
UDFS	45.41 ± 3.37	90.39 ± 2.34	36.67 ± 6.72
NDFS	43.31 ± 3.13	89.56 ± 2.69	43.10 ± 6.10
RUFS	54.26 ± 3.22	91.13 ± 0.00	48.89 ± 5.89
EUFS	57.42 ± 1.81	91.13 ± 0.51	51.99 ± 2.21
Ours	**59.40** ± **3.53**	**91.43** ± **0.25**	**52.87** ± **2.85**

Table 3
Clustering results (NMI% ± std) of different feature selection algorithms on six benchmark datasets. The best results are highlighted in bold.

Dataset	COIL20	JAFFE	Yale
All-Fea	71.33 ± 2.01	81.60 ± 2.80	55.22 ± 3.61
LS	73.43 ± 2.01	82.58 ± 2.70	55.41 ± 2.29
MCFS	76.07 ± 2.11	82.87 ± 3.86	57.01 ± 2.16
UDFS	71.78 ± 2.44	84.94 ± 3.72	57.74 ± 2.13
NDFS	74.96 ± 1.86	85.75 ± 2.80	53.72 ± 2.61
RUFS	76.53 ± 1.54	85.65 ± 3.85	57.39 ± 3.44
EUFS	72.03 ± 1.93	85.88 ± 4.08	56.78 ± 2.41
Ours	**76.73** ± **1.47**	**86.30** ± **4.60**	**58.87** ± **1.78**
Dataset	MNIST	lung	TOX_171
All-Fea	34.85 ± 2.87	52.15 ± 5.05	24.79 ± 5.67
LS	34.94 ± 1.79	53.44 ± 5.28	24.91 ± 0.82
MCFS	45.79 ± 1.62	68.30 ± 6.61	25.26 ± 6.25
UDFS	37.10 ± 2.24	58.89 ± 3.28	11.43 ± 9.25
NDFS	36.72 ± 2.89	56.52 ± 4.81	19.04 ± 9.95
RUFS	45.83 ± 2.48	64.32 ± 2.90	24.56 ± 6.19
EUFS	47.51 ± 1.06	67.51 ± 5.77	28.45 ± 3.02
Ours	**49.64** ± **2.71**	**69.50** ± **5.67**	**31.88** ± **4.96**

3. EUFS obtains good performance on the whole datasets, and achieves the second best performance based on both AC and NMI metrics for most cases. It can attribute to its embedding model. Besides, UDFS shows a very poor performance on the TOX_171 dataset in terms of both AC and NMI.
4. Our proposed method is able to achieve the best performance on the whole datasets, which shows the effectiveness and robustness of our method. In particular, our method has more than 3% improvement on the TOX_171 data set based on both AC and NMI metrics. Therefore, our method is more capable of selecting the representative features and dealing with the noise in the raw data.

4.4. Parameter sensitivity and convergence study

First, we study the sensitiveness of parameters α and β in our model. Figs. 1 and 2 report the experimental results based on AC and NMI metrics for all the datasets. The logarithms (base 10) of parameters are taken. As we can see from Figs. 1 and 2, the performance is relatively sensitive to the parameters, which is still an open problem in feature selection.

An alternative iterative algorithm is utilized to solve the proposed problem (10) with the theoretical analysis on its convergence. Next, we experimentally investigate the speed of its convergence. Fig. 3 shows convergence curves of our method on different datasets. Note that our optimization algorithm is efficient and converges very fast, which ensures the speed of our proposed method.

5. Conclusion

In this paper, we propose a novel unsupervised feature selection method based on self-expression model. It can not only explore the relationships between the features, but also deal with the noise underlying in the raw data. Then a structure preserved constraint is incorporated into our model for keeping the local manifold structure of the data. Therefore, our proposed model can perform local structure learning and feature selection simultaneously. Furthermore, an efficient alternative iterative algorithm is utilized to optimize our objective function with the theoretical analysis on its convergence. Extensive experiments on real applications show that our method outperforms seven representative methods.

Note that our proposed model utilizes the Frobenius norm to deal with the noise in the data. Therefore, our model is more suitable for handling Gaussian noise. However, the noise in the real data may have a complicated statistical distribution. It is hard to model noise by simply using a certain norm. Our future work is to construct a Gaussian mixture model to estimate the distribution of noise in the raw data. Besides, we would like to use our feature selection model to deal with deep features in many deep learning algorithms [53,54].

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants 61761130079 and U1604153, in part by the Key Research Program of Frontier Sciences, CAS under Grant QYZDY-SSW-JSC044, in part by the Training Program for the Young-Backbone Teachers in Universities of Henan Province under Grant 2017GGJS065, and in part by the State Key Laboratory of Virtual Reality Technology and Systems under Grant BUAAVR-16KF-04.

References

[1] A.K. Jain, D.E. Zongker, Feature selection: evaluation, application, and small sample performance, IEEE Trans. Pattern Anal. Mach. Intell. 19 (2) (1997) 153–158.
[2] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, J. Mach. Learn. Res. 3 (2003) 1157–1182.
[3] P. Zhu, W. Zhu, Q. Hu, C. Zhang, W. Zuo, Subspace clustering guided unsupervised feature selection, Pattern Recognit. 66 (2017) 364–374.
[4] J. Gui, Z. Sun, S. Ji, D. Tao, T. Tan, Feature selection based on structured sparsity: a comprehensive study, IEEE Trans. Neural Netw. Learn. Syst. 28 (7) (2017) 1490–1507.
[5] Y. Saeys, I. Inza, P. Larrañaga, A review of feature selection techniques in bioinformatics, Bioinformatics 23 (19) (2007) 2507–2517.
[6] Y. Cong, S. Wang, J. Liu, J. Cao, Y. Yang, J. Luo, Deep sparse feature selection for computer aided endoscopy diagnosis, Pattern Recognit. 48 (3) (2015) 907–917.
[7] S. Alelyani, J. Tang, H. Liu, Feature selection for clustering: A review, in: Data Clustering: Algorithms and Applications, 2013, pp. 29–60.
[8] J. Tang, S. Alelyani, H. Liu, Feature selection for classification: A review, in: Data Classification: Algorithms and Applications, 2014, pp. 37–64.
[9] Y. Kim, J. Kim, Gradient LASSO for feature selection, in: Proceedings of the International Conference on Machine Learning, 2004.
[10] F. Nie, S. Xiang, Y. Jia, C. Zhang, S. Yan, Trace ratio criterion for feature selection, in: Proceedings of the Conference on Artificial Intelligence, 2008, pp. 671–676.
[11] Z. Zhao, L. Wang, H. Liu, Efficient spectral feature selection with minimum redundancy, in: Proceedings of the Conference on Artificial Intelligence, 2010.
[12] Y. Han, Y. Yang, Y. Yan, Z. Ma, N. Sebe, X. Zhou, Semisupervised feature selection via spline regression for video semantic recognition, IEEE Trans. Neural Netw. Learn . Syst. 26 (2) (2015) 252–264.
[13] Z. Zhao, H. Liu, Semi-supervised feature selection via spectral analysis, in: Proceedings of the International Conference on Data, 2007, pp. 641–646.
[14] F. Nie, W. Zhu, X. Li, Unsupervised feature selection with structured graph optimization, in: Proceedings of the Conference on Artificial Intelligence, 2016, pp. 1302–1308.
[15] 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.
[16] L. Shi, L. Du, Y. Shen, Robust spectral learning for unsupervised feature selection, in: Proceedings of the IEEE International Conference on Data Mining, 2014, pp. 977–982.
[17] Z. Li, J. Liu, Y. Yang, X. Zhou, H. Lu, Clustering-guided sparse structural learning for unsupervised feature selection, IEEE Trans. Knowl. Data Eng. 26 (9) (2014) 2138–2150.
[18] W. Zheng, H. Yan, J. Yang, J. Yang, Robust unsupervised feature selection by nonnegative sparse subspace learning, in: Proceedings of the International Conference on Pattern Recognition, 2016, pp. 3615–3620.
[19] Q. Cheng, H. Zhou, J. Cheng, The fisher-markov selector: Fast selecting maximally separable feature subset for multiclass classification with applications to high-dimensional data, IEEE Trans. Pattern Anal. Mach. Intell. 33 (6) (2011) 1217–1233.
[20] T. Hancock, H. Mamitsuka, Boosted network classifiers for local feature selection, IEEE Trans. Neural Netw. Learn. Syst. 23 (11) (2012) 1767–1778.
[21] L. Laporte, R. Flamary, S. Canu, S. Déjean, J. Mothe, Nonconvex regularizations for feature selection in ranking with sparse SVM, IEEE Trans. Neural Netw. Learn. Syst. 25 (6) (2014) 1118–1130.
[22] D. Chakraborty, N.R. Pal, Selecting useful groups of features in a connectionist framework, IEEE Trans. Neural Netw. 19 (3) (2008) 381–396.
[23] W. Yang, Y. Gao, Y. Shi, L. Cao, Mrm-lasso: a sparse multiview feature selection method via low-rank analysis, IEEE Trans. Neural Netw. Learn. Syst. 26 (11) (2015) 2801–2815.
[24] P. Zhu, W. Zuo, L. Zhang, Q. Hu, S.C.K. Shiu, Unsupervised feature selection by regularized self-representation, Pattern Recognit. 48 (2) (2015) 438–446.
[25] X. He, D. Cai, P. Niyogi, Laplacian score for feature selection, in: Proceedings of the Advances in Neural Information Processing Systems, 2005, pp. 507–514.
[26] A. Rakotomamonjy, Variable selection using svm-based criteria, J. Mach. Learn. Res. 3 (2003) 1357–1370.
[27] D. Cai, C. Zhang, X. He, Unsupervised feature selection for multi-cluster data, in: Proceedings of the International Conference on Knowledge Discovery and Data Mining, 2010, pp. 333–342.
[28] F. Nie, H. Huang, X. Cai, C.H.Q. Ding, Efficient and robust feature selection via joint $l_{2,1}$-norms minimization, in: Proceedings of the Advances in Neural Information Processing Systems, 2010, pp. 1813–1821.
[29] 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.
[30] Z. Li, Y. Yang, J. Liu, X. Zhou, H. Lu, Unsupervised feature selection using nonnegative spectral analysis, in: Proceedings of the Conference on Artificial Intelligence, 2012.
[31] M. Qian, C. Zhai, Robust unsupervised feature selection, in: Proceedings of the International Joint Conference on Artificial Intelligence, 2013, pp. 1621–1627.
[32] X. Zhu, X. Li, S. Zhang, C. Ju, X. Wu, Robust joint graph sparse coding for unsupervised spectral feature selection, IEEE Trans. Neural Netw. Learn. Syst. 28 (6) (2017) 1263–1275.
[33] Q. Lu, X. Li, Y. Dong, D. Tao, Subspace clustering by capped l_1 norm, in: Proceedings of the Chinese Conference on Pattern Recognition, 2016, pp. 663–674.
[34] C. Lu, H. Min, Z.-Q. Zhao, L. Zhu, D.-S. Huang, S. Yan, Robust and efficient subspace segmentation via least squares regression, in: Proceedings of the European Conference on Computer Vision, 2012, pp. 347–360.
[35] C. Lu, J. Feng, Z. Lin, S. Yan, Correlation adaptive subspace segmentation by trace lasso, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 1345–1352.
[36] B. Li, Y. Zhang, Z. Lin, H. Lu, Subspace clustering by mixture of gaussian regression, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2094–2102.
[37] H. Peng, F. Long, C. Ding, Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy, IEEE Trans. Pattern Anal. Mach. Intell. 27 (8) (2005) 1226–1238.
[38] M. Masaeli, G. Fung, J.G. Dy, From transformation-based dimensionality reduction to feature selection, in: Proceedings of the International Conference on Machine Learning, 2010, pp. 751–758.
[39] J.G. Dy, C.E. Brodley, Feature selection for unsupervised learning, J. Mach. Learn. Res. 5 (2004) 845–889.
[40] L. Wolf, A. Shashua, Feature selection for unsupervised and supervised inference: the emergence of sparsity in a weight-based approach, J. Mach. Learn. Res. 6 (2005) 1855–1887.
[41] S. Maldonado, R. Weber, A wrapper method for feature selection using support vector machines, Inf. Sci. 179 (13) (2009) 2208–2217.
[42] Y. Yang, H.T. Shen, Z. Ma, Z. Huang, X. Zhou, $l_{2,1}$-norm regularized discriminative feature selection for unsupervised learning, in: Proceedings of the International Joint Conference on Artificial Intelligence, 2011, pp. 1589–1594.
[43] S. Wang, J. Tang, H. Liu, Embedded unsupervised feature selection, in: Proceedings of the Conference on Artificial Intelligence, 2015, pp. 470–476.
[44] E. Elhamifar, R. Vidal, Sparse subspace clustering: Algorithm, theory, and applications, IEEE Trans. Pattern Anal. Mach. Intell. 35 (2013) 2765–2781.
[45] G. Liu, Z. Lin, Y. Yu, Robust subspace segmentation by low-rank representation., in: Proceedings of the International Conference on Machine Learning, 2010, pp. 663–670.
[46] M. Lee, J. Lee, H. Lee, N. Kwak, Membership representation for detecting block--diagonal structure in low-rank or sparse subspace clustering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1648–1656.
[47] X. Peng, H. Tang, L. Zhang, Z. Yi, S. Xiao, A unified framework for representation-based subspace clustering of out-of-sample and large-scale data, IEEE Trans. Neural Netw. Learn. Syst. 27 (2016) 2499–2512.
[48] D. Cai, X. He, J. Han, T.S. Huang, Graph regularized nonnegative matrix factor-

ization for data representation, IEEE Trans. Pattern Anal. Mach. Intell. 33 (8) (2011) 1548–1560.

[49] X. Li, G. Cui, Y. Dong, Graph regularized non-negative low-rank matrix factorization for image clustering, IEEE Trans. Cybern. 47 (2016) 3840–3853.

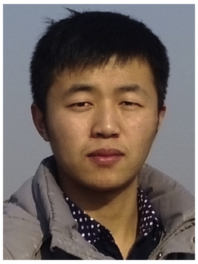
[50] X. Li, Q. Lu, Y. Dong, D. Tao, SCE: A manifold regularized set-covering method for data partitioning, IEEE Trans. Neural Netw. Learn. Syst.. 10.1109/TNNLS.2017.2682179

[51] X. Cao, C. Zhang, H. Fu, S. Liu, H. Zhang, Diversity-induced multi-view subspace clustering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 586–594.

[52] A. Strehl, J. Ghosh, Cluster ensembles—a knowledge reuse framework for combining multiple partitions, J. Mach. Learn. Res. 3 (2002) 583–617.

[53] N. Zeng, H. Zhang, B. Song, W. Liu, Y. Li, A.M. Dobaie, Facial expression recognition via learning deep sparse autoencoders, Neurocomputing 273 (2018) 643–649.

[54] N. Zeng, Z. Wang, H. Zhang, W. Liu, F.E. Alsaadi, Deep belief networks for quantitative analysis of a gold immunochromatographic strip, Cognit. Comput. 8 (4) (2016) 684–692.



Quanmao Lu is currently working toward the Ph.D. degree in the Center for Optical Imagery Analysis and Learning, Xi’an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi’an, China. His current research interests include machine learning and computer vision.

Xuelong Li is a full professor with the Center for Optical Imagery Analysis and Learning, Xi’an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi’an, China.

Yongsheng Dong received his Ph. D. degree in applied mathematics from Peking University in 2012. He was a postdoctoral research fellow with the Center for Optical Imagery Analysis and Learning, Xi’an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi’an, China from 2013 to 2016. From 2016 to 2017, he was a visiting research fellow at the School of Computer Science and Engineering, Nanyang Technological University, Singapore. He is currently an associate professor with the School of Information Engineering, Henan University of Science and Technology. His current research interests include pattern recognition, machine learning, and computer vision.

He has authored and co-authored over 30 papers at famous journals and conferences, including IEEE TIP, IEEE TNNLS, IEEE TCYB, IEEE SPL and ACM TIST. He has served as a reviewer for over 30 international prestigious journals and conferences, such as IEEE TNNLS, IEEE TIP, IEEE TCYB, IEEE TIE, IEEE TSP, IEEE TKDE, IEEE TCDS and ACM TIST. He has also served as a Program Committee Member for more than 10 international conferences. He is a member of the IEEE, ACM and CCF.