Unsupervised Feature Selection via Nonnegative Spectral Analysis and Redundancy Control

Zechao Li, *Member, IEEE*, and Jinhui Tang, *Senior Member, IEEE*

***Abstract*— In many image processing and pattern recognition problems, visual contents of images are currently described by high-dimensional features, which are often redundant and noisy. Toward this end, we propose a novel unsupervised feature selection scheme, namely, nonnegative spectral analysis with constrained redundancy, by jointly leveraging nonnegative spectral clustering and redundancy analysis. The proposed method can directly identify a discriminative subset of the most useful and redundancy-constrained features. Nonnegative spectral analysis is developed to learn more accurate cluster labels of the input images, during which the feature selection is performed simultaneously. The joint learning of the cluster labels and feature selection matrix enables to select the most discriminative features. Row-wise sparse models with a general $\ell_{2,p}$-norm ($0 < p \leq 1$) are leveraged to make the proposed model suitable for feature selection and robust to noise. Besides, the redundancy between features is explicitly exploited to control the redundancy of the selected subset. The proposed problem is formulated as an optimization problem with a well-defined objective function solved by the developed simple yet efficient iterative algorithm. Finally, we conduct extensive experiments on nine diverse image benchmarks, including face data, handwritten digit data, and object image data. The proposed method achieves encouraging the experimental results in comparison with several representative algorithms, which demonstrates the effectiveness of the proposed algorithm for unsupervised feature selection.**

***Index Terms*— Feature selection, nonnegative spectral clustering, constrained redundancy, row-sparsity.**

I. Introduction

In MANY image processing and multimedia problems, images are usually represented by high-dimensional visual features, such as local features (such as SIFT [1]). In practice, it is well known that all features that characterize images are not usually equal important for a given task and most of them are often correlated or redundant to each other, and sometimes noisy [2]. Besides, it is hard to discriminate images of different classes from each other in the high-dimensional space of visual features. That is, these high-dimensional features may bring some disadvantages, such as over-fitting, low efficiency and poor performance, to the traditional learning models [3]. As a consequence, it is necessary and challenging to select an optimal feature subset from high-dimensional image to remove irrelevant and redundant features, increase learning accuracy and improve the performance comprehensibility.

The task of selecting the "best" feature subset is known as *feature selection*, which is an important and widely used method. The importance of feature selection in improving both the efficiency and accuracy of image processing is threehold. First, it can result in computationally efficient algorithms since the dimensionality of selected feature subset is much lower. Second, it enables to provide a better understanding of the underlying structure of the data. Finally, it can improve the performance by removing noisy and redundant features. Therefore, many feature selection methods have been proposed and studied [4]–[16]. These algorithms can be categorized as supervised algorithms, semi-supervised algorithms and unsupervised algorithms, according to the way of utilizing label information. Since the discriminative information is encoded in the labels, supervised and semi-supervised approaches can generally achieve good performance. However, the labels of data annotated by human experts are typically expensive and time-consuming and there is usually no shortage of unlabeled data in many real-world applications. Consequently, it is quite promising and demanding to develop unsupervised feature selection techniques, which may be more practical.

In unsupervised feature selection, features are selected based on a frequently used criterion which evaluates features by their capability of keeping certain properties of the data, such as the data distribution, the redundancy of features or local structure. The methods with controlled redundancy discard those features which do not change or only slightly change the performance [17]–[19], while the methods considering local structure exploit different types of structures [5], [8], [11], [15], [20]–[22]. In [5] and [8], features are selected one by one based on the importance of each feature individually. To handle the disadvantage of selecting features individually, approaches selecting features jointly across all data points are proposed [11], [21]. Besides, sparsity-based feature selection is proposed to choose features jointly [22], [23]. On the other hand, the whole features contain necessary features (which are essential for the task), redundant features (which are useful but dependent on each other. Thus, not all of the redundant features are not necessary.), noisy features (which degrade the performance) and indifferent features (which do not matter for the task). The goal of feature selection is to select necessary features, discard noisy or indifferent features and control the

Manuscript received July 13, 2015; revised September 6, 2015; accepted September 8, 2015. Date of publication September 17, 2015; date of current version October 6, 2015. This work was supported in part by the 973 Program under Grant 2014CB347600 and in part by the National Natural Science Foundation of China under Grant 61522203 and Grant 61402228. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Ling Shao. *(Corresponding author: Jinhui Tang.)*

The authors are with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail: zechao.li@njust.edu.cn; jinhuitang@njust.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2015.2479560

use of redundant features. The above methods do not jointly consider these four features. Besides, they fail to exploit discriminative information from data.

In light of all these factors, in this work we propose a novel unsupervised feature selection algorithm to exploit discriminative information from data, select necessary features, discard noisy or indifferent features and control the use of redundant features simultaneously. Towards this end, a new method named Nonnegative Spectral analysis with Constrained Redundancy (NSCR) is developed by integrating nonnegative spectral analysis and redundancy control into a joint framework. Due to the importance of discriminative information, it is necessary and beneficial to exploit discriminative information in unsupervised feature selection. As a consequence, we propose a novel nonnegative spectral analysis scheme to uncover discriminative information by learning more accurate cluster indicators. With nonnegative and orthogonality constraints, the learned cluster indicators are much closer to the ideal ones and can be readily utilized to obtain more accurate cluster labels, which can be utilized to guide feature selection. The joint learning of the cluster labels and feature selection matrix enables to select the most discriminative features. For the sake of feature selection, the predictive matrix is constrained to be sparse in rows, which is formulated as a general $\ell_{2,p}$-norm ($0 < p \leq 1$) minimization term. Besides, the redundancy between features is explicitly exploited to control the redundancy of the selected features. The proposed problem is formulated as an optimization problem with a well-defined objective function. To solve the proposed problem, a simple yet efficient iterative algorithm is proposed. To verify the effectiveness of the proposed method, extensive experiments are conducted on 9 widely used real-world datasets. The proposed method achieves encouraging experimental results in comparison with several representative algorithms.

Contributions: The main contributions of this paper are highlighted as follows.

- We propose a novel unsupervised feature selection framework by exploiting nonnegative cluster analysis and redundancy control with row sparsity simultaneously. An effective and efficient algorithm is developed to solve the proposed formulation.
- We develop nonnegative spectral analysis to learn more accurate cluster indicators by imposing nonnegative and orthogonal constraints on the cluster indicator matrix.
- The redundancy between features is explicitly exploited to control the redundancy of the selected subset. To facilitate feature selection, the sparse feature selection models with the general $\ell_{2,p}$-norm ($0 < p \leq 1$) are exerted on the regularization term.
- We discuss the relationships between the proposed NSCR and several well-known feature selection algorithms.

The rest of this paper is structured as follows. We briefly overview the related work in Section II. Then we present our proposed formulation in Section III followed with the developed solution and discussions in Section IV. Extensive experiments are conducted and analyzed in Section V. Section VI concludes this work with future work.

II. Related Work

A. Unsupervised Feature Selection

According to the availability of label information, feature selection algorithms can be classified into three broad categories: supervised, semi-supervised and unsupervised approaches. More details can be obtained in [24] and [25]. In this section, we will elaborate unsupervised feature selection methods.

From the perspective of selection strategy, the unsupervised feature selection approaches can be broadly categorized as the *filter*, *wrapper* and *embedded* ones. For filter methods [5], [8], [20], [26], a proxy measure is utilized to score a feature subset instead of the error rate. The simplest measure may be the variance score with the assumption that larger variance means better representation ability. However, there is no reason to assume that these features are useful for discriminating data in different classes. Laplacian Score [5] selects features which can best reflect the underlying manifold structure. However, the redundancy among features is not exploited, which may result in redundant features and compromise the performance. In [14], the most informative features and instances are selected simultaneously from data. Filters are usually less computationally intensive than wrappers, but produce a feature set which is not tuned to a specific type of predictive model. Wrapper methods [3], [21], [27] score feature subsets using use a predictive model. They wrap feature search around the learning algorithms and utilize the learned results to select features. Clustering is a commonly utilized learning algorithm [11], [21], [27]. The clusterability of the input data points is measured by analyzing the spectral properties of the affinity matrix. MCFS [11] uses a two-step spectral regression approach to unsupervised feature selection. Embedded methods [28], [29] perform feature selection as a part of the model construction process, which fall in between filters and wrappers in terms of computational complexity.

State-of-the-art algorithms exploit discriminative information and feature correlation to select features [16], [22], [30]–[33]. Unsupervised Discriminative Feature Selection (UDFS) [22] considers the manifold structure to select the most discriminative features. It imposes an orthogonal constraint on the feature selection, which is unreasonable since feature weight vectors are not necessarily orthogonal with each other in nature. Nonnegative Discriminative Feature Selection (NDFS) [16], [30] proposes nonnegative spectral clustering to guide feature selection and selects features over the whole feature space. In [31], a global and a set of locally linear regression model are integrated into a unified learning framework. Qian and Zhai [32] extended NDFS to handle outliers or noise data. The graph embedding and sparse spectral regression are improved in [33]. In [34], a robust space learning framework is proposed for multiple image understanding tasks, which is general and effective. NDFS is a special case of this framework. However, the above methods do not explicitly control the redundancy between features, which may lead to redundancy existing in the selected features.

Different from previous work, our algorithm exploits nonnegative spectral clustering and explicitly controls the redundancy between features in a joint framework for unsupervised feature learning. One general sparse model with $\ell_{2,p}$-norm ($0 < p \leq 1$) is adopted to learn a better sparsity matrix than $\ell_{2,1}$-norm.

B. Consideration of Feature Dependency

Some methods have been designed to consider the dependency between features [35]. In [36], the redundancy between selected features is removed using a correlation-based filter. Unsupervised correlation-based feature selection approach is proposed in [37] based on the decision-dependent correlation between feature. Peng et al. [17] proposed a mutual information based two-stage feature selection approach to choose features with least redundancy by minimizing the mutual information among the selected features.

The feature selection problem is formulated as constrained 0-1 linear fractional program to avoid the redundancy between selected features [38]. Song et al. [39] proposed to detect the relevant features as well as remove the redundancy between features. A multilayer perceptron neural network is designed for feature selection with consideration a measure of linear dependency to control the redundancy in [19]. However, they only focus on considering the dependency between features, and fail to select discriminative features. In this work, we select features by considering the dependency between features and the discriminant information simultaneously. The most discriminant features with controlled redundancy are selected.

III. The Proposed NSCR Method

In this section, we will elaborate the proposed unsupervised feature selection method by integrating nonnegative spectral analysis and redundancy control into a unified framework.

A. Preliminary

Throughout this paper, we use bold uppercase characters to denote matrices, bold lowercase characters to denote vectors. For an arbitrary matrix $\mathbf{A} \in \mathbb{R}^{r \times t}$, $\mathbf{a}_i$ means the i-th row vector of $\mathbf{A}$, $\mathbf{a}^j$ means the j-th column vector of $\mathbf{A}$, A_{ij} denotes the (i, j)-th entry of $\mathbf{A}$, $\|\mathbf{A}\|_F$ is Frobenius norm of $\mathbf{A}$ and $\mathrm{Tr}[\mathbf{A}]$ is the trace of $\mathbf{A}$ if $\mathbf{A}$ is square. The $\ell_{2,p}$-norm ($p \in (0, 1]$) of $\mathbf{A}$ is defined as

$$\|\mathbf{A}\|_{2,p} = (\sum_{i=1}^{r}(\sqrt{\sum_{j=1}^{t} A_{ij}^2})^p)^{\frac{1}{p}} = (\sum_{i=1}^{r} \|\mathbf{a}_i\|_2^p)^{\frac{1}{p}}. \quad (1)$$

Assume that we have n images $\mathbb{X} = \{\mathbf{x}^i\}_{i=1}^n$. Let $\mathbf{X} = [\mathbf{x}^1, \cdots, \mathbf{x}^n]$ denote the data matrix, in which $\mathbf{x}^i \in \mathbb{R}^d$ is the feature descriptor of the i-th image. Suppose these n images are sampled from c classes. Denote $\mathbf{Y} = [\mathbf{y}_1^T, \cdots, \mathbf{y}_n^T]^T \in \{0, 1\}^{n \times c}$, where $\mathbf{y}_i \in \{0, 1\}^{1 \times c}$ is the cluster indicator vector for $\mathbf{x}^i$. That is, $Y_{ij} = 1$ if the image $\mathbf{x}^i$ is assigned to the j-th cluster, and $Y_{ij} = 0$ otherwise. The scaled cluster indicator matrix $\mathbf{F}$ is defined as:

$$\mathbf{F} = [\mathbf{f}^1, \mathbf{f}^2, \cdots, \mathbf{f}^c] = \mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-\frac{1}{2}}, \quad (2)$$

It turns out that

$$\mathbf{F}^T\mathbf{F} = (\mathbf{Y}^T\mathbf{Y})^{-\frac{1}{2}}\mathbf{Y}^T\mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-\frac{1}{2}} = \mathbf{I}_c, \quad (3)$$

where $\mathbf{I}_c \in \mathbb{R}^{c \times c}$ is an identity matrix.

B. The Proposed Framework

To select the discriminative features with controlled redundancy, we propose to exploit clustering analysis and explicitly consider the redundancy between features simultaneously. Clustering techniques are adopted to learn the cluster indicators (which can be regarded as pseudo class labels), which are used to guide the process of inferring the feature selection matrix. Hence, the scaled cluster indicator matrix $\mathbf{F} \in \mathbb{R}^{n \times c}$ and the feature selection matrix $\mathbf{W} \in \mathbb{R}^{d \times c}$ are learned simultaneously. Besides, there may exist redundancy between features. Intuitively, we explicitly control the redundancy between features. Therefore, our framework is formulated as

$$\begin{aligned} &\min_{\mathbf{F},\mathbf{W}} \mathcal{J}(\mathbf{F}) + \alpha l(h(\mathbf{W}; \mathbf{X}), \mathbf{F}) + \beta\Omega(\mathbf{W}) + \lambda g(\mathbf{W}) \\ &\text{s.t.} \quad \mathbf{F} = \mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-\frac{1}{2}}, \end{aligned} \quad (4)$$

where $\mathcal{J}(\mathbf{F})$ is a clustering criterion, $l(\cdot, \cdot)$ is the loss function, $h(\cdot)$ is a predictive function, $\Omega(\cdot)$ is a regularization function with row sparsity and $g(\cdot)$ is a function to control the redundancy. α, β and λ are three nonnegative trade-off parameter.

C. Nonnegative Spectral Analysis

To exploit the discriminant information, we propose a novel nonnegative spectral clustering algorithm to identify the cluster indicators of images. In cluster analysis, graph-theoretic methods have been well studied and utilized in many applications. As one of graph-theoretic methods, spectral clustering has been verified to be effective to detect the cluster structure of data and has received significant research attention [40], [41]. Therefore, we adopt spectral clustering as the cluster analysis technique.

Clearly, an effective cluster indicator matrix is more capable to reflect the discriminative information of the input data. The local geometric structure of data plays an important role in data clustering, which has been exploited by many spectral clustering algorithms [40], [41]. There are many strategies to construct a graph to uncover local data structure. In this work, we focus on proposing a new spectral analysis method instead of the graph construction. Thus, we use the strategy proposed in [40] to be the criterion for its simplicity. The local geometric structure can be effectively modeled by a nearest neighbor graph on a scatter of data points. To construct the affinity graph $\mathbf{S}$, we define

$$S_{ij} = \begin{cases} \exp(-\frac{\|\mathbf{x}^i - \mathbf{x}^j\|^2}{\sigma^2}) & \mathbf{x}^i \in \mathcal{N}_k(\mathbf{x}^j) \text{ or } \mathbf{x}^j \in \mathcal{N}_k(\mathbf{x}^i) \\ 0 & \text{otherwise,} \end{cases}$$

where $\mathcal{N}_k(\mathbf{x})$ is the set of k-nearest neighbors of $\mathbf{x}$. The local geometrical structure can be exploited by

$$\min_{\mathbf{F}} \frac{1}{2}\sum_{i,j=1}^{n} S_{ij}\|\frac{\mathbf{f}_i}{\sqrt{E_{ii}}} - \frac{\mathbf{f}_j}{\sqrt{E_{jj}}}\|_2^2 = \mathrm{Tr}[\mathbf{F}^T\mathbf{L}\mathbf{F}], \tag{5}$$

where $\mathbf{E}$ is a diagonal matrix with $E_{ii} = \sum_{j=1}^{n} S_{ij}$ and $\mathbf{L} = \mathbf{E}^{-1/2}(\mathbf{E}-\mathbf{S})\mathbf{E}^{-1/2}$ is the normalized graph Laplacian matrix. Therefore $\mathcal{J}(\mathbf{F})$ is defined as

$$\mathcal{J}(\mathbf{F}) = \mathrm{Tr}[\mathbf{F}^T\mathbf{L}\mathbf{F}], \quad \text{s.t.} \quad \mathbf{F} = \mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-\frac{1}{2}}. \tag{6}$$

From the definition of the cluster indicator matrix $\mathbf{F}$, we can see that each element of $\mathbf{F}$ is constrained to be a discrete value. It makes the proposed problem (4) an NP-hard problem [40]. To deal with this problem, we employ a well-known solution [40] to relax it from discrete values to continuous ones while keeping the property of Eq. 3. Thus, Eq. 6 is relaxed to

$$\mathcal{J}(\mathbf{F}) = \mathrm{Tr}[\mathbf{F}^T\mathbf{L}\mathbf{F}], \quad \text{s.t.} \quad \mathbf{F}^T\mathbf{F} = \mathbf{I}_c. \tag{7}$$

According to the definition of the cluster indicator matrix $\mathbf{F}$, each element F_{ij} indicates the relationship between the i-th image and the j-th cluster, which is nonnegative in nature. Unfortunately, the optimal $\mathbf{F}$ only with the constraint $\mathbf{F}^T\mathbf{F} = \mathbf{I}_c$ has mixed signs, which violates its definition. Moreover, the mixed signs make it difficult to get the cluster labels. Discrete process, such as spectral rotation or Kmeans, is performed in previous works to obtain the cluster labels. However, our work is a one-step model and contains no discrete process, which makes the learned $\mathbf{F}$ severely deviate from the ideal cluster indicators. To address this problem, it is natural and reasonable to impose nonnegative constraints on $\mathbf{F}$. When both nonnegative and orthogonal constraints are satisfied, only one element in each row of $\mathbf{F}$ is greater than zero and all of the others are zeros, which makes the learned $\mathbf{F}$ more accurate. Thus, we have the following objective function.

$$\mathcal{J}(\mathbf{F}) = \mathrm{Tr}[\mathbf{F}^T\mathbf{L}\mathbf{F}], \quad \text{s.t.} \quad \mathbf{F}^T\mathbf{F} = \mathbf{I}_c, \ \mathbf{F} \geq 0. \tag{8}$$

Here $\mathbf{F} \geq 0$ denotes that each element of $\mathbf{F}$ is nonnegative.

D. Sparse Prediction Model

In our framework, the features which are most discriminative to the cluster indicators are selected. To this end, we assume that there is a linear transformation between features and the cluster indicators and adopt a linear model to predict the cluster indicators. Therefore, we have the following function:

$$h(\mathbf{W}; \mathbf{X}) = \mathbf{X}^T\mathbf{W} \tag{9}$$

where $\mathbf{W} = [\mathbf{w}^1, \cdots, \mathbf{w}^c] \in \mathbb{R}^{d\times c}$ is the linear transformation matrix to predict the cluster indicators. To learn a more discriminative predictors for more reliable results and make our method robust to noisy features, we impose a more general and better sparse model on $\mathbf{W}$. It has been verified by extensive computational studies that ℓ_p-norm $(0 < p < 1)$ can lead to sparser solution than using ℓ_1-norm [42], [43], and $\ell_{2,p}$-norm based minimization can also achieve a better sparsity solution than $\ell_{2,1}$-norm [44]. Thus, we introduce a $\ell_{2,p}$-norm based regularization for $\mathbf{\Omega}$ to guarantee that $\mathbf{W}$ is sparse in rows. It can discard noisy or indifferent features.

$$\Omega(\mathbf{W}) = \sum_{i=1}^{d}\|\mathbf{w}_i\|_2^p = \|\mathbf{W}\|_{2,p}^p \tag{10}$$

The proposed problem in (4) can be rewritten as

$$\min_{\mathbf{F},\mathbf{W}} \mathcal{J}(\mathbf{F}) + \alpha l(\mathbf{X}^T\mathbf{W}, \mathbf{F}) + \beta\|\mathbf{W}\|_{2,p}^p + \lambda g(\mathbf{W})$$
$$\text{s.t.} \quad \mathbf{F}^T\mathbf{F} = \mathbf{I}_c, \ \mathbf{F} \geq 0. \tag{11}$$

E. Redundancy Control

Under the guide of nonnegative spectral clustering, the feature selection matrix with $\ell_{2,p}$-norm regularization can select necessary features and discard noisy or indifferent features. However, correlated features may be selected simultaneously since currently we do not penalize the proposed method for redundant features. For example, if the i-th feature is highly correlated to the j-th feature, we do not need to select both of them simultaneously. Towards this end, we introduce a penalty factor $g(\mathbf{W})$ into our feature selection scheme to control the redundancy while selecting features. Many strategies can be used to define the penalty for using redundant features. In this work, we adopt the correlation between features to define $g(\mathbf{W})$.

$$g(\mathbf{W}) = \frac{1}{d(d-1)}\sum_{i=1}^{d}\|\mathbf{w}_i\|_2 \sum_{j=1, j\neq i}^{d}\|\mathbf{w}_j\|_2 C_{ij} \tag{12}$$

$C_{ij} \geq 0$ is a measure of correlation between the i-th feature and the j-th feature. $\|\mathbf{w}_i\|_2$ is a weight to measure the importance of the i-th feature. The correlation can be measured linearly or nonlinearly. For a linear measure, the Pearsons correlation coefficient between the i-th feature and the j-th feature can be used. The mutual information between the i-th feature and the j-th feature can be used to measure the nonlinear correlation. In this work, the mutual information is adopted. If we set $C_{ii} = 0$, we have

$$g(\mathbf{W}) = \frac{1}{d(d-1)}\sum_{i,j=1}^{d}\|\mathbf{w}_i\|_2\|\mathbf{w}_j\|_2 C_{ij} \tag{13}$$

The normalized factor $\frac{1}{d(d-1)}$ is used just to make the regularization term independent of the number of features. By taking the redundancy into account, our method can avoid selected many members of a redundant set of features.

F. The Objective Function

By incorporating the nonnegative spectral clustering, sparse prediction model and the redundancy control into a unified framework, we obtain the following optimization problem.

$$\min_{\mathbf{F},\mathbf{W}} \mathrm{Tr}[\mathbf{F}^T\mathbf{L}\mathbf{F}] + \alpha l(\mathbf{X}^T\mathbf{W}, \mathbf{F}) + \beta\|\mathbf{W}\|_{2,p}^p$$
$$+ \frac{\lambda}{d(d-1)}\sum_{i,j=1}^{d}\|\mathbf{w}_i\|_2\|\mathbf{w}_j\|_2 C_{ij}$$
$$\text{s.t.} \quad \mathbf{F}^T\mathbf{F} = \mathbf{I}_c, \ \mathbf{F} \geq 0. \tag{14}$$

To solve the optimization problem in (14), we first decide which loss function is chosen for $l(\cdot, \cdot)$. In this work, we utilize the least square loss $l(x, y) = \frac{1}{2}(x - y)^2$ for simplicity and set $\gamma = \frac{\lambda}{d(d-1)}$. Hence we have

$$\begin{aligned} \min_{\mathbf{F},\mathbf{W}} \mathrm{Tr}[\mathbf{F}^T\mathbf{L}\mathbf{F}] + \frac{\alpha}{2}\|\mathbf{F} - \mathbf{X}^T\mathbf{W}\|_F^2 + \beta\|\mathbf{W}\|_{2,p}^p \\ + \gamma \sum_{i,j=1}^{d} \|\mathbf{w}_i\|_2\|\mathbf{w}_j\|_2 C_{ij} \\ \text{s.t.} \quad \mathbf{F}^T\mathbf{F} = \mathbf{I}_c, \ \mathbf{F} \geq 0. \end{aligned} \tag{15}$$

The joint minimization of the regression model and $\ell_{2,p}$-norm regularization term enables $\mathbf{W}$ to evaluate the correlation between pseudo labels and features, making it particularly suitable for feature selection. More specifically, $\mathbf{w}_i$, the i-th row of $\mathbf{W}$, shrinks to zero if the i-th feature is less discriminative to the pseudo labels $\mathbf{F}$. It can guarantee that the necessary features are selected and the noisy or indifferent features are discarded. The consideration of the redundancy can explicitly control the redundancy between the selected features. Once $\mathbf{W}$ is learned, we can select the top r ranked features by sorting all d features according to $\|\mathbf{w}_i\|_2$ $(i = 1, \cdots, d)$ in descending order. Therefore, the features corresponding to zero rows of $\mathbf{W}$ will be discarded when performing feature selection.

IV. Optimization

Since ℓ_p $(0 < p < 1)$ vector norm is neither convex nor Lipschitz continuous, $\ell_{2,p}$ matrix pseudo norm is not convex or Lipschitz continuous yet. The optimization problem (15) involves the $\ell_{2,p}$-norm which is not convex and non-smooth. Consequently, we propose an iterative optimization algorithm to solve the optimization problem (15). For the ease of representation, let us define

$$\begin{aligned} \mathscr{L}(\mathbf{F}, \mathbf{W}) = \mathrm{Tr}[\mathbf{F}^T\mathbf{L}\mathbf{F}] + \frac{\alpha}{2}\|\mathbf{F} - \mathbf{X}^T\mathbf{W}\|_F^2 + \beta\|\mathbf{W}\|_{2,p}^p \\ + \gamma \sum_{i,j=1}^{d} \|\mathbf{w}_i\|_2\|\mathbf{w}_j\|_2 C_{ij}. \end{aligned} \tag{16}$$

A. Update W as Given F

First, by computing the derivative of $\mathscr{L}$ with respect to $\mathbf{w}_i$,[1] we obtain:

$$\frac{\partial \mathscr{L}}{\partial \mathbf{w_i}} = \alpha\mathbf{X}(\mathbf{X}^T\mathbf{w}_i - \mathbf{f}_i) + \beta\frac{\mathbf{w}_i}{\|\mathbf{w}_i\|_2} + \gamma\frac{\sum_j \|\mathbf{w}_j\|_2 C_{ij}}{\|\mathbf{w}_i\|_2}\mathbf{w}_i. \tag{17}$$

The following equation can be easily induced.

$$\frac{\partial \mathscr{L}}{\partial \mathbf{W}} = \alpha\mathbf{X}(\mathbf{X}^T\mathbf{W} - \mathbf{F}) + \beta\mathbf{D}\mathbf{W} + \gamma\,\mathbf{H}\mathbf{W}, \tag{18}$$

where $\mathbf{D}$ is a diagonal matrix with $D_{ii} = \frac{p}{2\|\mathbf{w}_i\|_2^{2-p}}$ and $\mathbf{H}$ is another diagonal matrix with $H_{ii} = \frac{\sum_j \|\mathbf{w}_j\|_2 C_{ij}}{2\|\mathbf{w}_i\|_2}$. Setting $\frac{\partial \mathscr{L}(\mathbf{F},\mathbf{W})}{\partial \mathbf{W}} = 0$, we have

$$\begin{aligned} \alpha\mathbf{X}(\mathbf{X}^T\mathbf{W} - \mathbf{F}) + \beta\mathbf{D}\mathbf{W} + \gamma\,\mathbf{H}\mathbf{W} = 0 \\ \Rightarrow \mathbf{W} = \alpha(\alpha\mathbf{X}\mathbf{X}^T + \beta\mathbf{D} + \gamma\,\mathbf{H})^{-1}\mathbf{X}\mathbf{F} = \mathbf{G}^{-1}\mathbf{X}\mathbf{F} \end{aligned} \tag{19}$$

Here $\mathbf{G} = \mathbf{X}\mathbf{X}^T + \frac{\beta}{\alpha}\mathbf{D} + \frac{\gamma}{\alpha}\mathbf{H}$.

B. Update F as Given W

Owing to $\|\mathbf{A}\|_F^2 = \mathrm{Tr}(\mathbf{A}^T\mathbf{A})$ for any arbitrary matrix $\mathbf{A}$, we can rewrite Eq. 16 as follows.

$$\begin{aligned} \mathscr{L} &= \mathrm{Tr}[\mathbf{F}^T\mathbf{L}\mathbf{F}] + \alpha\mathrm{Tr}[(\mathbf{X}^T\mathbf{W} - \mathbf{F})^T(\mathbf{X}^T\mathbf{W} - \mathbf{F})] \\ &\quad + \beta\mathrm{Tr}[\mathbf{W}^T\mathbf{D}\mathbf{W}] + \gamma\,\mathrm{Tr}[\mathbf{W}^T\mathbf{H}\mathbf{W}] \\ &= \mathrm{Tr}[\mathbf{F}^T\mathbf{L}\mathbf{F}] + \alpha\mathrm{Tr}[\mathbf{F}^T\mathbf{F}] - 2\alpha\mathrm{Tr}[\mathbf{W}^T\mathbf{X}\mathbf{F}] \\ &\quad + \mathrm{Tr}[\mathbf{W}^T(\alpha\mathbf{X}\mathbf{X}^T + \beta\mathbf{D} + \gamma\,\mathbf{H})\mathbf{W}] \end{aligned} \tag{20}$$

By substituting the expression for $\mathbf{W}$ in Eq. 19 into the above equation, we have

$$\mathscr{L}(\mathbf{F}) = \mathrm{Tr}[\mathbf{F}^T(\mathbf{L} + \alpha\mathbf{I}_n - \alpha\mathbf{X}^T\mathbf{G}^{-1}\mathbf{X})\mathbf{F}] \tag{21}$$

Thus, we obtain the following optimization problem *w.s.r.* $\mathbf{F}$

$$\begin{aligned} \min_{\mathbf{F}} \mathrm{Tr}[\mathbf{F}^T\mathbf{M}\mathbf{F}] \\ \text{s.t.} \quad \mathbf{F}^T\mathbf{F} = \mathbf{I}_c; \ \mathbf{F} \geq 0 \end{aligned} \tag{22}$$

Here $\mathbf{M} = \mathbf{L} + \alpha\mathbf{I}_n - \alpha\mathbf{X}^T\mathbf{G}^{-1}\mathbf{X}$. Then we relax the orthogonal constraint by incorporating the orthogonal constraint of $\mathbf{F}$ into the objective function via Langrange multiplier and obtain the optimization problem as follows.

$$\min_{\mathbf{F}\geq 0} \mathrm{Tr}[\mathbf{F}^T\mathbf{M}\mathbf{F}] + \frac{\mu}{2}\|\mathbf{F}^T\mathbf{F} - \mathbf{I}_c\|_F^2 \tag{23}$$

$\mu > 0$ is a parameter to control the orthogonality condition. In practice, λ should be large enough to insure the orthogonality satisfied. Let ϕ_{ij} be the Lagrange multiplier for constraint $F_{ij} \geq 0$ and $\Phi = [\phi_{ij}]$. Since $\|\mathbf{A}\|_F^2 = \mathrm{Tr}(\mathbf{A}^T\mathbf{A})$, the Lagrange function is

$$\mathrm{Tr}[\mathbf{F}^T\mathbf{M}\mathbf{F}] + \frac{\mu}{2}\mathrm{Tr}[(\mathbf{F}^T\mathbf{F} - \mathbf{I}_c)^T(\mathbf{F}^T\mathbf{F} - \mathbf{I}_c)] + \mathrm{Tr}[\Phi\mathbf{F}^T]. \tag{24}$$

Setting its derivative with respect to $\mathbf{F}$ to 0, we have

$$2\mathbf{M}\mathbf{F} + 2\mu\mathbf{F}(\mathbf{F}^T\mathbf{F} - \mathbf{I}_c) + \Phi = 0. \tag{25}$$

Using the Karush-Kuhn-Tuckre (KKT) condition [45], [46] $\phi_{ij}F_{ij} = 0$, we obtain the updating rules:

$$\begin{aligned} 2[\mathbf{M}\mathbf{F} + \lambda\mathbf{F}(\mathbf{F}^T\mathbf{F} - \mathbf{I}_c)]_{ij}F_{ij} + \Phi_{ij}F_{ij} = 0 \\ \Rightarrow [\mathbf{M}\mathbf{F} + \lambda\mathbf{F}(\mathbf{F}^T\mathbf{F} - \mathbf{I}_c)]_{ij}F_{ij} = 0. \end{aligned} \tag{26}$$

There may exist mix-signed elements in $\mathbf{M}$. To guarantee the nonnegative property of $\mathbf{F}$, by introducing $\mathbf{M} = \mathbf{M}^+ - \mathbf{M}^-$, where $M_{ij}^+ = (|M_{ij}| + M_{ij})/2$ and $M_{ij}^- = (|M_{ij}| - M_{ij})/2$, the above equation is equivalent to

$$[(\mathbf{M}^+ - \mathbf{M}^-)\mathbf{F} + \mu\mathbf{F}(\mathbf{F}^T\mathbf{F} - \mathbf{I}_c)]_{ij}F_{ij} = 0. \tag{27}$$

[1] In practice, $\|\mathbf{w}_i\|_2$ could be close to zero but not zero. Theoretically, it could be zeros. For this case, we can regularize $\|\mathbf{w}_i\|_2 \leftarrow \|\mathbf{w}_i\|_2 + \epsilon$, where ϵ is a very small constant. When $\epsilon \to 0$, we can see that $\|\mathbf{w}_i\|_2 + \epsilon$ approximates $\|\mathbf{w}_i\|_2$.

Algorithm 1 The Proposed NSCR Method

```
Input:
    Data matrix $\mathbf{X} \in \mathcal{R}^{d\times n}$;
    Parameters $\alpha$, $\beta$, $\gamma$, $\mu$, $k$, $c$ and $p$
Construct the $k$-nearest neighbor graph and calculate $\mathbf{L}$;
Construct the correlation matrix between features $\mathbf{C}$;
The iteration step $t = 1$; Initialize $\mathbf{F}_t \in \mathcal{R}^{n\times c}$, set $\mathbf{D}_t \in \mathcal{R}^{d\times d}$ as an identity matrix and $\mathbf{H}_t \in \mathcal{R}^{d\times d}$ as a zero matrix;
repeat
    $\mathbf{G}_t = \mathbf{X}\mathbf{X}^T + \frac{\beta}{\alpha}\mathbf{D}_t + \frac{\gamma}{\alpha}\mathbf{H}_t$;
    $\mathbf{M}_t = \mathbf{L} + \alpha\mathbf{I}_n - \alpha\mathbf{X}^T\mathbf{G}_t^{-1}\mathbf{X}$;
    $\mathbf{M}_t^+ = (|\mathbf{M}_t| + \mathbf{M}_t)/2$ and $\mathbf{M}_t^- = (|\mathbf{M}_t| - \mathbf{M}_t)/2$;
    $(F_{t+1})_{ij} = (F_t)_{ij}\frac{(\mathbf{M}_t^-\mathbf{F}_t+\mu\mathbf{F}_t)_{ij}}{(\mathbf{M}_t^+\mathbf{F}_t+\mu\mathbf{F}_t\mathbf{F}_t^T\mathbf{F}_t)_{ij}}$;
    $\mathbf{W}_{t+1} = \mathbf{G}_t^{-1}\mathbf{X}\mathbf{F}_{t+1}$;
    Update the diagonal matrix $\mathbf{D}$ as
        $\mathbf{D}_{t+1} = \begin{bmatrix} \frac{p}{2\|(\mathbf{w}_{t+1})_1\|_2^{2-p}} & & \\ & \cdots & \\ & & \frac{p}{2\|(\mathbf{w}_{t+1})_d\|_2^{2-p}} \end{bmatrix}$;
    Update the diagonal matrix $\mathbf{H}$ as
        $\mathbf{H}_{t+1} = \begin{bmatrix} \frac{\sum_j \|(\mathbf{w}_{t+1})_j\|_2 C_{1j}}{2\|(\mathbf{w}_{t+1})_1\|_2} & & \\ & \cdots & \\ & & \frac{\sum_j \|(\mathbf{w}_{t+1})_j\|_2 C_{dj}}{2\|(\mathbf{w}_{t+1})_d\|_2} \end{bmatrix}$;
    t=t+1;
until Convergence criterion satisfied
Output:
    Sort all $d$ features according to $\|(\mathbf{w}_t)_i\|_2$ in descending order and select the top $r$ ranked features.
```

Here $|\cdot|$ denotes the absolute value function. The following updating rule is obtained.

$$F_{ij} \leftarrow F_{ij}\frac{(\mathbf{M}^-\mathbf{F} + \mu\mathbf{F})_{ij}}{(\mathbf{M}^+\mathbf{F} + \mu\mathbf{F}\mathbf{F}^T\mathbf{F})_{ij}}. \tag{28}$$

Then we normalize $\mathbf{F}$ with $(\mathbf{F}^T\mathbf{F})_{ii} = 1, i = 1, \cdots, c$.

From the above analysis, we can see that $\mathbf{D}$ and $\mathbf{H}$ related to $\mathbf{W}$ is required to solve $\mathbf{F}$ and it is still not straightforward to obtain $\mathbf{W}$ and $\mathbf{F}$. To this end, we design an iterative algorithm to solve the proposed formulation, which is summarized in Algorithm 1.

The alternative updating rules in Algorithm 1 monotonically decrease the objective function value of (15) in each iteration. That is, the proposed iterative procedure in Algorithm 1 can be verified to be convergent. The convergence of Algorithm 1 can be proved following the work in [7] and [30]. The convergence is also experimentally verified in our experiments. Besides, the proposed optimization algorithm is efficient. In the experiments, we observe that our algorithm usually converges around only 20 iterations.

C. Computational Complexity Analysis

Now, we briefly analyze the computational complexity. In our case, $c \ll n$ and $c \ll d$. It takes $O(nd^2)$ to obtain $\mathbf{C}$. The complexity of calculating the inverse of a matrix is $O(d^3)$. In each iteration step, the cost for updating $\mathbf{G}$ based on $\mathbf{W}$ and $\mathbf{C}$ is $O(cd + d^2)$. It needs $O(d^3)$ to obtain $\mathbf{G}^{-1}$. The cost for updating $\mathbf{M}$ is $O(nd^2 + n^2d)$. It takes $O(cn^2)$ to update $\mathbf{F}$ and $O(nd^2)$ to update $\mathbf{W}$, respectively. Thus the overall cost for the proposed NSCR is $O(T(d^3+nd^2+dn^2))$, where T is the number of iterations.

D. Discussions

In this section, we discuss the relationships between the proposed method and several algorithms, including MCFS [11], UDFS [22], SPFS [15] and NDFS [30].

Connection With MCFS: MCFS [11] uses a two-step strategy to select features according to spectral analysis and is formulated as the following form.

$$\min_{\mathbf{F}^T\mathbf{F}=\mathbf{I}_c} \text{Tr}[\mathbf{F}^T\mathbf{L}\mathbf{F}] \tag{29}$$

$$\min_{\mathbf{w}_i} \|\mathbf{f}_i - \mathbf{X}^T\mathbf{w}_i\| + \beta\|\mathbf{w}_i\|_1 \tag{30}$$

In our method, if we set $\gamma = 0$, $p = 1$ and remove the nonnegative constraint, when $\alpha \to 0$ and $\beta \to 0$, our method leads to a two-step algorithm, which has similar formulation to MCFS with different regularization forms on $\mathbf{W}$. Different from MCFS, NSCR is an one-step algorithm and more general. Second, $\mathbf{F}$ is constrained to be nonnegative. When both nonnegative and orthogonal constraints are satisfied, the learned $\mathbf{F}$ is much closer to the ideal result, and the solution can be directly obtained without discretization. Finally, in our framework, we perform clustering and redundancy control simultaneously, making the results more compact and accurate.

Connection With UDFS: UDFS [22] was propsoed to select discriminative features by optimizing the following objective function.

$$\min_{\mathbf{W}^T\mathbf{W}=\mathbf{I}_c} \text{Tr}[\mathbf{W}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{W}] + \beta\|\mathbf{W}\|_{2,1} \tag{31}$$

However, it is unreasonable to require the the feature selection projection matrix to be orthogonal since feature weight vectors are not necessarily orthogonal with each other in nature. If we set $\alpha \to +\infty$, $\gamma = 0$, $p = 1$ and do not consider the nonnegative and orthogonal constraint, we have $\mathbf{F} = \mathbf{X}^T\mathbf{W}$ and the proposed formulation becomes the problem of UDFS without the orthogonal constraint. In this extreme case, $\mathbf{F}$ is enforced to be linear, i.e., $\mathbf{F} = \mathbf{X}^T\mathbf{W}$. However, as indicated in [40], it is likely that $\mathbf{F}$ is nonlinear in many applications. Hence, NSCR is superior to UDFS due to its flexibility of linearity and the general sparse model. Additionally, $\mathbf{F}$ is constrained to be nonnegative, making it more accurate than the one with mixed signs. Therefore, NSCR is more capable of selecting discriminative features.

Connection With SPFS: SPFS [15] performs feature selection by preserving sample similarity, which is formulated as:

$$\min_{\|\mathbf{W}\|_{2,1}\le\tau} \sum_{i,j=1}^{n} (\mathbf{x}_i^T\mathbf{W}\mathbf{W}^T\mathbf{x}_j - S_{ij})^2 \tag{32}$$

Here $\tau(\tau > 0)$ is a hyper-parameter. In the proposed NSCR, when $\alpha \to +\infty$, and the orthogonal and nonnegative constraints are removed, we have $\mathbf{F} = \mathbf{X}^T\mathbf{W}$. Then, with $\gamma = 0$ and $p = 1$, NSCR becomes the above optimization problem.

Connection With NDFS: NDFS [30] is our preliminary version, which does not explicitly exploit the redundancy between features. NDFS is a special case of the proposed NSCR algorithm with $\lambda = 0$ and $p = 1$.

V. Experiments

In this section, we experimentally evaluate the performance of the proposed NSCR method for unsupervised feature selection, which can be applied to many applications, such as clustering and classification. Following previous unsupervised feature selection work [11], [22], we only evaluate the performance of NSCR and compared with representative algorithms in terms of clustering. In our experiments, we first select the top r features and then utilize Kmeans algorithm to cluster images based on the selected features.

A. Data Sets

The experiments are conducted on 9 publicly available image datasets, including four face image data sets, i.e., UMIST [47], AT&T [48], JAFFE [49] and Pointing4 [50], three handwritten digit data sets, i.e., MNIST used in [50], Binary Alphabet (BA) [47] and a subset of USPS with 40 samples randomly selected for each class [47], and two object image databases, i.e., COIL20 [51] and Caltech101 [52]. Data sets from different areas serve as a good test bed for a comprehensive evaluation.

The UMIST face image database [47] contains 575 multi-view gray scale face images with the size of 28 × 23 belonging to 20 different people, covering a broad range of poses from profile to frontal views. In the AT&T face image dataset [48], there are 10 gray scale images for each of the 40 human objects. There were taken at different times, varying the lighting, facial expressions and facial details. The image size is 32 × 32. The Japanese Female Facial Expression (JAFFE) database has 213 images with the size 26 × 26 of different facial expressions conducted by 10 Japanese female models [49]. For the Pointing4 face database used in [50], there are 15 sets of images in total with the size of 40 × 28. Each set contains two series of 93 images of the same person. In our experiment, each face is represented by a 1120D normalized feature vector. For the MNIST database, the first part of the test set of the MNIST database used in [50] is utilized in our experiments, which consists of 5000 images of handwritten numbers with each digital number having 500 images. The Binary Alphadigits [47] contains 1404 images, which are binary 20 × 16 digits of "0" through "9" and capital "A" through "Z", and each class has 39 images. The USPS database [47] contains gray-scale handwritten digit images of "0" through "9". A subset with 40 images randomly selected for each class is used. The image size is 16 × 16. The COIL20 [51] database contains 32 × 32 gray scale images of 20 objects viewed from varying angles and each object has 72 images. The Caltech101 dataset [52] contains 9144 images of 101 classes and an additional class of background images. In our experiments, we select the 10 largest categories, except the BACKGROUND_GOOGLE category. The SIFT descriptor is extracted and then 1000-dimensional bag of visual word is generated to represent each image. Table I summarizes the details of these 9 benchmark data sets used in the experiments in terms of the total number n of images, the total number c of clusters and the feature dimension d.

TABLE I

Image Dataset Description. n Is the Number of Images; d Denotes the Dimension of Features; c Is the Number of Clusters

Domain	Dataset	n	d	c
Face	UMIST	575	644	20
	AT&T	400	1024	40
	JAFFE	213	676	10
	Poingting4	2790	1120	15
Handwritten Digits	MNIST	5000	784	10
	BA	1404	320	36
	USPS	400	256	10
Object	Coil20	1440	1024	20
	Caltech101	3379	1000	10

B. Compared Scheme

To validate the effectiveness of the proposed NSCR for feature selection, we compare it with one baseline and several unsupervised feature selection methods. The compared algorithms are enumerated as follows.

1) **Baseline**: All original visual features are adopted.
2) **MaxVar**: Features corresponding to the maximum variance are selected to obtain the expressive features.
3) **LS** [5]: Features consistent with Gaussian Laplacian matrix are selected to preserve the local manifold structure.
4) **SPEC** [8]: Features are selected using spectral regression based on pairwise image similarity.
5) **SPFS-SFS** [15]: The traditional forward search strategy is utilized for similarity preserving feature selection in the SPFS framework.
6) **MCFS** [11]: Features are selected based on spectral analysis and sparse regression in a two-step scheme;
7) **UDFS** [22]: Features are selected by exploiting the local structure for local discriminative information and row-sparse models for feature correlations simultaneously.
8) **NDFS** [30]: Discriminative features are selected by a joint framework of nonnegative spectral analysis and linear regression with $\ell_{2,1}$-norm regularization.
9) **SCR**: A special case of the proposed method without considering the redundancy constraint for unsupervised feature selection, i.e., $\gamma = 0$.
10) **NSCR**: The proposed method with Nonnegative Spectral analysis and Controlled Redundancy for unsupervised feature selection.

C. Evaluation Metrics

With the selected features, we evaluate the performance in terms of clustering by two widely used evaluation metrics, i.e., Accuracy (ACC) and Normalized Mutual Information (NMI).

TABLE II

THE RANK OF FEATURES BY DIFFERENT METHODS ON THE CORRAL DATA. FEATURES ARE SELECTED FROM LEFT TO RIGHT, TOP TO BOTTOM

	MaxVar	LS	SPEC	SPFS-SFS	MCFS	UDFS	NDFS	SCR	NSCR
Rank	R, $A0$, $A1$, $B0$, $B1$, I	R, $A0$, $B0$, $A1$, $B1$, I	R, $A0$, $A1$, $B0$, $B1$, I	R, $B1$, $A0$, $B0$, $A1$, I	R, $A0$, $B0$, $A1$, $B1$, I	$A1$, R, $B0$, $A0$, $B1$, I	$B1$, $B0$, R, $A0$, $A1$, I	$B0$, $B1$, $A1$, R, $A0$, I	$B1$, $B0$, $A1$, $A0$, R, I

TABLE III

CLUSTERING RESULTS COMPARISON ON THE FACE DATA SETS. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD

Dataset	Baseline	MaxVar	LS	SPEC	SPFS-SFS	MCFS	UDFS	NDFS	SCR	NSCR
	ACC± std (%)									
UMIST	41.3 ± 2.9	45.6 ± 4.9	46.1 ± 1.9	48.5 ± 4.4	47.7 ± 2.9	47.5 ± 3.6	49.5 ± 3.2	52.4 ± 2.6	54.6 ± 1.1	$\mathbf{56.7 \pm 2.6}$
AT&T	50.8 ± 3.1	46.5 ± 3.2	47.6 ± 2.2	50.7 ± 3.2	51.9 ± 2.5	52.9 ± 2.9	52.3 ± 2.1	54.1 ± 2.7	53.7 ± 1.9	$\mathbf{56.2 \pm 1.3}$
JAFFE	73.6 ± 9.2	72.2 ± 5.1	74.9 ± 6.4	75.6 ± 5.8	76.6 ± 6.3	77.4 ± 6.1	77.9 ± 4.9	81.0 ± 7.1	82.2 ± 4.2	$\mathbf{82.9 \pm 4.0}$
Pointing4	35.6 ± 1.7	44.8 ± 2.0	38.0 ± 1.7	38.9 ± 2.5	39.3 ± 1.1	47.2 ± 1.9	45.4 ± 3.0	49.5 ± 2.6	50.1 ± 1.8	$\mathbf{52.3 \pm 1.1}$
	NMI± std (%)									
UMIST	62.7 ± 1.8	63.7 ± 3.7	64.1 ± 1.2	67.6 ± 1.9	62.7 ± 2.2	67.1 ± 2.3	69.3 ± 3.5	70.4 ± 1.8	71.6 ± 1.5	$\mathbf{73.4 \pm 2.5}$
AT&T	71.2 ± 1.8	68.9 ± 2.4	69.5 ± 1.4	71.4 ± 1.6	72.6 ± 1.5	73.5 ± 1.5	72.0 ± 1.9	73.7 ± 1.8	71.3 ± 0.6	$\mathbf{74.5 \pm 1.7}$
JAFFE	80.8 ± 5.9	74.7 ± 4.1	81.2 ± 4.4	82.3 ± 4.4	83.1 ± 3.8	81.7 ± 4.5	82.1 ± 3.8	85.5 ± 5.5	86.2 ± 2.2	$\mathbf{87.7 \pm 3.1}$
Pointing4	41.4 ± 1.0	51.6 ± 1.6	55.3 ± 1.5	56.3 ± 1.4	42.7 ± 1.2	55.8 ± 1.6	52.4 ± 1.6	56.4 ± 1.4	57.0 ± 1.5	$\mathbf{57.9 \pm 1.0}$

The larger ACC and NMI are, the better performance is. ACC is defined by

$$\mathrm{ACC} = \frac{1}{n}\sum_{i=1}^{n} \delta(c_i, \mathrm{map}(g_i)), \tag{33}$$

where c_i is the clustering label and g_i is the ground truth label of $\mathbf{x}^i$. $\mathrm{map}(g_i)$ is the optimal mapping function that permutes clustering labels and the ground truth labels. The optimal mapping can be obtained by using the Kuhn-Munkres algorithm. $\delta(c_i, g_i)$ is an indicator function that equals to 1 if $c_i = g_i$ and equals to 0 otherwise. NMI is defined as

$$\mathrm{NMI} = \frac{\sum_{l,h=1}^{c} t_{l,h} \log(\frac{n \times t_{l,h}}{t_l \hat{t}_h})}{\sqrt{(\sum_{l=1}^{c} t_l \log \frac{t_l}{n})(\sum_{h=1}^{c} \hat{t}_h \log \frac{\hat{t}_h}{n})}}, \tag{34}$$

where t_l is the number of samples in the l-th cluster $\mathcal{C}_l$ according to clustering results and $\hat{t}_h$ is the number of samples in the h-th ground truth class $\mathcal{G}_h$. $t_{l,h}$ is the number of overlap between $\mathcal{C}_l$ and $\mathcal{G}_h$.

To better indicate the effectiveness of the proposed redundancy control scheme, methods are also compared on redundancy rate (R), which is measured by

$$\mathrm{R} = \frac{1}{r(r-1)} \sum_{i,j \in \mathcal{S}, i<j} \rho_{i,j}, \tag{35}$$

where $\mathcal{S}$ is the set of the selected r features, and $\rho_{i,j}$ is the correlation between the i-th and j-th features. If many selected features are strongly correlated, the corresponding value of R is large and redundancy exists in the selected features.

D. Parameter Setting

In the compared methods, there are some hyper-parameters to be set in advance. For LS, SPEC, MCFS, UDFS, NDFS and NSCR, the k-nn graph should be constructed and k is set to 5 for all the datasets to specify the size of neighborhoods. To guarantee the orthogonality satisfied, we fix $\lambda = 10^8$ for NDFS and NSCR in our experiments. To fairly compare different unsupervised feature selection algorithms, we tune the parameters for all methods by a "grid-search" strategy from $\{10^{-6}, 10^{-4}, \cdots, 10^{6}\}$. The numbers of selected features are set as $\{5, 10, 20, 30, 50, 100, 150, 200, 300\}$ for all the datasets except USPS. Due to that the total number of features in USPS is 256, we set the number of selected features as $\{5, 10, 20, 30, 50, 80, 110, 140, 170, 200\}$. For all the algorithms, we report the the best clustering results from the optimal parameters. Different parameters may be used for different databases. In our experiments, we adopt Kmeans algorithm to cluster samples based on the selected features. The performance of Kmeans clustering depends on initialization. Following [11], [22], we repeat the clustering 20 times with random initialization for each setup. The average results with standard deviation (std) are reported. In real applications, it is impossible to tune parameters using the "grid-search" strategy. But it is an acceptable method to tune parameters for experimental comparisons since all the compared methods are with the well-chosen parameter values. The parameter sensitivity study and convergence study for NSCR will be shown in the following subsection.

E. Results on Synthetic Data

To well evaluate the effectiveness of the proposed NSCR method on the feature selection task, we conduct experiments on one widely used synthetic dataset, i.e., Corral [36]. It contains six Boolen features ($A0$, $A1$, $B0$, $B1$, I, R), in which the relevant features, irrelevant features and redundant features are provided. Specifically, the class labels of data points are defined by $(A0 \wedge A1) \vee (B0 \wedge B1)$ while $A0$, $A1$, $B0$ and $B1$ are independent to each other. Feature R is redundant by matching the class label 75% of the time, while feature I is uniformly random. That is, features $A0$, $A1$, $B0$ and $B1$ are necessary features, feature R is the redundant feature while feature I is the noisy feature. The results of features ranked by different methods are presented in Table II. For each method, features are selected from left to right, top to bottom. It can be seen that if the top 4 features are selected, only the proposed method can remove the noisy feature and

TABLE IV

CLUSTERING RESULTS COMPARISON ON THE HANDWRITTEN DIGIT DATA SETS. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD

Dataset	Baseline	MaxVar	LS	SPEC	SPFS-SFS	MCFS	UDFS	NDFS	SCR	NSCR
	ACC± std (%)									
MNIST	50.1 ± 5.3	54.2 ± 3.1	52.9 ± 3.4	53.3 ± 4.7	54.9 ± 4.7	55.8 ± 3.7	57.3 ± 2.5	58.5 ± 3.0	58.7 ± 1.9	**60.8 ± 2.7**
BA	40.7 ± 1.7	41.7 ± 1.1	42.6 ± 2.3	41.8 ± 1.4	42.5 ± 2.1	42.3 ± 1.7	42.8 ± 1.9	43.3 ± 1.7	42.1 ± 1.0	**45.7 ± 1.6**
USPS	62.8 ± 5.1	64.1 ± 1.3	64.5 ± 5.0	65.9 ± 4.0	63.4 ± 3.3	65.9 ± 2.8	65.2 ± 3.6	67.6 ± 1.3	68.1 ± 1.5	**72.0 ± 2.6**
	NMI± std (%)									
MNIST	47.2 ± 2.4	48.1 ± 1.8	48.2 ± 1.5	49.1 ± 2.2	49.6 ± 2.1	50.4 ± 1.6	51.1 ± 1.4	52.3 ± 2.5	50.9 ± 1.3	**53.9 ± 2.4**
BA	56.4 ± 0.9	57.3 ± 0.6	58.1 ± 0.6	57.4 ± 0.8	57.6 ± 1.2	57.9 ± 1.0	58.1 ± 1.0	58.6 ± 1.0	57.3 ± 0.5	**60.0 ± 1.3**
USPS	58.5 ± 3.1	60.2 ± 0.8	58.9 ± 2.8	59.2 ± 1.8	56.5 ± 1.7	59.4 ± 2.4	60.1 ± 3.4	61.5 ± 2.6	61.8 ± 1.1	**63.9 ± 1.1**

TABLE V

CLUSTERING RESULTS COMPARISON ON THE OBJECT IMAGE DATA SETS. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD

Dataset	Baseline	MaxVar	LS	SPEC	SPFS-SFS	MCFS	UDFS	NDFS	SCR	NSCR
	ACC± std (%)									
COIL20	59.0 ± 5.7	58.4 ± 4.0	57.3 ± 3.0	61.0 ± 2.1	62.5 ± 2.6	61.7 ± 4.3	62.9 ± 2.6	63.8 ± 2.8	65.3 ± 3.6	**67.2 ± 1.5**
Caltech101	47.7 ± 3.8	40.1 ± 1.9	47.6 ± 2.3	48.6 ± 3.2	49.2 ± 2.7	50.2 ± 5.0	50.8 ± 2.9	51.8 ± 3.1	52.3 ± 1.8	**54.3 ± 2.4**
	NMI± std (%)									
COIL20	72.9 ± 2.8	70.5 ± 0.9	70.4 ± 1.1	74.1 ± 2.4	76.2 ± 1.6	74.7 ± 2.3	75.9 ± 1.1	77.1 ± 1.8	77.3 ± 1.7	**78.9 ± 1.2**
Caltech101	34.5 ± 4.5	35.0 ± 2.0	37.5 ± 2.7	36.3 ± 2.2	33.4 ± 2.3	38.0 ± 5.3	38.4 ± 2.5	39.4 ± 2.9	50.9 ± 1.3	**42.2 ± 1.4**

the redundant feature simultaneously while other methods fail to filter out the redundant feature, which demonstrates the effectiveness of the proposed method for feature selection.

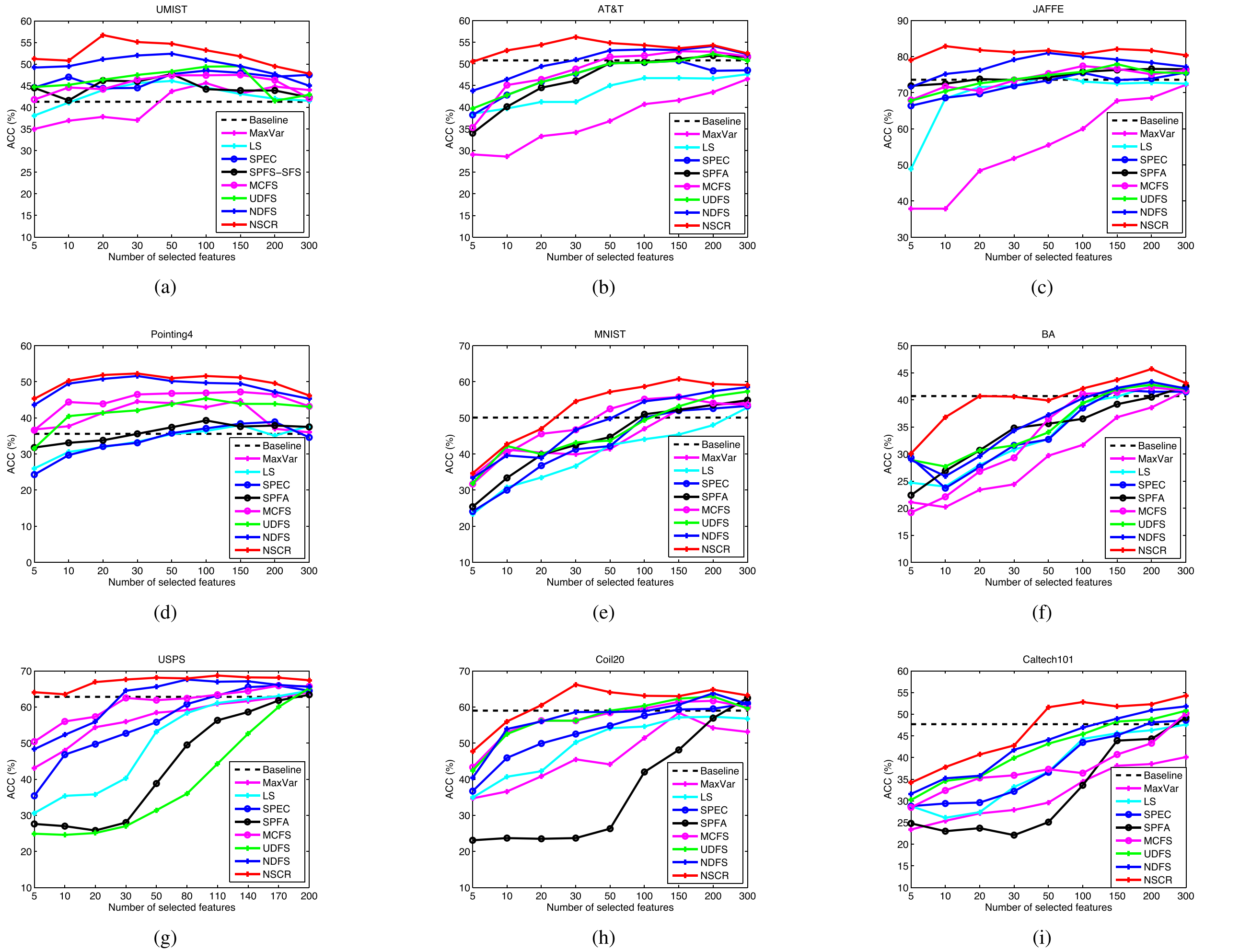
F. Performance Comparison on Benchmark Data

We now empirically evaluate the performance of these nine feature selection algorithms for clustering in terms of ACC and NMI. The detailed results on the face, handwritten digit and object data sets are summarized in Table III, Table IV and Table V, respectively. The results demonstrate that NSCR achieves the best performance on all the 9 image sets compared to other 8 feature selection algorithms, which validates its effectiveness.

From the above experimental results, we have the following observations. First, by comparing NDFS and NSCR, it is observed that NSCR achieves better performance than SCR by considering the nonnegative constraint while NSCR outperforms NDFS by explicitly considering the redundancy between features. It demonstrates that it is necessary and effective to introduce the nonnegative constraint and control the redundancy among the selected features. The improved NSCR enbles to remove the redundant features while preserving the necessary features. Second, NSCR and NDFS are both superior to MCFS by introducing the nonnegative constraint, which makes the scaled cluster indicators more accurate. They can select more necessary features. Third, NSCR, NDFS, UDFS and MCFS achieve larger values of ACC and NMI by exploiting discriminative information from data. It demonstrates that it is crucial to uncover the discriminative information in the unsupervised case, which can remove noisy features and indifferent features. Fourth, NSCR and NDFS achieve more accurate clustering performance than SPEC and MCFS. SPEC and MCFS adopt a two-step approach to introduce spectral analysis into feature selection while NDFS and NSCR are an one-step framework and perform spectral analysis and feature selection simultaneously. Fifth, by exploiting the local geometric structure of data distribution, LS, SPEC, MCFS, UDFS, NDFS and NSCR usually yield superior performance. Besides, it can be seen that it is necessary to select features jointly rather than one by one. The joint feature selection algorithms, such as MCFS, UDFS, NDFS and NSCR are always superior to the methods selecting features one after another, such as MaxVar and SPEC. Finally, compared with the baseline, it can be observed that feature selection is necessary and effective by removing the noise. It can not only reduce the number of features and make the algorithms more efficient, but also improve the performance. In conclusion, NSCR achieves the best performance on all data sets by exploiting nonnegative spectral analysis and redundancy between features simultaneously for feature selection, which can select necessary features, control the use of redundant features and discard noisy or indifferent features.

Besides, we also conduct experiments to evaluate the performance of the compared methods with different numbers of selected features and present the corresponding results in Fig. 1. From these results, the above conclusions can also be observed. Besides, we can see that the proposed NSCR method always achieves the best results with the smallest number of selected features. It is worth noting that the best results on the Caltech101 dataset are achieved with $r = 300$ when the number of the selected features is within $\{5, 10, 20, 30, 50, 100, 150, 200, 300\}$. It is because that the feature redundancy of this dataset is less than other datasets. NSCR directly controls the redundancy between features while exploits the discriminative information, which can guarantee that it is able to choose necessary features, control the redundancy between features and remove noisy or indifferent features. It is consistent with our motivation. Finally, the performance is comparatively sensitive to the number r of selected features, which is consistent with the observations in previous work [5], [8], [11], [22].

Experiments are also carried out to demonstrate the improvement of the proposed NSCR over the preliminary

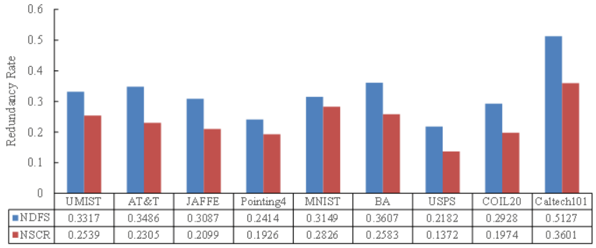


(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Fig. 1. The clustering performance in terms of ACC (%) with respect to the number of selected features on all the nine datasets of the compared nine methods. (a)–(i): the performance on the UMIST, AT&T, JAFFE, Pointing4, MNIST, BA, USPS, Coil20 and Caltech101 datasets, respectively. (Best viewed in color.)



	UMIST	AT&T	JAFFE	Pointing4	MNIST	BA	USPS	COIL20	Caltech101
■NDFS	0.3317	0.3486	0.3087	0.2414	0.3149	0.3607	0.2182	0.2928	0.5127
■NSCR	0.2539	0.2305	0.2099	0.1926	0.2826	0.2583	0.1372	0.1974	0.3601

Fig. 2. The averaged redundancy rate of NSCR and NDFS. (The lower the better.)

NDFS method in terms of redundancy. Figure 2 presents the averaged redundancy rates of the selected $r(r = 50)$ features by NSCR and NDFS. The results demonstrate that the feature subset selected by NSCR contain much less redundancy than the preliminary NDFS method, since the improved NSCR method explicitly controls the redundancy when selects features. This coincides with our motivation to remove redundancy and suggests that the redundancy control scheme in NSCR is effective.

G. Parameter Sensitivity

Like many other feature selection algorithms, our proposed NSCR method also requires several parameters α, β, γ and p to be set in advance. In our experiments, we observe that the parameters β and γ have more effect on the performance than the parameter α on the given datasets. Therefore, we focus on discussing the parameters β and γ. We will conduct the parameter sensitivity study in terms of β, γ and p in the rest of this subsection.

Let us first study the parameters β and γ. These two parameters are tuned within $\{10^{-6}, 10^{-4}, \cdots, 10^{4}, 10^{6}\}$. The results in terms of clustering accuracy on all the 9 image datasets are shown in Fig. 3. The results illustrate that the performance changes differently with respect to different data sets. It is well known that how to identify the optimal values of the hyper-parameters is data dependent and still an open problem. In the proposed method, the parameter β controls the row sparsity of the feature selection matrix. Noisy and correlative features can not be reduced with small β, while the informative features can be removed when β is very large. These can be verified with the observation in Fig. 3 that small β and large β both degrade the performance of the proposed method on the given datasets. With the suitable value of β, the learned feature selection matrix with row sparsity can reduce the noisy or indifferent features while select necessary or redundant features. On the other hand, the parameter γ

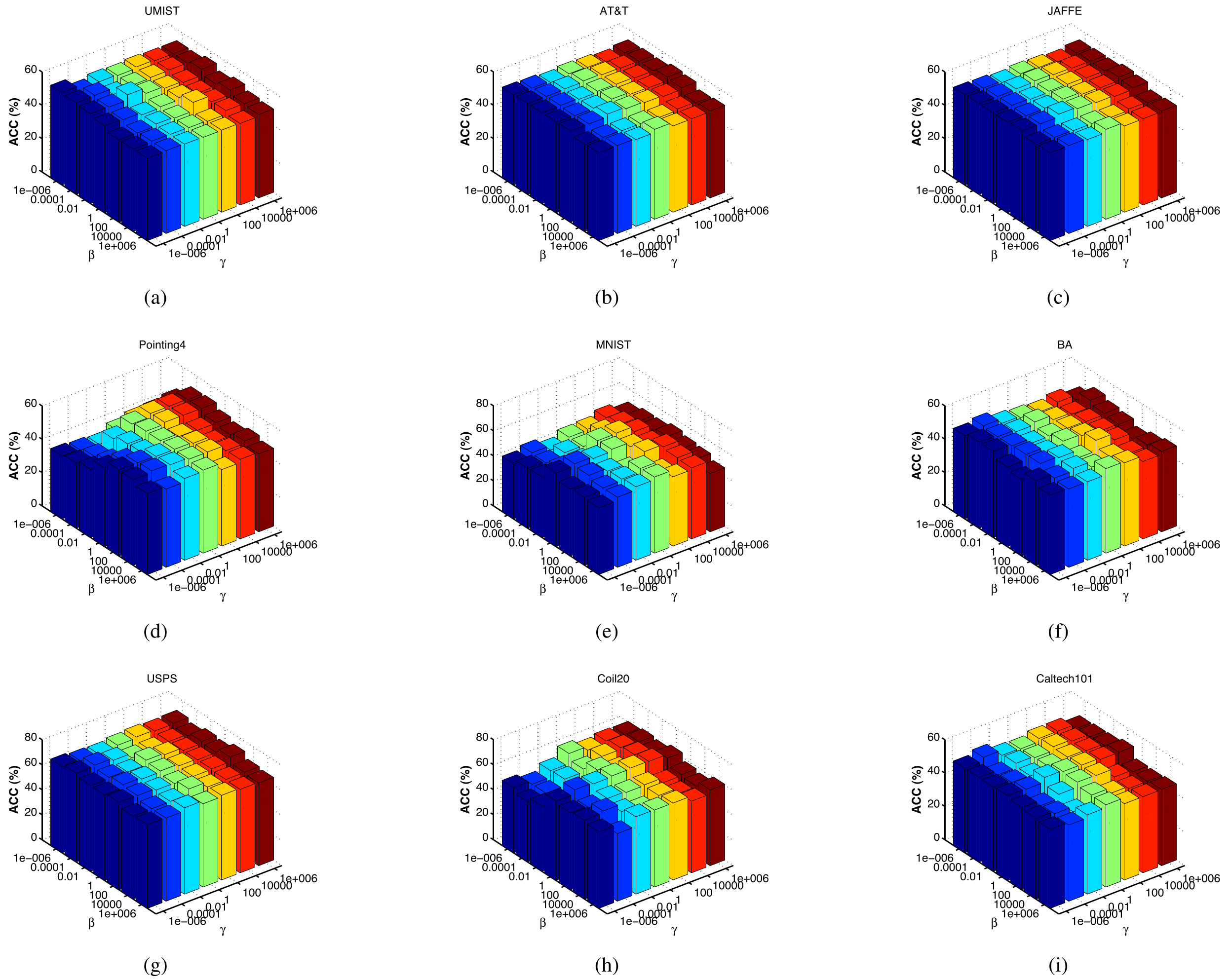


Fig. 3. Performance variation of the proposed method with respect to different values of the parameters β and γ. (a)–(i): the parameter sensitiveness on the UMIST, AT&T, JAFFE, Pointing4, MNIST, BA, USPS, Coil20 and Caltech101 datasets, respectively. (Best viewed in color.)

actually controls the redundancy between features. From the results in Fig. 3, it is observed that the best results are always achieved on the given datasets when γ is in the middle interval of the tuned range. When it is not too large or small, the performance is empirically good, which indicates that it is necessary to control the redundancy between the selected features.

Next, we study the performance variation of the proposed approach with respect to different values of p in the mixed norm $\ell_{2,p}$ of the sparse regularization term. p is tuned within $\{0.1, 0.25, 0.5, 0.75, 1\}$. The experimental results are shown in Fig. 4. The clustering accuracy comparisons demonstrate that the mixed norm $\ell_{2,p}$ $(0 < p \leq 1)$ matrix norms provide alternatives to $\ell_{2,1}$-norm. Besides, the best results are always achieved when $p = 0.5$, and $p = 0.5$ empirically outperforms $p = 1$ for choosing better sparse patterns in various situations, which is consistent with the observations in [44].

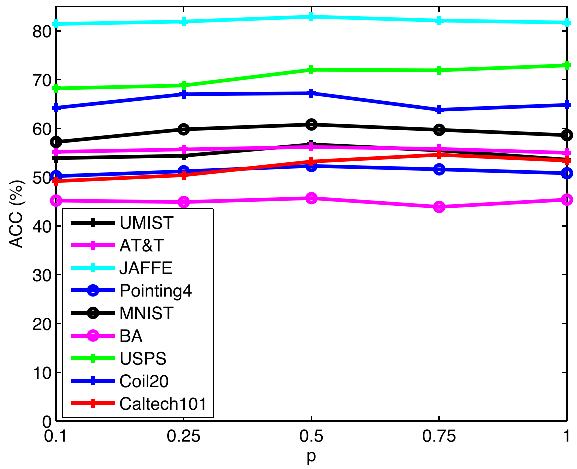
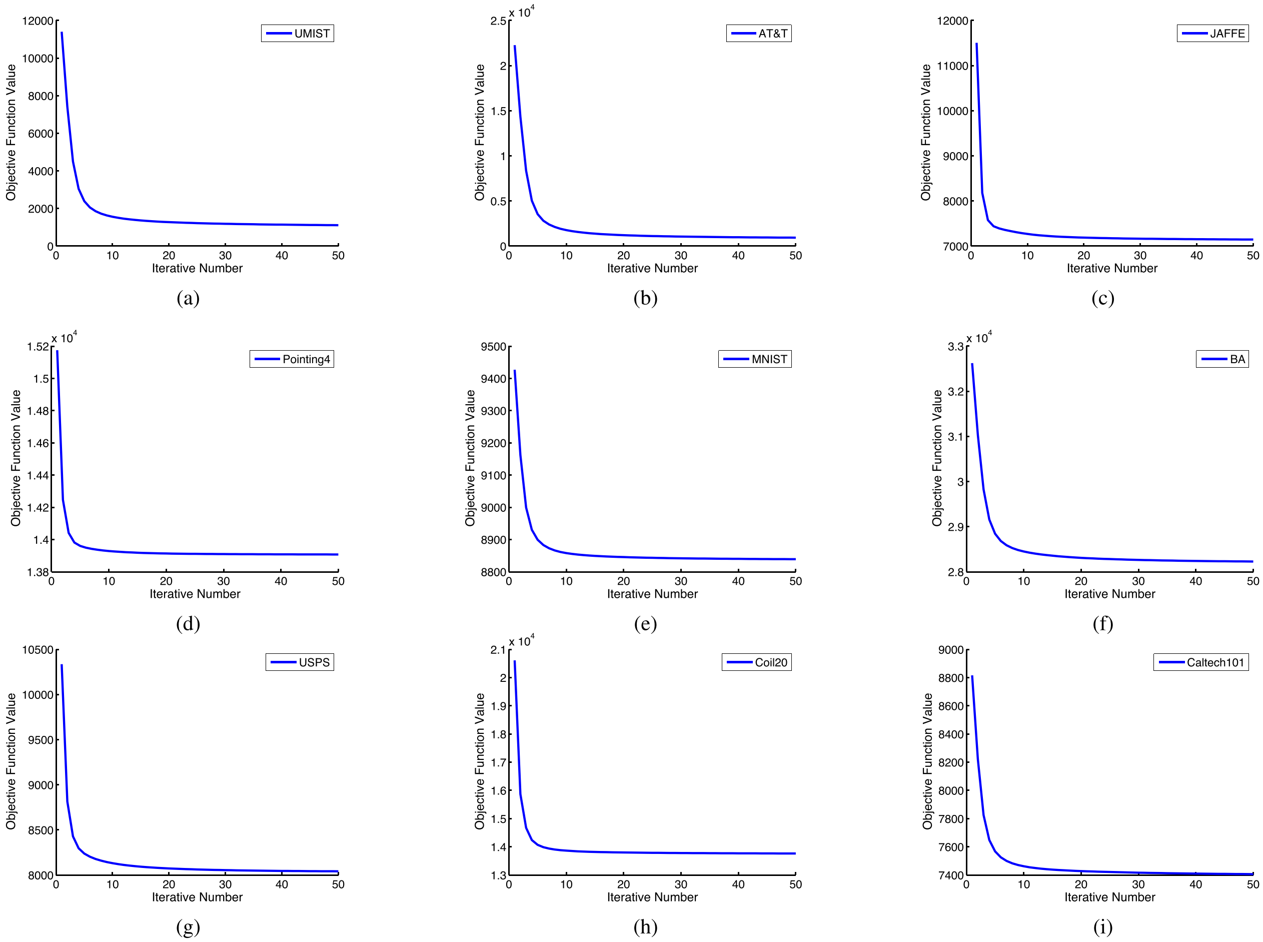


Fig. 4. Performance variation of the proposed method *w.r.t.* different values of p in the mixed norm $\ell_{2,p}$. (Best viewed in color.)

H. Convergence Study

To solve the proposed formulation, we propose an iterative update algorithm to optimize the proposed formulation. Now we experimentally study the speed of convergence of the proposed NSCR. The convergence curves on all the 9 image data sets are shown in Fig. 5. From these figures, we can see that our algorithm converges within 20 iterations for all the data sets, demonstrating that the proposed optimization algorithm is effective and converges quickly.



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Fig. 5. Convergence curve of the proposed feature selection algorithm NSCR on all the nine image datasets. (a)–(i): the convergence curve on the UMIST, AT&T, JAFFE, Pointing4, MNIST, BA, USPS, Coil20 and Caltech101 datasets, respectively.

VI. Conclusion

In this paper, we propose a novel unsupervised feature selection approach, which jointly exploits nonnegative spectral analysis and explicitly control the redundancy between features while a general sparse model with the $\ell_{2,p}$-norm is introduced. The proposed method can select necessary features, remove noisy or indifferent features while control the redundancy between the selected features. The cluster indicators learned by nonnegative spectral clustering are used to provide label information for unsupervised feature selection. To facilitate feature selection, the predictive matrix is constrained to be sparse in rows. By imposing the $\ell_{2,p}$-norm regularization, our methods jointly selects the most discriminative features across the entire feature space. To solve the proposed formulation, we develop an iterative optimization algorithms. Extensive experiments on 9 real-world image data sets are conducted to validate the effectiveness of the proposed method. For future work, we will focus on extending our methods in the kernel learning framework and the local learning framework. Besides, how to select the adaptive hyper-parameters and the number of selected features are also our directions for future research.

References

[1] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[2] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York, NY, USA: Wiley, 2001.

[3] L. Wolf and A. Shashua, "Feature selection for unsupervised and supervised inference: The emergence of sparsity in a weight-based approach," *J. Mach. Learn. Res.*, vol. 6, pp. 1855–1887, Jan. 2005.

[4] J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio, and V. Vapnik, "Feature selection for SVMs," in *Proc. Adv. Neural Inf. Process. Syst.*, 2000, pp. 668–674.

[5] X. He, D. Cai, and P. Niyogi, "Laplacian score for feature selection," in *Proc. Adv. Neural Inf. Process. Syst.*, 2005, pp. 507–514.

[6] W. Jiang, G. Er, Q. Dai, and J. Gu, "Similarity-based online feature selection in content-based image retrieval," *IEEE Trans. Image Process.*, vol. 15, no. 3, pp. 702–712, Mar. 2006.

[7] F. Nie, H. Huang, X. Cai, and C. H. Ding, "Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2010, pp. 1813–1821.

[8] Z. Zhao and H. Liu, "Spectral feature selection for supervised and unsupervised learning," in *Proc. Int. Conf. Mach. Learn.*, 2007, pp. 1151–1157.

[9] Z. Zhu, Y.-S. Ong, and M. Dash, "Markov blanket-embedded genetic algorithm for gene selection," *Pattern Recognit.*, vol. 40, no. 11, pp. 3236–3248, Nov. 2007.

[10] K.-Q. Shen, C.-J. Ong, X.-P. Li, and E. P. V. Wilder-Smith, "Feature selection via sensitivity analysis of SVM probabilistic outputs," *Mach. Learn.*, vol. 70, no. 1, pp. 1–20, Jan. 2008.

[11] D. Cai, C. Zhang, and X. He, "Unsupervised feature selection for multi-cluster data," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2010, pp. 333–342.

[12] J.-B. Yang and C.-J. Ong, "Feature selection using probabilistic prediction of support vector regression," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 22, no. 6, pp. 954–962, Jun. 2011.

[13] L. Song, A. Smola, A. Gretton, J. Bedo, and K. Borgwardt, "Feature selection via dependence maximization," *J. Mach. Learn. Res.*, vol. 13, no. 1, pp. 1393–1434, Jan. 2012.

[14] L. Zhang, C. Chen, J. Bu, and X. He, "A unified feature and instance selection framework using optimum experimental design," *IEEE Trans. Image Process.*, vol. 21, no. 5, pp. 2379–2388, May 2012.

[15] Z. Zhao, L. Wang, H. Liu, and J. Ye, "On similarity preserving feature selection," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 3, pp. 619–632, Mar. 2013.

[16] Z. Li, J. Liu, Y. Yang, X. Zhou, and H. Lu, "Clustering-guided sparse structural learning for unsupervised feature selection," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 9, pp. 2138–2150, Sep. 2014.

[17] H. C. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.

[18] Z. Zhao, L. Wang, and H. Liu, "Efficient spectral feature selection with minimum redundancy," in *Proc. Nat. Conf. Artif. Intell.*, 2010, pp. 673–678.

[19] R. Chakraborty and N. R. Pal, "Feature selection using a neural framework with controlled redundancy," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 1, pp. 35–50, Jan. 2015.

[20] P. Mitra, C. A. Murthy, and S. K. Pal, "Unsupervised feature selection using feature similarity," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 3, pp. 301–312, Mar. 2002.

[21] H. Zeng and Y.-M. Cheung, "Feature selection and kernel learning for local learning-based clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1532–1547, Aug. 2011.

[22] Y. Yang, H. T. Shen, Z. Ma, Z. Huang, and X. Zhou, "$\ell_{2,1}$-norm regularized discriminative feature selection for unsupervised learning," in *Proc. Int. Joint Conf. Artif. Intell.*, 2011, pp. 1589–1954.

[23] Z. Ma, F. Nie, Y. Yang, J. R. R. Uijlings, N. Sebe, and A. G. Hauptmann, "Discriminating joint feature analysis for multimedia data understanding," *IEEE Trans. Multimedia*, vol. 14, no. 6, pp. 1662–1672, Dec. 2012.

[24] A. Jain and D. Zongker, "Feature selection: Evaluation, application, and small sample performance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 2, pp. 153–158, Feb. 1997.

[25] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, Jan. 2003.

[26] Y.-M. Cheung and H. Zeng, "Local kernel regression score for selecting features of high-dimensional data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 12, pp. 1798–1802, Dec. 2009.

[27] J. G. Dy and C. E. Brodley, "Feature selection for unsupervised learning," *J. Mach. Learn. Res.*, vol. 5, pp. 845–889, Jan. 2004.

[28] M. H. C. Law, M. A. T. Figueiredo, and A. K. Jain, "Simultaneous feature selection and clustering using mixture models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 9, pp. 1154–1166, Sep. 2004.

[29] C. Constantinopoulos, M. K. Titsia, and A. Likas, "Bayesian feature and model selection for Gaussian mixture models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 6, pp. 1013–1018, Jun. 2006.

[30] Z. Li, Y. Yang, J. Liu, X. Zhou, and H. Lu, "Unsupervised feature selection using nonnegative spectral analysis," in *Proc. Nat. Conf. Artif. Intell. (AAAI)*, 2012, pp. 1026–1032.

[31] L. Du, Z. Shen, X. Li, P. Zhou, and Y.-D. Shen, "Local and global discriminative learning for unsupervised feature selection," in *Proc. IEEE Int. Conf. Data Mining*, Dec. 2013, pp. 131–140.

[32] M. Qian and C. Zhai, "Robust unsupervised feature selection," in *Proc. Int. Joint Conf. Artif. Intell.*, 2014, pp. 1621–1627.

[33] L. Shi, L. Du, and Y.-D. Shen, "Robust spectral learning for unsupervised feature selection," in *Proc. IEEE Int. Conf. Data Mining*, Dec. 2014, pp. 977–982.

[34] Z. Li, J. Liu, J. Tang, and H. Lu, "Robust structured subspace learning for data representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 10, pp. 2085–2098, Oct. 2015.

[35] J. Tang, Z. Li, M. Wang, and R. Zhao, "Neighborhood discriminant hashing for large-scale image retrieval," *IEEE Trans. Image Process.*, vol. 24, no. 9, pp. 2827–2840, Sep. 2015.

[36] L. Yu and H. Liu, "Efficient feature selection via analysis of relevance and redundancy," *J. Mach. Learn. Res.*, vol. 5, no. 10, pp. 1205–1224, 2004.

[37] G. Qu, S. Hariri, and M. Yousif, "A new dependency and correlation analysis for features," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 9, pp. 1199–1207, Sep. 2005.

[38] L. Zhou, L. Wang, and C. Shen, "Feature selection with redundancy-constrained class separability," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 21, no. 5, pp. 853–858, May 2010.

[39] Q. Song, J. Ni, and G. Wang, "A fast clustering-based feature subset selection algorithm for high-dimensional data," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 1, pp. 1–14, Jan. 2013.

[40] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, Aug. 2000.

[41] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Proc. Adv. Neural Inf. Process. Syst.*, 2001, pp. 849–856.

[42] R. Chartrand, "Exact reconstruction of sparse signals via nonconvex minimization," *IEEE Signal Process. Lett.*, vol. 14, no. 10, pp. 707–710, Oct. 2007.

[43] X. Chen, F. Xu, and Y. Ye, "Lower bound theory of nonzero entries in solutions of ℓ_2-ℓ_p minimization," *SIAM J. Sci. Comput.*, vol. 32, no. 5, pp. 2832–2852, 2010.

[44] L. Wang, S. Chen, and Y. Wang, "A unified algorithm for mixed $\ell_{2,p}$-minimizations and its application in feature selection," *Comput. Optim. Appl.*, vol. 58, no. 2, pp. 409–421, Jun. 2014.

[45] H. Kuhn and A. Tucker, "Nonlinear programming," in *Proc. Berkeley Symp. Math. Statist. Probab.*, 1951, pp. 481–492.

[46] Z. Yang, T. Hao, O. Dikmen, X. Chen, and E. Oja, "Clustering by nonnegative matrix factorization using graph random walk," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1079–1087.

[47] *Data for MATLAB Hackers*. [Online]. Available: http://cs.nyu.edu/~roweis/data.html, accessed Oct. 16, 2012.

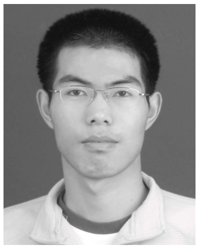
[48] F. S. Samaria and A. C. Harter, "Parameterisation of a stochastic model for human face identification," in *Proc. IEEE Workshop Appl. Comput. Vis.*, Dec. 1994, pp. 138–142.

[49] M. J. Lyons, J. Budynek, and S. Akamatsu, "Automatic classification of single facial images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, no. 12, pp. 1357–1362, Dec. 1999.

[50] Y. Yang, D. Xu, F. Nie, S. Yan, and Y. Zhuang, "Image clustering using local discriminant models and global integration," *IEEE Trans. Image Process.*, vol. 19, no. 10, pp. 2761–2773, Oct. 2010.

[51] S. A. Nene, S. K. Nayar, and H. Murase, "Columbia object image library (COIL-20)," Dept. Comput. Sci., Columbia Univ., New York, NY, USA, Tech. Rep. CUCS-005-96, 1996.

[52] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories," in *Proc. IEEE Comput. Vis. Pattern Recognit. Workshop*, Jun. 2004, pp. 178–186.



Zechao Li (M'10) received the B.E. degree from the University of Science and Technology of China, Anhui, China, in 2008, and the Ph.D. degree in pattern recognition and intelligent systems from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, in 2013. He is currently an Assistant Professor with the School of Computer Science and Engineering, Nanjing University of Science and Technology, China. His research interests include machine learning, subspace learning, and multimedia understanding. He is a member of ACM. He received the 2013 President Scholarship of the Chinese Academy of Science and the 2015 Excellent Doctoral Dissertation of the Chinese Academy of Sciences.

Jinhui Tang (SM'14) received the B.E. and Ph.D. degrees from the University of Science and Technology of China, in 2003 and 2008, respectively. From 2008 to 2010, he was a Research Fellow with the School of Computing, National University of Singapore. During that period, he visited the School of Information and Computer Science, UC Irvine, in 2010, as a Visiting Research Scientist. From 2011 to 2012, he visited Microsoft Research Asia, as a Visiting Researcher. He is currently a Professor with the School of Computer Science and Engineering, Nanjing University of Science and Technology. He has authored over 80 journal and conference papers in his research areas. His current research interests include large-scale multimedia search, social media mining, and computer vision. He was a co-recipient of the best paper award in ACM Multimedia 2007, PCM 2011, and ICIMCS 2011. He is a member of ACM. He serves as an Editorial Board Member of *Pattern Analysis and Applications*, *Multimedia Tools and Applications*, *Information Sciences*, and *Neurocomputing*, a Technical Committee Member for about 30 international conferences, and a Reviewer for about 30 prestigious international journals.