IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 12, 2018, accepted October 28, 2018, date of publication October 31, 2018, date of current version December 3, 2018.

*Digital Object Identifier 10.1109/ACCESS.2018.2878855*

# Unsupervised Feature Selection With Ordinal Preserving Self-Representation

**JIANGYAN DAI[1], YUQI CHEN[2], YUGEN YI[2], JINING BAO[3], LEI WANG[1], WEI ZHOU[3], AND GANG LEI[2]**

[1]School of Computer Engineering, Weifang University, Weifang 261061, China
[2]School of Software, Jiangxi Normal University, Nanchang 330022, China
[3]Northeastern University, Shenyang 110004, China

Corresponding authors: Yugen Yi (yiyg510@jxnu.edu.cn) and Jining Bao (yiyinbaobao@126.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61602221, Grant 41661083, Grant 61602222 and Grant 61762050, in part by the Project of Shandong Province Higher Educational Science and Technology Program under Grant J16LN68, Grant J16LB10, and Grant J18KA130, in part by the Shandong Province Natural Science Foundation under Grant ZR2017QF011, in part by the Weifang Science and Technology Development Plan Project under Grant 2017GX006 and Grant 2018GX009, in part by the Project of Doctoral Foundation of Weifang University under Grant 2015BS10, and in part by the Natural Science Foundation of Jiangxi Province under Grant 20171BAB212009.

**ABSTRACT** Unsupervised feature selection is designed to select an optimal feature subset without any label information from high-dimensional data, which is implemented by eliminating the irrelevant and redundant features and has been attracted widespread attention in recent years. Specifically, the obtained low-dimensional representation is interpretable that is useful to machine learning applications. In this paper, we propose a novel unsupervised feature selection algorithm, namely ordinal preserving self-representation (OPSR) for image classification and clustering. First, each feature in high-dimensional data is represented by the linear combination of other features. Then, the topology information is introduced into the objective function for utilizing the ordinal locality of high-dimensional data adequately. At last, an efficient iteratively update algorithm is designed to solve the proposed OPSR, and its convergence is proved in detail. Extensive experimental results on six benchmark databases demonstrate that the effectiveness of the OPSR and its superiority also is verified by comparing with some state-of-the-art feature selection algorithms.

**INDEX TERMS** Dimensionality reduction, unsupervised feature selection, ordinal locality-preserving, self-representation.

## I. INTRODUCTION

With the rapid development of biological science and information technology, data dimensionality involved in data mining, machine learning, computer vision and bioinformatics is growing explosively [1]. High-dimensional data can not only degenerate the performance of many algorithms because of the dimensionality curse and the existence of irrelevant, redundant and noisy features, but also significantly aggravate the time and memory burden [2]. Therefore, effectively handling the high-dimensional data and accurately selecting useful information have become a challenging issue. Dimensionality reduction is one of the most popular techniques to learn a low-dimensional compact representation by reducing the dimensions of high-dimensional data, which can be classified into two categories, feature extraction and feature selection [3]–[6]. Feature extraction constructs the new feature space with lower dimensionality by reducing the original feature space. However, it is difficult to analyze the new features due to the fact that the transformed features generated by feature extraction have no physical meanings. In contrast, feature selection extracts a subset of most relevant features from the original feature set, maintaining the readability and interpretability of features effectively [7].

In the past decades, feature selection has drawn much attention [8]–[14]. According to the availability of label information, feature selection is divided into three categories, including supervised, semi-supervised and unsupervised. Supervised feature selection determines the discriminative features in terms of the correlations of features with label information, such as Fisher score (FS) [15], Spectral feature selection (SPEC) [16], robust feature selection via joint $\ell_{2,1}$-norms minimization (RFS) [17] and self-weighted supervised discriminative feature selection (SSD-FS) [18]. However, the supervised feature selection methods could

not work well without sufficient labels, which are obtained expensively and time-consuming for many practical applications. Semi-supervised feature selection makes the best of both limited labeled data and unlabeled data to select relevant features [19]–[22], such as semi-supervised local discriminant analysis (SELD) [19] and semi-supervised feature selection via rescaled linear regression (RLSR) [20]. Unsupervised feature selection relies on some criteria to select features that can accurately preserve the intrinsic structure of high-dimensional data without any label information. Recent study has demonstrated that preserving the local geometrical data structure is obviously more important than preserving global structure in unsupervised feature selection [23]–[26]. Yang *et al.* [23] selected the most discriminative features in batch mode with the manifold structure. Li *et al*. [24] utilized the discriminative information from spectral clustering and feature correlation simultaneously. Guo *et al.* [25] introduced the triplet-based loss function to enforce selected feature groups for preserving ordinal locality of original data. Luo *et al*. [26] characterized the local structure of each neighborhood by using adaptive reconstruction nonnegative weights graph. Chang *et al.* [27] proposed the convex sparse principal component analysis and applied it to feature learning task. Although most of traditional methods preserve the inherent structure by removing irrelevant or redundant features and keeping the better interpretation, they neglect the intrinsic relationship of features and are sensitive to noisy features.

In recent years, self-representation has been investigated for unsupervised feature selection and widely been developed in machine learning and computer vision applications [28]–[36]. The concept of self-representation is that each feature can be well approximated by a linear combination of other features which are relevant with each other. Hu *et al*. [28] suggested the unified framework by simultaneously exploiting feature-level self-representation, K-means method, low-rank subspace selection and row-sparsity. Li *et al.* [29] used the property of the data to construct self-representation coefficient matrix and discovered its sparse structure by sparse representation. Tang *et al.* [30] obtained the feature representation coefficient matrix by feature self-representation and preserved the local geometrical structure by learning the sample similarity graph automatically. Lei and Zhu [31] exploited the global manifold structure using low-rank constraint on self-representation coefficient matrix. Tang *et al.* [32] maintained both the global and local structures of data well owning to the unified unsupervised feature selection framework derived from feature self-representation and robust graph regularization. Seen from the aforementioned self-representation dimensionality reduction methods, the majority of them are robust to noisy data, and the local geometric structure of data is preserved very well. However, the ordinal locality of original data is not taken into account in them sufficiently.

In order to address the aforementioned issues, we present a novel unsupervised feature selection approach using the ordinal locality-preserving and the feature self-representation. Specifically, the main contributions of this paper are summarized as follows:

1) We introduce the topology information namely the ordinal locality into the self-representation model, which could well preserve the local geometrical structure and the ordinal locality information simultaneously.

2) We develop an efficient iterative optimization algorithm to optimize the objective function, and prove its convergence theoretically.

3) Compared with some well-known methods, extensive experimental results demonstrate the effectiveness and the superiority of the proposed OPSR approach on six benchmark databases.

The remaining of this paper is organized as follows. Section II describes the proposed unsupervised feature selection approach in detail and gives the corresponding objective function. Then, we design an iterative optimization algorithm to solve the objective function and prove its convergence in Section III. Section IV analyzes the experimental results adequately and the conclusions are given in Section V.

## II. THE PROPOSED METHOD

In this section, we first define the used notations in this paper and then elaborate the proposed OPSR model in detail.

### A. NOTATIONS

In this paper, the uppercase letters and the lowercase letters are used to represent matrices and vectors, respectively. $Tr(X)$, $X^T$ and $X^{-1}$ stand for the trace of a matrix $X$, the transpose of $X$ and the inverse of $X$, respectively. Given the data matrix $X = [x_1; x_2; \ldots; x_N] \in R^{N\times d}, x_i \in R^{1\times d}$ denotes the $i$-th sample consisting of d features, $N$ and $d$ represent the numbers of samples and features, respectively. Given a response matrix $Y = [y_1, y_2, \ldots, y_d] \in R^{N\times c}$, $c$ denotes the number of classes or targets.

### B. THE OPSR MODEL

In recent popular methods [33], the feature selection problem is regarded as a multi-output regression problem:

$$\min_{W} f(Y - XW) + \beta\phi(W) \tag{1}$$

where $f(Y - XW)$ is the loss function, $W \in R^{d\times c}$ indicates the feature weight matrix, $\phi(W)$ defines the regularization term which is represented with $W$, and $\beta$ denotes a positive constant.

Generally, the response matrix $Y$ is defined as the class labels for supervised feature selection. However, the class labels are difficult to obtain for some reasons, such as budget limitation or unavailable message. In contrast, unsupervised feature selection algorithms set the response matrix $Y$ with pseudo-class labels acquired by some criteria, and the $Y$ is calculated differently for different methods. Obviously, it is hard to generate the proper response matrix in many feature selection methods.

Feature self-representation has been explored in unsupervised feature selection recently, which assumes that each feature can be well characterized by the linear combination with all the other features and the representation coefficient matrix is sparse that is considered as feature weights. Specifically, the data matrix $X$ is applied to the response matrix, i.e., $Y = X$, which is more physical and explanatory. Therefore, the feature selection problem is formulated as:

$$\min_{W} \|X - XW\|_{2,1} + \beta \|W\|_{2,1} \tag{2}$$

where $W$ denotes the self-representation coefficient matrix, the first term is used to calculate the feature reconstruction error and is robust to outliers, and the second term is the regularization on $W$ which can guide to select the relevant feature subset.

Thus, we can obtain the relevant features effectively by the model in (2). However, the local geometric structure of data is neglected, which is especially important for dimensionality reduction. Although many researches have preserved the local geometric structure of data by improving their model, such as the SPNFSR suggested by Zhou *et al.* [34], they yet ignored the ranking information in each sample's neighborhood that is crucial for feature selection. The literatures [25] and [37] illustrated that the ordinal locality plays a very important role in describing the local data structure, which reflected the ranking information of each sample's neighborhood as well as the neighborhood relationship between data samples simultaneously. Inspired by the above descriptions, we incorporate the ordinal locality into the self-representation model as shown in (2).

Given an original sample $x_i$, its corresponding selected feature group is represented by $z_i = W^T x_i^T$, then $Z = W^T X^T$. The sample $x_i$ with its neighbors $x_u$ and $x_v$ form a triplet $(x_i, x_u, x_v)$, and their new selected features also comprise of a triplet $(z_i, z_u, z_v)$. Suppose that $dist(\cdot, \cdot)$ is taken as a distance metric, the ordinal locality preserving loss function is optimized in (3) to acquire appropriate features for each data sample.

$$\max_{Z} \sum_{i=1}^{N} \sum_{u \in \Omega_i} \sum_{v \in \Omega_i} C_{uv}^{i} \left[dist(z_i, z_u) - dist(z_i, z_v)\right] \tag{3}$$

where $\Omega_i$ is an index set of the $k$ nearest neighbors of $x_i$. $C^i$ represents an antisymmetric matrix consisting of $(u, v)^{th}$ element that is computed by $dist(x_i, x_u) - dist(x_i, x_v)$. Particularly, the weight matrix $S \in R^{N \times N}$ is defined as:

$$S_{ij} = \begin{cases} \sum\limits_{u \in \Omega_i} C_{uj} & j \in \Omega_i \\ 0 & j \notin \Omega_i \end{cases} \tag{4}$$

Combining (3) and (4) together, (3) is formulated as:

$$\min_{z} \sum_{i=1}^{N} \sum_{j=1}^{N} S_{ij} dist(z_i, z_j) \tag{5}$$

For the sake of convenience, squared Euclidean distance is utilized to compute each pairwise distance. Consequently, (5) is rewritten as:

$$\min_{z} \sum_{i=1}^{N} \sum_{j=1}^{N} S_{ij} \left\|z_i - z_j\right\|^2 \tag{6}$$

Therefore, the regularization of the ordinal locality [18] is briefly formed as:

$$\min_{Z} tr(ZLZ^T) \tag{7}$$

where $Z = [z_1, z_2, \ldots, z_N] \in R^{d \times N} = W^T X^T$ signifies the new selected feature matrix, $L \in R^{N \times N}$ is the Laplacian matrix with $L = D - S$, and $D$ is the diagonal matrix with $D_{ii} = \sum_j S_{ij}$.

As a result, combining (2) and (7), we generate the overall objective function, which is defined as:

$$\begin{aligned} &\min_{W} \|X - XW\|_{2,1} + \alpha tr(W^T X^T L X W) + \beta \|W\|_{2,1} \\ &s.t.\ W \geq 0 \end{aligned} \tag{8}$$

where the first term is the feature self-representation for describing the property that each sample can be well approximated by all other samples through the linear combination and it is robust to outliers. The second term is the regularization that adequately preserves the ordinal locality of original samples including the ranking information of each sample's neighborhood and neighborhood relationships between data samples. And the third term reflects the row-sparsity on the feature coefficient matrix $W$ for selecting the appropriate feature subset. $\alpha > 0$ and $\beta > 0$ are the two trade off parameters.

## III. OPTIMATION AND CONVERGENCE ANALYSIS

Seen from (8), the OPSR algorithm is non-smooth and we cannot obtain a closed-form solution because of the $l_{2,1}$-norm. Therefore, an efficient iterative updating algorithm is developed to optimize our proposed OPSR model and its convergence is proved accordingly.

### A. ITERATIVE UPDATING ALGORITHM

For a matrix $U \in R^{n \times d}$, there is $\|U\|_{2,1} = tr\left(U^T V U\right)$, where $V$ represents a diagonal matrix and its diagonal elements is denoted as $v_{ii} = 1/(2\|u_i\|_2)$[33]. Let $R = X^T L X$, then the (8) can be rewritten as the following:

$$\begin{aligned} \min_{W} tr\left((X - XW)^T P (X - XW)\right) \\ + \alpha tr\left(W^T R W\right) + \beta tr\left(W^T Q W\right) \end{aligned} \tag{9}$$

where $P$ and $Q$ signify two diagonal metrics and their corresponding diagonal elements are defined as:

$$p_{ii} = \frac{1}{2\|x_i - x_i W\|_2} \tag{10}$$

$$q_{ii} = \frac{1}{2\left\|w^i\right\|_2} \tag{11}$$

Next, the Lagrange multiplier namely $\lambda = [\lambda]_{ij}$ is adopted to constrain $W \geq 0$. Let

$$\psi(W) = tr\left((X - XW)^T P (X - XW)\right) + \alpha tr\left(W^T R W\right) + \beta tr\left(W^T Q W\right) \quad (12)$$

Then, the Lagrangian function is formulated as:

$$L(W) = \psi(W) + tr(\lambda W) \quad (13)$$

After taking the derivatives of (13) with respect to $W$ and utilizing the Karush-Kuhn-Tucker (KKT) condition $\lambda_{ij} W_{ij} = 0$ [38], we can obtain

$$\frac{\partial L(W)}{\partial W} = X^T P X W + \alpha R W + \beta Q W - X^T P X + \lambda = 0 \quad (14)$$

In order to ensure the non-negativity of $W$, we define $R = R^+ - R^-$ as well as the literature [39], where

$$R^+ = \frac{|R_{ij}| + R_{ij}}{2} \quad (15)$$

$$R^- = \frac{|R_{ij}| - R_{ij}}{2} \quad (16)$$

Consequently, bringing the decomposed positive and negative parts of (15) and (16) into (14), we can obtain the update rule with respect to $W$ that is defined as:

$$W_{ij} \leftarrow W_{ij} \frac{\left(X^T P X + \alpha R^- W\right)_{ij}}{\left(\left(X^T P X + \beta Q + \alpha R^+\right) W\right)_{ij}} \quad (17)$$

Furthermore, $p_{ii}$ and $q_{ii}$ are redefined in the following because that the values of $\|x_i - x_i W\|_2$ and $\|w^i\|_2$ may be zero.

$$p_{ii} = \frac{1}{\max\left(2\|x_i - x_i W\|_2, \varepsilon\right)} \quad (18)$$

$$q_{ii} = \frac{1}{\max\left(2\|w^i\|_2, \varepsilon\right)} \quad (19)$$

where $\varepsilon$ is a very small constant.

As mentioned above, we can solve the (9) by alternately updating $W$, $P$ and $Q$. The whole optimization procedure of the proposed OPSR is summarized as follows:

### B. CONVERGENCE ANALYSIS

Under the proposed updating rule in (17), the value of the objective function in (9) monotonically decreases, which is seriously proved in this subsection. Meanwhile, the convergence of the iterative updating algorithm is analyzed in detail.

*Theorem 1:* For $W \geq 0$, the value of the objective function in (9) is non-increasing and has a lower boundary under the updating rule in (17).

*Proof of Theorem 1:* Firstly, an auxiliary function which is useful to the following lemmas is introduced and defined as:

*Definition 1:* $\varphi\left(u, u'\right)$ is an auxiliary function $\rho(u)$ supposed that $\varphi\left(u, u'\right) \geq \rho(u)$ and $\varphi(u, u) = \rho(u)$ are satisfied.

*Lemma 1:* If $\varphi$ is an auxiliary function of $\rho$, then $\rho$ is non-increasing with the following updating rule:

$$u^{(t+1)} = \arg\min_u \varphi\left(u, u^{(t)}\right) \quad (20)$$

where $t$ represents the $t$-th iteration.

*Proof:* $\rho\left(u^{(t+1)}\right) \leq \varphi\left(u^{(t+1)}, u^{(t)}\right) \leq \varphi\left(u^{(t)}, u^{(t)}\right) = \rho\left(u^{(t)}\right)$.

To prove the updating rule in (17) is absolutely as well as in (20) with an auxiliary function, the following function is designed as:

$$\varphi\left(W_{ij}, W_{ij}^t\right) = \rho_{ij}\left(W_{ij}^t\right) + \rho'_{ij}\left(W_{ij}^t\right)\left(W_{ij} - W_{ij}^t\right) + \frac{\left[\left(X^T P X + \beta Q + \alpha R^+\right) W\right]_{ij}}{W_{ij}^t}\left(W_{ij} - W_{ij}^t\right)^2 \quad (21)$$

*Lemma 2:* The function $\varphi\left(W_{ij}, W_{ij}^t\right)$ is an auxiliary function of $\rho_{ij}\left(W_{ij}\right)$.

*Proof:* Note that $\varphi\left(W_{ij}, W_{ij}\right) = \rho\left(W_{ij}\right)$, we only need to certify that $\varphi\left(W_{ij}, W_{ij}^t\right) \geq \rho_{ij}\left(W_{ij}\right)$.

The Taylor series of $\rho_{ij}\left(W_{ij}\right)$ is expressed as:

$$\rho_{ij}\left(W_{ij}\right) = \rho_{ij}\left(W_{ij}^t\right) + \rho'_{ij}\left(W_{ij}^t\right)\left(W_{ij} - W_{ij}^t\right) + \frac{1}{2}\rho''_{ij}\left(W_{ij}^t\right)\left(W_{ij} - W_{ij}^t\right)^2 \quad (22)$$

where $\rho'_{ij}\left(W_{ij}^t\right)$ and $\rho''_{ij}\left(W_{ij}^t\right)$ denote the first-order and second-order derivatives, respectively, which are formulated as:

$$\rho'_{ij}\left(W_{ij}^t\right) = 2\left(\left(X^T P X + \alpha R + \beta Q\right) W - X^T P X\right)_{ij} \quad (23)$$

$$\rho''_{ij}\left(W_{ij}^t\right) = 2\left(X^T P X + \alpha R + \beta Q\right)_{ii} \quad (24)$$

Note that $\left[\left(X^T P X + \beta Q + \alpha R^+\right) W\right]_{ij} \geq W_{ij}^t [X^T P X + \beta Q + \alpha R^+]_{ii}$, then, we have:

$$\frac{\left[\left(X^T P X + \beta Q + \alpha R^+\right) W\right]_{ij}}{W_{ij}^t} \geq \left[X^T P X + \beta Q + \alpha R^+\right]_{ii} \quad (25)$$

Therefore, $\varphi\left(W_{ij}, W_{ij}^t\right) \geq \rho_{ij}\left(W_{ij}\right)$ holds. Lemma 2 is proved. Apparently, since (21) is the auxiliary function of $\rho_{ij}$, $\rho_{ij}$ is non-increasing with the update criterion in (17). Replacing $\varphi\left(u, u^{(t)}\right)$ in (20) by the auxiliary function in (21) and setting it as zero, we can get the iterative updating rule of $W$ as follows:

$$W_{ij}^t \leftarrow W_{ij}^{t-1} \frac{\left(X^T P X + \alpha R^- W^{t-1}\right)_{ij}}{\left(\left(X^T P X + \beta Q + \alpha R^+\right) W^{t-1}\right)_{ij}} \quad (26)$$

Subsequently, the proof of the convergence of **Theorem 1** is given. For any non-zero vectors, $h \in R^d$ and $w \in R^d$, we have the following inequality holds:

$$\|h\|_2 - \frac{\|h\|_2^2}{2\|w\|_2} \leq \|w\|_2 - \frac{\|w\|_2^2}{2\|w\|_2} \quad (27)$$

The proof in (27) is similar to that in literature [23].

Let $P_t$ and $Q_t$ be the $t$-th iteration, after fixing them, $W_{t+1}$ is updated by solving the inequality:

$$\Psi\left(W_{t+1}, P_t, Q_t\right) \leq \Psi\left(W_t, P_t, Q_t\right) \tag{28}$$

Hence, we have

$$\begin{aligned} & tr\left(\left(X-XW_{t+1}\right)^T P_t\left(X-XW_{t+1}\right)\right) \\ & \quad +\alpha tr\left(W_{t+1}^T R W_{t+1}\right)+\beta tr\left(W_{t+1}^T Q_t W_{t+1}\right) \\ & \leq tr\left(\left(X-XW_t\right)^T P_t\left(X-XW_t\right)\right) \\ & \quad +\alpha tr\left(W_t^T R W_t\right)+\beta tr\left(W_t^T Q_t W_t\right) \end{aligned} \tag{29}$$

Note that, $\|W\|_{2,1} = \sum_{i=1}^{d} \left\|w^i\right\|_2$ and $\|X - XW\|_{2,1} = \sum_{i=1}^{n} \|x_i - x_i W\|_2$, the above inequality is rewritten as:

$$\begin{aligned} & \sum_i \frac{\|x_i - x_i W_{t+1}\|_2^2}{2\|x_i - x_i W_t\|_2} + \alpha tr\left(W_{t+1}^T R W_{t+1}\right) + \beta \sum_i \frac{\left\|w_{t+1}^i\right\|_2^2}{2\left\|w_{t+1}^i\right\|_2} \\ & \leq \sum_i \frac{\|x_i - x_i W_t\|_2^2}{2\|x_i - x_i W_t\|_2} + \alpha tr\left(W_t^T R W_t\right) + \beta \sum_i \frac{\left\|w_t^i\right\|_2^2}{2\left\|w_t^i\right\|_2} \end{aligned} \tag{30}$$

Then, we have the following inequality:

$$\begin{aligned} & \sum_i \|x_i - x_i W_{t+1}\|_2 \\ & \quad - \left(\sum_i \|x_i - x_i W_{t+1}\|_2 - \sum_i \frac{\|x_i - x_i W_{t+1}\|_2^2}{2\|x_i - x_i W_t\|_2}\right) \\ & \quad + \beta \sum_i \left\|w_{t+1}^i\right\|_2 - \beta \left(\sum_i \left\|w_{t+1}^i\right\|_2 - \sum_i \frac{\left\|w_{t+1}^i\right\|_2^2}{2\left\|w_t^i\right\|_2}\right) \\ & \quad + \alpha tr\left(W_{t+1}^T R W_{t+1}\right) \\ & \leq \sum_i \|x_i - x_i W_t\|_2 \\ & \quad - \left(\sum_i \|x_i - x_i W_t\|_2 - \sum_i \frac{\|x_i - x_i W_t\|_2^2}{2\|x_i - x_i W_t\|_2}\right) \\ & \quad + \beta \sum_i \left\|w_t^i\right\|_2 - \beta \left(\sum_i \left\|w_t^i\right\|_2 - \sum_i \frac{\left\|w_t^i\right\|_2^2}{2\left\|w_t^i\right\|_2}\right) \\ & \quad + \alpha tr\left(W_t^T R W_t\right) \end{aligned} \tag{31}$$

According to the property in (27), for each $i$, we have

$$\left\|w_{t+1}^i\right\|_2 - \frac{\left\|w_{t+1}^i\right\|_2^2}{2\left\|w_t^i\right\|_2} \leq \left\|w_t^i\right\|_2 - \frac{\left\|w_t^i\right\|_2^2}{2\left\|w_t^i\right\|_2} \tag{32}$$

$$\begin{aligned} & \|x_i - x_i W_{t+1}\|_2 - \frac{\|x_i - x_i W_{t+1}\|_2^2}{2\|x_i - x_i W_t\|_2} \\ & \qquad \leq \|x_i - x_i W_t\|_2 - \frac{\|x_i - x_i W_t\|_2^2}{2\|x_i - x_i W_t\|_2} \end{aligned} \tag{33}$$

Then, the following inequality holds:

$$\begin{aligned} & \sum_i \left\|w_{t+1}^i\right\|_2 - \sum_i \frac{\left\|w_{t+1}^i\right\|_2^2}{2\left\|w_t^i\right\|_2} \\ & \qquad \leq \sum_i \left\|w_t^i\right\|_2 - \sum_i \frac{\left\|w_t^i\right\|_2^2}{2\left\|w_t^i\right\|_2} \end{aligned} \tag{34}$$

$$\begin{aligned} & \sum_i \|x_i - x_i W_{t+1}\|_2 - \sum_i \frac{\|x_i - x_i W_{t+1}\|_2^2}{2\|x_i - x_i W_t\|_2} \\ & \qquad \leq \sum_i \|x_i - x_i W_t\|_2 - \sum_i \frac{\|x_i - x_i W_t\|_2^2}{2\|x_i - x_i W_t\|_2} \end{aligned} \tag{35}$$

Combining (31), (34) and (35), we can obtain:

$$\begin{aligned} & \|X - XW_{t+1}\|_{2,1} + \alpha tr\left(W_{t+1}^T R W_{t+1}\right) + \beta \|W_{t+1}\|_{2,1} \\ & \qquad \leq \|X - XW_t\|_{2,1} + \alpha tr\left(W_t^T R W_t\right) + \beta \|W_t\|_{2,1} \end{aligned} \tag{36}$$

The (36) implies that the objective function in (9) monotonically decreases in each iteration and has the lower boundary which makes it converge. Overall, Theorem 1 is proven.

### C. COMPUTATIONAL COMPLEXITY ANALYSIS

The computational complexity analysis of the proposed OPSR is analyzed in this section. Obviously, the computation complexity of calculating the weighting matrix $S$ among samples is $O(d^2n)$. Besides, the cost of each iteration in **Algorithm 1** is equal to $O(\min(n, d)dn)$. Therefore, the total computation complexity of our algorithm is equal to $O(d^2n + t(\min(n, d)dn))$, where $t$ is the number of iterations.

**Algorithm 1** OPSR

1: **Input:** The data matrix $X \in R^{n \times d}$, parameters $\alpha$ and $\beta$, and number of selected features $m$.
2: **Initialize:** Let $t = 1$, $P_1$ and $Q_1$ be the identity matrices.
3: Compute the weight matrix $S$.
4: Compute the diagonal matrix $D$, the Laplacian matrix $L$ and $R$.
5: **Repeat**
6: Compute matrix

$$W_{t+1} \leftarrow W_t \frac{\left(X^T P_t X + \alpha R^- W_t\right)_{ij}}{\left(X^T P_t X + \beta Q_t + \alpha R^+\right) W_t};$$

7: Update matrices $P_{t+1}$ and $Q_{t+1}$ using (18) and (19);
8: $t = t + 1$;
9: **Until** Convergence
10: Compute the weight of each feature by $\left\|w^i\right\|_2 (i = 1, 2, \cdots, d)$ and sort all of the weight values in descending order.
11: **Output**: The top $m$ ranked features.

## IV. EXPERIMENTAL RESULTS AND ANALYSIS

To evaluate the effectiveness of the proposed OPSR, a series of experiments including classification and

clustering are constructed particularly in this section. Meanwhile, to verify the superiority of OPSR, extensive experiments are carried out via comparing with some state-of-the-art unsupervised feature selection algorithms (SPEC [16], MCFS [12], RUFS [40], MFFS [41], RSR [33], L1-UFS [32], UFSOL [25], SCUFS [14], NNSRD [42] and DSRMR [30]) on six benchmark databases including Extended YaleB [43], CMU PIE [44], AR [45], JAFFE [46], ORL [47] and COIL20 [48].

In our work, the classification rate (CR) as the metric is utilized to evaluate the performance of the proposed OPSR quantitatively, which is defined as:

$$\mathrm{CR} = \frac{T}{N} \tag{37}$$

where $T$ is the number of test samples which are correctly classified and $N$ is the total number of test samples.

### A. DATABASES

The description of six benchmark databases are illustrated in Table 1, and some image samples from the databases are displayed in Figure 1. Subsequently, the details of these databases are summarized as following:

**TABLE 1. Details of the six benchmark databases.**

| Databases | Images | Features | Classes |
|---|---|---|---|
| Extended YaleB | 2432 | 1024 | 38 |
| CMU PIE | 1632 | 1024 | 68 |
| AR | 1400 | 1024 | 100 |
| JAFFE | 213 | 1024 | 10 |
| ORL | 400 | 1024 | 40 |
| COIL20 | 1440 | 1024 | 20 |

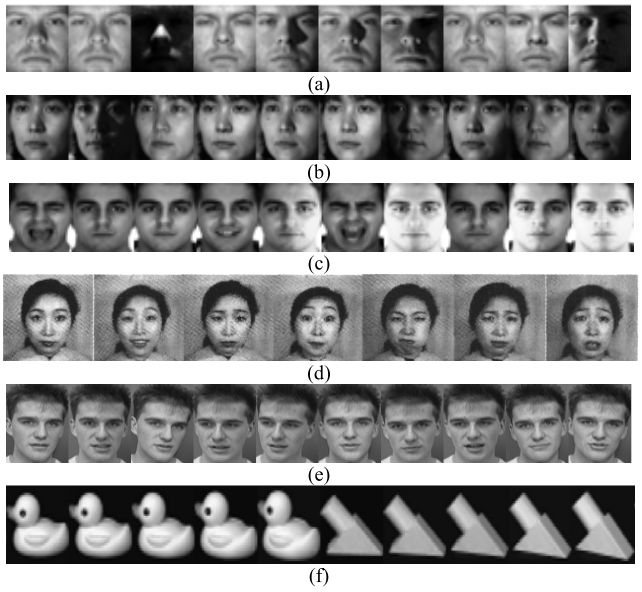


**FIGURE 1. Some image samples from different databases. (a) Extended YaleB. (b) CMU PIE. (c) AR. (d) JAFFE. (e) ORL. (f) COIL20.**

(1) Extended YaleB database [43] is consisted of 2414 frontal cropped facial images from 38 subjects. Each subject has 64 images with the size of 32 × 32 pixels, which were obtained with small illumination changes in head pose and facial expression.

(2) CMU PIE database [44] is consisted of 41368 face images from 68 subjects, and the face images of each subject were captured with 13 different poses, 43 different illumination changes and 4 different expressions. A subset named C29 which contains 24 images of each subject with only illumination changes is selected for our experiments.

(3) AR database [45] is consisted of 4000 face images from 126 distinct subjects (70 male and 56 female). Each subject has 26 facial images which were acquired under the condition of several expressions, illumination changes and sun glasses and scarf occlusions.

(4) JAFFE database [46] is consisted of 213 facial images from 10 Japanese female models which were captured with 7 kinds of facial expressions (6 basic facial expressions and 1 neutral).

(5) ORL database [47] is consisted of 400 images from 40 distinct subjects which were taken under the condition of different times, varying illumination and several facial expressions.

(6) COIL20 database [48] is consisted of 1440 images from 20 objects, which were obtained by varying object angels at intervals of 5 degree.

### B. CLASSIFICATION

In this subsection, Extended YaleB, CMU PIE and AR databases are utilized for the classification experiments. In particular, $l$ image samples ($l = 20$, 12 and 7) are chosen randomly for each subject from each database for training and the reminding samples are applied to testing. The process of the sample selection is repeated with 10 times and the calculated average classification rates and standard deviations are illustrated in Table 2. Besides, the running time of different algorithms are revealed in Table 2.

Obviously, several observations can be found from Table 2, which are summarized as follows: (1) We can see that the listed feature selection algorithms imply better classification performance since the feature selection procedure plays an important role in improving the classification rates by eliminating redundancy and noisy features significantly. (2) The performance of SPEC is superior to the Baseline algorithm but is inferior to the other algorithms, which results from selecting the features in one-by-one manner and ignoring the correlations among features. (3) In contrast, MCFS, RUFS, MFFS, RSR, L1-UFS, UFSOL, SCUFS, NNSRD and DSRMR have better performance by selecting features in a batch model and exploiting the correlations among features as much as possible. Specially, L1-UFS, SCUFS, NNSRD and DSRMR are superior to most of the compared algorithms which take advantage of the local structure of data. UFSOL also achieves better performance owning to preserve the ordinal locality structure of data sufficiently. Besides, RSR

**TABLE 2.** The average classification rates ± standard deviations (CR ± std) of different algorithms on three databases. The best results are highlighted in bold.

| Methods | AR | Extended YaleB | CMU PIE |
|---|---|---|---|
| Baseline | 0.6206 ±0.0162(1024) | 0.6193±0.0081(1024) | 0.8563±0.0072(1024) |
| SPEC | 0.6456±0.0154(500,8.45) | 0.6418±0.0096(500,5.11) | 0.8749±0.0082(470,7.71) |
| MCFS | 0.6521±0.0158(500,0.23) | 0.6589±0.0178(200,0.31) | 0.8791±0.0084(490,0.28) |
| RUFS | 0.6661±0.0171(480,246) | 0.6697±0.0132(480,30.1) | 0.8899±0.0091(490,263) |
| MFFS | 0.6741±0.0147(470,64.6) | 0.6722±0.0093(330,62.6) | 0.8862±0.0088(270,68.1) |
| RSR | 0.6671±0.0147(440,32.4) | 0.6883±0.0106(500,30.3) | 0.8937±0.0085(440,34.7) |
| L1-UFS | 0.6859±0.0165(350,131) | 0.7096±0.0116(500,128) | 0.9133±0.0114(500,138) |
| UFSOL | 0.6805±0.014(320,305) | 0.7120±0.0110(270,318) | 0.9125±0.0154(450,331) |
| SCUFS | 0.6856±0.0088(480,1271) | 0.7147±0.0107(410,1309) | 0.9140±0.0097(370,1383) |
| NNSRD | 0.6979±0.0152(430,5.54) | 0.7202±0.0119(440,5.49) | 0.9185±0.0096(410,0.96) |
| DSRMR | 0.7000±0.0194(480,710) | 0.7258±0.0097(340,594) | 0.9210±0.0114(390,975) |
| OPSR | **0.7086±0.0157(320,102)** | **0.7344±0.0114(260,107)** | **0.9311±0.0113(470,108)** |

Note that the numbers in parentheses are the number of the selected features that correspond to the best classification result and the running time (s).

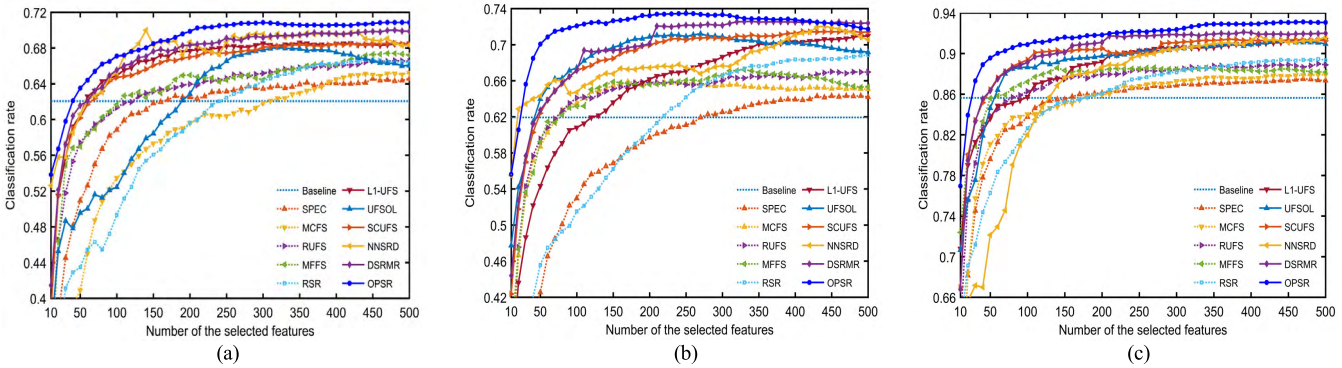


**FIGURE 2.** The classification rates of different algorithms with different numbers of selected features on three databases. (a) AR. (b) Extended YaleB. (c) CMU PIE.

indicates that the self-representation ability of features during the procedure of selecting the most representative features. (4) The proposed OPSR outperforms the compared algorithms on the three databases by introducing the ordinal locality structure of data into the self-representation model that can preserve the local geometrical structure and the ordinal locality features effectively. (5) The running time of the proposed OPSR is less than L1-UFS, UFSOL, SCUFS and DSRMR on the three databases and less than RUFS on the AR and CMU PIE databases. This is because the iterative updating strategy is designed to optimize the proposed OPSR. Nevertheless, the computational performance is slightly inferior to other methods, such as SPEC, MCFS, MFFS, RSR and NNSRD.

To illustrate the impact of the number of selected features, Figure 2 provides the curves about classification rate of different compared algorithms with different numbers of selected features on three databases. From Figure 2, we can clearly see that with the increasing number of the selected features, the classification rates gradually rise for all the listed algorithms. Whereafter, the trend towards relatively stable when the classification rates reach the peak at which all the algorithms achieve the best classification performance. For the AR and CMU PIE databases, the curves of OPSR are almost above the curves of all compared algorithms. For the Extended YaleB database, we can see that the classification rate of OPSR is slightly lower than DSRMR with continuing to increase the number of the selected features after achieving the best performance. Overall, the proposed OPSR is excellent compared to the other algorithms with the increasing number of selected features.

In order to analyze the influence of parameter values in the OPSR model, Figure 3 displays the classification rates of OPSR with various values of parameters $\alpha$ and $\beta$ on three databases. Apparently, it can be seen that the performance of OPSR is not very sensitive to various values of parameters $\alpha$ and $\beta$ from Figure 3. Moreover, OPSR achieves the best performance when the values of $\alpha$ and $\beta$ are neither too large nor too small. More specially, the classification rates increase gradually with the increasing values of $\alpha$ and $\beta$. However, the classification rates decrease dramatically with

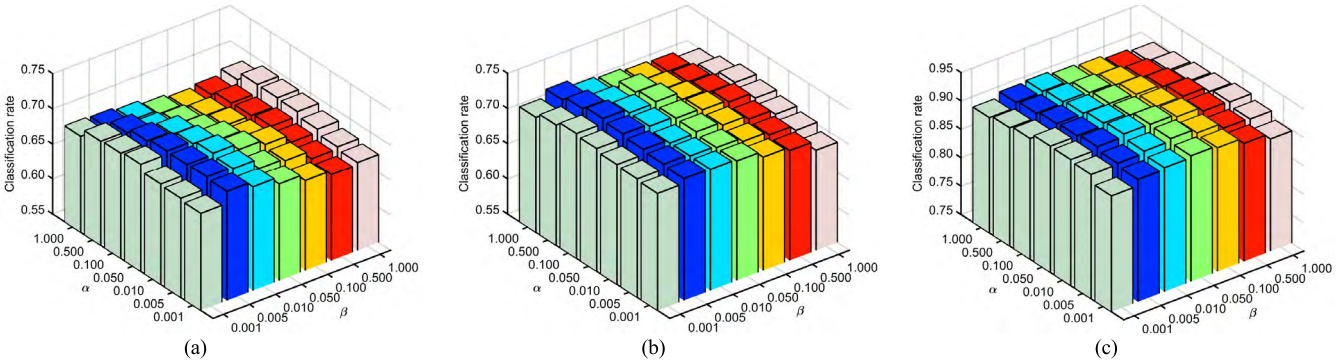


**FIGURE 3.** The classification rates of OPSR with various values of $\alpha$ and $\beta$ on three databases. (a) AR. (b) Extended YaleB. (c) CMU PIE.

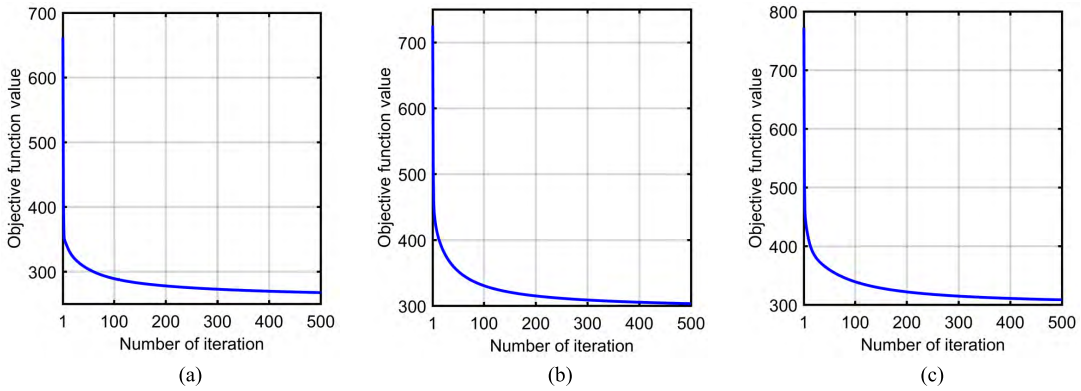


**FIGURE 4.** The convergence curves of OPSR on three databases. (a) AR. (b) Extended YaleB. (c) CMU PIE.

the increasing values of $\alpha$ and $\beta$ continuously after achieving the best performance. The cause of this phenomenon is that relative small $\alpha$ and $\beta$ values will weaken the role of second and third terms and make the first term play a dominant role in the objective function as shown in (8), which will neglect the ordinal locality of original samples and the sparsity of selected features. Likewise, relative large $\alpha$ and $\beta$ values can reduce the impact of the first term, which will not retain the ability of feature self-representation well.

Figure 4 shows the convergence curves of the proposed OPSR on the three databases, which intuitively indicate that the OPSR algorithm converges very fast (generally within 500 iterations) and demonstrates the effectiveness and efficiency of the proposed optimal algorithm.

### C. CLUSTERING

In this subsection, ORL, JAFEE and COIL20 databases are employed for the clustering experiments to demonstrate the effectiveness and superiority of OPSR. To evaluate the performance from a quantitative point of view, two metrics namely clustering accuracy (ACC) and normalized mutual information (NMI) [49] are used for comparative analysis of compared algorithms with OPSR. The definition of ACC is formulated as following:

$$\mathrm{ACC} = \frac{1}{n}\sum_{i=1}^{n}\psi(map(c_i), l_i) \tag{38}$$

where $n$ denotes the number of samples, $\psi(a, b)$ is the indicator function that $\psi(a, b) = 1$ if $a = b$, and $\psi(a, b)$=0 otherwise. $c_i$ and $l_i$ stand for the clustering label and the corresponding groundtruth label of $i$-th sample, respectively.$map(\cdot)$ is the optimal mapping function that matches the obtained clustering label and the groundtruth well in terms of the Kuhn-Munkres algorithm. Clearly, the higher ACC means the better clustering performance.

In addition, NMI is defined as:

$$\mathrm{NMI} = \frac{I(C, G)}{\sqrt{H(C)H(G)}} \tag{39}$$

where $C$ and $G$ denote the clustering label and the groundtruth of $i$-th input sample, respectively. $I(C, G)$ is the mutual information between $C$ and $G$. $H(C)$ and $H(G)$ are the entropies corresponding to $C$ and $G$. Similarly, the higher NMI means the better clustering performance.

For convenience, a simple algorithm namely $k$-means is made use of clustering the selected features for various feature selection algorithms. Due to the initialization plays an important role in the performance of $k$-means algorithms, 50 times with different random initializations are repeated during clustering experiments. The calculated average clustering results with standard deviations are used to evaluate the performance of the proposed OPSR quantitatively.

Table 3 and Table 4 depict the ACC and NMI with standard deviations obtained by the optimal fixed parameters as well

**TABLE 3.** Clustering results (ACC ±std) of different algorithms on three databases.

| Methods | ORL | JAFFE | COIL20 |
|---|---|---|---|
| BaseLine | 0.7526±0.0439(1024) | 0.7873±0.0228(1024) | 0.5527±0.0271(1024) |
| SPEC | 0.8030±0.0756(170,4.07) | 0.8521±0.0708(470,2.09) | 0.6128±0.0476(340,15.5) |
| MCFS | 0.8210±0.0555(240,0.65) | 0.8709±0.0871(240,0.03) | 0.6214±0.0512(250,0.78) |
| RUFS | 0.8300±0.0542(140,53.5) | 0.8864±0.0781(470,10.3) | 0.6408±0.0484(360,461) |
| MFFS | 0.8390±0.0523(40,53.2) | 0.8958±0.0298(500,26.7) | 0.6460±0.0286(300,90.5) |
| RSR | 0.8310±0.0378(270,24.4) | 0.8728±0.0518(500,5.91) | 0.6486±0.0272(470,47.8) |
| L1-UFS | 0.8622±0.0589(170,103) | 0.8994±0.0609(420,24.4) | 0.6657±0.03.54(220,186) |
| UFSOL | 0.8660±0.0358(430,111) | 0.9042±0.0346(490,76.7) | 0.6730±0.0202(490,415) |
| SCUFS | 0.8717±0.0548(220,1040) | 0.9104±0.0631(220,256) | 0.6753±0.0290(360,1752) |
| NNSRD | 0.8730±0.0459(200,3.42) | 0.9138±0.0543(250,1.53) | 0.6793±0.0280(480,7.88) |
| DSRMR | 0.8770±0.0529(180,1429) | 0.9151±0.0635(360,180) | 0.6825±0.0229(490,4773) |
| OPSR | 0.8900±0.0367(230,75.4) | 0.9321±0.0150(420,23.1) | 0.6988±0.0212(480,138) |

Note that the numbers in parentheses are the number of the selected features that correspond to the best clustering result and the running time (s).

**TABLE 4.** Clustering results (NMI ±std) of different algorithms on different databases.

| Methods | ORL | JAFFE | COIL20 |
|---|---|---|---|
| BaseLine | 0.7964±0.0310(1024) | 0.8213±0.0143(1024) | 0.7035±0.0131(1024) |
| SPEC | 0.8318±0.0228(450,4.07) | 0.8828±0.0614(390,2.09) | 0.7245±0.0128(370,15.5) |
| MCFS | 0.8383±0.0390(230,0.65) | 0.8887±0.0500(160,0.03) | 0.7289±0.0247(370,0.78) |
| RUFS | 0.8588±0.0316(140,53.5) | 0.8914±0.0448(500,10.3) | 0.7486±0.0096(290,461) |
| MFFS | 0.8701±0.0236(50,53.2) | 0.8960±0.0264(500,26.7) | 0.7486±0.0136(260,90.5) |
| RSR | 0.8600±0.0298(230,24.4) | 0.8952±0.0276(500,5.91) | 0.7463±0.0163(470,47.8) |
| L1-UFS | 0.8987±0.0254(140,103) | 0.9074±0.0401(460,24.4) | 0.7592±0.0202(420,186) |
| UFSOL | 0.9047±0.0325(190,111) | 0.9052±0.0209(380,76.7) | 0.7660±0.0135(440,415) |
| SCUFS | 0.9086±0.0329(220,1040) | 0.9188±0.0324(230,256) | 0.7671±0.0172(310,1752) |
| NNSRD | 0.9125±0.0287(260,3.42) | 0.9179±0.0292(240,1.53) | 0.7696±0.0130(500,7.88) |
| DSRMR | 0.9115±0.0300(210,1429) | 0.9214±0.0371(360,180) | 0.7704±0.0152(500,4773) |
| OPSR | 0.9290±0.0239(150,75.4) | 0.9313±0.0129(450,23.1) | 0.7826±0.0117(420,138) |

Note that the numbers in parentheses are the number of the selected features that correspond to the best classification result and the running time (s).

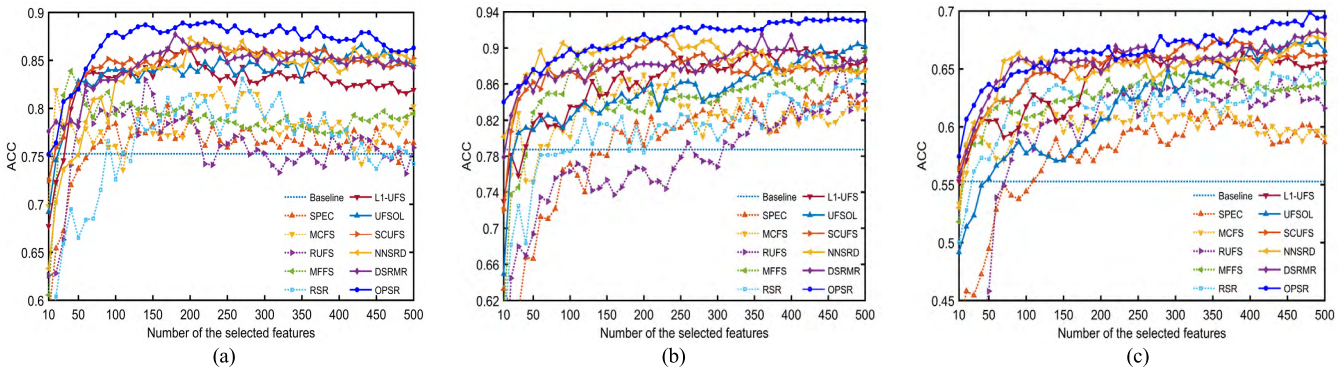


**FIGURE 5.** The ACC of different feature selection algorithms. (a) ORL. (b) JAFFE. (c) COIL20.

as the running time on three databases, respectively. Obviously, we can see that our OPSR performs better than all compared algorithms with relatively few of selected features. The phenomenon sufficiently reveals the effectiveness and superiority of the proposed OPSR for clustering task. Meanwhile, we also can see that all feature selection algorithms

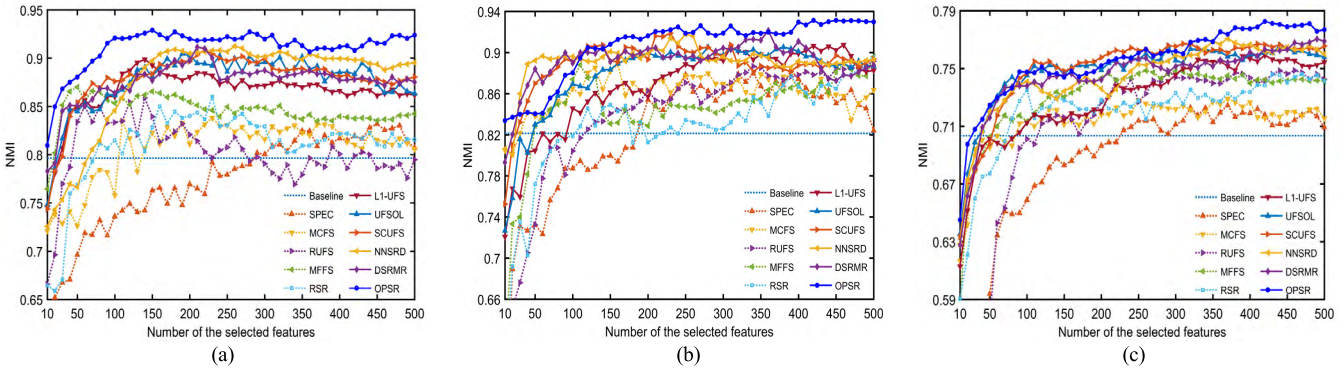


**FIGURE 6.** The NMI of different feature selection algorithms. (a) ORL. (b) JAFFE. (c) COIL20.

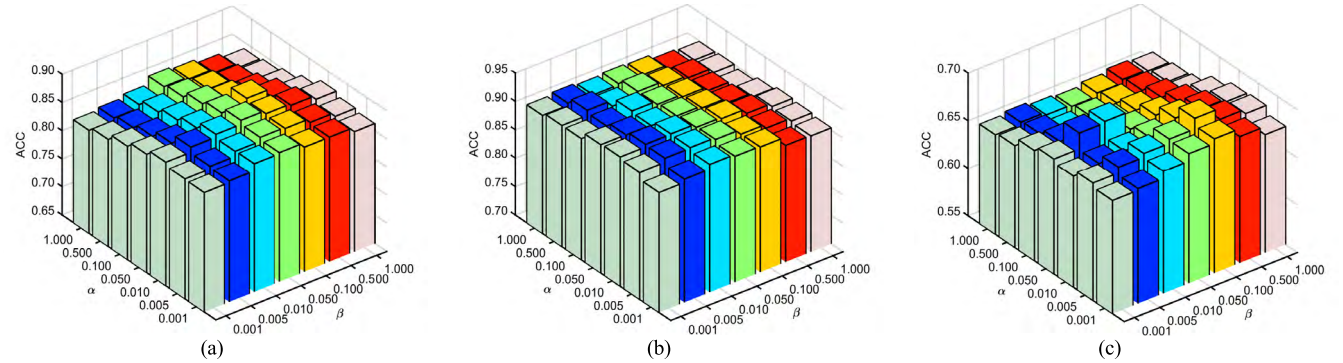


**FIGURE 7.** The ACC of OPSR vs. parameters $\alpha$ and $\beta$ on three databases. (a) ORL. (b) JAFFE. (c) COIL20.

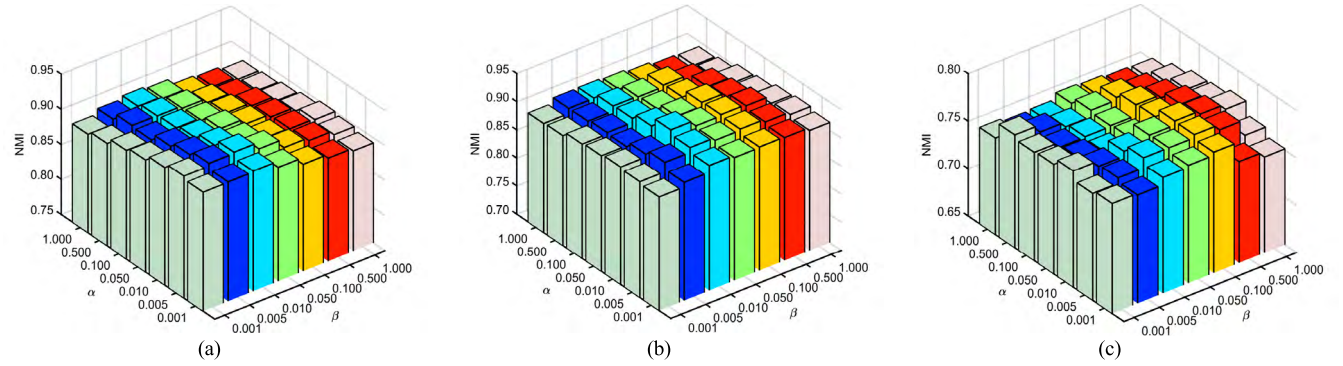


**FIGURE 8.** The NMI of OPSR vs. parameters $\alpha$ and $\beta$ on three databases. (a) ORL. (b) JAFFE. (c) COIL20.

are superior to the BaseLine, which indicates that the procedure of feature selection plays an important role in clustering. Besides, the computational performance of the proposed OPSR is consistent with the classification experiments described in Section IV.B.

Furthermore, the change curves of ACC and NMI with the number of selected features on three databases are shown in Figure 5 and Figure 6, respectively. In the same way, it can be seen that the curves of OPSR is higher than compared algorithms in most cases with the increasing number of selected features. The observations illustrate that our OPSR is highly competitive with the state-of-the-art compared algorithms.

To discuss the sensitivity of parameters $\alpha$ and $\beta$ in our OPSR for clustering task, Figure 7 and Figure 8 display the ACC and NMI changes with varying values of parameters $\alpha$ and $\beta$ on three databases, respectively. Evidently, seen from these figures, the proposed OPSR is not sensitive to the parameters and achieves the best performance with moderate parameter values on three databases.

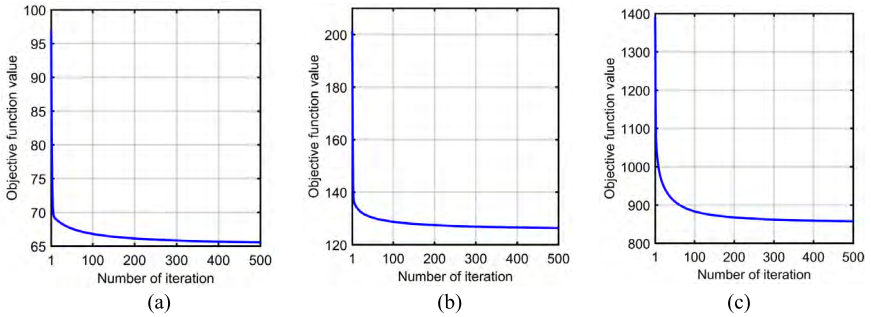


**FIGURE 9.** The convergence curves of OPSR on the three databases. (a) ORL. (b) JAFFE. (c) COIL20.

At last, the convergence curves of the proposed OPSR with approximately 500 iterations on ORL, JAFFE and COIL20 databases are shown in Figure 9. Visibly, the value of the objective function declines for each iteration and converges very fast on the three databases.

## V. CONCLUSIONS

This paper suggests a novel Ordinal Locality-preserving and Feature Self-representation algorithm for unsupervised feature selection. In detail, the ordinal locality is brought into the self-representation model, which can maintain the local geometrical structure and the ordinal locality of original data simultaneously. In addition, an efficient iterative optimization algorithm is designed to solve the objective function. Finally, extensive classification and clustering experiments indicate that the performance of the proposed OPSR outperforms some well-known algorithms in terms of classification rate, ACC and NMI metrics.

In the future, our work will focus on incorporating other regularized constraints for the structure preserving and developing our algorithm with semi-supervised learning method that can utilize the labeled samples adequately for improving classification rates. In addition, we try to apply the proposed OPSR to other fields, such as the scene classification.

## ACKNOWLEDGMENT

The authors would like to thank the editor, the associate editor, and referees for comments and suggestions that have greatly improved our paper.

## REFERENCES

[1] Y. Yuan, X. Zheng, and X. Lu, "Discovering diverse subset for unsupervised hyperspectral band selection," *IEEE Trans. Image Process.*, vol. 26, no. 1, pp. 51–64, Jan. 2017.

[2] J. Song, L. Gao, F. Nie, H. T. Shen, Y. Yan, and N. Sebe, "Optimized graph learning using partial tags and multiple features for image and video annotation," *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 4999–5011, Nov. 2016.

[3] Y. Yi, Y. Shi, H. Zhang, J. Wang, and J. Kong, "Label propagation based semi-supervised non-negative matrix factorization for feature extraction," *Neurocomputing*, vol. 149, pp. 1021–1037, Feb. 2015.

[4] Y. Yi, S. Qiao, W. Zhou, C. Zheng, Q. Liu, and J. Wang, "Adaptive multiple graph regularized semi-supervised extreme learning machine," *Soft Comput.*, vol. 22, no. 11, pp. 3545–3562, 2018.

[5] J. Wang, R. Zhao, Y. Wang, C. Zheng, J. Kong, and Y. Yi, "Locality constrained graph optimization for dimensionality reduction," *Neurocomputing*, vol. 245, pp. 55–67, Jul. 2017.

[6] W. Zhou, C. Wu, J. Wang, X. Yu, and Y. Yi, "Double regularized matrix factorization for image classification and clustering," *EURASIP J. Image Video Process.*, vol. 2018, no. 1, p. 49, 2018.

[7] J. Cai, J. Luo, S. Wang, and S. Yang, "Feature selection in machine learning: A new perspective," *Neurocomputing*, vol. 300, pp. 70–79, Jul. 2018.

[8] R. Sheikhpour, M. A. Sarram, S. Gharaghani, and M. A. Z. Chahooki, "A survey on semi-supervised feature selection methods," *Pattern Recognit.*, vol. 64, pp. 141–158, Apr. 2017.

[9] K. Sechidis and G. Brown, "Simple strategies for semi-supervised feature selection," *Mach. Learn.*, vol. 107, no. 2, pp. 357–395, 2018.

[10] J. Gui, Z. Sun, S. Ji, D. Tao, and T. Tan, "Feature selection based on structured sparsity: A comprehensive study," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 7, pp. 1490–1507, Jul. 2017.

[11] J. Li *et al.*, "Feature selection: A data perspective," *ACM Comput. Surv.*, vol. 50, no. 6, 2017, Art. no. 94.

[12] D. Cai, C. Zhang, and X. He, "Unsupervised feature selection for multi-cluster data," in *Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2010, pp. 333–342.

[13] C. Hou, F. Nie, X. Li, D. Yi, and Y. Wu, "Joint embedding learning and sparse regression: A framework for unsupervised feature selection," *IEEE Trans. Cybern.*, vol. 44, no. 6, pp. 793–804, Jun. 2014.

[14] P. Zhu, W. Zhu, Q. Hu, C. Zhang, and W. Zuo, "Subspace clustering guided unsupervised feature selection," *Pattern Recognit.*, vol. 66, pp. 364–374, Jun. 2017.

[15] D. Du, O. Richard, E. H. Peter, and G. S. David, *Pattern Classification*. Hoboken, NJ, USA: Wiley, 2012.

[16] Z. Zhao and H. Liu, "Spectral feature selection for supervised and unsupervised learning," in *Proc. 24th Int. Conf. Mach. Learn.*, 2007, pp. 1151–1157.

[17] 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.

[18] R. Zhang, F. Nie, and X. Li, "Self-weighted supervised discriminative feature selection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 8, pp. 3913–3918, Aug. 2018.

[19] W. Liao, A. Pizurica, P. Scheunders, W. Philips, and Y. Pi, "Semisupervised local discriminant analysis for feature extraction in hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 1, pp. 184–198, Jan. 2013.

[20] X. Chen, F. Nie, G. Yuan, and J. Z. Huang, "Semi-supervised feature selection via rescaled linear regression," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, 2017, pp. 1525–1531.

[21] M. Luo, X. Chang, L. Nie, Y. Yang, A. G. Hauptmann, and Q. Zheng, "An adaptive semisupervised feature analysis for video semantic recognition," *IEEE Trans. Cybern.*, vol. 48, no. 2, pp. 648–660, Feb. 2018.

[22] X. Chen, G. Yuan, W. Wang, F. Nie, X. Chang, and J. Z. Huang, "Local adaptive projection framework for feature selection of labeled and unlabeled data," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published, doi: 10.1109/TNNLS.2018.2830186.

[23] 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. (IJCAI)*, 2011, pp. 1589–1594.

[24] Z. Li, Y. Yang, J. Liu, X. Zhou, and H. Lu, "Unsupervised feature selection using nonnegative spectral analysis," in *Proc. AAAI*, vol. 2, 2012, pp. 1026–1032.

[25] J. Guo, Y. Quo, X. Kong, and R. He, "Unsupervised feature selection with ordinal locality," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2017, pp. 1213–1218.
[26] M. Luo, F. Nie, X. Chang, Y. Yang, A. G. Hauptmann, and Q. Zheng, "Adaptive unsupervised feature selection with structure regularization," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 4, pp. 944–956, Apr. 2017.
[27] X. Chang, F. Nie, Y. Yang, C. Zhang, and H. Huang, "Convex sparse PCA for unsupervised feature learning," *ACM Trans. Knowl. Discovery Data*, vol. 11, no. 1, pp. 3:1–3:16, 2016.
[28] R. Hu *et al.*, "Self-representation dimensionality reduction for multi-model classification," *Neurocomputing*, vol. 253, pp. 154–161, Aug. 2017.
[29] Y. Li, C. Lei, Y. Fang, R. Hu, Y. Li, and S. Zhang, "Unsupervised feature selection by combining subspace learning with feature self-representation," *Pattern Recognit. Lett.*, vol. 109, pp. 35–43, Jul. 2018.
[30] C. Tang *et al.*, "Robust unsupervised feature selection via dual self-representation and manifold regularization," *Knowl. Based Syst.*, vol. 145, pp. 109–120, Apr. 2018.
[31] C. Lei and X. Zhu, "Unsupervised feature selection via local structure learning and sparse learning," *Multimedia Tools Appl.*, vol. 77, no. 22, pp. 29605–29622, Nov. 2018.
[32] C. Tang, X. Zhu, J. Chen, P. Wang, X. Liu, and J. Tian, "Robust graph regularized unsupervised feature selection," *Expert Syst. Appl.*, vol. 96, pp. 64–76, Apr. 2018.
[33] P. Zhu, W. Zuo, L. Zhang, Q. Hu, and S. C. K. Shiu, "Unsupervised feature selection by regularized self-representation," *Pattern Recognit.*, vol. 48, no. 2, pp. 438–446, 2015.
[34] W. Zhou, C. Wu, Y. Yi, and G. Luo, "Structure preserving non-negative feature self-representation for unsupervised feature selection," *IEEE Access*, vol. 5, pp. 8792–8803, 2017.
[35] M. Qi, T. Wang, F. Liu, B. Zhang, J. Wang, and Y. Yi, "Unsupervised feature selection by regularized matrix factorization," *Neurocomputing*, vol. 273, pp. 593–610, Jan. 2018.
[36] Y. Yi, W. Zhou, C. Bi, G. Luo, Y. Cao, and Y. Shi, "Inner product regularized nonnegative self representation for image classification and clustering," *IEEE Access*, vol. 5, pp. 14165–14176, 2017.
[37] Z. Zhang, L. Shao, Y. Xu, L. Liu, and J. Yang, "Marginal representation learning with graph structure self-adaptation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 10, pp. 4645–4659, Oct. 2018.
[38] C. Ding, T. Li, and M. I. Jordan, "Convex and semi-nonnegative matrix factorizations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 1, pp. 45–55, Jan. 2010.
[39] Y. Yi, C. Bi, X. Li, J. Wang, and J. Kong, "Semi-supervised local ridge regression for local matching based face recognition," *Neurocomputing*, vol. 167, pp. 132–146, Nov. 2015.
[40] M. Qian and C. Zhai, "Robust unsupervised feature selection," in *Proc. 27th AAAI Conf. Artif. Intell.*, 2013, pp. 1621–1627.
[41] S. Wang, W. Pedrycz, Q. Zhu, and W. Zhu, "Subspace learning for unsupervised feature selection via matrix factorization," *Pattern Recognit.*, vol. 48, no. 1, pp. 10–19, 2015.
[42] R. Shang, W. Wang, R. Stolkin, and L. Jiao, "Non-negative spectral learning and sparse regression-based dual-graph regularized feature selection," *IEEE Trans. Cybern.*, vol. 48, no. 2, pp. 793–806, Feb. 2018.
[43] K.-C. Lee, J. Ho, and D. Kriegman, "Acquiring linear subspaces for face recognition under variable lighting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 5, pp. 684–698, May 2005.
[44] T. Sim, S. Baker, and M. Bsat, "The CMU pose, illumination, and expression database," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 12, pp. 1615–1618, Dec. 2003.
[45] A. Martinez, "The AR face database," CVC, Luxembourg, Tech. Rep. 24, 1998.
[46] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with Gabor wavelets," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, Apr. 1998, pp. 200–205.
[47] F. S. Samaria and A. C. Harter, "Parameterisation of a stochastic model for human face identification," in *Proc. 2nd IEEE Workshop Appl. Comput. Vis.*, vol. 22, Dec. 1995, pp. 138–142.
[48] S. Nene, S. Nayar, and H. Murase, "Columbia object image library (COIL-20)," Columbia Univ., New York, NY, USA, Tech. Rep. CUCS-005-96, 1996.
[49] X. Fang, Y. Xu, X. Li, Z. Lai, S. Teng, and L. Fei, "Orthogonal self-guided similarity preserving projection for classification and clustering," *Neural Netw.*, vol. 88, pp. 1–8, Apr. 2017.

**JIANGYAN DAI** received the B.S. and M.S. degrees from the Computer School, Northeast Normal University, China, in 2008 and 2010, respectively, and the Ph.D. degree from the School of Mathematics and Statistics, Northeast Normal University, in 2014. She is currently a Lecturer with the School of Computer Engineering, Weifang University. Her main research interests are digital image processing, computer vision, biometrics, and information security.

**YUQI CHEN** was born in Huizhou, Guangdong, China, in 1993. She received the B.S. degree from the College of Software, Jiangxi Normal University, China, in 2017, where she is currently pursuing the M.S. degree with the Software College. Her research interests include dimensionality reduction and feature extraction.

**YUGEN YI** was born in Pingxiang, Jiangxi, China, in 1986. He received the B.S. degree from the College of Humanities and Sciences, Northeast Normal University, China, in 2009, the M.S. degree from the College of Computer Science and Information Technology, Northeast Normal University, in 2012, and the Ph.D. degree from the School of Mathematics and Statistics, Northeast Normal University, in 2015. He is currently a Lecturer with the School of Software, Jiangxi Normal University. His research interests include dimensionality reduction and feature extraction.

**JINING BAO** received the B.S. degree in communication engineering and the M.S. degree in measurement and control engineering from the Shenyang University of Technology, Shenyang, China. She is currently pursuing the Ph.D. degree in pattern recognition and intelligent systems with Northeastern University, Shenyang. Her research interests include image processing, wireless sensor networks, intelligent robot, and artificial intelligence.

**LEI WANG** received the M.S. degree from the School of Software, Shandong University, China, in 2006, and the Ph.D. degree from the School of Computer Science and Technology, Shandong University, in 2010. He is currently an Associate Professor with Weifang University. His research interests are in the areas of graphics, vision, and human–computer interaction.

**WEI ZHOU** received the M.S. degree in computer science and technology from Northeast Normal University, Changchun, in 2015. She is currently pursuing the Ph.D. degree with Northeastern University, Shenyang, China. Her research interests include medical imaging processing, dimensionality reduction, and feature selection.

**GANG LEI** is currently an Associate Professor and a Master Supervisor with Jiangxi Normal University. His research interests include data mining and machine learning. He is also a member of the China Computer Federation.