Contents lists available at ScienceDirect

Pattern Recognition

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

Unsupervised feature selection by regularized self-representation

Pengfei Zhu [a,*], Wangmeng Zuo [a,b], Lei Zhang [a], Qinghua Hu [c], Simon C.K. Shiu [a]

[a] Department of Computing, The Hong Kong Polytechnic University, Hong Kong, China
[b] School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China
[c] School of Computer Science and Technology, Tianjin University, Tianjin 300072, China

ARTICLE INFO

Article history:
Received 8 January 2014
Received in revised form
16 July 2014
Accepted 9 August 2014
Available online 19 August 2014

Keywords:
Self-representation
Unsupervised feature selection
Sparse representation
Group sparsity

ABSTRACT

By removing the irrelevant and redundant features, feature selection aims to find a compact representation of the original feature with good generalization ability. With the prevalence of unlabeled data, unsupervised feature selection has shown to be effective in alleviating the curse of dimensionality, and is essential for comprehensive analysis and understanding of myriads of unlabeled high dimensional data. Motivated by the success of low-rank representation in subspace clustering, we propose a regularized self-representation (RSR) model for unsupervised feature selection, where each feature can be represented as the linear combination of its relevant features. By using $L_{2,1}$-norm to characterize the representation coefficient matrix and the representation residual matrix, RSR is effective to select representative features and ensure the robustness to outliers. If a feature is important, then it will participate in the representation of most of other features, leading to a significant row of representation coefficients, and vice versa. Experimental analysis on synthetic and real-world data demonstrates that the proposed method can effectively identify the representative features, outperforming many state-of-the-art unsupervised feature selection methods in terms of clustering accuracy, redundancy reduction and classification accuracy.

1. Introduction

The explosive use of electronic sensors and social media produces a huge amount of high-dimensional data [1,2], and the high dimensionality greatly increases the time and space complexities for data processing, making the clustering and classification methods, which are analytically or computationally manageable in low-dimensional space, completely intractable [3]. Feature selection is an important step to remove the irrelevant and redundant features from the original data [4], alleviating the curse of dimensionality, reducing the storage space and time complexity, and building a compact data representation with good generalization ability [5,6]. In recent years, continuous efforts have been made to develop new feature selection algorithms [3,6–12].

Feature selection methods can be categorized into unsupervised and supervised ones [4,5,13]. Supervised feature selection methods include wrapper models and filter models. Wrapper models search in the space of feature subset, and employ one classifier to repeatedly evaluate the goodness of the selected feature subsets, making it computationally intensive and intractable for large-scale problems [5]. Filter models are independent of certain classifiers and they use some feature evaluation indices to rank features or evaluate feature subsets, e.g., Fisher score.

In many data mining applications, sample labels are unknown, therefore making unsupervised feature selection indispensable [14]. Early unsupervised feature selection methods mainly use some evaluation indices to evaluate each individual feature or feature subset, and then select the top *K* features or the best feature subset. These indices evaluate the clustering performance, redundancy, information loss, sample similarity or manifold structure, e.g., variance [5], Laplacian score [7], or trace ratio [15]. These methods, however, are computationally expensive in searching. To reduce the computational cost, a feature clustering method is proposed in [14] to find the representative features based on feature similarity without searching. Recently, a series of algorithms have been developed based on spectral clustering techniques to select a feature subset that best preserves the sample similarity [3,6,7,15–17]. In [7,15,16], features are selected one by one and the correlation between features is totally ignored [9], while in [3,6,17], the importance of features is evaluated jointly and features are selected in batch.

On the other hand, sparsity regularization has been widely used in feature selection and shown good effectiveness, robustness and efficiency, e.g., L_1-SVM [18] and sparse logistic regression [19]. Group sparsity, which is often used in multi-task learning [20] and joint representation [21], has also been applied to feature selection. By modeling feature selection as a loss minimization problem, in [8,9,6,17,22] group sparsity is imposed on the feature weights matrix to select features. The $L_{2,1}$-norm group sparsity

[*] Corresponding author. Tel.: +85267610177.
E-mail address: zhupengfeifly@gmail.com (P. Zhu).

http://dx.doi.org/10.1016/j.patcog.2014.08.006
0031-3203/© 2014 Elsevier Ltd. All rights reserved.

regularization has been adopted and shown good performance to remove the redundancy in feature selection [6,23].

Unlike supervised feature selection, in unsupervised feature selection the class label information is unavailable to guide the selection of minimal feature subset. In this paper, we find that the self-representation property of redundant features, which characterizes the property that each feature can be well approximated by the linear combination of its relevant features, also provides some insights on unsupervised feature selection. In nature, self-similarity widely exists, i.e., a part of an object is similar to other parts of itself, e.g., coastlines [24], stock market movements [25] and images [26]. Taking images for example, patches at different locations in an image perhaps are similar to each other, which is called non-local self-similarity. In image processing, the so-called non-local self-similarity has been successfully used in high performance image restoration and denoising [26]. Based on self-similarity of objects in nature, self-representation property generally holds for most high dimensional data, and has been extensively used in machine learning and computer vision fields. Just as sparsity leads to sparse representation, self-similarity results in self-representation.

With the above considerations, in this paper we propose a simple yet very effective unsupervised feature selection method by exploiting the self-representation ability of features. The feature matrix is represented over itself to find the representative feature components. The representation residual is minimized by $L_{2,1}$-norm loss to reduce the effect of outlier samples. Different from the other applications, in unsupervised feature selection, our goal is to identify a representative feature subset so that all the features can be well reconstructed by them. Thus, $L_{2,1}$-norm regularization is imposed on the representation coefficients to enforce group sparsity. With the proposed regularized self-representation model, if a feature is important, it will participate in the representation of other features and hence produce a significant row of representation coefficients and vice versa. To the best of our knowledge, this work is the first attempt to conduct unsupervised feature selection from the viewpoint of feature self-representation. Extensive experiments have been performed on synthetic and real-world data sets, and the results validate the leading performance of the proposed method in terms of clustering, redundancy and classification evaluation measures.

The rest of this paper is organized as follows: Section 2 introduces the unsupervised feature selection task; in Section 3, regularized self-representation is proposed; Section 4 presents the optimization and algorithms; Section 5 discusses the relationships between RSR and low rank representation; Section 6 conducts experiments and Section 7 concludes this paper.

2. Problem statement

The objective of unsupervised feature selection is to select a desired feature subset from a given dataset without label information. The real-world data are often very redundant in features and can have outlier samples. Fig. 1(a) illustrates a corrupted data matrix. Each row vector is a sample and each column vector represents one feature of all samples. The shaded central column represents a redundant feature, and the shaded central row represents an outlier sample. As shown in Fig. 1(b) and (c), a robust and effective feature selection algorithm should eliminate the effect of the outlier samples and indicate the redundant features.

Let $\boldsymbol{X} \in \Re^{n \times m}$ be a data matrix, where n and m are the numbers of samples and features, respectively. We use $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ to represent the n samples, $\boldsymbol{x}_i \in \Re^m$ and $\boldsymbol{X} = [\boldsymbol{x}_1; \boldsymbol{x}_2; \ldots; \boldsymbol{x}_n]$. We use $f_1, f_2, \ldots, f_m$ to denote the m features, and $\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_m$ are the corresponding feature vectors, where $\boldsymbol{f}_i \in \Re^n$ and $\boldsymbol{X} = [\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_m]$.

Early unsupervised feature selection methods use some metrics (e.g, variance, Laplacian score [7]) to evaluate each feature, and then rank the features by the evaluated metric values. The recently developed methods [17,6,9,3] usually first calculate the sample similarity or sample manifold structure, and then build a response matrix $\boldsymbol{Y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_m]$. The feature selection problem is then converted into a multi-output regression problem:

$$\min_{\boldsymbol{W}} l(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{W}) + \lambda R(\boldsymbol{W}) \tag{1}$$

where $\boldsymbol{W}$ is the feature weight matrix, $l(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{W})$ is the loss item, $R(\boldsymbol{W})$ is the regularization imposed on $\boldsymbol{W}$ and λ is a positive constant.

In Eq. (1), the response matrix $\boldsymbol{Y}$ is known before the optimization phase and $\boldsymbol{W}$ is the variable. $\boldsymbol{Y}$ contains the sample similarity information and it is calculated differently in different methods. Taking minimum redundancy feature selection (MRFS) [6] for example, the sample similarity matrix $\boldsymbol{S}$ is first calculated, and then the elements of $\boldsymbol{Y}$ are determined as $\boldsymbol{y}_k = \lambda_k^{1/2} \boldsymbol{\xi}_k$, where λ_k and $\boldsymbol{\xi}_k$ are the kth eigenvalue and eigenvector of normalized similarity matrix $\hat{\boldsymbol{S}}$.

3. Regularized self-representation

The model in Eq. (1) considers the data similarity and selects features jointly. Though it is widely used in many feature selection methods, it is difficult to choose the proper response matrix. Thanks to self-representation property of features, in this section we propose a regularized self-representation (RSR) model for unsupervised feature selection. The proposed RSR model simply uses the data matrix $\boldsymbol{X}$ as the response matrix , i.e., $\boldsymbol{Y} = \boldsymbol{X}$, which is more natural and can be well interpreted by the self-representation principle, i.e., each feature can be well represented by all features. For each feature $\boldsymbol{f}_i$ in $\boldsymbol{X}$, we represent it as a linear combination of other features (including itself):

$$\boldsymbol{f}_i = \sum_{j=1}^{m} \boldsymbol{f}_j w_{ji} + \boldsymbol{e}_i \tag{2}$$

Then for all the features, we have

$$\boldsymbol{X} = \boldsymbol{X}\boldsymbol{W} + \boldsymbol{E} \tag{3}$$

where $\boldsymbol{W} = [w_{ji}] \in \Re^{m \times m}$ is the representation coefficients matrix. The above representation model is a kind of self-representation of features.

Clearly, the matrix $\boldsymbol{W}$ to be learned should reflect the importance of different features while making the representation residual $\boldsymbol{E}$ small. One may use the Frobenius norm to measure the residual, i.e., $\min_{\boldsymbol{W}} \|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{W}\|_F^2$. However, as illustrated in Fig. 1, there can be some outlier samples in the data matrix $\boldsymbol{X}$, while the Frobenius norm is sensitive to outliers. Considering that an outlier sample is a row of the matrix $\boldsymbol{X}$, and its representation residual is a row in the matrix $\boldsymbol{E} = \boldsymbol{X} - \boldsymbol{X}\boldsymbol{W}$, we propose to use the $L_{2,1}$−norm to characterize $\boldsymbol{E}$; that is, we impose row-sparsity on $\boldsymbol{E}$ to enforce robustness to outlier samples. Meanwhile, if we let $\boldsymbol{W}$ be an $m \times m$ identity matrix, a trivial solution will be obtained with the residual $\boldsymbol{E} = 0$. Thus, a regularization item $R(\boldsymbol{W})$ must be introduced to avoid the trivial solution of $\boldsymbol{W}$ and guide the selection of feature subset. Then we have the following minimization problem:

$$\hat{\boldsymbol{W}} = \arg\min_{\boldsymbol{W}} \|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{W}\|_{2,1} + \lambda R(\boldsymbol{W}) \tag{4}$$

Let $\boldsymbol{W} = [\boldsymbol{w}_1; \ldots; \boldsymbol{w}_i; \ldots; \boldsymbol{w}_m]$, where $\boldsymbol{w}_i$ is ith row of $\boldsymbol{W}$. $\|\boldsymbol{w}_i\|_2$ can be used as the feature weight because it reflects the importance of the ith feature in representation. For example, if $\|\boldsymbol{w}_i\|_2 = 0$, it means that the ith feature will contribute nothing to the representation of other features. If the ith feature take part in the representation of all features, then $\|\boldsymbol{w}_i\|_2$ must be significant. Therefore, the row-sparsity is expected for regularizing the coefficients matrix $\boldsymbol{W}$. We let

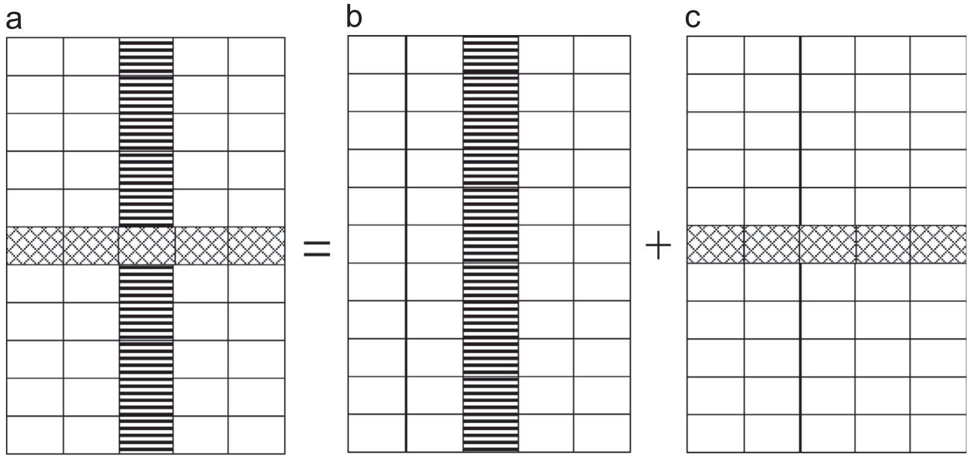

Fig. 1. A data matrix with outliers and redundant features. (a) Corrupted data matrix (b) Redundant features and (c) Outliers.

$R(\boldsymbol{W}) = \|\boldsymbol{W}\|_{2,1} = \sum_{i=1}^{m} \|\boldsymbol{w}_i\|_2$. Consequently, Eq. (4) becomes

$$\hat{\boldsymbol{W}} = \arg\min_{\boldsymbol{W}} \|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{W}\|_{2,1} + \lambda \|\boldsymbol{W}\|_{2,1} \tag{5}$$

We call the above model as the regularized self-representation (RSR) for unsupervised feature selection.

4. Optimization and algorithms

The RSR model in Eq. (5) is convex, but both the loss and the regularization terms are non-smooth. In this section, we first propose an iterative reweighted least-squares (IRLS) algorithm to solve the RSR model, and then prove its convergence in the bound optimization framework [27].

For the IRLS algorithm, given the current estimation $\boldsymbol{W}^t$, we define the diagonal weighting matrices $\boldsymbol{G}_L^t$ and $\boldsymbol{G}_R^t$ by $g_{L,i}^t = 1/2\|\boldsymbol{x}_i - \boldsymbol{x}_i\boldsymbol{W}^t\|$ and $g_{R,j}^t = 1/2\|\boldsymbol{w}_j^t\|_2$, and then $\boldsymbol{W}^{t+1}$ is updated by solving the following weighted least squares problem:

$$\begin{aligned} \boldsymbol{W}^{t+1} &= \arg\min_{\boldsymbol{W}} Q(\boldsymbol{W}|\boldsymbol{W}^t) \\ &= \arg\min_{\boldsymbol{W}} \left\{ \begin{array}{l} \operatorname{tr}((\boldsymbol{X} - \boldsymbol{X}\boldsymbol{W})^T \boldsymbol{G}_L^t (\boldsymbol{X} - \boldsymbol{X}\boldsymbol{W})) \\ +\lambda \operatorname{tr}(\boldsymbol{W}^T \boldsymbol{G}_R^t \boldsymbol{W}) \end{array} \right\} \end{aligned} \tag{6}$$

The closed form solution of $\boldsymbol{W}^{t+1}$ can be obtained by

$$\boldsymbol{W}^{t+1} = ((\boldsymbol{G}_R^t)^{-1} \boldsymbol{X}^T \boldsymbol{G}_L^t \boldsymbol{X} + \lambda \boldsymbol{I})^{-1} (\boldsymbol{G}_R^t)^{-1} \boldsymbol{X}^T \boldsymbol{G}_L^t \boldsymbol{X}$$

where $\boldsymbol{I} \in \Re^{n \times n}$ is the identity matrix. To avoid the overflow error, a sufficiently small value ε is introduced by defining $g_{L,i}^t = 1/\max(2\|\boldsymbol{x}_i - \boldsymbol{x}_i\boldsymbol{W}^t\|_2, \varepsilon)$ and $g_{R,j}^t = 1/\max(2\|\boldsymbol{w}_j^t\|_2, \varepsilon)$.

Table 1 summarizes the IRLS algorithm for solving RSR. The features can be ranked according to the values of $\|\boldsymbol{w}_j\|_2$ and then the top K features f^s can be selected. The selected feature subset can be used for clustering and classification. In RSR, we mainly need to update $\boldsymbol{W}$ in each iteration, whose computational complexity is basically $O(m^3 + m^2 n)$, where m and n are the number of features and samples, respectively. Hence, the time complexity of RSR is $O(T(m^3 + m^2 n))$, where T is the total number of iterations.

Based on the property that the trace function is invariant on cyclic permutation, we have

$$Q(\boldsymbol{W}|\boldsymbol{W}^t) = \operatorname{tr}(\boldsymbol{G}_L^t (\boldsymbol{X} - \boldsymbol{X}\boldsymbol{W})(\boldsymbol{X} - \boldsymbol{X}\boldsymbol{W})^T) + \lambda \operatorname{tr}(\boldsymbol{G}_R^t \boldsymbol{W}\boldsymbol{W}^T) \tag{7}$$

Note that $\boldsymbol{G}_L^t$ and $\boldsymbol{G}_R^t$ are diagonal. $Q(\boldsymbol{W}|\boldsymbol{W}^t)$ can be rewritten as

$$Q(\boldsymbol{W}|\boldsymbol{W}^t) = \sum_i \frac{\|\boldsymbol{x}_i - \boldsymbol{x}_i\boldsymbol{W}\|_2^2}{2\|\boldsymbol{x}_i - \boldsymbol{x}_i\boldsymbol{W}^t\|_2} + \lambda \sum_j \frac{\|\boldsymbol{w}_j\|_2^2}{2\|\boldsymbol{w}_j^t\|_2} \tag{8}$$

Table 1
Algorithm of regularized self-representation (RSR) based unsupervised feature selection.

```
Input: Data matrix $\boldsymbol{X} \in \Re^{n \times m}$ and $\lambda$
Output: Feature weights vector $\boldsymbol{v} = [v_1, v_2, \ldots, v_m]$

Set $t$=0. Initialize $\boldsymbol{G}_L^t$ and $\boldsymbol{G}_R^t$ as the identity matrices
Repeat
    $\boldsymbol{W}^{t+1} = ((\boldsymbol{G}_R^t)^{-1} \boldsymbol{X}^T \boldsymbol{G}_L^t \boldsymbol{X} + \lambda \boldsymbol{I})^{-1} ((\boldsymbol{G}_R^t)^{-1} \boldsymbol{X}^T \boldsymbol{G}_L^t \boldsymbol{X})$;
    update$\boldsymbol{G}_L^{t+1}$ and $\boldsymbol{G}_R^{t+1}$;
    $t = t+1$;
Until Convergence
Calculate feature weights $v_i = \|\boldsymbol{w}_i\|_2$, $i = 1, 2, \ldots, m$
```

Let

$$L(\boldsymbol{W}) = \|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{W}\|_{2,1} + \lambda \|\boldsymbol{W}\|_{2,1} \tag{9}$$

We get the following two theorems:

Theorem 1. *$Q(\boldsymbol{W}|\boldsymbol{W}^t)$ is a surrogate function, i.e., $L(\boldsymbol{W}) - Q(\boldsymbol{W}|\boldsymbol{W}^t)$ attains its maximum when $\boldsymbol{W} = \boldsymbol{W}^t$.*

Proof. Let $F(\boldsymbol{W}) = L(\boldsymbol{W}) - Q(\boldsymbol{W}|\boldsymbol{W}^t)$. In the following, we will prove that for any $\boldsymbol{W}$, there is $F(\boldsymbol{W}^t) - F(\boldsymbol{W}) \geq 0$. First, $F(\boldsymbol{W}^t)$ can be rewritten as

$$\begin{aligned} F(\boldsymbol{W}^t) &= L(\boldsymbol{W}^t) - Q(\boldsymbol{W}^t|\boldsymbol{W}^t) \\ &= \sum_i \|\boldsymbol{x}_i - \boldsymbol{x}_i\boldsymbol{W}^t\|_2 + \lambda \sum_j \|\boldsymbol{w}_j^t\|_2 \\ &\quad - \left(\sum_i \frac{\|\boldsymbol{x}_i - \boldsymbol{x}_i\boldsymbol{W}^t\|_2^2}{2\|\boldsymbol{x}_i - \boldsymbol{x}_i\boldsymbol{W}^t\|_2} + \lambda \sum_j \frac{\|\boldsymbol{w}_j^t\|_2^2}{2\|\boldsymbol{w}_j^t\|_2} \right) \\ &= \frac{1}{2} \left(\sum_i \|\boldsymbol{x}_i - \boldsymbol{x}_i\boldsymbol{W}^t\|_2 + \lambda \sum_j \|\boldsymbol{w}_j^t\|_2 \right) \end{aligned}$$

Then, $F(\boldsymbol{W}^t) - F(\boldsymbol{W})$ can be rewritten as

$$\begin{aligned} &F(\boldsymbol{W}^t) - F(\boldsymbol{W}) \\ &= \frac{1}{2} \left(\sum_i \|\boldsymbol{x}_i - \boldsymbol{x}_i\boldsymbol{W}^t\|_2 + \lambda \sum_j \|\boldsymbol{w}_j^t\|_2 \right) \\ &\quad - \left(\begin{array}{c} \sum_i \|\boldsymbol{x}_i - \boldsymbol{x}_i\boldsymbol{W}\|_2 + \lambda \sum_j \|\boldsymbol{w}_j\|_2 \\ - \sum_i \frac{\|\boldsymbol{x}_i - \boldsymbol{x}_i\boldsymbol{W}\|_2^2}{2\|\boldsymbol{x}_i - \boldsymbol{x}_i\boldsymbol{W}^t\|_2} - \lambda \sum_j \frac{\|\boldsymbol{w}_j\|_2^2}{2\|\boldsymbol{w}_j^t\|_2} \end{array} \right) \\ &= \sum_i \frac{1}{2\|\boldsymbol{x}_i - \boldsymbol{x}_i\boldsymbol{W}^t\|_2} (\|\boldsymbol{x}_i - \boldsymbol{x}_i\boldsymbol{W}^t\|_2 - \|\boldsymbol{x}_i - \boldsymbol{x}_i\boldsymbol{W}\|_2)^2 \end{aligned}$$

$$+\sum_j \frac{\lambda}{2\|\boldsymbol{w}_j^t\|_2}(\|\boldsymbol{w}_j^t\|_2-\|\boldsymbol{w}_j\|_2)^2$$

Because $1/2\|\boldsymbol{x}_i-\boldsymbol{x}_i\boldsymbol{W}^t\|_2 \geq 0$ and $1/2\|\boldsymbol{w}_j^t\|_2 \geq 0$, one can easily see that $F(\boldsymbol{W}^t)-F(\boldsymbol{W}) \geq 0$, and $Q(\boldsymbol{W}|\boldsymbol{W}^t)$ is a surrogate function. □

In the bound optimization framework, the loss function $L(\boldsymbol{W})$ can be minimized by iteratively minimizing the surrogate function $Q(\boldsymbol{W}|\boldsymbol{W}^t)$, and we can obtain the following theorem:

Theorem 2. *Let* $\boldsymbol{W}^{t+1}=\arg\min_{\boldsymbol{W}} Q(\boldsymbol{W}|\boldsymbol{W}^t)$. *We have* $L(\boldsymbol{W}^{t+1}) \leq L(\boldsymbol{W}^t)$.

Proof. It is easy to see that

$$\begin{aligned}
L(\boldsymbol{W}^{t+1}) &= L(\boldsymbol{W}^{t+1})-Q(\boldsymbol{W}^{t+1}|\boldsymbol{W}^t)+Q(\boldsymbol{W}^{t+1}|\boldsymbol{W}^t)\\
&\quad (\text{Note}: \boldsymbol{W}^t=\arg\max_{\boldsymbol{W}} L(\boldsymbol{W})-Q(\boldsymbol{W}|\boldsymbol{W}^t))\\
&\leq L(\boldsymbol{W}^t)-Q(\boldsymbol{W}^t|\boldsymbol{W}^t)+Q(\boldsymbol{W}^{t+1}|\boldsymbol{W}^t)\\
&\quad (\text{Note}: \boldsymbol{W}^{t+1}=\arg\min_{\boldsymbol{W}} Q(\boldsymbol{W}|\boldsymbol{W}^t))\\
&\leq L(\boldsymbol{W}^t)-Q(\boldsymbol{W}^t|\boldsymbol{W}^t)+Q(\boldsymbol{W}^t|\boldsymbol{W}^t)\\
&= L(\boldsymbol{W}^t)
\end{aligned}$$

where the first inequality which stems from that $Q(\boldsymbol{W}|\boldsymbol{W}^t)$ is a surrogate function, and the second inequality results from that $\boldsymbol{W}^{t+1}=\arg\min_{\boldsymbol{W}} Q(\boldsymbol{W}|\boldsymbol{W}^t)$. □

Thus, the proposed IRLS algorithm can decrease the loss function in each iteration, and $L(\boldsymbol{W})$ can be optimized by iteratively minimizing $Q(\boldsymbol{W}|\boldsymbol{W}^t)$. Note that $\boldsymbol{W}^{t+1}$ has a closed form solution in each iteration and to improve the stability of the computation, a regularized version of G_L^t and G_R^t is used. Based on the proof above, IRLS can converge to a stationary point very efficiently.

We use an example to better illustrate the mechanism of the proposed feature selection algorithm. We extract 700 face images of 100 subjects from the AR face database. Each face image is resized to 60 × 43. Each image is stretched to a vector and input as a row of the data matrix $\boldsymbol{X}$. By applying the proposed RSR algorithm to $\boldsymbol{X}$ (λ is set to 0.1), $\boldsymbol{W}$ is obtained. The learned feature weights are shown in Fig. 2(a). In Fig. 2(b), (c), and(d), we show an original face, the reconstructed face by using the learned weight matrix $\boldsymbol{W}$, and the representation residual, respectively. We can see from $\boldsymbol{W}$ that the proposed RSR algorithm can identify the most informative parts of a face, e.g., eyes, nose and mouth. Besides, the reconstructed face image is close to the original face image, and the representation residual is random noise like.

To show the insensitiveness of RSR to outliers, we select 100 object images from Caltech101 database. Then 700 face images and 100 object images are combined together to yield a mixed dataset. After $\boldsymbol{W}$ is learned by RSR, the error matrix $\boldsymbol{E}=[\boldsymbol{e}_1;\ldots,\boldsymbol{e}_i;\ldots;\boldsymbol{e}_n]$ is obtained. Then $\|\boldsymbol{e}_i\|_2$ is used to recognize the outlier samples, that is, the outlier samples have larger errors. The errors of 800 samples are shown in Fig. 3. We can see that the last 100 object images from Caltech101 get much larger errors than the first 700 face images from AR. Besides, the learned feature map is similar to Fig. 2(a), which shows that $\boldsymbol{W}$ is almost not affected by the outlier samples.

To further validate the impact of outliers on RSR, we randomly select one object image and combine it with 700 face images to obtain a mixed dataset. As shown in Fig. 4, the final sample is the outlier object image, which achieves the largest error for four cases. The result show whether a set of object images or a single object image exist in the face image sets, it can be detected by RSR, and thus the impact of outliers on RSR can be alleviated.

5. Discussions

In this section, we reveal the relationships among self-representation, sparse representation and low-rank representation. Besides, the bottleneck of RSR in feature selection is also discussed.

Relationships with SR and LRR: Recently, representation based techniques, such as sparse representation (SR) and low-rank representation (LRR), have been successfully applied to image restoration [28], face recognition [29] and subspace segmentation [30]. In this section, we analyze the relationships among RSR, SR and LRR.

SR aims to represent a test sample $\boldsymbol{z}$ over a dictionary $\boldsymbol{D}=[\boldsymbol{d}_1;\boldsymbol{d}_2;\ldots;\boldsymbol{d}_k]\in\Re^{k\times m}$: $\boldsymbol{z}^T=\sum_{j=1}^n a_j\boldsymbol{d}_j^T+\boldsymbol{e}=\boldsymbol{D}^T\boldsymbol{a}+\boldsymbol{e}$, where $\boldsymbol{a}=[a_1,a_2\ldots,a_k]\in\Re^k$ is the representation coefficient and $\boldsymbol{e}\in\Re^m$ is the representation residual. $\boldsymbol{D}$ is often over-complete, and a sparse solution of $\boldsymbol{a}$ can be obtained by solving the following optimization problem:

$$\hat{\boldsymbol{a}}=\arg\min_{\boldsymbol{a}}\left\|\boldsymbol{z}^T-\boldsymbol{D}^T\boldsymbol{a}\right\|_2^2+\lambda\left\|\boldsymbol{a}\right\|_1 \tag{10}$$

The dictionary $\boldsymbol{D}$ can be the original training samples or learned by dictionary learning methods. In signal reconstruction, the sample $\boldsymbol{z}$ can be reconstructed as $\boldsymbol{D}^T\hat{\boldsymbol{a}}$. For example, sparse subspace clustering utilizes the sparse coding coefficients $\hat{\boldsymbol{a}}$ for subspace

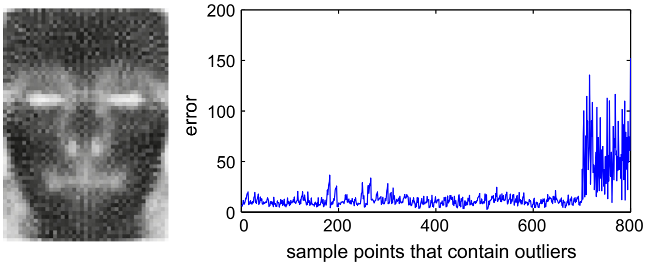

Fig. 3. The learned feature weight map and the error $\|\boldsymbol{e}_i\|_2$.

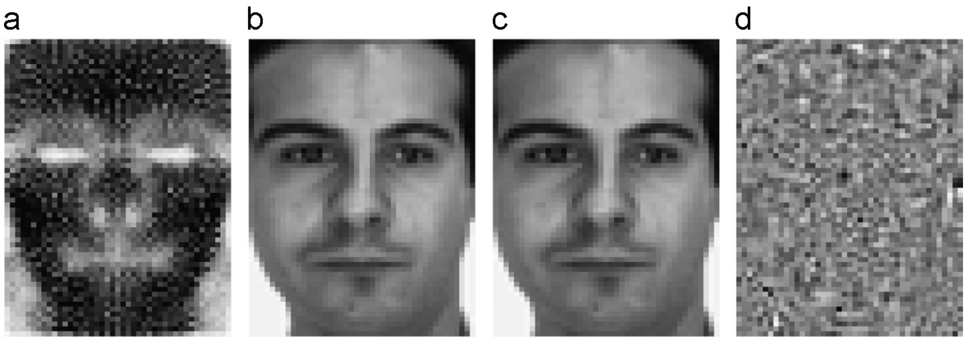

Fig. 2. (a) The learned feature weight map; (b) a raw face; (c) reconstructed face by the learned weights; (d) representation residual.

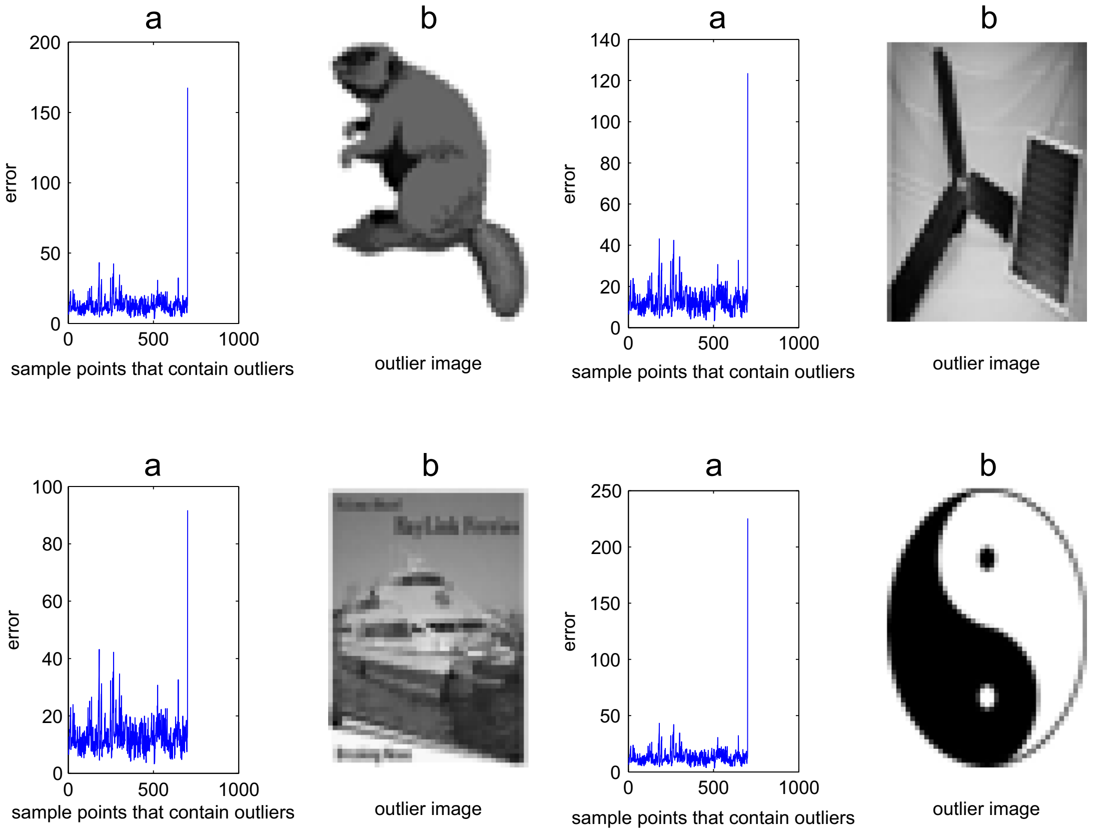

Fig. 4. (a) The error $\|\boldsymbol{e}_i\|_2$ of face images and single randomly selected object image; (b) outlier object image.

segmentation [31]. In signal classification, the label information of the atoms in $\boldsymbol{D}$ is available, and the reconstruction error of each class can be used to classify $\boldsymbol{z}$, as done in the sparse representation based classifier [29].

In LRR [30], the goal is to do subspace segmentation. The original training samples in $\boldsymbol{X}$ are used as the dictionary to represent $\boldsymbol{X}$ itself: $\boldsymbol{X}^T = \boldsymbol{X}^T\boldsymbol{A} + \boldsymbol{E}$. LRR minimizes the representation error $\boldsymbol{E}$ with low-rank regularization imposed on the coefficient matrix $\boldsymbol{A}$:

$$\hat{\boldsymbol{A}} = \arg\min_{\boldsymbol{A}} \|\boldsymbol{X}^T - \boldsymbol{X}^T\boldsymbol{A}\|_{2,1} + \lambda\|\boldsymbol{A}\|_* \tag{11}$$

where $\|\boldsymbol{A}\|_*$ is the nuclear norm of $\boldsymbol{A}$.

Both SR and LRR belong to sample-level representation while the proposed RSR is feature-level representation. According to different tasks, self-representation can also be modeled as a sparse- or a low-rank representation model. The goal in this paper is to select features in unsupervised tasks, and we propose the $L_{2,1}$-norm RSR model in Eq. (5).

Limitation of RSR: Self-representation is based on the widely existing fact in nature that there is redundancy in features and features are usually linearly correlated. However, in some cases, all features are independent, or the correlation between features is quite weak, or the correlation between features is non-linear. Then self-representation perhaps does not work very well.

6. Experiments

6.1. Experiment setup

Datasets: We evaluate the performance of RSR on synthetic and real-world datasets. A diversity of six real-world benchmark datasets is selected to compare RSR with different unsupervised feature selection algorithms. The six datasets include one handwritten digit dataset (i.e., USPS[1]), one spoken letter dataset (i.e., ISOLET[2]), one face dataset (i.e., AR10P[3]), and three microarray datasets (i.e., CLL-SUB-1116,[4] TOX-1717[5] and Prostate-GE[6]). The detailed information of the six datasets is summarized in Table 2. The number of features varies from 256 to 11,340 and the feature types include image and microarray.

Comparison methods: We compare the proposed RSR method with the following representative and state-of-the-art unsupervised feature selection methods:

C 1. *FSFS*[7] [14]: Feature selection using feature similarity.
C 2. *Laplacian score*[8] [7]: Select features which are most consistent with the Gaussian Laplacian matrix.
C 3. *MCFS*[9] [3]: Select features using spectral clustering with L_1-norm regularization.
C 4. *UDFS*[10] [9]: Discriminative feature selection with $L_{2,1}$-norm regularization.
C 5. *SPEC*[11] [16]: Select features using spectral clustering.

[1] http://www.csie.ntu.edu.tw/cjlin/libsvmtools/datasets/multiclass.html#usps
[2] http://archive.ics.uci.edu/ml/datasets/ISOLET
[3] http://featureselection.asu.edu/datasets.php
[4] http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE2466
[5] http://www.ncbi.nlm.nih.gov/sites/GDSbrowser
[6] http://www.ncbi.nlm.nih.gov/pubmed/12381711
[7] http://cse.iitkgp.ac.in/pabitra/paper.html
[8] http://www.cad.zju.edu.cn/home/dengcai/Data/ReproduceExp.html#LaplacianScore
[9] http://www.cad.zju.edu.cn/home/dengcai/Data/ReproduceExp.html#MCFS
[10] http://www.cs.cmu.edu/yiyang/Publications.html
[11] https://docs.google.com/file/d/0BwtPEpCafuLyWm1hZWdXUnNYZE0/edit

Table 2
Summary of the benchmark datasets.

Data	Instances	Features	Classes	Keywords
USPS	9298	256	10	Image, Hand-written digit
ISOLET	1560	617	26	Mixed feature, Spoken letter
AR10P	130	2400	10	Image, Face
CLL-SUB-111	111	11,340	3	Microarray, Bio
TOX-171	171	5748	4	Microarray, Bio
Prostate-GE	102	5966	2	Microarray, Bio

C 6. *MRFS*[12] [6]: Feature selection via minimum redundancy with $L_{2,1}$-norm regularization.

Evaluation metrics: Following the experiment setting in [16,6,9], we evaluate the unsupervised feature selection algorithms from three perspectives: clustering performance, classification performance and redundancy.

Two clustering evaluation metrics, clustering accuracy (ACC) and Normalized Mutual Information (NMI), are used to measure the clustering performance. Denote by q_i the clustering results and by p_i the true label of $\boldsymbol{x}_i$. ACC is defined as follows:

$$ACC = \frac{\sum_{i=1}^{n} \delta(p_i, map(q_i))}{n} \tag{12}$$

where $\delta(x, y) = 1$ if $x = y$; otherwise $\delta(x, y) = 0$. $map(q_i)$ is the best mapping function that permutes clustering labels to match the true labels using the Kuhn–Munkres algorithm. In clustering, we expect a large ACC. Given two variables P and Q, NMI is defined as

$$NMI(P, Q) = \frac{I(P, Q)}{\sqrt{H(P)H(Q)}} \tag{13}$$

where $H(P)$ and $H(Q)$ are the entropies of P and Q, respectively, and $I(P, Q)$ is the mutual information between P and Q. For clustering, P and Q are the clustering results and the true labels, respectively. NMI reflects the consistency between clustering results and ground truth labels. Hence, a large NMI is expected.

Assume that f^s is the set of selected features, the redundancy rate is measured by [6]

$$RED(f^s) = \frac{1}{m(m-1)} \sum_{f_i, f_j \in f^s, i > j} \rho_{i,j} \tag{14}$$

where $\rho_{i,j}$ is the correlation between feature f_i and feature f_j. A large value of $RED(f^s)$ means that many selected features are still significantly correlated. For classification accuracy, we use the nearest neighbor classifier to evaluate the classification performance.

Parameter setting: In methods Laplacian Score, MCFS and UDFS, the size of neighborhood k is set as 5 on all the datasets. In methods MCFS, UDFS and the proposed RSR, the regularization parameter needs to be chosen, while in Laplacian score and SPEC, the bandwidth parameter for Gaussian kernel needs to be chosen. For fair comparison, following the parameter setting in UDFS [9], we tune the regularization and bandwidth parameters from $\{0.001, 0.01, 0.1, 1, 10, 100, 1000\}$ and record the best result. For feature dimension, we set the number of features as $\{10, 20, 40, \ldots, 200\}$ and report the average results over different dimensions. The K-means clustering algorithm is performed on the selected features by different algorithms. Because the K-means clustering result varies with initialization, the experiment is run for 20 times with different random initializations and the average results are reported for all the competing algorithms.

6.2. Experiments on synthetic data

In this experiment, we use synthetic data to test whether the proposed RSR method can find the representative features. We first extract from the IRIS[13] dataset 150 samples and 4 features. Then 16 features are artificially generated as the linear combination of the 4 features. The combination coefficients are randomly generated with their sum being 1. To increase the difficulty in feature selection, the Gaussian white noise is added to the synthetic 16 features (generated by matlab function 'randn'). Finally we get a synthetic data matrix with 150 samples and 20 features (the first 4 features are the original features).

We apply the proposed RSR to this dataset and resolve the coefficients matrix $\boldsymbol{W}$ and the feature weights $\|\boldsymbol{w}_i\|_2$. The coefficients matrix $\boldsymbol{W}$ is shown in the left part of Fig. 5. We can clearly see that the coefficients of the four raw features are much larger than the other features, which are generated from the four features. The feature weights are shown in the right part of Fig. 5. The four raw features have much higher weights than other features, whose weights are close to zero. This validates that RSR can select the most representative features and the $L_{2,1}$-norm regularization is very useful in feature selection.

6.3. Experiments on real-world data

Comparison with feature selection methods: We then compare RSR with the other six unsupervised feature selection methods on the six real-world benchmark datasets. The clustering accuracy, NMI, classification accuracy and redundancy rate are shown in Tables 3, 4, 5 and 6, respectively. From these results we can see than RSR achieves the best performance among all the competing methods. The method FSFS [14] uses feature clustering to select the representative features, and it only considers the relationship between two features and selects features one by one. In contrast, the proposed RSR uses the self-representation property of features and considers their relationships jointly. For Laplacian score, SPEC, MCFS and MRFS, they all try to preserve the data similarity of the original feature space. MCFS and MRFS use a regression model to select features while Laplacian score and SPEC select features independently. The performance of MCFS and MRFS is better than Laplacian score and SPEC, which validates the superiority of converting unsupervised feature selection to a regression problem. The method UDFS introduces discriminative information and can also be considered as a regression model. Compared with RSR, UDFS achieves comparable clustering performance while it is inferior to RSR in terms of classification accuracy and redundancy rate.

Comparison with SSC and LRR: We then compare the performance of RSR with sparse subspace clustering (SSC) and low rank representation (LRR). Note that SSC and LRR are proposed for subspace segmentation rather than feature selection. Hence, we only evaluate the clustering performance and do not consider classification accuracy and redundancy. We collect the source code of SSC[14] [31] and LRR[15] [30]. The experiment results are shown in Tables 7 and 8. Note that SSC and LRR use all the features while for RSR we report the average clustering accuracy and NMI with a different number of features. Compared with SSC and LRR, the performance of RSR is totally much better although on some datasets SSC gets better NMI.

Comparison between raw and reconstructed features: After self-representation, all features are reconstructed, i.e., $\boldsymbol{X}$ becomes $\boldsymbol{XW}$.

[12] https://docs.google.com/file/d/0BwtPEpCafuLyRHBTNUxvTnNBMk0/edit

[13] http://archive.ics.uci.edu/ml/datasets/Iris

[14] http://www.vision.jhu.edu/code/

[15] https://sites.google.com/site/guangcanliu/

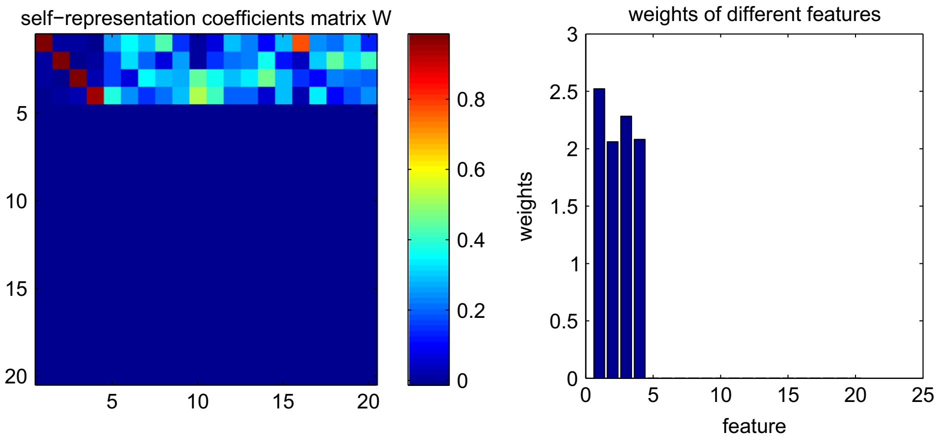

Fig. 5. Weight matrix and feature weights.

Table 3
Clustering results (ACC) of different feature selection algorithms.

Data	FSFS	Laplacian	MCFS	UDFS	SPEC	MRFS	RSR
USPS	51.5	60.2	62.6	63.1	40.6	59.8	**65.2**
ISOLET	36.5	50.8	53.1	54.2	45.8	49.5	**54.7**
AR10P	35.7	21.1	22.7	**46.3**	45.0	35.5	38.6
CLL-SUB-111	54.0	37.4	50.9	50.3	50.0	53.7	**54.3**
TOX-171	33.6	40.4	40.3	40.3	39.0	49.0	**49.3**
Prostate-GE	57.6	58.3	58.5	58.7	58.1	61.0	**61.4**
Average	44.8	44.7	48.0	52.1	46.4	51.4	**53.9**

Bold values are the best feature selection result.

Table 4
Clustering results (NMI) of different feature selection algorithms.

Data	FSFS	Laplacian	MCFS	UDFS	SPEC	MRFS	RSR
USPS	47.3	55.9	58.9	**60.2**	35.3	55.5	60.0
ISOLET	53.2	67.0	67.2	67.9	58.4	64.8	**69.1**
AR10P	38.9	20.2	20.0	**51.5**	47.1	39.1	40.9
CLL-SUB-111	22.9	2.9	19.7	14.9	19.0	22.0	**23.4**
TOX-171	6.9	11.9	11.8	11.4	10.2	**28.0**	26.3
Prostate-GE	3.0	2.4	2.2	**8.3**	2.2	5.3	5.6
Average	28.7	26.7	30.0	35.7	28.7	35.8	**37.5**

Bold values are the best feature selection result.

Table 5
Classification rates (%) of different feature selection algorithms.

Data	FSFS	Laplacian	MCFS	UDFS	SPEC	MRFS	RSR
USPS	89.5	87.5	**92.8**	71.3	60.2	88.6	**92.8**
ISOLET	46.8	69.5	71.1	44.7	62.8	68.2	**77.6**
AR10P	63.7	66.2	74.4	84.2	76.1	83.3	**89.1**
CLL-SUB-111	56.0	62.3	58.1	61.8	59.8	63.2	**66.3**
TOX-171	59.1	54.7	65.5	56.9	55.0	63.9	**64.5**
Prostate-GE	72.3	67.3	80.2	**81.9**	80.1	78.2	79.5
Average	64.6	67.9	73.7	66.8	65.7	74.2	**78.3**

Bold values are the best feature selection result.

Then we may wonder whether there is a performance gap between the raw features $\boldsymbol{X}$ and the reconstructed features $\boldsymbol{XW}$. We evaluate the performance on $\boldsymbol{X}$ and $\boldsymbol{XW}$ and the experiment result is shown in Table 9. RSR and RSR(R) represent using the original features and the reconstructed features, respectively. From the result, we can see that the difference between RSR and RSR

Table 6
Average redundancy rates of different feature selection algorithms.

Data	FSFS	Laplacian	MCFS	UDFS	SPEC	MRFS	RSR
USPS	14.6	21.8	15.9	12.1	11.3	**10.3**	16.5
ISOLET	47.4	47.1	23.6	38.5	26.2	13.4	**13.3**
AR10P	31.4	50.3	26.7	38.7	29.8	23.9	**18.9**
CLL-SUB-111	46.5	64.8	40.4	44.5	**10.6**	36.3	35.3
TOX-171	28.1	45.0	27.8	25.8	32.9	23.2	**22.5**
Prostate-GE	18.7	84.6	32.9	28.2	20.5	22.2	**19.7**
Average	31.1	52.3	27.9	31.3	21.9	21.6	**21.0**

Bold values are the best feature selection result.

Table 7
Clustering results (ACC) of SSC, LRR and RSR.

Data	SSC [31]	LRR [30]	RSR
USPS	47.5	48.6	65.2
ISOLET	19.2	19.2	54.7
AR10P	30.8	21.5	38.6
CLL-SUB-111	48.7	38.7	54.3
TOX-171	50.3	48.5	49.3
Prostate-GE	35.3	35.3	61.4

Table 8
Clustering results (NMI) of SSC, LRR and RSR.

Data	SSC [31]	LRR [30]	RSR
USPS	51.4	54.9	60.0
ISOLET	62.6	56.1	69.1
AR10P	36.8	19	40.9
CLL-SUB-111	35.4	26	23.4
TOX-171	29.3	25.2	26.3
Prostate-GE	27.8	16.3	5.5

Table 9
Clustering results (NMI) of SSC, LRR and RSR.

Data	Cluster accuracy		NMI		Classification accuracy		Redundancy	
	RSR	RSR(R)	RSR	RSR(R)	RSR	RSR(R)	RSR	RSR(R)
USPS	65.2	65.4	60.0	60.2	92.8	91.9	16.5	16.6
ISOLET	54.7	51.8	69.1	66.1	77.6	74.5	13.3	14.0
AR10P	38.6	36.4	40.9	38.5	89.1	87.7	18.9	19.1
CLL-SUB-111	54.3	54.5	23.4	23.5	66.3	65.2	35.3	35.3
TOX-171	49.3	49.3	26.3	26.4	64.5	62.7	22.5	25.6
Prostate-GE	61.4	60.2	5.5	8.7	79.5	78.1	19.7	19.7

(R) is quite little in terms of four different evaluation indexes. Compared with the raw feature, the reconstructed feature removes the error item $\boldsymbol{E}$. The result shows that the error matrix $\boldsymbol{E}$ has little impact on the clustering and classification performance. The key point is that from self-representation matrix $\boldsymbol{W}$, the most representative features are found. Hence, we can simply use the original selected features.

7. Conclusions

Motivated by the fact that a feature can be well represented as the linear combination of its relevant features in a redundant feature vector, in this paper we proposed a novel regularized self-representation (RSR) model for unsupervised feature selection by representing the data matrix over itself. The $L_{2,1}$-norm is used to measure the representation residual and to regularize the representation coefficients. As a result, the most representative features which can be used to reconstruct other features are selected. Our extensive experiments on synthetic and real datasets clearly demonstrated that RSR can effectively identify the most representative features. It can reduce much the feature redundancy while leading to high clustering and classification accuracies.

Conflict of interest

None declared.

Acknowledgments

This work was partly supported by HK PolyU ICRG under grant no. G-YK79, National Program on Key Basic Research Project (973 Program) under Grant 2013CB329304, National Natural Science Foundation of China under Grant 61271093.

References

[1] J.G. Dy, C.E. Brodley, A. Kak, L.S. Broderick, A.M. Aisen, Unsupervised feature selection applied to content-based retrieval of lung images, IEEE Trans. Pattern Anal. Mach. Intell. 25 (2003) 373–378.

[2] J. Tang, H. Liu, Unsupervised feature selection for linked social media data, in: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 904–912.

[3] D. Cai, C. Zhang, X. He, Unsupervised feature selection for multi-cluster data, in: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 333–342.

[4] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, J. Mach. Learn. Res. 3 (2003) 1157–1182.

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

[6] Z. Zhao, L. Wang, H. Liu, Efficient spectral feature selection with minimum redundancy, in: Proceedings of the 24th AAAI Conference on Artificial Intelligence, pp. 673–678.

[7] X. He, D. Cai, P. Niyogi, Laplacian score for feature selection, in: Advances in Neural Information Processing Systems, pp. 507–514.

[8] F. Nie, H. Huang, X. Cai, C. H. Ding, Efficient and robust feature selection via joint $l_{2,1}$-norms minimization, in: Advances in Neural Information Processing Systems, pp. 1813–1821.

[9] Y. Yang, H. T. Shen, Z. Ma, Z. Huang, X. Zhou, $l_{2,1}$-Norm regularized discriminative feature selection for unsupervised learning, in: Proceedings of the 22nd International Joint Conference on Artificial Intelligence, pp. 1589–1594.

[10] Y. Hong, S. Kwong, Y. Chang, Q. Ren, Unsupervised feature selection using clustering ensembles and population based incremental learning algorithm, Pattern Recognit. 41 (2008) 2742–2756.

[11] M. Breaban, H. Luchian, A unifying criterion for unsupervised clustering and feature selection, Pattern Recognit. 44 (2011) 854–865.

[12] D. Zhang, S. Chen, Z.-H. Zhou, Constraint score: a new filter method for feature selection with pairwise constraints, Pattern Recognit. 41 (2008) 1440–1451.

[13] H. Liu, J. Ye, On similarity preserving feature selection, IEEE Trans. Knowl. Data Eng. 25 (2013) 619–632.

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

[15] F. Nie, S. Xiang, Y. Jia, C. Zhang, S. Yan, Trace ratio criterion for feature selection, in: Proceedings of the 23rd National Conference on Artificial intelligence, pp. 671–676.

[16] Z. Zhao, H. Liu, Spectral feature selection for supervised and unsupervised learning, in: Proceedings of the 24th International Conference on Machine Learning, pp. 1151–1157.

[17] Z. Li, Y. Yang, J. Liu, X. Zhou, H. Lu, Unsupervised feature selection using nonnegative spectral analysis, in: Proceedings of the 26th AAAI Conference on Artificial Intelligence, pp. 1026–1032.

[18] J. Zhu, S. Rosset, T. Hastie, R. Tibshirani, 1-Norm support vector machines, in: Advances in Neural Information Processing Systems, vol. 16, pp. 49–56.

[19] A.Y. Ng, Feature selection, l_1 vs. l_2 regularization, and rotational invariance, in: Proceedings of the 21st International Conference on Machine Learning, pp. 78–85.

[20] A. Evgeniou, M. Pontil, Multi-task feature learning, in: Advances in Neural Information Processing Systems, vol. 19, pp. 41–48.

[21] X.-T. Yuan, S. Yan, Visual classification with multi-task joint sparse representation, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2010, pp. 3493–3500.

[22] C. Hou, F. Nie, D. Yi, Y. Wu, Feature selection via joint embedding learning and sparse regression, in: Proceedings of the 22nd International Joint Conference on Artificial Intelligence, pp. 1324–1329.

[23] R. He, T. Tan, L. Wang, W.-S. Zheng, $l_{2,1}$ Regularized correntropy for robust feature selection, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Orlando, Florida, United States, pp. 2504–2511.

[24] B.B. Mandelbrot, How long is the coast of britain, Science 156 (1967) 636–638.

[25] J.Y. Campbell, The Econometrics of Financial Markets, Princeton University Press, Princeton, NJ, United States, 1997.

[26] A. Buades, B. Coll, J.-M. Morel, A non-local algorithm for image denoising, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, 2005, pp. 60–65.

[27] K. Lange, D.R. Hunter, I. Yang, Optimization transfer using surrogate objective functions, J. Comput. Graph. Stat. 9 (2000) 1–20.

[28] J. Mairal, M. Elad, G. Sapiro, Sparse representation for color image restoration, IEEE Trans. Image Process. 17 (2008) 53–69.

[29] J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry, Y. Ma, Robust face recognition via sparse representation, IEEE Trans. Pattern Anal. Mach. Intell. 31 (2009) 210–227.

[30] G. Liu, Z. Lin, Y. Yu, Robust subspace segmentation by low-rank representation, in: Proceedings of the 27th International Conference on Machine Learning, pp. 663–670.

[31] E. Elhamifar, R. Vidal, Sparse subspace clustering, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2009, pp. 2790–2797.

Pengfei Zhu received his B.S. and M.S. from Harbin Institute of Technology, Harbin, China in 2009 and 2011, respectively. He is now a Ph.D. candidate with the Hong Kong Polytechnic University. His research interests are focused on machine learning and computer vision.

Wangmeng Zuo (M'09) received the Ph.D. degree in computer application technology from the Harbin Institute of Technology, Harbin, China, in 2007. From July 2004 to December 2004, from November 2005 to August 2006, and from July 2007 to February 2008, he was a Research Assistant in the Department of Computing, Hong Kong Polytechnic University. From August 2009 to February 2010, he was a Visiting Professor in Microsoft Research Asia. He is currently an Associate Professor in the School of Computer Science and Technology, Harbin Institute of Technology. His current research interests include sparse representation, biometrics, pattern recognition, and computer vision. He is an Associate Editor of the IET Biometrics.

Lei Zhang (M'04) received the B.S. degree, in 1995, from Shenyang Institute of Aeronautical Engineering, Shenyang, PR China, the M.S. and Ph.D. degrees in Automatic Control Theory and Engineering from Northwestern Polytechnical University, Xi'an, PR China, respectively, in 1998 and 2001. From 2001 to 2002, he was a Research Associate in the Department of Computing, The Hong Kong Polytechnic University. From January 2003 to January 2006 he worked as a Postdoctoral Fellow in the Department of Electrical and Computer Engineering, McMaster University, Canada. In 2006, he joined the Department of Computing, The Hong Kong Polytechnic University, as an Assistant Professor.

Since September 2010, he has been an Associate Professor in the same department. His research interests include Image and Video Processing, Biometrics, Computer Vision, Pattern Recognition, Multisensor Data Fusion and Optimal Estimation Theory. He is an Associate Editor of IEEE Transactions on CSVT and Image and Vision Computing Journal. He was awarded the Faculty Merit Award in Research and Scholarly Activities, in 2010 and 2012, and the Best Paper Award of SPIE VCIP2010. More information can be found in his homepage http://www4.comp.polyu.edu.hk/~cslzhang/.

Qinghua Hu (M'11) received B.S., M.S. and Ph.D. degrees from Harbin Institute of Technology, Harbin, China, in 1999, 2002 and 2008, respectively. He was an Associate Professor with Harbin Institute of Technology from 2008 to 2011. Now he is a Full Professor with School of Computer Science and Technology, Tianjin University. His research interests are focused on intelligent modeling, data mining, knowledge discovery for classification and regression. He is a PC Co-Chair of RSCTC 2010 and serves as a Referee for a great number of journals and conferences. He has published more than 70 journals and conference papers in the areas of pattern recognition and fault diagnosis.

Simon C.K. Shiu (M'90) received the M.Sc. degree in computing science from the University of Newcastle Upon Tyne, Newcastle Upon Tyne, UK, the M.Sc. degree in business systems analysis and design from City University London, London, UK, and the Ph.D. degree in computing from the Hong Kong Polytechnic University, Hong Kong, in 1985, 1986, and 1997, respectively. He is currently an Assistant Professor with the Department of Computing, the Hong Kong Polytechnic University. From 1985 to 1990, he was a System Analyst and the Project Manager with several business organizations at Hong Kong. His current research interests include case-based reasoning, machine learning, and soft computing. He co-authored the book Foundations of Soft Case-Based Reasoning (Hoboken, NJ, USA: Wiley, 2004). He was a Guest Co-Editor of a special issue on soft case-based reasoning of the Applied Intelligence. He is a member of the British Computer Society.