SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

September 2021, Vol. 64 192108:1–192108:12
https://doi.org/10.1007/s11432-020-3063-0

Semi-supervised local feature selection for data classification

Zechao LI & Jinhui TANG*

School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094*, China*

Received 29 February 2020/Revised 29 April 2020/Accepted 16 July 2020/Published online 23 August 2021

Abstract Conventional feature selection methods select the same feature subset for all classes, which means that the selected features might work better for some classes than the others. Towards this end, this paper proposes a new semi-supervised local feature selection method (S2LFS) allowing to select different feature subsets for different classes. According to this method, class-specific feature subsets are selected by learning the importance of features considering each class separately. In particular, the class labels of all available data are jointly learned under a consistent constraint over the labeled data, which enables the proposed method to select the most discriminative features. Experiments on six data sets demonstrate the effectiveness of the proposed method compared to some popular feature selection methods.

Keywords local feature selection, label-specific feature, semi-supervised learning, data classification, discriminative feature

Citation Li Z C, Tang J H. Semi-supervised local feature selection for data classification. Sci China Inf Sci, 2021, 64(9): 192108, https://doi.org/10.1007/s11432-020-3063-0

1 Introduction

Feature selection is an important problem for many tasks, including machine learning and pattern recognition, which is considered with the aim to improve the performance and reduce the computational cost [1, 2]. Feature selection methods are often applied when analyzing high-dimensional data such as images and gene expression data [3–5]. Feature selection methods aim at finding features useful for the task at hand and removing redundant or noisy features. Carefully selecting a subset of feature can speed up the learning process and provide insights into the nature of data [3, 6, 7].

Conventional feature selection methods select a single feature subset for all the classes. Unsupervised feature selection methods are designed to find the optimal subset by exploring the data structure. The feature similarity was explored for unsupervised feature selection in [8], while the Laplacian score was learned to select features in [9]. Supervised feature selection methods explore the discriminative information encoded in class labels [10, 11]. In [12], features were selected one by one based on spectral graph theory. Features were selected in a supervised manner based on linear models with sparsity regularization in [6, 13]. Semi-supervised feature selection methods have been proposed to leverage unlabeled and labeled data [14, 15]. Features were selected by maximizing the classification margin and exploring the manifold regularization over both labeled and unlabeled data in [14]. Constrained Laplacian score was proposed for semi-supervised feature selection in [15]. To alleviate the need for labeled data that are typically expensive to obtain, this study focuses on semi-supervised feature selection to jointly consider labeled and unlabeled data.

Existing methods learn from data using the same feature subset for all the samples from different classes. The underlying assumption is that the feature space of all classes can be optimally characterized by a single feature subset [16, 17]. However, in fact, such an assumption may not be useful in some cases, considering that people recognize samples from different classes using different features. Figure 1 illustrates one such example, where the shape feature is useful for distinguishing a banana from an orange,

* Corresponding author (email: jinhuitang@njust.edu.cn)

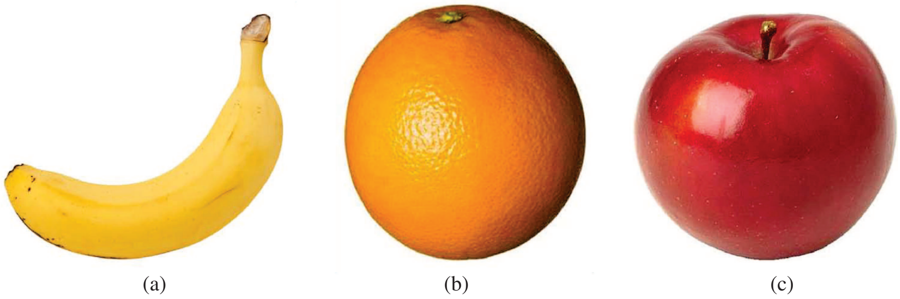

Figure 1 (Color online) Illustration of recognizing samples from different classes using different features. It can be observed that images from different classes can be well recognized using different visual features such as shape and color.

while the color feature is useful to distinguish an orange from an apple. That is, it may be not optimal to describe data of different classes using a common feature subset.

Towards this end, this work proposes a new semi-supervised local feature selection method (S2LFS) for data classification by jointly exploring the discriminative information and class-specific feature selection. According to the proposed method, an indicator vector learned for each class represents whether a specific feature is chosen or not for the class. To leverage the discriminative information for classification, a predicted class matrix is learned from all available data using discrete and nonnegative spectral analysis with consistent constraints over the labeled data. As a result, the optimal feature subset minimizing the predicted error with its corresponding class is selected. The proposed method is formulated as an optimization problem and an algorithm to solve is presented. The experimental results over several real data sets demonstrated the effectiveness of the proposed method for feature selection compared to several state-of-the-art methods.

The main contributions of this study are summarized as follows.

- A novel semi-supervised local feature selection method is proposed to select different feature subsets for different classes.
- The discriminative information and class-specific feature selection are jointly explored to guarantee the effectiveness of the selected feature subsets.
- A predicted class matrix is learned using all available data by leveraging a discrete and nonnegative spectral analysis model.

The rest of this paper is organized as follows. Related methods are briefly discussed in Section 2. The proposed S2LFS method is detailed in Section 3. Section 4 presents the experimental results and their analysis, including sensitiveness and convergence analysis. Section 5 concludes the paper and outline future research directions.

2 Related work

Many feature selection methods have been proposed for finding the desired feature subset by exploring different criteria. The widely studied methods are the global feature selection ones, which select a common subset for all classes [18–20]. These methods can be grouped into three categories: unsupervised, supervised and semi-supervised methods. Unsupervised feature selection methods explore the data structure to select the desired feature subset [5, 8, 9, 21–24]. For example, feature similarity was explored in [8], while He et al. [9] proposed to learn the Laplacian score of data. The multi-cluster feature selection (MCFS) method [21] was proposed for unsupervised feature selection based on a two-step spectral regression approach. Spectral analysis was employed to learn cluster indicators for unsupervised feature selection [22–25]. The redundancy for feature selection was explored in [1, 5]. To improve the discriminative ability of selected feature subsets, supervised feature selection methods were proposed to explore the available label information [10–12]. Linear prediction models with sparsity regularization were proposed to select the discriminative feature subsets [6, 13]. The class correlation information was preserved for supervised feature selection in [26]. In [27], supervised feature selection was performed using self-weighted orthogonal linear discriminant analysis. The cost of obtaining labeled data for supervised learning is high [28, 29]. To address this problem, semi-supervised feature selection methods were proposed by jointly exploring unlabeled and labeled data [14, 15, 30–33]. In [31], features were ranked based on an extended least square regression model. This study focuses on semi-supervised feature selection considering both the labeled and unlabeled data. Several survey papers investigated global feature selection methods, e.g., further details on these methods can be found in [34].

Methods such as co-clustering [35] and subspace clustering [36–38] were proposed to explore local information for feature selection. In co-clustering, features and samples are both clustered using a feature matrix to find co-clusters. Subspace clustering finds clusters within different subspaces assuming that valid clusters are defined by only a subset of dimensions. In contrast, local feature selection aims to find different feature subsets for different classes. In [17], localized feature selection was proposed to select a feature subset for each region of the sample space. In [39], a unified probabilistic method was proposed to perform global and local feature selection for clustering. An embedded method was developed in [40] to locally weight variables for global feature selection. These methods mainly explore the local information to select feature subsets.

In contrast to previous methods, the method proposed in this paper aims to select label-specific feature subsets, generating different feature subsets for different classes. The discriminative ability of the selected features is also guaranteed to well predict the labels of data.

3 Semi-supervised local feature selection

This section presents the proposed S2LFS method for finding class-specific discriminative feature subsets.

3.1 Preliminary

Throughout this paper, bold uppercase and lowercase characters are used to denote matrices and vectors, respectively. Scalars are denoted using lowercase italic characters. Given a matrix $\boldsymbol{A}$, its i-th column vector and j-th row vector are expressed as $\boldsymbol{a}_i$ and $\boldsymbol{a}^j$, respectively. A_{ij} is the (i,j)-th element of $\boldsymbol{A}$. $\mathrm{Tr}[\boldsymbol{A}]$ denotes the trace of the square $\boldsymbol{A}$, while $\boldsymbol{A}^{\mathrm{T}}$ denotes the transposed matrix of $\boldsymbol{A}$. The Frobenius norm of $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ can be obtained as $\|\boldsymbol{A}\|_F^2 = \sum_{i=1}^{m}\sum_{j=1}^{n} A_{ij}^2 = \mathrm{Tr}[\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}]$. $\mathbf{1}_m = [1,\ldots,1]^{\mathrm{T}} \in \mathbb{R}^{m\times 1}$ and $\mathbf{I}_c \in \mathbb{R}^{c\times c}$ is an identity matrix.

Semi-supervised learning considers a data set $\boldsymbol{X} = [\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n] \in \mathbb{R}^{d\times n}$ with c classes. The data set has two parts: l labeled samples $\boldsymbol{X}_L = [\boldsymbol{x}_1,\ldots,\boldsymbol{x}_l]$ with their corresponding labels $\boldsymbol{Y}_L = [\boldsymbol{Y}_1,\ldots,\boldsymbol{Y}_l]^{\mathrm{T}} \in \mathbb{R}^{l\times c}$, and $u = n - l$ unlabeled samples $\boldsymbol{X}_U = [\boldsymbol{x}_{l+1},\ldots,\boldsymbol{x}_n]$ with their unknown labels $\boldsymbol{Y}_U = [\boldsymbol{Y}_{l+1},\ldots,\boldsymbol{Y}_n]^{\mathrm{T}} \in \mathbb{R}^{u\times c}$. For the labeled data, $Y_{ij} = 1$ if $\boldsymbol{x}_i$ is labeled with the j-th class and 0 otherwise. Let $\boldsymbol{G} \in \mathbb{R}^{n\times c}$ denote the learned label matrix. Semi-supervised feature selection can be formulated as follows:

$$\min_{\boldsymbol{W}} \inf_{f\in\mathbb{F}} \mathrm{E}_{\boldsymbol{X},\boldsymbol{Y}} L(\boldsymbol{Y}, f(\boldsymbol{X},\boldsymbol{W})), \tag{1}$$

where $\boldsymbol{Y} = [\boldsymbol{Y}_L;\boldsymbol{Y}_U]$, $\boldsymbol{W}$ denotes the feature selection matrix, $\mathbb{F}$ denotes the class of the predicted function, L denotes the loss function, and E denotes the expectation.

3.2 Problem formulation

The proposed S2LFS method aims to find a feature subset containing the most discriminative features for each class separately, by simultaneously exploring labeled and unlabeled data. The discriminative ability of the selected subsets is leveraged jointly.

A discriminative analysis model is leveraged to learn the label information of unlabeled data. Inspired from semi-supervised learning, the discriminative information is explored using a label propagation model, which enables to guide the local feature selection procedure. A label matrix $\boldsymbol{G}$ is learned for unlabeled data under consistent constraints on labeled data. That is, the learned label matrix for labeled data is constrained to be consistent with their groundtruth label matrix. Moreover, local information is considered to preserve the structure information of the data. The manifold structure smoothness constraints are introduced to explore local information for simplicity [41]. Therefore, the optimization problem of label propagation can be formulated as the following problem:

$$\min_{\boldsymbol{G}} \mathrm{Tr}[\boldsymbol{G}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{G}] \qquad \text{s.t.} \ \ \boldsymbol{G}_L = \boldsymbol{Y}_L, \tag{2}$$

where $\boldsymbol{L} = \boldsymbol{D} - (\boldsymbol{S}+\boldsymbol{S}^{\mathrm{T}})/2$ is the Laplacian matrix, $\boldsymbol{S}$ is the affinity matrix of data, and $\boldsymbol{D}$ is a diagonal matrix with $D_{ii} = \sum_{j=1}^{n}(S_{ij}+S_{ji})/2$. Each element G_{ij} of $\boldsymbol{G}$ denotes the binary relationship between the i-th sample and the j-th class. According to the definition of $\boldsymbol{G}$, $G_{ij}=1$ or $G_{ij}=0$, and $\boldsymbol{G}\mathbf{1}_c=\mathbf{1}_n$. Thus, the above optimization problem becomes

$$\min_{\boldsymbol{G}} \operatorname{Tr}[\boldsymbol{G}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{G}]+\operatorname{Tr}[(\boldsymbol{G}-\boldsymbol{Y})^{\mathrm{T}}\boldsymbol{U}(\boldsymbol{G}-\boldsymbol{Y})] \qquad \text{s.t. } \boldsymbol{G}\in\{0,1\}^{n\times c},\ \boldsymbol{G}\mathbf{1}_c=\mathbf{1}_n, \tag{3}$$

where $\boldsymbol{U}$ is a diagonal matrix representing consistent constraints imposed on the labeled data, with U_{ii} equaling to a very large number (e.g., 10^{10}) if $i=1,\ldots,l$ and 0 if $i=l+1,\ldots,n$.

With the learned label matrix $\boldsymbol{G}$, the feature subset that has the most discriminative ability is selected for each class. For this purpose, a regularized linear regression model is introduced with an indicator vector for each class. The induced prediction model is formulated as follows:

$$\min_{\boldsymbol{W},\boldsymbol{Z}} \sum_{k=1}^{c} L(\boldsymbol{g}_k, f(\boldsymbol{X},\boldsymbol{z}_k,\boldsymbol{w}_k))+\lambda\|\boldsymbol{W}\|_F^2 \qquad \text{s.t. } \boldsymbol{z}_k\in\{0,1\}^d,\ \boldsymbol{z}_k^{\mathrm{T}}\mathbf{1}_d=m, \tag{4}$$

where $\boldsymbol{z}_k$ is an indicator vector representing whether a feature is chosen or not for the k-th class, $\boldsymbol{w}_k$ is the prediction function for the k-th class based on the selected features, m denotes the number of the desired features, and λ is the regularization parameter.

To jointly explore discriminative information and local feature selection, the discriminative and regularized linear regression models are considered together to formulate the following problem:

$$\begin{aligned}&\min_{\boldsymbol{G},\boldsymbol{W},\boldsymbol{Z}} \sum_{k=1}^{c} L(\boldsymbol{g}_k, f(\boldsymbol{X},\boldsymbol{z}_k,\boldsymbol{w}_k))+\lambda\|\boldsymbol{W}\|_F^2+\beta(\operatorname{Tr}[\boldsymbol{G}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{G}]+\operatorname{Tr}[(\boldsymbol{G}-\boldsymbol{Y})^{\mathrm{T}}\boldsymbol{U}(\boldsymbol{G}-\boldsymbol{Y})])\\ &\text{s.t. } \boldsymbol{G}\in\{0,1\}^{n\times c},\ \boldsymbol{G}\mathbf{1}_c=\mathbf{1}_n,\ \boldsymbol{z}_k\in\{0,1\}^d,\ \boldsymbol{z}_k^{\mathrm{T}}\mathbf{1}_d=m,\end{aligned} \tag{5}$$

where β is a trade-off parameter that balances feature selection and discriminative analysis. Owing to the constraints imposed on $\boldsymbol{G}$ and $\boldsymbol{Z}$, the above problem is NP-hard. The constraints imposed on $\boldsymbol{G}$ ensure that only one element in each row of $\boldsymbol{G}$ is equal to 1 and all the others are 0. Similar to [5], the nonnegative and orthogonal constraints imposed on $\boldsymbol{G}$ have the same properties. In fact, features are selected according to their importance. Thus, the integer constraints imposed on $\boldsymbol{Z}$ are also relaxed to real nonnegative constraints. To select several features without introducing an additional hyper-parameter, a simplex constraint $\boldsymbol{z}_k^{\mathrm{T}}\mathbf{1}_d=1$ is introduced. Consequently, the above problem can be re-formulated as

$$\begin{aligned}&\min_{\boldsymbol{G},\boldsymbol{W},\boldsymbol{Z}} \sum_{k=1}^{c} L(\boldsymbol{g}_k, f(\boldsymbol{X},\boldsymbol{z}_k,\boldsymbol{w}_k))+\lambda\|\boldsymbol{W}\|_F^2+\beta(\operatorname{Tr}[\boldsymbol{G}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{G}]+\operatorname{Tr}[(\boldsymbol{G}-\boldsymbol{Y})^{\mathrm{T}}\boldsymbol{U}(\boldsymbol{G}-\boldsymbol{Y})])\\ &\text{s.t. } \boldsymbol{G}\geqslant 0,\ \boldsymbol{G}^{\mathrm{T}}\boldsymbol{G}=\mathbf{I}_c,\ \boldsymbol{z}_k\geqslant 0,\ \boldsymbol{z}_k^{\mathrm{T}}\mathbf{1}_d=1.\end{aligned} \tag{6}$$

To optimize problem (6), the loss function L and function class $\mathbb{F}$ should be defined. For simplicity, the squared error $L(x,y)=(x-y)^2$ and linear functions are used.

$$\begin{aligned}&\min_{\boldsymbol{G},\boldsymbol{W},\boldsymbol{Z}} \sum_{k=1}^{c} \|\boldsymbol{g}_k-\boldsymbol{X}^{\mathrm{T}}\operatorname{diag}(\sqrt{\boldsymbol{z}_k})\boldsymbol{w}_k\|^2+\lambda\|\boldsymbol{W}\|_F^2+\beta(\operatorname{Tr}[\boldsymbol{G}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{G}]+\operatorname{Tr}[(\boldsymbol{G}-\boldsymbol{Y})^{\mathrm{T}}\boldsymbol{U}(\boldsymbol{G}-\boldsymbol{Y})])\\ &\text{s.t. } \boldsymbol{G}\geqslant 0,\ \boldsymbol{G}^{\mathrm{T}}\boldsymbol{G}=\mathbf{I}_c,\ \boldsymbol{z}_k\geqslant 0,\ \boldsymbol{z}_k^{\mathrm{T}}\mathbf{1}_d=1.\end{aligned} \tag{7}$$

To simplify the model, $\operatorname{diag}(\sqrt{\boldsymbol{z}_k})\boldsymbol{w}_k$ is substituted with $\boldsymbol{w}_k$.

$$\begin{aligned}&\min_{\boldsymbol{G},\boldsymbol{W},\boldsymbol{Z}} \sum_{k=1}^{c} \|\boldsymbol{g}_k-\boldsymbol{X}^{\mathrm{T}}\boldsymbol{w}_k\|^2+\lambda\sum_{k=1}^{c}\boldsymbol{w}_k^{\mathrm{T}}\operatorname{diag}(\boldsymbol{z}_k^{-1})\boldsymbol{w}_k+\beta(\operatorname{Tr}[\boldsymbol{G}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{G}]+\operatorname{Tr}[(\boldsymbol{G}-\boldsymbol{Y})^{\mathrm{T}}\boldsymbol{U}(\boldsymbol{G}-\boldsymbol{Y})])\\ &\text{s.t. } \boldsymbol{G}\geqslant 0,\ \boldsymbol{G}^{\mathrm{T}}\boldsymbol{G}=\mathbf{I}_c,\ \boldsymbol{z}_k\geqslant 0,\ \boldsymbol{z}_k^{\mathrm{T}}\mathbf{1}_d=1.\end{aligned} \tag{8}$$

The value of each Z_{jk} represents the importance of the j-th feature for the k-th class. If the j-th feature is irrelevant to the k-th class, the learned value of Z_{jk} is small, which leads to a large penalty being applied to W_{jk}.

3.3 Optimization

The proposed problem is solved using an iterative algorithm. Before optimization, the closed form solution of $\boldsymbol{W}$ can be easily obtained by setting the derivative of the above objective function with respect to $\boldsymbol{W}$ to 0.

$$\boldsymbol{w}_k = (\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}} + \lambda \mathrm{diag}(\boldsymbol{z}_k^{-1}))^{-1}\boldsymbol{X}\boldsymbol{g}_k. \tag{9}$$

By substituting $\boldsymbol{w}_k$ by the above equation, the problem (8) can be rewritten as

$$\begin{aligned} &\min_{\boldsymbol{G},\boldsymbol{Z}} -\sum_{k=1}^{c} \boldsymbol{g}_k^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}} + \lambda \mathrm{diag}(\boldsymbol{z}_k^{-1}))^{-1}\boldsymbol{X}\boldsymbol{g}_k + \beta(\mathrm{Tr}[\boldsymbol{G}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{G}] + \mathrm{Tr}[(\boldsymbol{G}-\boldsymbol{Y})^{\mathrm{T}}\boldsymbol{U}(\boldsymbol{G}-\boldsymbol{Y})]) \\ &\quad \text{s.t.}\ \ \boldsymbol{G} \geqslant 0,\ \boldsymbol{G}^{\mathrm{T}}\boldsymbol{G} = \mathbf{I}_c,\ \boldsymbol{z}_k \geqslant 0,\ \boldsymbol{z}_k^{\mathrm{T}}\mathbf{1}_d = 1. \end{aligned} \tag{10}$$

For a given $\boldsymbol{Z}$, $\boldsymbol{G}$ is optimized by using the Lagrange multiplier method and the Karush-Kuhn-Tucker (KKT) condition.

$$\begin{aligned} &\min_{\boldsymbol{G},\boldsymbol{Z}} -\sum_{k=1}^{c} \boldsymbol{g}_k^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}} + \lambda \mathrm{diag}(\boldsymbol{z}_k^{-1}))^{-1}\boldsymbol{X}\boldsymbol{g}_k + \beta(\mathrm{Tr}[\boldsymbol{G}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{G}] + \mathrm{Tr}[(\boldsymbol{G}-\boldsymbol{Y})^{\mathrm{T}}\boldsymbol{U}(\boldsymbol{G}-\boldsymbol{Y})]) \\ &\quad +\frac{\gamma}{2}\|\boldsymbol{G}^{\mathrm{T}}\boldsymbol{G} - \mathbf{I}_c\|_F^2 + \mathrm{Tr}[\Phi^{\mathrm{T}}\boldsymbol{G}^{\mathrm{T}}], \end{aligned} \tag{11}$$

where γ is a very large positive parameter to ensure that the orthogonality is satisfied and ϕ_{ij} is a Lagrange multiplier for $G_{ij} > 0$. The updated rules are obtained as follows:

$$\boldsymbol{g}_k = \boldsymbol{g}_k \otimes [(\boldsymbol{T} + \beta\boldsymbol{U} + \gamma\mathbf{I}_n)\boldsymbol{g}_k] \oslash [(\beta(\boldsymbol{L}+\boldsymbol{U}) + \gamma\boldsymbol{G}\boldsymbol{G}^{\mathrm{T}})\boldsymbol{g}_k], \tag{12}$$

where $\boldsymbol{T} = \boldsymbol{X}^{\mathrm{T}}(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}} + \lambda\mathrm{diag}(\boldsymbol{z}_k^{-1}))^{-1}\boldsymbol{X}$. $\otimes$ and $\oslash$ denote the element-wise multiplication and division operators, respectively.

With the learned $\boldsymbol{G}$, the optimal value of $\boldsymbol{W}$ can be obtained using (9). Then, the solution of $\boldsymbol{Z}$ can be reduced to the following sub-problem:

$$\min_{\boldsymbol{z}_k} \boldsymbol{w}_k^{\mathrm{T}}\mathrm{diag}(\boldsymbol{z}_k^{-1})\boldsymbol{w}_k \quad \boldsymbol{z}_k \geqslant 0,\ \boldsymbol{z}_k^{\mathrm{T}}\mathbf{1}_d = 1. \tag{13}$$

The solution of $\boldsymbol{z}_k$ can be obtained using the Lagrange multiplier and KKT conditions as $Z_{jk} = |W_{jk}|/\sum_{h=1}^{d}|W_{hk}|$.

The most important feature subset for each class can be obtained by finding solutions of $\boldsymbol{G}$ and $\boldsymbol{z}_k$ for each class. The class labels of unlabeled data can be obtained using the learned $\boldsymbol{G}_U$. For any new data sample $\boldsymbol{x}$, the found features are first chosen by using the learned $\boldsymbol{z}_k$ for each class, and then, the probability score is predicted using $\boldsymbol{w}_k$. The class label is assigned to the one with the highest predicted probability.

As the aforementioned analysis, it may be impossible to directly solve all the desired variables. Therefore, an iterative algorithm summarized in Algorithm 1 is used to update them. The convergence criterion employed in this study is that $|\mathcal{L}_{t-1} - \mathcal{L}_t|/\mathcal{L}_{t-1} < 10^{-6}$, where $\mathcal{L}_t$ is the value of the objective function in the t-th iteration. The objective function monotonically decreases in each updating step, and the formulated objective function has the lower bound. This proposed optimization algorithm converges to a local minimum. The convergence rate is evaluated in Section 4.

Algorithm 1 The proposed S2LFS algorithm

```
Input:
    Data feature matrix $\boldsymbol{X}$;
    Labeled matrix $\boldsymbol{Y}_L$;
    Parameters $\lambda$ and $\beta$;
Output:
    Feature selection matrix $\boldsymbol{W}$.
Compute the image similarity $\boldsymbol{S}$ and Laplacian matrix $\boldsymbol{L}$;
Initialize $\boldsymbol{U}$ and $\boldsymbol{Z}$;
repeat
    Update $\boldsymbol{W}$ according to (9);
    Update $\boldsymbol{G}$ according to (12);
    Update $\boldsymbol{Z}$ by optimizing problem (13);
until Convergence criterion satisfied.
```

Table 1 Dataset description

Dataset	# of sample (n)	# of feature (d)	# of class (c)
USPS	9298	256	10
COIL20	1440	1024	20
ORL	400	1024	10
Binary Alphabet	1404	320	36
Pointing4	2790	490	15
YaleB	2414	1024	38

3.4 Computational complexity analysis

This subsection will discuss the computational cost of the proposed S2LFS method. The complexity of the employed algorithm is described using the big O notation.

It can be seen from Algorithm 1 that the affinity graph is constructed first using the Euclidean distance with cost of $O(dn^2)$, where n is the number of data points and d is the dimensionality of features. The desired variables are updated iteratively. It takes $O(d^3 + d^2n)$ to obtain $(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}} + \lambda\mathrm{diag}(\boldsymbol{z}_k^{-1}))^{-1}\boldsymbol{X}$. The computational cost to update $\boldsymbol{W}$ is $O(dnc^2)$, where c is the number of classes. It takes $O(cn^2)$ to calculate $\boldsymbol{G}\boldsymbol{G}^{\mathrm{T}}$ and $O(cdn^2)$ to update $\boldsymbol{G}$. The cost of updating $\boldsymbol{Z}$ is $O(cd)$. Therefore, the total cost of the proposed method is $O(dn^2 + T(cd^3 + cd^2n + c^2dn + cn^2 + cdn^2 + cd))$, where T is the number of iterations. Since $n \gg c$, the total cost is $O(T(cd^3 + cd^2n + cn^2))$.

It is worth noting that the proposed method is scalable to larger data sets, although the computational cost of the proposed method would also become larger in this case. Once the desired variables are obtained, the learned $\boldsymbol{Z}$ can be used to select features for all data, and then the label information can be predicted based on the selected features. If the size of the used data set is too large, a subset can be randomly sampled to learn the feature selection model. The learned feature selection model is used to choose a feature subset for all data, including the sampled subset and the remaining data. That is, the proposed method is scalable to larger data sets.

4 Experiments

This section will investigate whether the proposed method can select reasonable local features for each class. The experiments are conducted on six real-world data sets. In the experiments, the data are normalized with zero mean and unit variance.

4.1 Data set description

Experiments are conducted on six publicly available data sets including USPS [42], COIL20 [43], ORL[1], Binary Alphabet[2], Pointing4 [44] and YaleB [45]. The details of these data sets are listed in Table 1 and briefly described as follows.

- **USPS [42].** This data set contains 9298 images of 10 handwritten digits. The size of each image is 16 × 16. Each image is described by a 256 dimensional feature vector.
- **COIL20 [43].** This data set contains 1440 grayscale images of 20 object classes. Each class has 72 images. The size of each image is 32 × 32. And each image is described by a 1024 dimensional feature vector.
- **ORL.** This data set is composed of face images from 40 classes. There are 400 samples and each sample is represented by a 1024 dimensional feature vector.
- **Binary Alphabet.** This data set contains 1404 samples from 36 classes. Each sample is represented by a 320 dimensional feature vector.
- **Pointing4 [44].** This data set has 2790 samples from 15 classes. Each sample is represented by a 490 dimensional feature vector.
- **YaleB [45].** This data set has 2414 images from 38 classes. The size of each image is 32 × 32. A vector with 1024 dimension is introduced to describe each data sample.

[1] The ORL database of faces. http://www.face-rec.org/databases/.

[2] Handwritten digits. https://cs.nyu.edu/ roweis/data.html.

Table 2 Classification accuracy (CA%±std) of different feature selection methods over the USPS and COIL20 data sets with selected 80 and 100 features, respectively[a)]

Data set		USPS			COIL20		
		$s = 10$	$s = 20$	$s = 50$	$s = 10$	$s = 20$	$s = 50$
SAFS	Semi	82.6 ± 1.6	84.7 ± 1.3	85.7 ± 0.7	60.2 ± 2.3	66.9 ± 1.4	74.6 ± 1.2
CLS	Semi	81.5 ± 0.9	84.3 ± 1.2	85.9 ± 1.0	62.0 ± 3.6	70.4 ± 2.0	75.7 ± 1.2
RSSL	Semi	82.9 ± 1.6	86.1 ± 1.2	87.3 ± 1.3	68.6 ± 2.0	75.5 ± 1.6	82.6 ± 1.6
RLFS	Semi	82.3 ± 1.8	85.3 ± 1.0	86.9 ± 0.5	67.4 ± 1.6	76.6 ± 1.5	85.2 ± 1.7
S2FS	Semi	80.7 ± 1.0	82.4 ± 2.3	83.6 ± 2.1	69.1 ± 2.4	77.9 ± 1.2	85.6 ± 1.8
S2LFS	Semi	$\mathbf{84.5 \pm 0.5}$	$\mathbf{87.7 \pm 1.0}$	$\mathbf{88.9 \pm 1.3}$	$\mathbf{71.5 \pm 1.3}$	$\mathbf{78.2 \pm 0.4}$	$\mathbf{87.8 \pm 0.6}$
SAFS	Test	81.7 ± 0.8	84.2 ± 0.7	85.6 ± 1.1	58.1 ± 2.9	66.2 ± 3.1	74.2 ± 2.3
CLS	Test	81.1 ± 0.3	83.4 ± 1.2	85.4 ± 0.8	59.9 ± 2.3	68.7 ± 2.4	75.4 ± 0.9
RSSL	Test	82.5 ± 1.1	85.6 ± 1.3	86.4 ± 1.0	67.1 ± 2.6	74.9 ± 1.3	80.9 ± 2.0
RLFS	Test	82.1 ± 1.7	85.2 ± 0.7	86.4 ± 2.2	66.3 ± 3.3	73.5 ± 2.2	83.1 ± 1.4
S2FS	Test	80.3 ± 2.6	82.2 ± 2.4	83.4 ± 2.1	68.1 ± 0.7	75.9 ± 1.7	85.2 ± 1.1
S2LFS	Test	$\mathbf{83.6 \pm 0.5}$	$\mathbf{87.2 \pm 0.9}$	$\mathbf{87.6 \pm 0.4}$	$\mathbf{69.4 \pm 0.5}$	$\mathbf{77.9 \pm 1.0}$	$\mathbf{87.1 \pm 0.8}$

a) The best results are highlighted in bold.

For each data set, 50% of samples were randomly chosen as training data and the remaining samples were used as test data. Furthermore, s% of training data were randomly sampled as labeled data. In experiments, s is set to 10, 20 and 50, respectively.

4.2 Comparison scheme

To demonstrate the effectiveness of the proposed method for feature selection, it was compared to the following representative semi-supervised feature selection methods and feature selection methods exploring local information.

- **SAFS.** The spectral analysis semi-supervised feature selection method learning feature's relevance using a regularization framework [30].
- **CLS.** The constrained laplacian score method evaluating the relevance of features by preserving their locality using constraints [15].
- **RSSL.** The robust structured subspace learning method for semi-supervised feature selection by exploring local information [37].
- **RLFS.** The rescaled linear regression-based semi-supervised feature selection method rescaling regression coefficients using a set of scale factors for feature selection in [31].
- **S2FS.** The proposed semi-supervised feature selection method using one feature selection indicator vector for all classes.
- **S2LFS.** The proposed semi-supervised local feature selection method exploring discriminative information and performing class-specific feature selection.

The above listed methods were compared using selected features with the same dimensionality. The grid search strategy was used for all the methods to tune their hyper-parameters. The search range is $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3\}$ for all the methods. The parameters are tuned based on the unlabeled learning data. Following previous studies on feature selection, the number of selected features was set to $\{50, 80, 110, 140, 170, 200\}$ for the USPS data set and $\{50, 100, 150, 200, 250, 300\}$ for the other data sets. The ridge regression model was used as the classifier to verify the performance of the methods in terms of their classification accuracy over unlabeled and test data.

4.3 Results and analysis

The training and test data sets were generated five times to alleviate the bias of the data partitioning process, and all the methods were run over these five sets. The average results over the five runs obtained by each method for the unlabeled and test data are listed in Tables 2–4.

From the results, it can be easily observed that the proposed S2LFS method achieves the best performance over the unlabeled and test data for all the six data sets, which verifies the effectiveness of the proposed method. There are some other observations. First, the proposed S2LFS method outperforms S2FS, which demonstrates the superiority of the local feature selection scheme. Second, the proposed S2LFS is better than RLFS, which focuses on rescaling the regression coefficients of a rigid regression

Table 3 Classification accuracy (CA%±std) of different feature selection methods over the ORL and Binary Alphabet data sets with 100 selected features[a)]

Data set		ORL			Binary Alphabet		
		$s = 10$	$s = 20$	$s = 50$	$s = 10$	$s = 20$	$s = 50$
SAFS	Semi	46.6 ± 6.5	49.6 ± 2.3	65.7 ± 1.9	13.9 ± 0.7	14.8 ± 1.6	32.4 ± 2.7
CLS	Semi	30.2 ± 4.3	31.9 ± 2.1	34.2 ± 3.8	15.9 ± 1.5	16.6 ± 1.3	31.8 ± 1.7
RSSL	Semi	47.5 ± 3.9	52.4 ± 4.0	70.8 ± 3.4	16.4 ± 1.2	19.6 ± 0.8	38.1 ± 1.6
RLFS	Semi	42.1 ± 2.8	43.5 ± 3.2	68.3 ± 5.0	16.6 ± 0.8	18.2 ± 1.8	35.6 ± 2.9
S2FS	Semi	49.4 ± 3.7	54.4 ± 2.7	72.5 ± 3.2	18.2 ± 0.7	20.2 ± 1.2	39.8 ± 1.1
S2LFS	Semi	**51.2 ± 2.5**	**55.8 ± 1.8**	**74.3 ± 2.6**	**19.8 ± 1.3**	**21.8 ± 1.1**	**41.3 ± 1.5**
SAFS	Test	44.7 ± 3.4	46.8 ± 2.9	60.7 ± 6.6	13.5 ± 1.3	13.4 ± 1.0	31.5 ± 1.5
CLS	Test	31.1 ± 3.5	32.3 ± 3.1	34.1 ± 4.6	14.5 ± 1.0	16.2 ± 1.4	31.9 ± 1.1
RSSL	Test	45.0 ± 0.5	49.4 ± 1.6	69.5 ± 3.2	15.6 ± 0.9	18.3 ± 0.8	35.7 ± 1.2
RLFS	Test	40.8 ± 4.7	42.6 ± 4.8	67.3 ± 3.8	15.4 ± 0.7	17.5 ± 1.9	34.2 ± 1.9
S2FS	Test	47.6 ± 1.2	51.2 ± 1.8	71.3 ± 1.4	16.2 ± 1.0	19.7 ± 0.6	36.4 ± 0.9
S2LFS	Test	**50.6 ± 1.8**	**52.5 ± 1.6**	**72.3 ± 1.3**	**17.5 ± 0.4**	**20.6 ± 1.2**	**38.1 ± 1.4**

a) The best results are highlighted in bold.

Table 4 Classification accuracy (CA%±std) of different feature selection methods over the Pointing4 and YaleB data sets with 100 selected features[a)]

Data set		Pointing4			YaleB		
		$s = 10$	$s = 20$	$s = 50$	$s = 10$	$s = 20$	$s = 50$
SAFS	Semi	62.1 ± 1.1	68.7 ± 2.5	74.7 ± 1.5	48.3 ± 1.5	61.5 ± 2.4	77.2 ± 1.1
CLS	Semi	63.3 ± 1.6	67.8 ± 2.1	72.4 ± 1.9	49.8 ± 3.5	66.0 ± 2.4	80.3 ± 1.0
RSSL	Semi	64.2 ± 2.9	72.9 ± 0.8	80.6 ± 2.3	56.8 ± 2.9	70.9 ± 1.7	81.6 ± 2.4
RLFS	Semi	65.7 ± 1.6	69.4 ± 0.9	77.2 ± 1.0	55.4 ± 1.4	71.3 ± 1.5	82.7 ± 1.2
S2FS	Semi	66.6 ± 1.2	74.7 ± 1.3	81.6 ± 1.1	57.2 ± 2.4	72.9 ± 2.3	84.6 ± 1.9
S2LFS	Semi	**68.4 ± 1.5**	**75.7 ± 0.8**	**83.8 ± 1.5**	**58.5 ± 1.9**	**74.1 ± 1.3**	**86.2 ± 1.3**
SAFS	Test	62.1 ± 2.4	64.7 ± 2.5	72.3 ± 1.4	46.1 ± 3.0	59.1 ± 2.5	75.9 ± 4.3
CLS	Test	62.1 ± 2.3	66.5 ± 1.3	71.2 ± 0.8	48.2 ± 4.2	65.5 ± 1.9	79.2 ± 1.4
RSSL	Test	62.7 ± 0.7	71.6 ± 1.3	78.3 ± 1.7	55.6 ± 2.0	68.4 ± 2.2	79.8 ± 2.1
RLFS	Test	63.6 ± 2.3	69.9 ± 1.5	75.2 ± 1.2	55.2 ± 3.5	71.5 ± 2.2	81.9 ± 1.2
S2FS	Test	65.6 ± 1.6	72.8 ± 2.4	80.4 ± 1.2	56.5 ± 1.7	72.5 ± 1.1	82.5 ± 0.9
S2LFS	Test	**67.3 ± 2.1**	**74.1 ± 1.5**	**82.2 ± 1.4**	**57.3 ± 0.8**	**73.5 ± 1.2**	**84.8 ± 1.6**

a) The best results are highlighted in bold.

model, which can well indicate the superiority of the proposed local feature selection method. Furthermore, S2LFS and RSSL, which are semi-supervised feature selection methods based on nonnegative spectral analysis, achieve better results in most cases compared to SAFS, CLS, and RLFS. This indicates that the underlying semantic information of unlabeled data can be uncovered, which can help guide the procedure of feature selection.

The experiments were conducted by varying the number of selected features. The results with $s = 20$ over the test data are illustrated in Figure 2, where the proposed S2LFS method is compared with SAFS, CLS, RSSL, and RLFS. It can be noticed from the figure that the proposed S2LFS method achieves the best classification performance over all the six data sets with different numbers of selected features in almost all cases. This verifies the effectiveness of the proposed method that jointly learns the class labels of unlabeled data and local feature selection indicators for feature selection. The anomalous results obtained for the Binary Alphabet data set can be explained by the feature distribution. When the top 150 features are selected, classifiers are being somewhat misled. This misleading information can be removed or rectified if selecting less or more features.

4.4 Sensitivity analysis

The proposed method relies on two hyper-parameters, β and λ, which balance the importance of different terms. Different values of these two hyper-parameters may lead to different performance. Hence, it is important to conduct the hyper-parameter sensitivity study with respective to β and λ using the six data

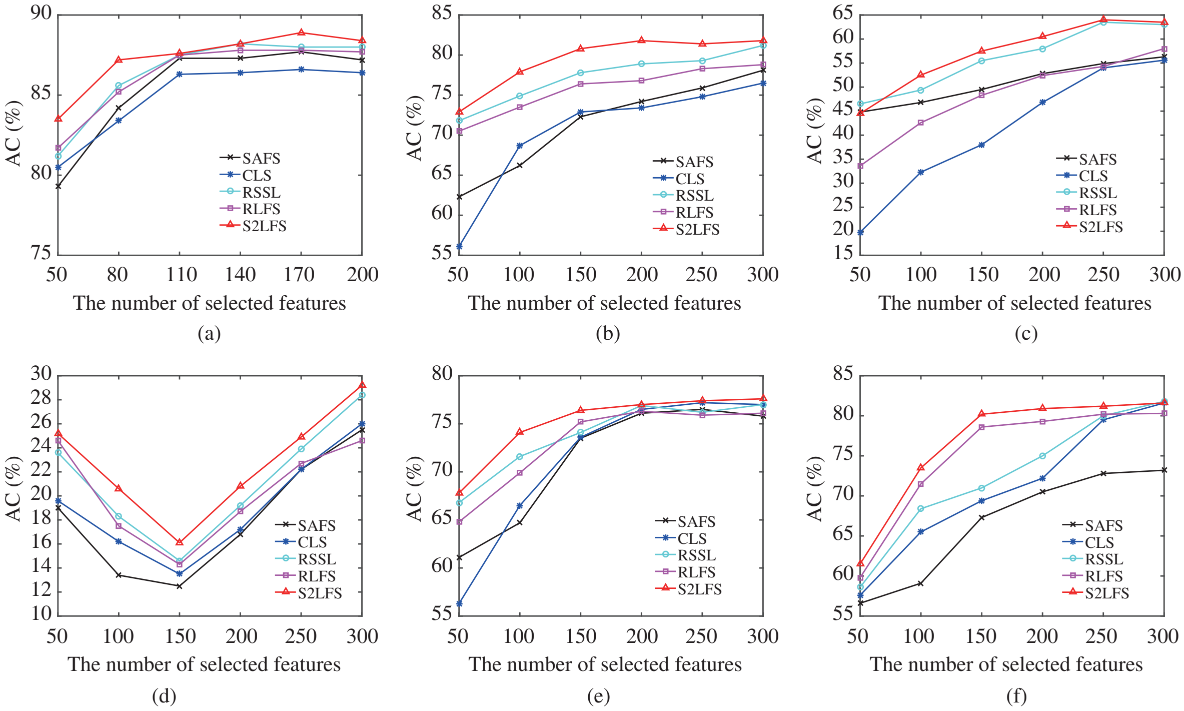

Figure 2 (Color online) The classification accuracy of different feature selection methods with respect to (w.r.t.) different numbers of selected features on the six data sets with $s = 20$. (a) USPS; (b) COIL20; (c) ORL; (d) Binary Alphabet; (e) Pointing4; (f) YaleB.

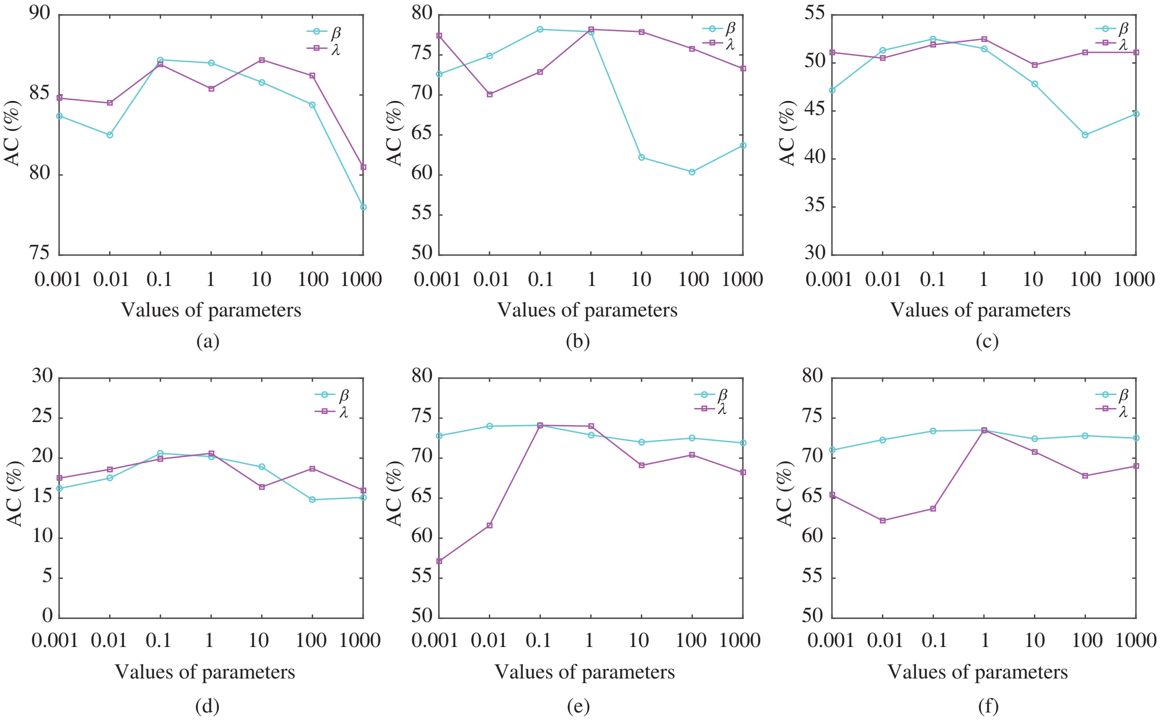

Figure 3 (Color online) Performance variation of the proposed method w.r.t. different values of the parameters β and λ over the (a) USPS, (b) COIL20, (c) ORL, (d) Binary Alphabet, (e) Pointing4, and (f) YaleB data sets with $s = 20$.

sets.

Figure 3 illustrates the experimental results for the test data of all the six data sets with $s = 20$. The number of selected features was set to 80 for the USPS data set and 100 for the other data sets. It can be noticed from the figure that the values of the hyper-parameters β and λ actually affect the classification

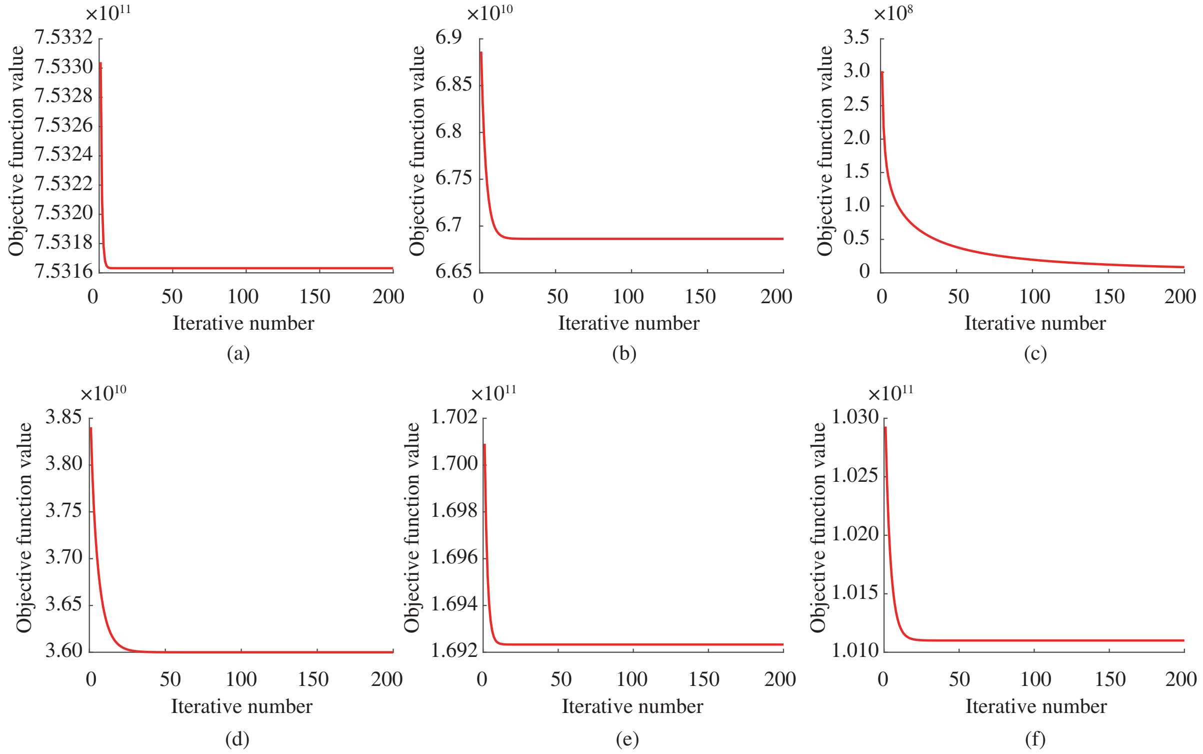

Figure 4 (Color online) Convergence curves for the proposed method over the (a) USPS, (b) COIL20, (c) ORL, (d) Binary Alphabet, (e) Pointing4, and (f) YaleB data sets.

performance, which indicates the importance of the corresponding terms. That is, it is necessary and helpful to introduce the corresponding terms. While the default values of these two hyper-parameters can be provided, it would be advisable to tune them. The suitable ranges for the default values are $[0.1, 1]$ for β and $[0.1, 10]$ for γ.

4.5 Convergence study

The convergence of an optimization algorithm is an important problem. Figure 4 illustrates the convergence curves for the proposed S2LFS over all the six data sets.

In this figure, the x-axis represents the number of iterations while the y-axis represents the values of the corresponding objective functions. It can be noticed from the convergence curves that the proposed optimization algorithm quickly converges to the local optimal solutions. That means that the proposed method enables efficient learning of the desired class-specific feature selection matrix.

5 Conclusion

This paper presented a new semi-supervised local feature selection method that selects different discriminative feature subsets to represent samples from different classes. According to the proposed method, the class-specific importance scores of features for different classes are learned simultaneously with the classes of unlabeled data, which allows the selected features to optimally adapt to the classes. The discriminative ability of the selected feature subsets was explored. Experiments on six real data sets demonstrated the effectiveness of the proposed local feature selection method. In the future, we plan to extend the proposed method to address the zero-shot learning problem and the cases where only a big amount of unsupervised data are available.

Acknowledgements This work was supported by National Key Research and Development Program of China (Grant No. 2017YFC0820601), National Natural Science Foundation of China (Grant No. 61720106004, 61732007), and Natural Science Foundation of Jiangsu Province (Grant No. BK20170033).

References

1 Zhuang Y T, Han Y H, Wu F, et al. Stable multi-label boosting for image annotation with structural feature selection. Sci China Inf Sci, 2011, 54: 2508–2521

2 Liu C W, Pei M T, Wu X X, et al. Learning a discriminative mid-level feature for action recognition. Sci China Inf Sci, 2014, 57: 052112

3 Chen J B, Stern M, Wainwright M J, et al. Kernel feature selection via conditional covariance minimization. In: Proceedings of Advances in Neural Information Processing Systems, Long Beach, 2017. 6946–6955

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

5 Li Z C, Tang J H. Unsupervised feature selection via nonnegative spectral analysis and redundancy control. IEEE Trans Image Process, 2015, 24: 5343–5355

6 Nie F P, Huang H, Cai X. et al. Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization. In: Proceedings of Advances in Neural Information Processing Systems, Vancouver, 2010. 1813–1821

7 Li Z C, Yang Y, Liu J, et al. Unsupervised feature selection using nonnegative spectral analysis. In: Proceedings of AAAI Conference on Artificial Intelligence, Toronto, 2012. 1026–1032

8 Mitra P, Murthy C A, Pal S K. Unsupervised feature selection using feature similarity. IEEE Trans Pattern Anal Machine Intell, 2002, 24: 301–312

9 He X F, Cai D, Niyogi P. Laplacian score for feature selection. In: Proceedings of Advances in Neural Information Processing Systems, Vancouver, 2005. 1813–1821

10 Kolar M, Liu H. Feature selection in high-dimensional classification. In: Proceedings of International Conference on Machine Learning, Atlanta, 2013. 329–337

11 Gao S Y, ver Steeg G, Galstyan A. Variational information maximization for feature selection. In: Proceedings of Advances in Neural Information Processing Systems, Barcelona, 2016. 487–495

12 Zhao Z, Liu H. Spectral feature selection for supervised and unsupervised learning. In: Proceedings of International Conference on Machine Learning, Corvallis, 2007. 1151–1157

13 Helleputte T, Dupont P. Partially supervised feature selection with regularized linear models. In: Proceedings of International Conference on Machine Learning, Montreal, 2009. 409–416

14 Xu Z L, Jin R, Lyu M R, et al. Discriminative semi-supervised feature selection via manifold regularization. In: Proceedings of International Joint Conference on Artificial Intelligence, Pasadena, 2009. 1303–1308

15 Benabdeslem K, Hindawi M. Constrained laplacian score for semi-supervised feature selection. In: Proceedings of European Conference on Machine Learning and Knowledge Discovery in Databases, Athens, 2011. 204–218

16 Li Y H, Dong M, Hua J. Localized feature selection for clustering. Pattern Recogn Lett, 2008, 29: 10–18

17 Armanfard N, Reilly J P, Komeili M. Local feature selection for data classification. IEEE Trans Pattern Anal Mach Intell, 2016, 38: 1217–1227

18 Bugata P, Drotar P. On some aspects of minimum redundancy maximum relevance feature selection. Sci China Inf Sci, 2020, 63: 112103

19 Huang T T, Xu Y C, Bai S, et al. Feature context learning for human parsing. Sci China Inf Sci, 2019, 62: 220101

20 Zhang Q, Li R, Chu T G. Kernel semi-supervised graph embedding model for multimodal and mixmodal data. Sci China Inf Sci, 2020, 63: 119204

21 Cai D, Zhang C Y, He X F. Unsupervised feature selection for multi-cluster data. In: Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, 2010. 333–342

22 Boutsidis C, Mahoney M W, Drineas P. Unsupervised feature selection for the k-means clustering problem. In: Proceedings of Advances in Neural Information Processing Systems, Vancouver, 2009. 153–161

23 Li C Z, Xu Z B, Qiao C, et al. Hierarchical clustering driven by cognitive features. Sci China Inf Sci, 2014, 57: 012109

24 An S, Wang J, Wei J M, et al. Unsupervised feature selection with joint clustering analysis. In: Proceedings of ACM Conference on Information and Knowledge Management, Singapore, 2017. 1639–1648

25 Li Z C, Liu J, Yang Y, et al. Clustering-guided sparse structural learning for unsupervised feature selection. IEEE Trans Knowl Data Eng, 2014, 26: 2138–2150

26 Wang J, Wei J M, Yang Z L. Supervised feature selection by preserving class correlation. In: Proceedings of ACM International Conference on Information and Knowledge Management, Indianapolis, 2016. 1613–1622

27 Zhang R, Nie F P, Li X L. Self-weighted supervised discriminative feature selection. IEEE Trans Neural Netw Learn Syst, 2018, 29: 3913–3918

28 Tang J H, Shu X B, Qi G J, et al. Tri-clustered tensor completion for social-aware image tag refinement. IEEE Trans Pattern Anal Mach Intell, 2017, 39: 1662–1674

29 Tang J H, Shu X B, Li Z C, et al. Social anchor-unit graph regularized tensor completion for large-scale image retagging. IEEE Trans Pattern Anal Mach Intell, 2019, 41: 2027–2034

30 Zhao Z, Liu H. Semi-supervised feature selection via spectral analysis. In: Proceedings of SIAM International Conference on Data Mining, Minneapolis, Minnesota, 2007. 641–646

31 Chen X J, Yuan G W, Nie F P, et al. Semi-supervised feature selection via sparse rescaled linear square regression. IEEE Trans Knowl Data Eng, 2020, 32: 165–176

32 Yuan G W, Chen X J, Wang C, et al. Discriminative semi-supervised feature selection via rescaled least squares regression-supplement. In: Proceedings of AAAI Conference on Artificial Intelligence, New Orleans, 2018. 8177–8178

33 Benabdeslem K, Hindawi M. Efficient semi-supervised feature selection: constraint, relevance, and redundancy. IEEE Trans Knowl Data Eng, 2014, 26: 1131–1143

34 Sheikhpour R, Sarram M A, Gharaghani S, et al. A survey on semi-supervised feature selection methods. Pattern Recogn, 2017, 64: 141–158

35 Dhillon I S. Co-clustering documents and words using bipartite spectral graph partitioning. In: Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, 2001. 269–274

36 Nakajima S, Takeda A, Babacan S D, et al. Global solver and its efficient approximation for variational Bayesian low-rank subspace clustering. In: Proceedings of Advances in Neural Information Processing Systems, Lake Tahoe, 2013. 1439–1447

37 Li Z C, Liu J, Tang J H, et al. Robust structured subspace learning for data representation. IEEE Trans Pattern Anal Mach Intell, 2015, 37: 2085–2098

38 Sun Y J, Todorovic S, Goodison S. Local-learning-based feature selection for high-dimensional data analysis. IEEE Trans Pattern Anal Mach Intell, 2010, 32: 1610–1626

39 Guan Y, Dy J G, Jordan M I. A unified probabilistic model for global and local unsupervised feature selection. In: Proceedings

of International Conference on Machine Learning, Bellevue, 2011. 1073–1080

40 Hindawi M, Benabdeslem K. Local-to-global semi-supervised feature selection. In: Proceedings of ACM International Conference on Information and Knowledge Management, San Francisco, 2013. 2159–2168

41 Zhu X J, Ghahramani Z B, Lafferty J D. Semi-supervised learning using gaussian fields and harmonic functions. In: Proceedings of International Conference on Machine Learning, Washington, 2003. 912–919

42 Hull J J. A database for handwritten text recognition research. IEEE Trans Pattern Anal Machine Intell, 1994, 16: 550–554

43 Nene S A, Nayar S K, Murase H. Columbia Object Image Library (COIL-20). Technical Report CUCS-005-96. 1996

44 Gourier N, Hall D, Crowley J L. Estimating face orientation from robust detection of salient facial features. In: Proceedings of Pointing 2004 ICPR International Workshop on Visual Observation of Deictic Gestures, Cambridge, 2004. 1–9

45 Georghiades A S, Belhumeur P N, Kriegman D J. From few to many: illumination cone models for face recognition under variable lighting and pose. IEEE Trans Pattern Anal Mach Intell, 2001, 23: 643–660