Semisupervised Discriminative Locally Enhanced Alignment for Hyperspectral Image Classification

Qian Shi, Student Member, IEEE, Liangpei Zhang, Senior Member, IEEE, and Bo Du, Member, IEEE

Abstract—This paper proposes a new semisupervised dimension reduction (DR) algorithm based on a discriminative locally enhanced alignment technique. The proposed DR method has two aims: to maximize the distance between different classes according to the separability of pairwise samples and, at the same time, to preserve the intrinsic geometric structure of the data by the use of both labeled and unlabeled samples. Furthermore, two key problems determining the performance of semisupervised methods are discussed in this paper. The first problem is the proper selection of the unlabeled sample set; the second problem is the accurate measurement of the similarity between samples. In this paper, multilevel segmentation results are employed to solve these problems. Experiments with extensive hyperspectral image data sets showed that the proposed algorithm is notably superior to other state-of-the-art dimensionality reduction methods for hyperspectral image classification.

Index Terms—Dimension reduction (DR), multilevel segmentation, semisupervised learning.

I. INTRODUCTION

THE imaging spectrometer, airborne or spaceborne, is a novel sensor that has been developed over the last 30 years. It can gather data with hundreds of bands covering a broad spectrum of wavelength range. Hyperspectral imagery contains rich spectral information and has proven to be very effective for discriminating the subtle differences in ground objects. However, as to hyperspectral classification, not every spectral band contributes to the material identification; thus, feature extraction methods are usually employed to establish more concentrated features for separating different materials [1], [2]. Among them, discriminative feature extraction methods learn a suitable subspace where the separability between the different classes is expected to be enhanced. Typical methods include linear discriminant analysis [3] and nonparametric weighted feature extraction (NWFE) [4], which design proper scatter matrices to effectively measure the class separability and have been widely used for hyperspectral imagery. Many extensions to these two methods have been proposed, such as modified Fisher’s linear discriminant analysis [5], regularized linear discriminant analysis [6], cosine-based nonparametric feature extraction [7], and double nearest proportion feature extraction [8]. Recently, manifold learning methods have proved effective in exploiting the intrinsic relationship between hyperspectral image samples. For example, marginal Fisher analysis (MFA) [9] constructs an intrinsic graph to describe the intraclass compactness and a penalty graph to characterize the interclass separability. Locality preserving projections (LPPs) [10] seek a linear mapping that preserves the local structure of neighboring samples in the input space. Local fisher discriminant analysis (LFDA) [11] utilizes the idea of LPP to preserve the underlying structure of the multimodal non-Gaussian class distribution. Discriminative locality alignment (DLA) [12], which is based on the patch alignment strategy [13], selects neighbors for a local patch from both intraclass and interclass to enlarge the margin between different classes.

However, one problem with DLA is that a fixed number of neighboring points from interclass and intraclass are used in the algorithm, implying that every pixel has the same separability, which does not correspond with reality. As a result, the more separable points may be overemphasized, whereas the severely mixed points may not be well discriminated. Therefore, the optimal objective function in part optimization [12] should aim to preserve the distances between well-separated local patches and, at the same time, place an adequate emphasis on local patches which are hard to discriminate.

In many supervised-learning-based classification methods, the collection of labeled training data requires expensive human labor [14]–[16]. Meanwhile, it is much easier to obtain unlabeled data. Utilizing the abundant unlabeled data, as well as the relatively limited amount of labeled data, usually presents better classification results. This is the basic idea underlying semisupervised learning methods such as transductive support vector machines (SVMs) [17], cotraining [18], and graph-based techniques [15], [19]–[21]. This paper focuses on the graph-based methods.

In order to get a more stable and credible solution, a smoothness constraint is imposed on the graph and can be represented by a regularization term. In fact, by introducing this regularization term, many conventional supervised dimension reduction (DR) methods can be extended to the semisupervised variants. For example, semisupervised discriminant analysis (SDA) [23] uses this term to alleviate the singular matrix problem of within-class covariance due to insufficient training data. Subspace
semisupervised LPP and subspace semisupervised MFA \[24\] regularize this term by incorporating the manifold structure as prior knowledge, since the data points lying on a densely linked subgraph are likely to have the same label.

However, in practical applications, two crucial problems still exist with the semisupervised methods. One problem comes from the method of choosing the unlabeled points \[25\]. A typical hyperspectral image usually consists of over a million pixels. However, these pixels are highly correlated and cannot be used to represent the manifold structure, so it is inappropriate to use all the pixels in the hyperspectral image as the unlabeled samples. In fact, only part of the representative pixels may be favorable and can reveal the intrinsic geometrical structure of the data set. As pointed out in \[26\] and \[27\], the selected unlabeled data can be beneficial only if the marginal data distribution is properly modeled. In this paper, efficient graph-based image segmentation \[28\] is utilized to oversegment the image into a large number of small patches. The roots of the oversegmented patches can then be used as unlabeled candidates.

Another crucial problem with the off-the-shelf semisupervised methods is how to accurately measure the similarity between samples, which is also the key to many manifold learning algorithms \[29\]–\[31\]. A similarity metric can determine the neighborhood relationship and adjacency degree between the neighborhood samples. Many approaches have been proposed to learn an appropriate similarity function for a data set \[32\]–\[39\]. For hyperspectral applications, Camps-Valls et al. defined a family of composite kernels to incorporate contextual information, which can make full use of the intrinsic spatial information of hyperspectral imagery \[20\] and presents a better performance than a single kernel from spectral information. This paper further exploits the spatial relationship between the samples to develop a suitable similarity measurement by a merging operation on the oversegmented results. The basic assumption is that, if two regions have a high similarity in the spectral feature space, and can also be merged in the following merging operations, they probably share the same label.

Based on the aforementioned consideration, this paper proposes a DR method based upon a patch alignment framework for hyperspectral imagery, named semisupervised discriminative locally enhanced alignment (DLEA) (SDLEA). The main contributions of this paper include the following.

1) SDLEA focuses on overcoming the aforementioned problem of DLA. What is more, the similarity measurement of samples within the same class is constructed so as to preserve the local structure in the class.

2) The oversegmented result is used to obtain a comprehensive and representative unlabeled set. Furthermore, the relevance between the labeled and unlabeled samples is also taken into consideration.

3) The spatial relationship from the multisegmentation results is combined with the spectral information as auxiliary information in the framework, preserving the intrinsic features of the hyperspectral image data set after the DR operation.

The rest of this paper is organized as follows. Section II details our proposed method. Section III describes the way of utilizing multisegmentation to solve the problems associated with a semisupervised method and introduces the implementation of our method on hyperspectral images, with a step-by-step description. Experimental results are presented in Section IV, and conclusions are given in Section V.

II. SDLEA

The general problem with linear feature extraction is first illustrated. Suppose that we have a labeled set \(X = [x_1, x_2, \ldots, x_n]\) and an unlabeled set \(X_{l+1}, x_{l+2}, \ldots, x_n\) in \(R^{n\times d}\), where \(I_n\) is the number of labeled samples, \(n\) is the number of samples, \(n - I_n\) is the number of unlabeled samples, and \(B\) is the dimension of the hyperspectral data. These samples belong to \(c\) classes. As to the hyperspectral classification, samples actually refer to the vectored pixels in the image data set. The purpose is to find a transformation matrix \(A \in R^{B \times d}\) that projects all the vectored pixels \(X = [x_1, x_2, \ldots, x_n]\) onto a lower dimensional space by \(Z = A^T X\), where \(Z = [z_1, z_2, \ldots, z_n]^T \in R^{n\times d}\) with \(d < B\).

The purpose of these projections is to project the original data set onto a subspace which allows for a better separability between different land-cover types.

The semisupervised dimensionality reduction framework \[16\], \[22\] contains two terms: a discrimination term and a regularization term. The global objective function is made up of these two terms. The labeled samples are used to construct the discrimination term. Meanwhile, all the samples are used to construct the regularization term, in the same way as other semisupervised learning methods.

Based on the patch alignment framework \[13\], this semisupervised dimensionality reduction framework can be divided into two succeeding procedures, comprising part optimization and global optimization. Each sample is associated with a local patch, which consists of its nearest neighbors. Part optimization refers to preserving the local data structure in the local patch after DR. The succeeding global optimization takes all the part optimization patches into consideration by preserving the data structure after DR. In our method, the discrimination term is employed to maximize the distance between the different kinds of neighbors in the part optimization. An example of samples with different degrees of separability is shown in Fig. 1. The left picture is the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) hyperspectral image. We cut out a part of the image which has three classes: corn-notill,
soybeans-min, and soybeans-notill. The right picture is the scatter plot of the three classes. Each sample is associated with a local patch, which consists of the nearest neighbors. Thus, from the label of each neighbor, we can determine the degree of separability of the given samples. If neighbors from the same class dominate the local patch, it implies high separability, like patch A. The number of neighbors from the same class is set as \( k_{i1} \), and the number of neighbors from different classes is set as \( k_{i2} \). If we set \( k_{i1} \) and \( k_{i2} \) to fixed values, as DLA does, the low-separability patches cannot be well characterized and ultimately not discriminated. In contrast, the high-separability patches may be overemphasized. To overcome this defect of DLA, \( k_{i1} \) and \( k_{i2} \) are dynamically selected in our method.

First, we just use the labeled samples to derive the subspace \( \mathbf{Z}^l = [\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n]^T \in \mathbb{R}^{d \times n} \) \((d < B)\). For a given labeled sample \( \mathbf{x}_i \in \mathbf{X}^l \), we can find its \( k \)-nearest neighbors (KNNs), in which neighboring samples from the same class and different classes can be found. The criterion for determining the two numbers is that, in a patch with high separability, more neighbors from the same class are chosen to preserve the local geometry structure. Meanwhile, in a patch with low separability, more neighbors from different classes are used to model the discriminative information.

In the low-dimensional space, for each patch, the distances \( D_{\text{same}} \) between \( \mathbf{z}_i \) and its neighboring samples from the same class should be preserved to model the local geometry, while the distances \( D_{\text{diff}} \) between \( \mathbf{z}_i \) and the neighboring samples from different classes should be maximized

\[
D_{\text{same}(i)} = \arg\min_{\mathbf{z}_j} \sum_{j=1}^{k_{i1}} \| \mathbf{z}_i^l - \mathbf{z}_j^l \|^2 \ w_{ij}
\]

\[
D_{\text{diff}(i)} = \arg\max_{\mathbf{z}_j} \sum_{j=1}^{k_{i2}} \| \mathbf{z}_i^l - \mathbf{z}_j^l \|^2
\]

where \( w_{ij} \) records the weight calculated by the heat kernel

\[
w_{ij} = \exp(-\| \mathbf{z}_i^l - \mathbf{z}_j^l \|^2 / 2\sigma^2).
\]

The aforementioned two equations can be combined as the part discriminator by using a linear manipulation

\[
D(\mathbf{z}_i^l) = \arg\min_{\mathbf{z}_i^l} \left( \sum_{j=1}^{k_{i1}} \| \mathbf{z}_i^l - \mathbf{z}_j^l \|^2 w_{ij} + \beta \sum_{j=1}^{k_{i2}} \| \mathbf{z}_i^l - \mathbf{z}_j^l \|^2 \right)
\]

where \( \beta \) is a scaling factor to balance the constraints of the within-class distance and the between-class distance. \( \mathbf{z}_i^l \) is the \( j \)th interclass point, and \( \mathbf{z}_j^l \) is the \( j \)th intraclass point. Eq. 2 shows the part optimization stage in DLEA (SDLEA). The center red points are the given points, and the other colored points represent their neighbors. After the DR process, the distance between the interclass points is minimized, and in contrast, the distance between the intraclass points is maximized.

Define the coefficient vector

\[
w_{\text{label}(i)} = \begin{bmatrix} k_{i1} & -\beta \cdots -\beta \\ w_{i1}, \ldots, w_{i_k1}, \ldots, -\beta \end{bmatrix}^T
\]

in (3), \( k_{i1} + k_{i2} = k \). Then, substitute (3) into (2)

\[
D(\mathbf{z}_i^l) = \arg\min_{\mathbf{z}_i^l} \left( \sum_{j=1}^{k_{i1}} \| \mathbf{z}_i^l - \mathbf{z}_j^l \|^2 w_{ij} + \beta \sum_{j=1}^{k_{i2}} \| \mathbf{z}_i^l - \mathbf{z}_j^l \|^2 \right)
\]

where \( \beta \) is the alignment matrix. It is obtained by an iterative procedure

\[
\mathbf{L}_{\text{label}}(F_i, F_i) \leftarrow \mathbf{L}_{\text{label}}(F_i, F_i) + \mathbf{L}_{\text{label}(i)}
\]

where \( F_i \) is the set of indices for the \( i \)th patch, which is built by the given sample \( x_i \) and its \( k \) related neighbors in the labeled sample set.

For linearization [19], we can impose \( \mathbf{U}_d^T \mathbf{U}_d = \mathbf{I}_d \) to determine the projection matrix \( \mathbf{U}_d \), such that \( \mathbf{Z}^l = \mathbf{U}_d^T \mathbf{X}^l \) is used and (7) is deformed as

\[
\arg\min_{\mathbf{U}_d} \left( \mathbf{U}_d^T \mathbf{X}^l \mathbf{L}_{\text{label}} \mathbf{X}^r \mathbf{U}_d \right) \quad \text{s.t.} \quad \mathbf{U}_d^T \mathbf{U}_d = \mathbf{I}_d.
\]
Equation (9) is a basic optimization problem, which can be solved by using a Lagrangian multiplier method, and its solution can be obtained by conducting a generalized or standard eigenvalue decomposition on $X^T L_{\text{label}} X^T$.

The regularization term is expected to maintain the neighboring relations between all the samples in the whole data set. However, in some cases, neighboring samples in the original data space may belong to different classes. This contradicts the assumption that samples which are close in the original feature space are likely to belong to the same class. In this case, the obtained weight matrix may not precisely represent the relationship of the points. To alleviate this problem, the regularization term is expected to maintain the neighboring relations between all the samples in the whole data set. Thus, the regularization term can be defined as follows:

$$R(Z_i) = \arg \min_{Z_i} \left( \sum_{j=1}^{k_s} \| z_i - z_j \|_2^2 w_{\text{unlabel}(i)} \right)$$

where $k_s$ is the number of neighbors of the given sample.

Define the coefficient vector

$$w_{\text{unlabel}(i)} = [s(i, i_1), s(i, i_2), \ldots, s(i, i_{k_s})]^T$$

where $s(i, j) = \exp(-\|z_i - z_j\|^2/2\sigma^2)$, $j = 1, \ldots, k_s$, and the weight is measured in the discriminant subspace. Then, substitute (11) into (10)

$$R(Z_i) = \arg \min_{Z_i} \text{tr} \left( Z_i \begin{bmatrix} -c_{k_s} & I_{k_s} \end{bmatrix} \text{diag}(w_{\text{unlabel}(i)}) \begin{bmatrix} -c_{k_s} & I_{k_s} \end{bmatrix}^T \right) \times \left\{ Z_i L_{\text{unlabel}(i)} Z_i^T \right\}$$

$$= \arg \min_{Z_i} \text{tr} \left( Z_i L_{\text{unlabel}(i)} Z_i^T \right)$$

where $L_{\text{unlabel}(i)} \in R^{(k_s+1) \times (k_s+1)}$. We define the global objective function as

$$Q(Z) = \arg \sum_{i=1}^{l} \min_{Z_i} \left( Z_i^T L_{\text{label}(i)} Z_i^T \right) + \frac{\gamma}{n} \sum_{i=1}^{n} \min_{Z_i} \left( Z_i^T L_{\text{unlabel}(i)} Z_i^T \right)$$

$$= \arg \min_Z \left( \sum_{i=1}^{l} S_{i}^L L_{\text{label}(i)} (S_{i}^L)^T + \gamma \sum_{i=1}^{n} S_{i}^U L_{\text{unlabel}(i)} (S_{i}^U)^T \right) Z^T$$

$$= \arg \min_Z \left( Z (L_{\text{label}} + \gamma L_{\text{unlabel}}) Z^T \right)$$

where $\gamma$ is a control parameter, $S_{i}^L \in R^{n \times (k_s+1)}$ and $S_{i}^U \in R^{n \times (k_s+1)}$ are the selection matrices, and $L_{\text{label}} \in R^{n \times n}$ is the alignment matrix constructed by

$$L_{\text{label}} (F_{i}^L, F_{i}^L) \leftarrow L_{\text{label}} (F_{i}^L, F_{i}^L) + L_{\text{label}(i)}$$

$$L_{\text{unlabel}} (F_{i}^U, F_{i}^U) \leftarrow L_{\text{unlabel}} (F_{i}^U, F_{i}^U) + L_{\text{unlabel}(i)}$$

We can impose $U^T U = I_d$ to determine the projection matrix $U$, such that $Z = U^T X$ is used and (13) is transformed as

$$\arg \min_U \text{tr}(U^T X L X^T U) \text{ s.t. } U^T U = I_d.$$  

Equation (14) can be solved by using a Lagrangian multiplier method, and its solution can be obtained by conducting a generalized or standard eigenvalue decomposition on $X L X^T$.

### III. Implementation of SDLEA

In order to perform the semisupervised method, two key problems should be solved before its application to the hyperspectral image. The first problem is the selection of the unlabeled samples, and the second problem is the accurate similarity measurement between samples, which will be incorporated in the regularization term.

To solve these problems, images are first oversegmented into many extremely small patches, which is termed bottom-level segmentation, and unlabeled samples can be selected from the roots of these small patches. The root points on the segmentation image can well represent the original image, with regard to spectral and spatial properties. For example, compared with the results of random sampling, the samples in the homogeneous regions will be reduced; on the other hand, the samples in the heterogeneous regions will be increased.

After the oversegmentation step, the small patches can be merged to produce a midlevel-segmentation result, which plays the role of a reference to construct the spatial relationship between the unlabeled samples. As a result, the spatial relationship can be constructed from a higher segmentation level. To ensure a high degree of efficiency, efficient graph-based image segmentation [28] is employed as the segmentation method. Graph-based image segmentation can capture perceptually important groupings or regions from the image, which helps to keep the spatial consistency.

#### A. Multilevel Graph-Based Segmentation

The first step of the graph-based segmentation method is the graph construction, which differs for bottom-level segmentation and midlevel segmentation. We define the bottom-level graph as $G_b = (V_b, E_b)$. Each vertex represents a pixel in the original image, and the edges are built by each pixel and its eight neighboring pixels, which can be seen in Fig. 3(a). Based on the merging criterion in [28], some edges can be merged, and the bottom-level segmentation result can be generated, which can be seen in Fig. 3(b). Pixels with the same color
belong to the same cluster, and pixels 2, 5, and 20 are the root points of these clusters. The remaining edges construct the midlevel graph \( G_m = (V_m, E_m) \), on which each vertex is the root point on the bottom-level segmentation. If there are still edges existing between the clusters, these clusters can be seen as adjacent regions. Based on these relationships, we can use the graph-based segmentation method to undertake another merging process and generate the midlevel-segmentation result. More details about how to find the adjacent similar regions on the higher segmentation can be found in [40].

**B. Generating Unlabeled Samples From the Bottom-Level Segmentation**

The roots of oversegmented patches can be used to generate unlabeled candidates. Fig. 4 shows the root points on the bottom-level segmentation result. We can see that the number of root points in these three labeled regions is different, which illustrates the phenomenon of sampling densely in an area with large variance and sampling sparsely in an area with small variance. In order to show this idea more intuitively, a scatter plot of the three regions of equal size in the original images is shown in Fig. 5. Clearly, the scatter points in region 1 (blue) are more decentralized than the points in regions 2 and 3. In addition, the unlabeled samples in region 1 are denser than those in the other two regions, as shown in Fig. 4.

At the same time, the relevance relationship between the existing labeled data set and the unlabeled candidates should also be taken into consideration. This is because, in the semisupervised framework, there is the assumption that the labeled and unlabeled samples should lie in the same submanifold. The relevant unlabeled candidates are more informative and should be selected in the final unlabeled sample data set. To achieve this goal, we seek the \( t \) spectrally nearest neighbors for each unlabeled sample. If labeled samples exist among the neighbors, the relevance exits; thus, it should be selected in the final unlabeled set. In contrast, if no labeled neighbors can be found in these neighbors, the relevance may be weak, and it should be discarded. Note that the number of the neighbors \( t \) controls this relevance criterion. In practice, we should set a larger \( t \) to relax the constraint on this selection procedure, so as to tolerate more spectral variability of the samples in the hyperspectral image data set. Fig. 6 shows the distribution of the final unlabeled sample set.

Note that, if the unlabeled samples and labeled samples are located in the same patch on the bottom level, the label of the unlabeled samples can be determined directly by the homogeneity property of the patch on the bottom-level segmentation, given the same labels as the corresponding labeled samples.

**C. Similarity Adjustment in the Regularization Term by the Multilevel Segmentation Results**

We can see from (3) and (11) that the coefficient vector \( w_i \) records the similarity degree of each sample. For an unlabeled sample, the weight is set according to the similarity, which is calculated by the radial basis function (RBF) kernel with the given sample

\[
s_{\text{spectral}}(z_i, z_j) = \exp \left( -\frac{\|z_i - z_j\|^2}{2\sigma^2} \right). \tag{15}
\]

If spatial adjacency exits, the kernel parameter \( \sigma \) can be set higher. We can define \( \sigma' = c\sigma (c > 1) \); then, the similarity can be represented as

\[
s_{\text{spe_spa}}(z_i, z_j) = \exp \left( -\frac{\|z_i - z_j\|^2}{2\sigma'^2} \right). \tag{16}
\]
In this paper, the multisegmentation results are used to exploit this spatial adjacency relationship. If two different regions from the bottom-level segmentation map can be merged on the midlevel-segmentation map, they have spatial adjacency.

To sum up, this paper proposes a way of considering both the spectral and spatial neighborhood relationships of the hyperspectral image data set while also making full use of the spectral properties by a manifold structure analysis. The flowchart of the proposed approach is illustrated in Fig. 7.

Algorithm: Procedure of the SDLEA algorithm

**Inputs:** Labeled training samples $X = [x_1, x_2, \ldots, x_{l_n}] \in \mathbb{R}^{l_n \times B}$ and class labels $Y = [y_1, y_2, \ldots, y_{l_n}]$, $y_i \in [1, 2, \ldots, C]$

**Step 1:** Construct the alignment matrix $L_{\text{label}(i)}$ for each labeled sample by (1)–(8);

**Step 2:** Compute the discriminant subspace $Z'$ by (9);

**Step 3:** Multisegmentation;

**Step 4:** Extract the unlabeled samples from the bottom-level segmentation;

**Step 5:** Construct the alignment matrix $L_{\text{unlabel}(i)}$ for each unlabeled sample
- Look for spectral neighbors in the discriminant subspace constructed by (9);
- Considering the spatial adjacency, adjust the similarity by (16);
- Construct the alignment matrix $L_{\text{unlabel}(i)}$ by (10)–(13);

**Step 6:** Calculate the transformation matrix $U$ by (14) and the reduced feature space $Z = U^T X$

**Output:** Reduced feature space $Z \in \mathbb{R}^{d \times n}$ and the transformation matrix $U \in \mathbb{R}^{B \times d}$.

IV. EXPERIMENTS

In this section, we validate our approach with several popular hyperspectral imagery data sets and present experimental results demonstrating the benefits of SDLEA-based dimensionality reduction with classifiers such as KNNs and SVMs. We also report on the performance of the classification systems, as measured by the overall classification accuracy [42], [43]. The primary objectives of the experimental results reported in the next two sections are as follows: 1) to tune the parameters in the classification system (dimensionality reduction and classification for the hyperspectral image data sets); 2) to test the validity of the discrimination term in SDLEA by visualizing the scatter plot after the DR; 3) to test the efficacy of the unlabeled samples in our semisupervised method; and 4) to quantify the efficacy of SDLEA-based dimensionality reduction for hyperspectral image classification and compare it to that of traditional state-of-the-art DR methods, as used in the hyperspectral imagery community.

A. Experimental Hyperspectral Data Set Description

The Indian Pine hyperspectral data set was acquired by the National Aeronautics and Space Administration’s AVIRIS sensor [41]. It contains $145 \times 145$ pixels, with each pixel having 220 spectral bands covering the range of 375–2200 nm. The corresponding spatial resolution is approximately 20 m. A total of 9425 pixels from ten classes were used for our experiments.
Fig. 8(a)–(c) shows information about the Indian Pine data, which includes the pseudocolor image, ground truth, and the corresponding spectral curves. Fig. 8(d) shows the midlevel-segmentation result, based on the bottom-level segmentation result. Fig. 6 shows the selected unlabeled sample set in the Indian Pine data, which includes 1209 pixels. Before using the
relevance criterion ($t$ is set as ten), it includes 2138 pixels. The number of samples in the Indian Pine data set used for the experiments is listed in Table I.

The Washington DC Mall data set is a Hyperspectral Digital Imagery Collection Experiment airborne hyperspectral image. This data set contains 1280 scan lines, and each line has 307 pixels. It includes 210 bands covering the 0.4–2.4-$\mu$m wavelength of the visible and IR spectra. Some water absorption channels were discarded in the data-processing procedure, and the remaining 191 channels were used in this paper. A total of 27282 pixels from nine classes were used in our experiments. Considering that the spectral variation of the roof classes is very significant, we divided the roofs into three subclasses, according to the spectral curves of the different materials. Fig. 9(a)–(c) shows the information about the Washington DC Mall data set, where 2867 pixels were utilized as the unlabeled samples. Before using the relevance criterion ($t$ is set as ten), it includes 3692 pixels. Fig. 9(d) shows the distribution of a part of the unlabeled samples, and Fig. 9(e) shows the midlevel-segmentation result of the data set. The number of samples in the Washington DC Mall data set used for the experiments is listed in Table II.

The third hyperspectral image data set was acquired by the airborne Reflective Optics System Imaging Spectrometer (ROSIS) at the urban test area of Pavia, northern Italy. The image scene, with a size of 610 × 340 pixels, is centered at the University of Pavia. After removing 12 bands due to noise and water absorption, it comprises 103 spectral channels. Nine ground-truth classes, with a total of 43 923 samples, were considered in the experiments. Fig. 10(a) shows a false-color composite of the image, while Fig. 10(b) shows the nine ground-truth classes of interest, which comprise urban features as well as soil and vegetation features. The number of samples in each class is shown in Table III. Moreover, 1967 pixels were utilized as the unlabeled samples. Before using the relevance criterion ($t$ is set as ten), it includes 2698 pixels.

<table>
<thead>
<tr>
<th>Classes</th>
<th>Samples</th>
<th>Classes</th>
<th>Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Road</td>
<td>5,139</td>
<td>Trails</td>
<td>2,401</td>
</tr>
<tr>
<td>Roof1</td>
<td>1,491</td>
<td>Trees</td>
<td>5,763</td>
</tr>
<tr>
<td>Roof2</td>
<td>2,069</td>
<td>Shadow</td>
<td>2,820</td>
</tr>
<tr>
<td>Roof3</td>
<td>1,553</td>
<td>Water</td>
<td>1,424</td>
</tr>
<tr>
<td>Grass</td>
<td>4,622</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

B. Parameter Analysis

The Indian Pine data set was used to assess the influence of parameters in our algorithm. According to our framework in Section II, there are four tuning parameters: $k$ and $\sigma$ are basic parameters in the LPP-based method, $\beta$ is a scaling factor in the discrimination term, parameter $c$ adjusts the similarity with spatial adjacency in the regularization term, $\gamma$ controls the weight of the discrimination term and regularization term, $t$ controls the number of the last unlabeled samples, and $d$ is the dimension of the reduced space.

Based on past experience from LPP, the choosing of $k$ is still an open question [29], [44], [45]. We experimentally set $k$ to the region [4, 14] and found an optimum value in this region to ensure that the classification overall accuracy was maximized. $\sigma$ is the parameter in the RBF kernel function which is used in both the discrimination and regularization terms. In order to facilitate the adjustment, we made the normalization in each local patch. We set $\sigma$ to the region $[0, 0.5]$. A larger value of $\sigma$ leads to a smoother similarity function and vice versa.

First of all, for simplicity, the parameters in the two terms were adjusted individually. The effect of the parameters is reflected by the classification accuracy rates, where our proposed DR method was combined with the KNN classifier, and the method used for comparison was SVM with an RBF kernel. There are two parameters associated with soft-margin SVM: $\gamma_{\text{SVM}}$ and $C$, where $\gamma_{\text{SVM}}$ is the parameter in the RBF kernel function and $C$ is the slack variable. The best free parameters were selected through threefold cross-validation. The soft-margin SVM with RBF kernel was performed using LIBSVM, a library for SVMs [24].

For the discrimination term, as shown in Fig. 11, when $k$ is small, the discriminative ability in the algorithm may not be well modeled. Therefore, the overall accuracy is lower, but a larger $\beta$ slightly improves the accuracy. With the increasing of $k$, more samples from different classes play a role in constructing the local patches, and the accuracy is also increased. However, when $k$ reaches 15, there is a peak in the accuracy curve. As $k$ continues to increase, the accuracy falls off quickly. In this case, many neighbors coming from different classes occupy the local patch, so the local geometry property may be ignored.

For the regularization term, Fig. 12 shows the effect of parameters $c$ and $\sigma$. SVM was used to test the performance. To narrow the range of $\sigma$, we normalized the samples in the local patch to $[0, 1]$. Clearly, when the value of $\sigma$ is larger than a critical point, the curve becomes stable. Parameter $c$ was used to adjust the value of $\sigma^{*}$. When $c > 1$, the neighborhood samples with spatial adjacency get a higher weight, which, as a result, leads to a better classification result. This beneficial effect can be clearly seen before the curve reaches the stable state. However, when $c$ continues to increase, the accuracy is unchanged or slightly increased. In contrast, if $c < 1$, the neighborhood samples with spatial adjacency get a lower weight, which leads to a worse classification result than before. According to these observations, we advise that the value of $c$ be selected within a limited range near one, i.e., one to twenty in our experiments, to get the optimum value.

The value of $\gamma$ is used to balance the contribution of the regularization term. When the value of $\gamma$ is zero, SDLLEA deactivates the regularization term and only utilizes the discrimination term. From the results in Fig. 13, the two curves have similar trends in that they rise quickly at first and then begin to descend, and the peak value always arises with an appropriate value of $\gamma$. In our experiments, the value of $\gamma$ can be selected between [0, 3] to find the optimum value.
Fig. 10. Details about the ROSIS University data set.

TABLE III
NUMBER OF SAMPLES IN THE ROSIS DATA SET
USED FOR THE EXPERIMENTS

<table>
<thead>
<tr>
<th>Classes</th>
<th>Samples</th>
<th>Classes</th>
<th>Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asphalt</td>
<td>6,631</td>
<td>Bare Soil</td>
<td>5029</td>
</tr>
<tr>
<td>Meadows</td>
<td>18,649</td>
<td>Bitmen</td>
<td>1330</td>
</tr>
<tr>
<td>Gravel</td>
<td>2,099</td>
<td>Bricks</td>
<td>3682</td>
</tr>
<tr>
<td>Trees</td>
<td>3064</td>
<td>Shadows</td>
<td>947</td>
</tr>
<tr>
<td>Metal_sheets</td>
<td>1345</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Fig. 11. Classification rates with respect to parameter $k$ and $\beta$.

For the other two data sets, the rough selection region was the same. More specifically, for the Washington data set, $k = 6$, $\beta = 0.8$, $c = 10$, $\sigma = 0.03$, and $\gamma = 0.6$, and for the ROSIS data set, $k = 6$, $\beta = 0.9$, $c = 10$, $\sigma = 0.04$, and $\gamma = 0.7$. For simplicity, the reduced dimension of the data set was set as $c - 1$ ($c$ is the number of classes).

C. Test of the Validity of the Discrimination Term in SDLEA

If we discard the regularization term, the SDLEA will be transformed to DLEA. Here, we test the validity of the discrimination term in SDLEA.

In order to observe the discriminative performance more intuitively, we chose three classes from the Indian Pine AVIRIS data set, with 80 samples in each class. They are corn-notill, soybeans-notill, and soybeans-clear. The corresponding spectral curves are plotted in Fig. 14. We compared our result with LPP, which is a classical unsupervised DR method. We also compared the result with other state-of-the-art supervised...
methods, including local discriminant analysis, LFDA, and DLA. For simplicity, all the algorithms were reduced to two (class-1) dimensions.

Fig. 15 shows the original data distribution on bands 27 and 60. Fig. 15(b)–(f) shows the data distribution after application of the different DR algorithms. LPP shows the poorest discriminative ability, for which a possible reason is that the assumption that neighborhood samples are likely to be the same class is untrue in this case. In addition, the neighborhood samples may come from different classes than the given one. Local discriminant analysis and Local Fisher Discriminant Analysis can separate the different classes well. LFDA can keep the multimodal structure in the data set, while LDA projects the data into subspaces, wherein a unimodal structure is imposed, which distorts the information contained in the multimodal distributions. Thus, it fails to produce a discriminant subspace in which the similarity between samples in the same class can still be accurately measured. DLA still presents the overlapping phenomenon, and the reason for this is that the same parameter is imposed on all the samples; as a result, the samples located in a low-separability patch may not be separated well. However, DLEA avoids this defect by dynamically selecting the parameter and can effectively separate different classes. Meanwhile, DLEA can preserve the rich structure in the same class, like LFDA, due to the introduction of the similarity measure between the samples in the same class.

D. Quantifying the Efficacy of the Unlabeled Samples Derived From the Oversegmented Result

We have shown that the classification accuracy of our method is improved significantly, compared with the methods that take no consideration of the regularization term (see Fig. 13). In this section, we further exploit the efficacy of the unlabeled samples derived from the over-segmented result. This section can be divided into two parts: 1) testing the effect of different ways of using the unlabeled samples in the unsupervised DR method and 2) testing the effect of the unlabeled samples as a regularization term in the semisupervised method. A VIRIS Indian Pines Data Set image data set is used in this section.

We compared the DR effect in two different ways of choosing the unlabeled samples: a segmentation method and random sampling. In order to ensure the stability of performance, the relevance was considered in both ways of selecting the samples. The mean of ten times random sampling is reported. Two unsupervised DR methods were used in this experiment: principal component analysis (PCA) and LPP with spatial adjacency (LPPS). Fig. 16 shows the effect of the two different ways of choosing unlabeled samples, by using different DR methods under different numbers of unlabeled samples. The number of unlabeled samples was determined by the segmentation algorithm with different parameters.

Fig. 16(a) shows the superiority of the utilization of segmentation results when choosing the unlabeled samples in the PCA subspace. This superiority is particularly obvious in the small sample data sets such as N229 and N458. As shown in Fig. 16(b), under a relatively small sample set, LPP shows poor accuracy, but with the increasing sample number in the sample set, LPP shows higher accuracy than PCA, which demonstrates that the result of LPP is sensitive to the sample size. Thus, the better performance of LPP once again shows the superiority of using segmentation for selecting the unlabeled samples, which is applied in our proposed SDLEA.

We also compared the effect of the unlabeled samples as a regularization term in our proposed semisupervised method. Optimized $\gamma$ for different methods in different numbers of the unlabeled samples was set according to the highest overall
accuracy in each trial. As with the previous experiment, the mean classification accuracies with these random samplings are reported in this experiment. As shown in Fig. 17, the quality of the unlabeled sample set directly influences the overall accuracy. This way of using the segmentation result to generate unlabeled samples can produce a better performance than the random sampling method.

**E. Quantifying the Efficacy of SDLEA-Based Dimensionality Reduction for Hyperspectral Image Classification and a Comparison With Other Regularization-Based Semisupervised Methods**

In order to investigate SDLEA-based dimensionality reduction, a group of comparative experiments designed to evaluate the performance with real hyperspectral images were conducted. First of all, as the traditional algorithms, LDA and NWFE were selected as the comparative algorithms. Meanwhile, LPP was also selected, whose idea is used for incorporating the local geometry in the regularization term. To ensure a fair comparison with our method, the unlabeled samples derived from the bottom-level segmentation results were used as the training samples for LPP. To investigate the efficiency of adding spatial adjacency, LPPS was used to compare with LPP. Furthermore, we compared two semisupervised algorithms: SDA and semisupervised discriminant local alignment (SDL). The regularization terms in these two algorithms were the same as for SDLEA.

Throughout our experiments, we used the 1-nearest neighbor and SVM classifier for all the DR methods. Since, in the case of small number of training samples, some algorithms, such as LDA, do not work well on a small sample set, we chose the classes with 60 samples, to guarantee that all the methods could work well, as shown in Table IV. The remainder of the data sets were divided into two parts: One is for cross-validation to train the optimal parameters, and the other was used for testing. Tables IV–VI detail the class-specific rates of all the methods for all the data sets. Figs. 18–21 show the classification maps of all the methods on all the data sets by SVM.

Several conclusions can be drawn from these results.

1) SDLEA performs better than the other methods and achieves the best classification.

2) LPPS performs better than LPP, and the best results are obtained in C3 (grass/pasture) and C4 (hay-windrowed) in the Indian Pine data set. The probable reason for this result is that spatial adjacency can be treated as another constraint added in the similarity measurement, which allows for a more accurate measure of the relationship between the samples.

3) After using unlabeled samples, SDA can effectively overcome the shortcomings of LDA and achieve a higher accuracy.

4) As a supervised traditional DR algorithm, NWFE is still comparable with some semisupervised algorithms such as SDA and SDL.

5) SDLEA always performs better than SDLA, which proves the efficiency of dynamically selecting interclass and intraclass neighborhood samples to construct the local patch.

To investigate the influence of the labeled samples, the three semisupervised algorithms (SDA, SDLA, and SDLEA) were used to test the sensitivity; note that the unlabeled samples are the same in all cases. Fig. 22 shows the testing errors with all three data sets. We can see that the proposed SDLEA method also performs well on a small sample set, has the smallest
V. CONCLUSION

This paper has proposed a novel semisupervised DR method for hyperspectral image classification. For the discrimination term, the distance between different classes is maximized according to the separability of pairwise samples in the local patch. For the regularization term, to further improve the performance of our method, in the process of selecting unlabeled samples, oversegmented regions are used to derive a more representative unlabeled set. Furthermore, multisegmentation results are employed to exploit the relationship between samples to improve the spectral kernel between pairwise samples. The proposed method was compared with other state-of-the-art methods on various hyperspectral data sets, with our proposed method presenting the highest classification accuracy.
TABLE VI
CLASS-SPECIFIC RATES IN PERCENTAGE FOR VARIOUS FEATURES IN THE ROSIS UNIVERSITY DATA SET

<table>
<thead>
<tr>
<th>DR + Classifier</th>
<th>Asphalt</th>
<th>Meadows</th>
<th>Gravel</th>
<th>Trees</th>
<th>Metal_sheets</th>
<th>Bare_soil</th>
<th>Bitumen</th>
<th>Bricks</th>
<th>Shadows</th>
<th>OA</th>
<th>(\kappa)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original SVM</td>
<td>79.85</td>
<td>84.62</td>
<td>83.80</td>
<td>87.59</td>
<td>99.02</td>
<td>76.07</td>
<td>75.76</td>
<td>66.28</td>
<td>100</td>
<td>81.74</td>
<td>0.7515</td>
</tr>
<tr>
<td>NN</td>
<td>82.12</td>
<td>86.20</td>
<td>64.67</td>
<td>89.23</td>
<td>98.14</td>
<td>58.08</td>
<td>70.08</td>
<td>59.76</td>
<td>94.73</td>
<td>79.87</td>
<td>0.7467</td>
</tr>
<tr>
<td>LDA SVM</td>
<td>84.27</td>
<td>86.26</td>
<td>72.40</td>
<td>89.38</td>
<td>98.95</td>
<td>76.99</td>
<td>83.02</td>
<td>71.90</td>
<td>100</td>
<td>84.18</td>
<td>0.7859</td>
</tr>
<tr>
<td>NN</td>
<td>74.01</td>
<td>79.86</td>
<td>61.32</td>
<td>92.01</td>
<td>99.55</td>
<td>49.48</td>
<td>46.29</td>
<td>51.10</td>
<td>83.42</td>
<td>73.64</td>
<td>0.6957</td>
</tr>
<tr>
<td>NWFE SVM</td>
<td>87.98</td>
<td>91.94</td>
<td>69.50</td>
<td>90.76</td>
<td>100</td>
<td>76.14</td>
<td>78.18</td>
<td>73.51</td>
<td>98.54</td>
<td>85.45</td>
<td>0.7676</td>
</tr>
<tr>
<td>NN</td>
<td>88.67</td>
<td>82.67</td>
<td>67.93</td>
<td>86.97</td>
<td>98.62</td>
<td>67.35</td>
<td>61.34</td>
<td>67.93</td>
<td>99.63</td>
<td>82.31</td>
<td>0.7996</td>
</tr>
<tr>
<td>LPP SVM</td>
<td>86.49</td>
<td>88.07</td>
<td>65.10</td>
<td>91.31</td>
<td>99.70</td>
<td>78.32</td>
<td>76.76</td>
<td>71.02</td>
<td>100</td>
<td>83.07</td>
<td>0.7990</td>
</tr>
<tr>
<td>NN</td>
<td>84.84</td>
<td>87.84</td>
<td>62.06</td>
<td>85.65</td>
<td>98.96</td>
<td>64.11</td>
<td>69.02</td>
<td>66.08</td>
<td>98.13</td>
<td>80.26</td>
<td>0.7734</td>
</tr>
<tr>
<td>LPPS SVM</td>
<td>74.19</td>
<td>84.14</td>
<td>73.83</td>
<td>84.20</td>
<td>99.70</td>
<td>70.20</td>
<td>70.00</td>
<td>71.19</td>
<td>100</td>
<td>84.33</td>
<td>0.8022</td>
</tr>
<tr>
<td>NN</td>
<td>87.84</td>
<td>86.82</td>
<td>63.99</td>
<td>87.18</td>
<td>99.63</td>
<td>62.83</td>
<td>64.07</td>
<td>67</td>
<td>98.57</td>
<td>81.68</td>
<td>0.7759</td>
</tr>
<tr>
<td>SDA SVM</td>
<td>87.50</td>
<td>89.06</td>
<td>66.75</td>
<td>91.97</td>
<td>99.92</td>
<td>78.16</td>
<td>86.09</td>
<td>73.54</td>
<td>100</td>
<td>85.19</td>
<td>0.8347</td>
</tr>
<tr>
<td>NN</td>
<td>73.59</td>
<td>82.47</td>
<td>63.59</td>
<td>93.37</td>
<td>99.62</td>
<td>60.35</td>
<td>48.91</td>
<td>60.63</td>
<td>86.76</td>
<td>79.28</td>
<td>0.7467</td>
</tr>
<tr>
<td>SDLA SVM</td>
<td>88.54</td>
<td>90.01</td>
<td>72.85</td>
<td>92.12</td>
<td>99.48</td>
<td>78.78</td>
<td>84.60</td>
<td>76.85</td>
<td>100</td>
<td>87.33</td>
<td>0.8403</td>
</tr>
<tr>
<td>NN</td>
<td>81.36</td>
<td>85.27</td>
<td>64.49</td>
<td>92.46</td>
<td>97.31</td>
<td>72.76</td>
<td>70.74</td>
<td>65.93</td>
<td>97.94</td>
<td>82.34</td>
<td>0.7931</td>
</tr>
<tr>
<td>SDLEA SVM</td>
<td>88.91</td>
<td>90.40</td>
<td>80.68</td>
<td>93.65</td>
<td>99.75</td>
<td>83.04</td>
<td>87.57</td>
<td>80.28</td>
<td>100</td>
<td>89.10</td>
<td>0.8605</td>
</tr>
<tr>
<td>NN</td>
<td>86.64</td>
<td>88.26</td>
<td>68.16</td>
<td>94.95</td>
<td>99.85</td>
<td>74.96</td>
<td>71.56</td>
<td>68.18</td>
<td>99.89</td>
<td>84.12</td>
<td>0.8226</td>
</tr>
</tbody>
</table>

Fig. 18. Classification maps for all the methods with the Indian Pine data set based on SVM. (a) Original. (b) LDA. (c) NWFE. (d) LPP. (d) LPPS. (d) SDA. (e) SDLA. (f) SDLEA.

Fig. 19 Classification maps for SDLEA with the Washington DC Mall data set based on SVM.
Fig. 20. Classification maps for all the methods with the Washington DC Mall data set based on SVM.

Fig. 21. Classification maps for all the methods with the ROSIS data set based on SVM.

Fig. 22. Testing errors with the three data sets. (a) Indian Pine data set. (b) Washington DC Mall data set. (c) ROSIS University data set.
ACKNOWLEDGMENT

The authors would like to thank Prof. Gamba from the Data Fusion Technical Committee of the IEEE Geoscience and Remote Sensing Society for providing the ROSIS data set and the handling editor and anonymous reviewers for their careful reading and helpful remarks.

REFERENCES

Liangpei Zhang (M’06–SM’08) received the B.S. degree in physics from Hunan Normal University, Changsha, China, in 1982, the M.S. degree in optics from Xi’an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi’an, China, in 1988, and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 1998.

He is currently the Head of the Remote Sensing Division, State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University. He is also a “Chang-Jiang Scholar” Chair Professor appointed by the Ministry of Education of China. He is currently a Principal Scientist for the China State Key Basic Research Project (2011–2016) appointed by the Ministry of National Science and Technology of China to lead the remote sensing program in China. He edits several conference proceedings, issues, and geoinformatics symposiums. He also serves as an Associate Editor of the International Journal of Ambient Computing and Intelligence, International Journal of Image and Graphics, International Journal of Digital Multimedia Broadcasting, Journal of Geo-spatial Information Science, and Journal of Remote Sensing. He has more than 260 research papers. He is the holder of five patents. His research interests include hyperspectral remote sensing, high-resolution remote sensing, image processing, and artificial intelligence.

Dr. Zhang is a Fellow of the Institution of Electrical Engineers, an Executive Member (Board of Governor) of the China National Committee of International Geosphere–Biosphere Programme, an Executive Member of the China Society of Image and Graphics, etc. He regularly serves as a Cochair of the series Society of Photo-Optical Instrumentation Engineers Conferences on Multispectral Image Processing and Pattern Recognition, Conference on Asia Remote Sensing, and many other conferences.

Bo Du (M’10) received the B.S. degree from Wuhan University, Wuhan, China, in 2005 and the Ph.D. degree in photogrammetry and remote sensing from the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, in 2010.

He is currently a Lecturer with the School of Computer, Wuhan University. His major research interests include pattern recognition, hyperspectral image processing, and signal processing.