Feature extraction for hyperspectral images based on semi-supervised local discriminant analysis

Wenzhi Liao, Aleksandra Pizurica, Wilfried Philips
Department of TELIN, iBBT
Ghent University
Ghent, Belgium
{wliao, sanja, philips}@telin.UGent.be

Youguo Pi
School of Automation Science and Engineering
South China University of Technology
Guangzhou, China
auygpf@scut.edu.cn

Abstract—We propose a novel semi-supervised local discriminant analysis (SELD) method for feature extraction in hyperspectral remote sensing imagery. The proposed method combines a supervised method (Linear Discriminant Analysis (LDA)) and an unsupervised method (Neighborhood Preserving Embedding (NPE)) without any free parameters. The underlying idea is to design optimal projection vectors, which can discover the global discriminant structure of the available labeled samples while preserving the local neighborhood spatial structure of the unlabeled samples. Furthermore, in our approach the number of extracted feature bands is no longer limited by the number of classes, which is a disadvantage of LDA. Experimental results demonstrate that the proposed method outperforms consistently other related semi-supervised methods and that it is also much more stable when the percentage of the labeled samples changes.

I. INTRODUCTION

Nowadays, it is possible to collect hyperspectral images with more than one hundred bands [1]. The complexity of hyperspectral image processing techniques usually depends on the number of spectral bands in the acquired data [2]. Therefore, it is necessary to find methods which can transform these high-dimensional data into a lower space with reduced dimensionality, while preserving as much information content as possible.

A number of approaches have been developed for feature extraction of hyperspectral images [2-4], ranging from unsupervised methods to supervised ones. Principle Component Analysis (PCA) has been widely used for hyperspectral images [2, 5], even though it fails to discover the class discriminant due to its unsupervised nature. A common supervised method is Linear Discriminant Analysis (LDA) [6]. When label information is sufficient, LDA can achieve significantly better performance than PCA for the classification task [7]. However, both PCA and LDA fail to discover the local spatial structure of the data. Comparisons in [2, 8], show that the local methods, which preserve the properties of spatial neighborhoods achieve a better classification performance. Moreover, when only a small number of labeled training samples is available (which is often the case in hyperspectral images), LDA (like all other supervised feature extraction methods) tends to perform poorly due to overfitting.

An attractive alternative is to modify the LDA so that it can use a small number of labeled samples and take local spatial information of the data into account. In this context, Semi-supervised Discriminant Analysis (SDA) [9] was proposed to regularize the LDA method by using a parameter to control the regularizer. The resulting method makes use of limited labeled samples to maximize the class discriminant and use both labeled and unlabeled samples to preserve the intrinsic geometric structure of the data. Semi-supervised local Fisher discriminant analysis (SELF) was proposed to linearly combine local Fisher discriminant analysis and PCA with a tunable parameter [10]. However, it is not easy to specify the optimal parameter values in these and similar semi-supervised techniques. Another semi-supervised method was recently proposed in [11].

In this paper, we propose a novel semi-supervised local discriminant analysis (SELD) that efficiently combines LDA and a local method Neighborhood Preserving Embedding (NPE) [12] without any tuning parameters. By making full use of the available labeled samples, the proposed method attempts at discovering the global discriminant structure of the data, while at the same time preserving local neighborhood spatial structure through the NPE. Another advantage is that our approach can extract as many feature bands as the number of dimensions. The main novelty of the proposed framework is how we combine the supervised and unsupervised methods. While we employ the NPE method [12], this novel framework can be applied in combination with other linear unsupervised methods too.

The organization of the paper is as follows. Section 2 provides a brief review of supervised and unsupervised methods. In Section 3, the proposed semi-supervised local discriminant analysis method is analyzed in detail. Section 4 presents the experimental results. Finally, the conclusions of the paper are drawn in section 5.

II. SUPERVISED METHOD AND UNSUPERVISED METHOD

Let \( \{x_i\}_{i=1}^N, x_i \in \mathbb{R}^d \) denote high-dimensional data, \( \{z_i\}_{i=1}^N \), and \( z_i \in \mathbb{R}^r \) the low-dimensional representations of the high-dimensional original data \( r \leq d \). In our application, \( d \) is the number of spectral bands of hyperspectral images, and \( r \) is the dimensionality of the feature space. The goal of feature extraction is to find a \( d \times r \) projection matrix \( W \), which can map every original data point \( x_i \) to \( z_i = W^T x_i \) such that most
information of the high-dimensional data is kept in a much lower dimensional feature space.

As a supervised method, LDA seeks directions on which the ratio of the between-class covariance to within-class covariance is maximized. The optimization problem of LDA is as follows:

$$w_{LDA} = \arg \max_w \frac{w^T S_b w}{w^T S_w w}$$  \hspace{1cm} (1)

$$S_b = \sum_{k=1}^{C} n_k (u^{(k)} - u)(u^{(k)} - u)^T$$  \hspace{1cm} (2)

$$S_w = \sum_{k=1}^{C} \left( \sum_{i=1}^{n_k} (x_i^{(k)} - u^{(k)})(x_i^{(k)} - u^{(k)})^T \right)$$  \hspace{1cm} (3)

where $n_k$ is the number of samples in the $k$th class, $u$ is the mean of the entire training samples, $u^{(k)}$ is the mean of the $k$th class, $x_i^{(k)}$ is the $i$th sample in the $k$th class. $S_b$ is called the between-class scatter matrix and $S_w$ the within-class scatter matrix.

NPE [12] is a linear approximation to the Locally Linear Embedding [8], which seeks a projection direction which preserves neighboring data structure in the low-dimensional feature space. The projection matrix of NPE can be optimized as follows:

$$w_{NPE} = \arg \max_w \frac{w^T XX^T w}{w^T XMX w}$$  \hspace{1cm} (4)

where $M = (I - Q)^T (I - Q)$, $I$ is identity matrix, $Q$ the reconstruction weights matrix [12], and $X$ is the original data.

The solution to (1) and (4) is equivalent to solving the following generalized eigenvalue problem:

$$S_w \lambda = S_b \lambda$$  \hspace{1cm} (5)

For LDA, $\bar{S} = S_b$ and $\bar{S} = S_w$. For NPE, $\bar{S} = XX^T$ and $\bar{S} = XMX^T$. The projection matrix $W = (w_1, w_2, \ldots, w_r)$ is made up by the $r$ eigenvectors of the matrix $\bar{S}^{-1} \bar{S}$ associated with the largest $r$ eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r$.

Fig.1 illustrates that LDA can preserve the class discriminant well but fails to preserve the spatial information of the data. Moreover, as the rank of the between-class scatter matrix $S_b$ is $C-1$, the LDA can extract at most $C-1$ features, which is not always sufficient to represent essential information of the original data. NPE method can preserve the spatial information well. However, NPE fails to discover the global discriminant structure in the original data; due to its unsupervised nature samples from different classes are mixed.

### III. SEMI-SUPERVISED LOCAL DISCRIMINANT ANALYSIS

Our approach combines supervised method (LDA) and unsupervised method (NPE) to make full use of labeled samples and unlabeled samples in a novel and efficient way.

Suppose a training data set $X = [X_{labeled}, X_{unlabeled}] = \{x_1, \ldots, x_n, x_{n+1}, \ldots, x_N\}$, with labeled set $X_{labeled} = \{(x_i, y_i)\}_{i=1}^{n}, y_i \in \{1, 2, \ldots, C\}$, $C$ is the number of classes, and unlabeled set $X_{unlabeled} = \{x_i\}_{i=n+1}^{N}$. The $k$th class has $n_k$ samples with $\sum_{k=1}^{C} n_k = n$. Without loss of generality, we center the data points by subtracting the mean vector from all the sample vectors, and assume that the labeled samples in $X_{labeled} = \{x_1, x_2, \ldots, x_n\}$ are ordered according to their labels, with data matrix of the $k$th class $X^{(k)} = [x_1^{(k)}, x_2^{(k)}, \ldots, x_{n_k}^{(k)}]$ where $x_i^{(k)}$ is the $i$th sample in the $k$th class. Then the labeled set can be expressed as $X_{labeled} = [X^{(1)}, X^{(2)}, \ldots, X^{(n)}]$. We have

$$S_b = \sum_{k=1}^{C} n_k (u^{(k)})(u^{(k)})^T$$  \hspace{1cm} (6)

By subtracting the between-class scatter $S_b$, $S_w$ can be obtained

$$S_w = X_{labeled}(X_{labeled})^T - X_{labeled}P_{n \times n}(X_{labeled})^T$$  \hspace{1cm} (7)

From the formulation of the LDA, we know that LDA employs only the labeled samples to estimate the projection matrix. When the number of labeled training samples is not sufficient, and this is the real situation in hyperspectral images,
the projection matrix may not be accurately estimated. In this situation, unlabeled training samples are used to estimate the local neighborhood spatial structure of the original data. Let
\[
P = \begin{bmatrix} P_{n \times n} & 0 \\ 0 & 0_{(N-n) \times (N-n)} \end{bmatrix}, \quad I = \begin{bmatrix} I_{n \times n} & 0 \\ 0 & 0_{(N-n) \times (N-n)} \end{bmatrix}
\]
\[
M = \begin{bmatrix} 0_{n \times n} & 0 \\ 0 & M_{(N-n) \times (N-n)} \end{bmatrix}, \quad I = \begin{bmatrix} 0_{n \times n} & 0 \\ 0 & I_{(N-n) \times (N-n)} \end{bmatrix}
\]
Now we can reformulate the generalized eigenvalue problems of the LDA and NPE respectively, as:
\[
XPX^T w = \lambda X(I - P)X^T w \quad (8)
\]
\[
XIX^T w = \lambda XMX^T w \quad (9)
\]
The main idea of our approach is to add the two formulations from (8) and (9) together, in this way, we get the generalized eigenvalue problem of the proposed SELD as:
\[
X(P + I)X^T w = \lambda X((I - P) + M)X^T w \quad (10)
\]
By fixing the matrix \( S \) and \( S \) as follows:
\[
S_{SELD} = X(P + I)X^T = [X_{labeled}, X_{unlabeled}] \cdot \cdots \begin{bmatrix} 0_{n \times n} & 0 \\ 0 & 0_{(N-n) \times (N-n)} \end{bmatrix} [X_{labeled}, X_{unlabeled}]^T
\]
\[
S_{SELD} = X((I - P) + M)X^T = [X_{labeled}, X_{unlabeled}] \cdot \cdots \begin{bmatrix} P_{n \times n} & 0 \\ 0 & M_{(N-n) \times (N-n)} \end{bmatrix} [X_{labeled}, X_{unlabeled}]^T
\]
the expression in (10) becomes:
\[
S_{SELD} w = \lambda S_{SELD} w \quad (11)
\]
By solving this generalized eigenvalue problem. We get the projection matrix \( W \). Features extracted by the proposed method can both preserve the class discriminant and local neighborhood spatial information, as shown in Fig.1. SELD magnifies the advantages of LDA and LLFE, and compensates for disadvantages of the two at the same time.

In many applications such as classification of hyperspectral images, the supervised method LDA sometimes confronts with the difficulty that the matrix \( S_w \) is singular. The fact is that sometimes the number of labeled training samples \( n \) is much smaller than the number of dimensions \( d \). In this situation, the rank of the \( S_w \) is at most \( n \) as clearly explains in (7), while the matrix \( X((I - P) + M)X^T \) in (8) is a \( d \times d \) matrix, which implies that the \( S_w \) matrix in LDA method is sometimes singular. Simultaneously, the between-class matrix \( S_b \) in LDA method uses only the labeled samples. The rank of \( S_b \) is \( C - 1 \) shown in (6), implying the LDA can extract at most \( C - 1 \) features. This may not be sufficient to represent essential information of the original data.

The proposed SELD method overcomes these problems, the matrices \( S_{SELD} \) and \( S_{SELD} \) in our approach are both symmetric and positive semi-definite, which makes sure that SELD can extract as much feature bands as the number of the bands and the corresponding eigenvalues are not negative.

IV. EXPERIMENTAL RESULTS

The dataset that we used in the experiments was collected with an airborne sensor system over the Washington DC Mall, with 1280×307 pixels and 210 spectral bands in the 0.4-2.4μm region. The DC Mall data set, which consists of 191 spectral bands after elimination of water absorption and noisy bands, is available at http://cobweb.ecn.purdue.edu/~biehl/, originally from Multispec®. The data set includes 7 land cover/use classes, with the number of labeled samples in the groundtruth roof (3834), street (416), path (175), grass (1928), trees (405), water (1224), and shadow (97). For the classification, we used support vector machines (SVMs) with codes available in [13], and 10% of the labeled samples per class selected from the groundtruth for training SVMs with a linear kernel. Then, the trained classifier is applied to the remaining 90% of the known ground pixels in the scene. All the parameters used in SDA and SELF are default as specified in the corresponding references [9, 10]. We keep the SVM classifiers the same in all experiments in order to fairly compare the performance of the feature extraction methods.

The experimental results for five methods are summarized in Fig.2-Fig.4 and Table 1. It can be seen that the features extracted by NPE are not influenced by the number of labeled samples (which is due to their unsupervised nature). The performance of the supervised method LDA is better when more labeled samples are used to train the projection matrix \( W \). When only 5 labeled samples per class from the groundtruth were used for feature extraction, the average classification accuracy (ACA) of LDA was less than 70%. When more than 40 samples per class were labeled, the ACA for LDA increased to nearly 80%, see Fig.3. One limitation of LDA is that the number of extracted feature bands depends on the number of classes. The unsupervised NPE method improves its performance by using more extracted feature bands (see Fig.4): with 6 extracted feature bands its OA is above 90%.
SDA [9] improved the performance of the LDA method by considering the intrinsic geometrical structure inferred from both labeled and unlabeled samples. SDA performs much better than LDA, and can extract more feature details as the number of the labeled samples increases. However, with relatively few labeled samples (25 labeled samples per class of groundtruth), SDA performs worse than NPE (see Table 1). Also, SDA has the same limitation as LDA in terms of the maximum number of extracted feature bands. SELF [10] performs poorly when labeled samples were limited.

The proposed SELD performs best even when very few labeled samples are available. We can see from Fig.3 that the overall classification accuracy (OCA) and ACA for our method are above 95% even with 5 labeled samples per class, which is much better than for other methods. Moreover, SELD can extract as much feature bands as the number of dimensions, and its performance is very stable when the number of labeled samples changes.

V. CONCLUSION

In this paper, we present a new semi-supervised method for hyperspectral feature extraction which combines LDA and NPE without any free parameters. When a small number of labeled samples is available, the performance of our method compared to other semi-supervised methods is not only consistently better but also much more stable as the number of labeled samples changes. Our future work will include development of a nonlinear version of this method.

ACKNOWLEDGMENT

This work was supported by the China Scholarship Council and FWO project “Spatial extension for classification of multispectral images”. Aleksandra Pizurica is a postdoctoral research of the Fund for Scientific Research in Flanders (FWO).

REFERENCES

[9] Deng Cai, Xiaofei He, Jiawei Han, “Semi-supervised discriminant analysis", IEEE 11th International Conference on Computer Vision 'ICCV07', 2007, pp. 1-7