Discriminative Supervised Neighborhood Preserving Embedding Feature Extraction for Hyperspectral-image Classification

Renbo Luo, Wenzhi Liao, and Youguo Pi
Key Laboratory of Autonomous Systems and Network Control, Ministry of Education College of Automation Science and Engineering, South China Univ. of Tech Wushan RD., Tianhe District, Guangzhou, P.R.China, 510641 e-mail: luo.rb@mail.scut.edu.cn

Abstract
A novel discriminative supervised neighborhood preserving embedding (DSNPE) method is proposed for feature extraction in classifying hyperspectral remote sensing imagery. DSNPE can preserve the local manifold structure and the neighborhood structure. What's more, for each data point, DSNPE aims at pulling the neighboring points with the same class label towards it as near as possible, while simultaneously pushing the neighboring points with different labels apart from it as far as possible. Experimental results on two real hyperspectral image datasets are reported to illustrate the performance of DSNPE and to compare it with a few competing methods.

Keywords: Classification, feature extraction, hyperspectral image

1. Introduction
With much more spectral bands, hyperspectral images contain much richer spectral information than regular RGB images and multi-spectral images. So it is widely used in the field of surface target identification, such as environmental monitoring, land crops analysis [1], weather analysis [2]. One of the major difficulties of hyperspectral data classification is usually that the number of training samples is limited and even not enough to estimate the parameters in the density model precisely and small training sets usually result in the Hughes phenomenon and singularity problem [3]. Feature extraction methods are developed to reduce the dimensionality of hyperspectral remote sensing image while keeping as much intrinsic information as possible, relatively few bands can represent most information of the hyperspectral data [4].

Principal component analysis (PCA) [5] is a popular unsupervised feature extraction method and has been widely used in hyperspectral images. It performs feature extraction through analyzing the covariance matrix of the original data. But as a unsupervised method, PCA does not utilize the class label information. Nonparametric weighted feature extraction (NWFE) [6] is a widely used supervised dimensionality reduction methods for hyperspectral image data. NWFE finds a projection matrix that maximizes the trace of the between-class scatter matrix and minimizes the trace of the within-class scatter matrix in the projected subspace simultaneously.

He et al. apply a linearization procedure to construct explicit maps over new measurements. Examples of this approach include locality preserving projections (LPP) [7], neighborhood preserving embedding (NPE) [8]. NPE mainly aims at preserving the local manifold structure, specifically, for each data point, it is represented as a linear combination of the neighboring data points and the combination coefficients are specified in the weight matrix. However, for classification problems, NPE ignores the differences between samples from the same classes and from the different classes.

In this paper, we propose a new algorithm, which we call Discriminative supervised neighborhood preserving embedding (DSNPE). DSNPE aims at preserving the local neighborhood structure on the data manifold. For each data point, it is represented as a linear combination of the neighboring data points from same class instead of whole neighboring data.
points and the combination coefficients are specified in the weight matrix. At the same time, it makes data points which have different labels as far as possible.

2. Neighborhood Preserving Embedding (NPE)

In this Section, we introduce a linear dimensionality reduction algorithm NPE. NPE preserves local manifold structure. For each data point, we find its K nearest neighbors. Let G denote a graph with m nodes. The i-th node corresponds to the data point \( x_i \). Put a directed edge from node \( i \) to \( j \) if \( x_j \) is among the K nearest neighbors of \( x_i \). Let \( W \) denote the weight matrix with \( W_{ij} \) having the weight of the edge from node \( i \) to node \( j \), and 0 if there is no such edge. The weights on the edges can be computed by minimizing the following objective function,

\[
\min \sum_i \left\| x_i - \sum_j W_{ij} x_j \right\|^2
\]

(1)

\[
\sum_j W_{ij} = 1, \ j = 1, 2, \ldots, m
\]

(2)

Solve the following generalized eigenvector problem:

\[
XMX^T a = \lambda X X^T a
\]

(3)

\[
X = (x_1, x_2 \cdots, x_m) \quad M = (I - W)^T (I - W) \quad I = \text{diag}(1, 1, \ldots, 1)
\]

It is easy to check that \( M \) is symmetric and semipositive definite. Let the column vectors \( a_1, a_2, \ldots, a_d \) be the solutions of equation (1), ordered according to their eigenvalues \( \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_d \). Thus, the embedding is as follows:

\[
x_i \rightarrow y_i = A^T x_i \quad A = (a_1, a_2, \ldots, a_d)
\]

where \( y_i \) is a d-dimensional vector, and \( A \) is an \( n \times d \) matrix.

3. DSNPE

Different from NPE, DSNPE assumes that each sample can be reconstructed using samples having the same label. That is, on the patch, \( x_i \) can be linearly reconstructed from \( x_{i1}, x_{i2}, \ldots, x_{il}, x_{i1}, x_{i2}, \ldots, x_{il} \) have same label as \( x_i \) and in its K nearest neighbors. As

\[
x_i = W_{i1} x_{i1} + W_{i2} x_{i2} + \cdots + W_{il} x_{il} + \epsilon_i
\]

(4)

where \( \epsilon_i \) is the reconstruction error. Minimizing the reconstruction error yields

\[
\min \sum_i \left\| \epsilon_i \right\|^2 = \min \sum_i \left\| x_i - \sum_j W_{ij} x_j \right\|^2
\]

(5)

\[
W_{ij} = \frac{\text{dist}(x_i, x_j)^{-1}}{\sum_k \text{dist}(x_i, x_k)^{-1}}
\]

(6)
where \( \text{dist}(a,b) \) denotes the Euclidean distance from \( a \) to \( b \). Weighting method \( W_j \) is used to emphasis the importance of \( x_i \)’s K nearest neighbors, if the distance between \( x_i \) and \( x_{ij} \) (\( x_{ij} \) is one of \( x_i \)’s K nearest neighbors) is small, then its weight will be close to 1; otherwise, will be close to 0. For the other points \( x_j, W_j = 0 \).

Consider the problem of mapping the original data points to a low dimensional subspace so that each data point can be represented as a linear combination of its neighbors with the coefficients \( W_j \). Let \( Y = (y_1, y_2, \ldots, y_m)^T \) be extracted features. A reasonable criterion for choosing a “good” map is to minimize the following cost function

\[
\min_i \sum \left\| x_j - \sum_j W_{ij} x_{ij} \right\|^2 \rightarrow \min_i \sum \left\| y_j - \sum_j W_{ij} y_{ij} \right\|^2
\]

(7)

Figure 1. Performance of each method

Figure 2. Performance of each method for AVIRIS Indian Pine (Ni=100) for KSC data set (Ni=100)

Meanwhile, we expect that neighborhood samples having different labels \( y_{i1}, y_{i2}, \ldots, y_{ic} \) \((c = k - l)\) are as far away as possible from sample \( y_i \). To achieve this, we maximize the sum of the distances between \( y_i \) and \( y_{i1}, y_{i2}, \ldots, y_{ic} \) so that we have

\[
\max \sum_i \sum_p \left\| y_i - y_p \right\|^2
\]

(8)

Combining the (7) and (8), the optimization problem can be resolved by converting them into the following Ratio problem:

\[
\min \frac{\sum \left\| y_j - \sum_j W_{ij} y_{ij} \right\|^2}{\sum \left\| y_j - \sum_j W_{ij} y_{ij} \right\|^2}
\]

(9)
Suppose the transformation is linear, that is, \( y^T = a^T X \), where the \( i \)-th column vector of \( X \) is \( x_i \), and assume that

\[
\frac{1}{2} \sum_{ij} (y_i - \sum_{j} W_{ij} y_{ij})^2 = \frac{1}{2} \sum_{ij} (y_i - \sum_{j} W_{ij} y_{ij})^2 = y^T (I - W)^T (I - W) y = y^T (D - B) y
\]

\[
= \frac{a^T X (I - W)^T (I - W) X a}{a^T X (D - B) X a} = \frac{a^T X M X a}{a^T X L X a}
\]

\[
B_{ij} = \begin{cases} 
1 & x_j \in N(x_i) \text{ and } c_i \neq c_j \\
0 & \text{otherwise}
\end{cases}
\]

where \( D \) is a diagonal matrix; its entries are column (or row) sum of \( B \), \( D = \sum j B_{ij} \), \( L = D - B \), \( c_i \) denotes the label of \( x_i \), \( N(x_i) \) denotes the \( k \) nearest neighbors of point \( x_i \).

We can now reformulate the objective functions of (9) as follows:

\[
a^* = \arg \min_{a} \frac{a^T X M X a}{a^T X L X a}
\]

Table 1. Highest accuracies (in percent) and corresponding kappa statistic (in percent) of the test data sets using k-NN classifier on Indian Pine and KSC data sets

<table>
<thead>
<tr>
<th>Data set</th>
<th>( N_i )</th>
<th>OA</th>
<th>PCA</th>
<th>NWFE</th>
<th>LPP</th>
<th>NPE</th>
<th>DSNPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indian Pine</td>
<td>40</td>
<td>OA</td>
<td>69.1(13)</td>
<td>73.8(11)</td>
<td>64.1(14)</td>
<td>70.4(15)</td>
<td>75.0(11)</td>
</tr>
<tr>
<td></td>
<td>k</td>
<td>65.5</td>
<td>70.4</td>
<td>60.2</td>
<td>67.0</td>
<td>71.8</td>
<td></td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>OA</td>
<td>76.2(16)</td>
<td>82.9(11)</td>
<td>76.0(18)</td>
<td>77.8(17)</td>
<td>84.3(11)</td>
</tr>
<tr>
<td></td>
<td>k</td>
<td>73.4</td>
<td>81.0</td>
<td>73.1</td>
<td>75.1</td>
<td>82.4</td>
<td></td>
</tr>
<tr>
<td>KSC</td>
<td>40</td>
<td>OA</td>
<td>80.5(17)</td>
<td>82.6(12)</td>
<td>68.0(16)</td>
<td>79.3(16)</td>
<td>83.9(15)</td>
</tr>
<tr>
<td></td>
<td>k</td>
<td>72.3</td>
<td>80.5</td>
<td>65.7</td>
<td>77.1</td>
<td>81.9</td>
<td></td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>OA</td>
<td>82.4(18)</td>
<td>87.4(12)</td>
<td>80.2(19)</td>
<td>84.3(16)</td>
<td>89.4(15)</td>
</tr>
<tr>
<td></td>
<td>k</td>
<td>80.5</td>
<td>86.0</td>
<td>78.5</td>
<td>83.1</td>
<td>88.1</td>
<td></td>
</tr>
</tbody>
</table>

The transformation vector \( a \) that minimizes the objective function is given by the minimum eigenvalue solution to the generalized eigenvalue problem:

\[
X M X a = \lambda X L X a
\]

The vectors \( a_i (i = 1,2,\ldots,d) \) that minimize the objective function are given by the minimum eigenvalue solutions to the generalized eigenvalue problem.

4. Experiment Results

In this paper, two real data sets, namely, Indian Pine and KSC are applied to explore the effects of different feature extraction methods. Other four feature extraction methods, namely PCA, NWFE, LPP and NPE, are included to evaluate the performance of the proposed DSNPE. The classification accuracies of the transformed test data by using different number of
features are computed. Two statistics, overall accuracy (OA) and Kappa coefficient ($k$) based on confusion matrix, are utilized to evaluate the classification performances. For the classifier, we used k-Nearest Neighbors classifier.

4.1. Hyperspectral Data Sets

The Indian Pine hyperspectral data set contains $145 \times 145$ pixels. Each pixel has 220 spectral bands, and the corresponding spatial resolution is approximately 20 m. The simulated infrared (IR) image is shown in Fig. 3(a). There are 16 different identified land-cover types in this region and 12 types are selected to do the experiments. For more details, the readers are referred to [9]. The Kennedy Space Center (KSC) data, acquired from an altitude of approximately 20 km, have a spatial resolution of 18 m and 224 bands. After removing water absorption and low-SNR bands, 176 bands were used for the analysis. For classification purposes, 13 classes representing the various land-cover types that occur in this environment were defined. For more information, visit http://www.csr.utexas.edu.

4.2. Result

In order to explore the influences of the training sample size on the feature-extraction methods, the number of training sample for each class ($N_i$) is set to be 40 and 100. The testing data is the remaining known ground pixels in the scene.

The experimental results are summarized in Figure 1, Figure 2 and Table 1. Figure 1 and Figure 2 indicate the performance of each method for Indian Pine, KSC respectively as the number of features increases when training sample is 100 in each class. The highest overall accuracies (OA) and the corresponding number of employed features (placed in parentheses), Kappa coefficient ($k$) are listed in Table 1. Particularly for Indian Pine data set, the corresponding classified maps associated with DSNPE and other methods are shown in Figure 3(b). Experiment results show that the DSNPE-related classification results have the highest accuracies in all the considered situations. Comparing with NPE, DSNPE considers not only the intraclass geometry but also the discriminative information derived from the interclass samples.

With the increasing number of features, the classification overall accuracies of NWFE-related are going up at first, but begin to decline after reach to the highest point. DSNPE approach also achieves better performance than PCA and LPP methods. This shows that our proposed method is able to extract more discriminative features than the both methods.

![Figure 3](image-url)

Figure 3. (a) Simulated grayscale IR of Indian Pine; (b) Classification maps for Indian Pine using k-NN classifier ($N_i=40$).
5. Conclusion

In this paper, we proposed a discriminative supervised neighborhood preserving embedding for the classification of land-cover types in hyperspectral images. By introducing prior class label information during neighbourhood selection at training stage, the proposed method can make data points which have different labels as far as possible and which have same labels as near as possible. The experimental results demonstrated the effectiveness of the proposed algorithm compared with representative dimensionality reduction algorithms, e.g., PCA, NWFE, LPP and NPE.

References


