Hyperspectral Image Classification based on Co-training

Zhijun Zheng¹ and Yanbin Peng¹

¹School of Information and Electronic Engineering, Zhejiang University of Science and Technology, Hangzhou 310023, China
zjzheng9999@163.com

Abstract

The abundant information available in hyperspectral image has provided important opportunities for land-cover classification and recognition. However, “Curse of dimensionality” and small training sample set are two difficulties which hinder the improvement of computational efficiency and classification precision. In this paper, we present a co-training based method on hyperspectral image classification. Firstly, two views of samples are generated through two kinds of dimensionality reduction methods. After that, the co-training process is viewed as combinative label propagation over two independent views. Experimental results on real hyperspectral image show that the proposed method has better performance than the other state-of-the-art methods.

Keywords: hyperspectral image classification, dimensionality reduction, Feature selection, Feature abstraction, co-training

1. Introduction

Hyperspectral sensors simultaneously collect hundreds of continuous spectral bands with the resolution of the nanoscale. Such a large number of spectral bands contain rich information, providing significant challenges to hyperspectral image classification. In general, the dimensionality of samples strongly affects the performance of many classification methods. For hyperspectral image, on the one hand, there are redundant bands, and some bands maybe has weak discriminative ability. On the other hand, as the number of dimension improved, the number of training sample points needed increase exponentially, which is so called “Curse of dimensionality” problem. In the meanwhile, due to the cost of labeling sample points, the number of training sample points is few. It is therefore advantageous to decrease the dimension of hyperspectral data and import more efficient classification method.

Nowadays, a considerable amount of research has been done on hyperspectral image classification using machine learning algorithms during the past decade. Chen [1] propose a novel dictionary learning method for hyperspectral image classification. Its proposed method, linear regression fisher discrimination dictionary learning, obtains a more discriminative dictionary and a classifier by incorporating linear regression term and the fisher discrimination into the objective function during training, which improves the performances of classification. Li [2] constructs a new family of generalized composite kernels which exhibit great flexibility when combining the spectral and the spatial information contained in hyperspectral data. And then propose a multinomial logistic regression classifier for hyperspectral image classification. Ji [3] propose a new method to address both the pixel spectral and spatial constraints, in which the relationship among pixels is formulated in a hypergraph structure. A spatial-based hyperedge is generated to model the layout among pixels. Both the learning on the combinational hypergraph
is conducted by jointly investigating the image feature and the spatial layout of pixels to seek their joint optimal partitions. Liu [4] present a post processing algorithm for a kernel sparse representation based hyperspectral image classifier, which is based on the integration of spatial and spectral information. Huang [5] propose a new sparse manifold learning method, called sparse manifold preserving for hyperspectral image classification. This method constructs the affinity weight using the sparse coefficients which reserves the global sparsity and manifold structure of hyperspectral image data, which it doesn’t need to choose any model parameters for the similarity graph. Qian [6] propose a hyperspectral feature extraction and pixel classification method based on structured sparse logistic regression and three-dimensional discrete wavelet transform texture features. However, there are still some spaces to improve precision of hyperspectral image classification.

In this paper, based on our former research works [7-10], we present a co-training based method on hyperspectral image classification. Firstly, two views of samples are generated through two kinds of dimensionality reduction methods. The research on the dimensionality reduction can be divided into two main directions. One is feature selection method, which seeks a subset of original feature space which contains most of the discriminative abilities. This paper adopts spectral clustering [11-16] based band selection method. The other is feature abstraction method, which transforms the original sample points into the destination feature space which has less number of features while keep the discriminative abilities. This paper adopts linear discriminant analysis (LDA for short) and manifold learning based feature abstraction method. After that, the co-training process is viewed as combinative label propagation over two independent views. Experimental results on real hyperspectral image show that the proposed method has better performance than the other state-of-the-art methods.

2. Dimensionality Reduction Methods

The whole hyperspectral image cube $H$ is shown in Figure 1. I, J and K corresponds to three dimensions of the data cube. I and J stand for width and length dimension and K stands for the spectral dimension. One band image is expressed as $H_k$, which is a data matrix with $I \times J$ dimensions. One sample point is a pixel vector $x_i$ with $K$ dimensions.
image increases the precision to classify and recognize the land-cover materials. However, the number of labeled sample points is few due to high labeling cost. At the same time, high dimensionality of hyperspectral image improves computational complexity. What’s worse is there are redundancy bands which may decrease classification precision. Therefore, it is an important issue to reduce the dimensionality without degrading the classification precision. Dimensionality reduction can reduce computational complexity and improve classification precision. There are two major kinds of dimensionality reduction methods: feature selection and feature abstraction. The former method seeks a subset of original feature space which contains most of the discriminative abilities. This paper adopts spectral clustering based band selection method. The latter transforms the original sample points into the destination feature space which has less number of features while keep the discriminative abilities. This paper adopts LDA and manifold learning based feature abstraction method. These two methods are detailed in the following two subsections.

2.1. First View: Spectral Clustering based Band Selection Method

Spectral clustering based band selection method is an unsupervised feature selection technique. In practice, unsupervised band selection technique can be considered as a clustering problem, which is a process of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups. In clustering based band selection method, each band is regarded as an object. The similarity between two bands is measured using a kind of measurement, based on which the bands are grouped into several clusters. The dimensionality reduction can be done by selecting a representative band from every band clusters. This paper adopts mutual information to measure similarity between bands, and spectral clustering algorithm to cluster band data.

Suppose we have similarity matrix \( W = \{w_{ij}\} \) of band data, wherein, \( w_{ij} \in \mathbb{R}^{L \times L} \) stands for similarity degree between band \( i \) and band \( j \). This paper adopts mutual information [17, 18] based method to calculate similarity degree. We assume that \( q(c_i) \) and \( q(c_j) \) are the probability distributions of spectral band \( i \) and \( j \) respectively. \( q(c_i, c_j) \) represents the joint probability distribution. The mutual information between band \( i \) and band \( j \) is defined as:

\[
MI(band_i, band_j) = \sum_{c_i=1}^{c} \sum_{c_j=1}^{c} (q(c_i, c_j)) \log \frac{q(c_i, c_j)}{q(c_i)q(c_j)}. 
\]

Based on the former definition, spectral clustering based band selection algorithm is as follows [19, 20]:

Algorithm 1: spectral clustering based band selection algorithm

Inputs: bands vector \( \text{band}_1, \text{band}_2, \ldots, \text{band}_m \); number \( k \) of clusters to construct

Outputs: \( k \) representative bands

1) Compute probability distribution of all bands: \( q(c_1), q(c_2), \ldots, q(c_n) \).

2) Compute similarity degree between bands using mutual information formula, obtaining similarity matrix \( W \).

3) Compute the unnormalized Laplacian matrix \( L \) of \( W \).

4) Compute the first \( m \) eigenvectors \( v_1, v_2, \ldots, v_m \) of \( L \).

5) Let \( V \in \mathbb{R}^{n \times m} \) be the matrix containing the vectors \( v_1, v_2, \ldots, v_m \) as columns.

6) For \( i = 1, 2, \ldots, n \), let \( x_i \in \mathbb{R}^m \) be the vector corresponding to the \( i \)-th row of \( V \).

7) Cluster the points \( x_1, x_2, \ldots, x_n \) with the \( k \)-means algorithm into \( k \) clusters: \( C_1, C_2, \ldots, C_k \).

8) Select a band from every cluster randomly, obtaining \( k \) representative bands.
2.2. Second View: LDA and Manifold Learning based Feature Abstraction Method

In hyperspectral image classification, unlabeled training samples are readily available but labeled ones are fairly expensive to obtain. Therefore, it is an interesting idea to extract features with the help of unlabeled data, which belongs to semi-supervised learning category. This section combines LDA and manifold learning to extract features through an optimization procedure. Linear discriminant analysis is a well-known dimensionality reduction method whose projection matrix is obtained by minimizing the within class covariance and at the same time maximizing the between class covariance \[2.1\]. LDA is a supervised method, whose performance depends on the number of labeled training sample points. Therefore, manifold learning is imported to improve the performance by make full use of the unlabeled data \[2.2\].

Given \(n_1\) labeled sample points \(\{(x_i, y_i)\}_{i=1}^{n_1}\) and \(n_2\) unlabeled sample points \(\{(x_j)\}_{j=1}^{n_2}\). Wherein, \(x_i \in \mathbb{R}^k\) is original feature vector of pixel, \(y_i\) is class label of pixel. Suppose \(w\) is the projection matrix which transforms original feature vector into new feature space with lower number of features. We obtain \(w\) by minimizing the objective function: 
\[ \text{arg min}_{w} \{ \alpha M(w) + (1 - \alpha) L(w) \} \]. Wherein, \(\alpha \in [0, 1]\) is weight coefficient. \(M(w)\) stands for manifold learning regularization, which incurs a heavy penalty if two neighboring points in the original space lie far away in the output space. The corresponding weight matrix is as follows:
\[ M_{ij} = \begin{cases} 1 & \text{if } x_i \in \text{knn}(x_j) \text{ or } x_j \in \text{knn}(x_i) \\ 0 & \text{otherwise} \end{cases} \]

Wherein, kNN() stands for k nearest neighbor algorithm. To apply kNN algorithm, we should construct a similarity degree calculating method to measure how well sample \(x_i\) is similar to sample \(x_j\). This section use mutual information based method to calculate similarity degree. Firstly, we assume that \(p(c_i)\) and \(p(c_j)\) represent the probability distributions of sample points \(x_i\) and \(x_j\) respectively. \(p(c_i, c_j)\) represents the joint probability distribution. The mutual information is defined as:
\[ MI(x_i, x_j) = \sum_{c_i} \sum_{c_j} (p(c_i, c_j) \log \frac{p(c_i, c_j)}{p(c_i)p(c_j)}) \]

Therefore, the corresponding manifold regularization item is as follows:
\[ M(w) = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} ||w^T x_i - w^T x_j||^2 M_{ij} = Y^T (D_{MM} - M M) Y = Y^T L_{MM} Y \]

To integrate LDA and manifold learning into a unified framework, we adjust the objective function of LDA according to the objective function of manifold learning. Within-class weight matrix is defined as follows:
\[ W_{ij} = \begin{cases} 1 & \text{if both point } x_i \text{ and } x_j \text{ have the same label} \\ 0 & \text{otherwise} \end{cases} \]

Between-class weight matrix is defined as follows:
\[ B_{E_{ij}} = \begin{cases} 1 & \text{if point } x_i \text{ and } x_j \text{ are different labeled} \\ 0 & \text{otherwise} \end{cases} \]

LDA tries to minimize the within class covariance to keep the compactness in class, therefore, the following objective function should be minimized:
\[ T_{uv} = \sum_{j=1}^{n_u} \sum_{j'=1}^{n_v} \| w^T x_j - w^T x_{j'} \|^2 W_{ij} \]

Meanwhile, LDA also tries to maximize the between class covariance to keep the separability between classes, therefore, the following objective function should be maximized:

\[ T_{uv} = \sum_{j=1}^{n_u} \sum_{j'=1}^{n_v} \| w^T x_j - w^T x_{j'} \|^2 B_{ij} \]

Therefore, the LDA objective function is defined as follows:

\[ L(w) = \beta T_{uv} - (1 - \beta) T_{ws} = Y^T (\beta L_{us} - (1 - \beta) L_{sd}) Y \]

Wherein, \( \beta \in [0,1] \) is weight coefficient.

3. Co-training based Hyperspectral Image Classification Method

In order to solve the problem of small labeled training sample set, based on the two views obtained in Section 2, this paper enlarges the scale of labeled training sample points through importing co-training technique, thereby improves the classification precision of hyperspectral image recognition.

Co-training algorithm employs two classifiers, each of which labels the unlabeled sample point for the other during the learning process. Suppose we have labeled training sample set \( L = \{(V_1, T_1, p_1), (V_2, T_2, p_2), \ldots, (V_\ell, T_\ell, p_\ell)\} \), wherein \( V_i \) is feature 1 of ith instance, \( T_i \) is the feature 2 of ith instance, \( p_i \) is class label of ith instance, \( |L| \) is the number of labeled samples. Unlabeled training sample set is \( U = \{(V_1, T_1), (V_2, T_2), \ldots, (V_U, T_U)\} \), wherein \( |U| \) is the number of unlabeled samples. This paper use two support vector machine (SVM for short) to train and predict instances expressed as feature 1 and feature 2 respectively. Two SVMs provide reliable new labeled sample points to each other iteratively. Eventually, they joint predict the class label of pixels on the enlarged labeled training sample set. In order to choose appropriate unlabeled sample points to label, the distance of sample point to the classification hyperplane is used to compute the labeling confidence. The co-training based hyperspectral image classification algorithm is as follows:

Algorithm 2. Co-training based hyperspectral image classification

Inputs: labeled sample set \( L \), unlabeled sample set \( U \),

Maximum number of iterations \( T \)

Outputs: classifier \( g_1(x) \) and \( g_2(x) \)

\[ L_1 = L; \quad L_2 = L; \quad U_1 = U; \quad U_2 = U; \]

\( g_1 \leftarrow SVM.Train(L_1, feature_1); //SVM trained in feature 1, returns classifier g_1(x) \)

\( g_2 \leftarrow SVM.Train(L_2, feature_2); //SVM trained in feature 2, returns classifier g_2(x) \)

Repeat for \( T \) rounds // Loop exit criteria (1)

For \( k \in \{1, 2\} \) do

\[ x^* = \arg \max_{x \in U_k} (g_k(x)) \] //find the most confident unlabeled sample point

If \( (|g_k(x^*)| > 1) \) //the confidence must be larger than 1

\[ y^* = \text{sgn}(g_k(x^*)) \]

\[ U_{(k-2)\%2+1} = U_{(k-2)\%2+1} - \{(x^*)\} \] //delete from the other’s unlabeled sample set

\[ L_{(k-2)\%2+1} = L_{(k-2)\%2+1} \cup \{(x^*, y^*)\} \] //add to the other’s labeled sample set

End if

End for

Copyright © 2015 SERSC
If neither \( L_1 \) or \( L_2 \) changes // Loop exit criteria (2)

   Exit

   Else

   \[ g_1 \leftarrow SVM\text{.Train}(L_1, feature_1); \]

   \[ g_2 \leftarrow SVM\text{.Train}(L_2, feature_2); \]

   End if

End repeat

4. Experiments and Results

Having presented the method of co-training based hyperspectral image classification (CHSIC for short) in the previous section; we now demonstrate the effect of our new method through several comparative experiments. These experiments are done in a real-world hyperspectral image data set which was acquired by AVIRIS sensor over the Kennedy Space Center (KSC for short), Florida, on March 23, 1996 [25]. KSC data set have original 224 bands, among which 48 bands are consider as water absorption and low SNR bands. These noise bands are removed from original data in this experiment. For classification purpose, 13 classes representing the various land cover materials were defined in this environment. The class number, class name, the number of training and testing samples of KSC data is listed in Table 1. Figure 2 shows the 31\textsuperscript{th} band image of KSC data set.

![Figure 2. Image of 31\textsuperscript{th} Band in KSC Data Set](image)

In order to assess the performance of the new method proposed in this paper, we choose two models for comparison: 1) ID algorithm proposed in literature [26] with SVM classifier; 2) MVPCA algorithm proposed in literature [27] with SVM classifier.

<table>
<thead>
<tr>
<th>Class number</th>
<th>Class name</th>
<th>Train</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Scrub</td>
<td>40</td>
<td>400</td>
</tr>
<tr>
<td>2</td>
<td>Willow swamp</td>
<td>20</td>
<td>200</td>
</tr>
<tr>
<td>3</td>
<td>Cabbage palm hammock</td>
<td>20</td>
<td>200</td>
</tr>
<tr>
<td></td>
<td>Classification</td>
<td>Number</td>
<td>Percentage</td>
</tr>
<tr>
<td>---</td>
<td>-------------------------</td>
<td>--------</td>
<td>------------</td>
</tr>
<tr>
<td>4</td>
<td>Cabbage palm/oak hammock</td>
<td>20</td>
<td>200</td>
</tr>
<tr>
<td>5</td>
<td>Slash pine</td>
<td>15</td>
<td>120</td>
</tr>
<tr>
<td>6</td>
<td>Oak/broadleaf hammock</td>
<td>20</td>
<td>200</td>
</tr>
<tr>
<td>7</td>
<td>Hardwood swamp</td>
<td>12</td>
<td>60</td>
</tr>
<tr>
<td>8</td>
<td>Graminoid marsh</td>
<td>30</td>
<td>300</td>
</tr>
<tr>
<td>9</td>
<td>Spartina marsh</td>
<td>35</td>
<td>300</td>
</tr>
<tr>
<td>10</td>
<td>Cattail marsh</td>
<td>30</td>
<td>300</td>
</tr>
<tr>
<td>11</td>
<td>Salt marsh</td>
<td>30</td>
<td>300</td>
</tr>
<tr>
<td>12</td>
<td>Mud flats</td>
<td>35</td>
<td>300</td>
</tr>
<tr>
<td>13</td>
<td>Water</td>
<td>40</td>
<td>400</td>
</tr>
</tbody>
</table>

Figure 3 displays the classification precision of algorithm ID, MVPCA and CHSIC. As is shown in Figure 3, the classification precision improves along with the increase number of features. Among them, CHSIC algorithm achieves the best performance.

![Figure 3. Comparison of Classification Precision](chart)

5. Conclusions

In this paper, we present a co-training based hyperspectral image classification method. In order to apply co-training algorithm, two independent views of original data were obtained through dimensionality reduction process. After that, two SVM classifiers iteratively label the unlabeled sample point for the other during the learning process. Based on the enlarged training sample set, classification precision is improved. A comparison with two popular hyperspectral image classification methods ID and MVPCA is conducted. SVM classifiers are utilized to evaluate the performance of relevance algorithms. The statistical experimental results on KSC data show that the proposed method has better performance than the other state-of-the-art methods.
Acknowledgments

Project supported by the Zhejiang Provincial Natural Science Foundation of China (No. LQ13F020015).

References


Authors

Zhijun Zheng, received his Ph.D. degree from Xi’an Jiaotong University. He has been associate professor at Zhejiang University of Science and Technology. His research interests include machine learning, computer architecture, software engineering, multimedia analysis retrieval, computer animation, image retrieval and statistical learning.

Yanbin Peng, received his Ph.D. degree from the College of Computer Science and Technology, Zhejiang University. He has been associate professor at Zhejiang University of Science and Technology. His research interests include artificial intelligence, machine learning, data mining, pattern recognition and image retrieval.