Computing and Informatics, Vol. 36, 2017, 205–222, doi: 10.4149/cai_2017_1_205

KERNEL FEATURE EXTRACTION FOR HYPERSPECTRAL IMAGE CLASSIFICATION USING CHUNKLET CONSTRAINTS

Haishi Zhao, Laijun Lu, Chen Yang*

College of Earth Science, Jilin University
Changchun, 130061, China
e-mail: yangc616@jlu.edu.cn

Renchun Guan

College of Computer Science and Technology, Jilin University
Changchun, 130012, China
e-mail: guanrenchu@jlu.edu.cn

Abstract. A novel semi-supervised kernel feature extraction algorithm to combine an efficient metric learning method, i.e. relevant component analysis (RCA), and kernel trick is presented for hyperspectral imagery land-cover classification. This method obtains projection of the input data by learning an optimal nonlinear transformation via a chunklet constraints-based FDA criterion, and called chunklet-based kernel relevant component analysis (CKRCA). The proposed method is appealing as it constructs the kernel very intuitively for the RCA method and does not require any labeled information. The effectiveness of the proposed CKRCA is successfully illustrated in hyperspectral remote sensing image classification. Experimental results demonstrate that the proposed method can greatly improve the classification accuracy compared with traditional linear and conventional kernel-based methods.

Keywords: Feature extraction, hyperspectral remote sensing image, chunklet constraints, kernel method

* corresponding author

1 INTRODUCTION

Dimensionality reduction is a widely used technique before further information processing and plays an increasingly important role in analysis of high-dimensional data. Hyperspectral remote sensing is a technology that can collect and process detailed spectral and spatial information from across the electromagnetic spectrum [1]. Comparing with the general image, the hyperspectral imagery divides the spectrum into dozens or hundreds of narrow and adjacent spectral bands, which results in a great number of bands, high correlation between neighbor bands, and a lot of redundancy information. High-dimensional hyperspectral data also makes the sample points sparse in a data space. Unlike general image analysis, the basic processing unit of hyperspectral image classification is high-dimensional spectral vector for categorizing each spectral pixel to one of the classes. Therefore, it often encounters the problem of insufficient sample labels because in traditional classification methods they are not suitable to be directly introduced in hyperspectral data. On the other hand, high dimensionality hyperspectral data will greatly increase the computation time. In order to overcome the curse of dimensionality of hyperspectral data [2, 3], it is necessary to transform the high dimension hyperspectral space to a low dimension subspace, i.e. dimension reduction before conventional analysis.

Generally, dimension reduction of hyperspectral remote sensing image can be divided into two areas: feature extraction and band selection (i.e. feature selection) [4]. In this paper, we focus on feature extraction. According to whether training samples are exploited, feature extraction can be divided into three categories: unsupervised, supervised and semi-supervised methods. In general, unsupervised methods do not use prior information and cannot get satisfactory results. Principal component analysis (PCA) [5] is one of the most common and representative unsupervised feature extraction method. The PCA gets a transformation matrix composed by chosen eigenvectors corresponding to larger eigenvalues of data covariance matrix on the original data to achieve the purpose of dimension reduction. Supervised methods identify the subset of original spectral bands based on the class separability transformation of labeled samples. Fisher discriminant analysis (FDA) [6] is a well-known supervised feature extraction technique. The FDA can learn a linear transformation matrix by means of maximizing covariance between classes and minimizing covariance within-class to reduce the dimension of original data. However, in real application, the collection of labeled samples in hyperspectral imagery will spend a lot of manpower and material resources. As a result, some semi-supervised methods [7, 8, 9] fall between the supervised and unsupervised learning, and has been applied to hyperspectral image dimension reduction. Compared with the unsupervised and supervised methods, semi-supervised methods make use of small amounts of prior information which are more accessible.

On the other hand, pairwise constraints, i.e., positive (must-link) and negative (cannot-link) constraints as another common types of prior knowledge can be de-

rived in an easier way and can be automatically collected by photo interpretation. Recently, a novel efficient and non-iterative metric learning method, i.e. relevant component analysis (RCA) [10, 11, 12], has been developed for learning a Mahalanobis metric in a semi-supervised fashion. The RCA utilizes prior information expressed as equivalence constraints which indicate positive relationship between data instances, but without knowing their labels. According to the given positive constraints, a new kind of prior information, i.e., chunklet constraints can be formed. The basic idea of RCA is to learn an optimal data transformation that leads to the optimal distance metric by minimizing the total variance of data instances within the same chunklets. Recently, many semi-supervised algorithms based on Chunklet information have been developed with good results in image retrieval [13], band selection [14], and face recognition [12].

However, the above-mentioned techniques are limited to learn linear transformations and cannot exploit the intrinsic nonlinear properties of the hyperspectral imagery [15]. Therefore, some non-linear methods such as neural networks (NN) [16] and kernel-based algorithms [17] have been introduced to remote sensing data analysis. Kernel methods provide a more powerful and unified framework and can be easily combined with the linear methods. Popular kernel-based feature extraction algorithms include unsupervised methods, such as kernel principal component analysis (KPCA) [18, 19] and kernel independent component analysis (KICA) [20], as well as supervised methods, such as kernel Fisher discriminant component analysis (KFDA) [21], which are the kernel version of linear methods, i.e. the PCA, ICA, and FDA, respectively.

In recent work, a kernel version of the RCA, i.e. kernel relevant component analysis (KRCA) [22], which can produce nonlinear transformation for learning distance metrics of structural objects, has proved to be effective for retrieval and clustering tasks. In this paper, we present a semi-supervised kernel feature extraction method based on the RCA, called chunklet-based kernel relevant component analysis (CKRCA). Different from [22], the goal of proposed CKRCA is to learn the optimal nonlinear transformation by the kernel trick for projecting a nonlinear feature subspace via a chunklet constraints-based FDA criterion. The feature subspace extracted with this method can be used as input for hyperspectral classification. The readily available prior information (chunklet constraints) and nonlinearity of chunklet-based KRCA are the key motivation behind studying its benefits in the hyperspectral analysis. The experiments are carried out on the real hyperspectral imagery. The comparisons between the classical linear and nonlinear methods are conducted to demonstrate that the chunklet-based KRCA improves results with the combination of chunklet constraints and kernel method.

The remaining parts of this paper are organized as follows. Section 2 reviews standard formulation of the RCA and kernel method. The proposed chunklet-based KRCA feature extraction method is detailed in Section 3. Section 4 presents the experiments on real hyperspectral imagery. Finally, Section 5 concludes this paper.

2 RELATED WORK

2.1 Relevant Component Analysis (RCA)

The RCA [10, 11, 12] method is a simple and efficient algorithm for learning a Mahalanobis metric. The RCA changes the feature space for data representation by learning a whitening transformation matrix using a new prior information, i.e. chunklet constraints (small subsets of data points that are known to belong to the same but unknown class).

Let $X = \{x_1, x_2, \ldots, x_N\} \subset R^{N \times M}$ be a data set where M is the number of features and each feature contains N data points. Assume K chunklets can be generated according to the given positive constraints, the k^{th} chunklet is termed as $C_k = \left\{x_1^{(k)}, x_2^{(k)}, \ldots, x_{n_k}^{(k)}\right\}$ where n_k is the number of data points in the k^{th} chunklet.

In order to perform the RCA, an approximation of the covariance matrix can be calculated using chunklet constraints. The estimated covariance matrix, i.e. within chunklet covariance matrix, is defined as

$$\hat{C}_c = \frac{1}{N_c} \sum_{k=1}^{K} \sum_{i=1}^{n_k} \left(x_i^{(k)} - d_k\right) \left(x_i^{(k)} - d_k\right)^T \tag{1}$$

where $x_i^{(k)}$ denotes the i^{th} data point of the k^{th} chunklet; d_k is mean of the k^{th} chunklet, $d_k = \frac{1}{n_k} \sum_{i=1}^{n_k} x_i^{(k)}$. $N_c = \sum_{k=1}^{K} \mid C_k \mid$, $\mid \cdot \mid$ denotes the cardinality of a set.

Then, data set is whitened with respect to the estimated covariance matrix. The whitening transformation assigns lower weights to the directions of large variability, as this variability is mainly due to in-class changes and irrelevant to the classification task. The whitening transformation can be computed from the within chunklet covariance matrix by using the following equation

$$W = \hat{C}_c^{-\frac{1}{2}}. \tag{2}$$

For high dimensional data, the estimated covariance matrix can be used for semi-supervised dimensionality reduction with a constraints based FDA criterion. The whitening transformation matrix can be obtained by eigenanalysis of $\hat{C}_t \cdot \hat{C}_c^{-1}$ leading to get the feature subspace, where $\hat{C}_t$ is the total chunklets covariance matrix and can be written as

$$\hat{C}_t = \frac{1}{N_c} \sum_{k=1}^{K} \sum_{i=1}^{n_k} \left(x_i^{(k)} - d\right) \left(x_i^{(k)} - d\right)^T \tag{3}$$

where d is the mean of all the samples in all chunklets, $d = \frac{1}{N_c} \sum_{k=1}^{K} \sum_{i=1}^{n_k} x_i^{(k)}$.

2.2 Kernel Method

Kernel method is a kind of non-linear technology that has been widely applied to machine learning and pattern recognition, and achieves good results in many fields.

In theory, a linearly inseparable pattern in low-dimensional space can be converted to linearly separable ones by mapping into a high-dimensional feature space with non-linear transformation [17]. There are many problems to be encountered when we directly process data in the high-dimensional feature space obtained by a non-linear transformation. First, the form and parameters of non-linear mapping function are difficult to determine. The following question is that we cannot get the desired result if a conventional linear processing method is applied to high-dimensional feature space due to the "curse of dimensionality". Fortunately, the kernel technology can achieve this non-linear mapping and get rid of the aforementioned intractable issues simultaneously.

The projecting principle of kernel function can be briefly described as follows. Assume $X \subset R^M$, with $x_i, x_j \in X$, where M is the number of dimensions of input space. Define a nonlinear function $\Phi(\cdot)$ mapping the input space to a high dimensional Hilbert space $H(X \to H)$. Based on the kernel trick [15], a kernel function K is defined as

$$K(x_i, x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle \tag{4}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product of two vectors. As it can be seen from (4), the inner product of the projection function of high-dimensional Hilbert space can be obtained via kernel function in low-dimensional input space.

The performance of kernel method is influenced by the type of kernel function. In practical application, the choice of kernel function relies on the characteristic of data set. The functions satisfying Mercer theorem can be used as the kernel function. Common kernel functions [23] include Gaussian, Polynomial, and Hyperbolic Tangent (Sigmoid) kernel function, etc.

The procedure of kernel method can usually be divided into following five basic steps [24].

1. Preprocess original data,
2. Select kernel function and set parameters,
3. Compute kernel matrix by kernel function, i.e. mapping input space to a high dimensional Hilbert space via non-linear transformation function,
4. Apply kernel matrix to linear algorithm,
5. Get non-linear model of input space.

It is observed that kernel method achieves the goal of nonlinear mapping by computing the inner products with kernel function. There is no need to work explicitly with the concrete form of non-linear transformation function. The nonlinear mapping can be implicitly changed by adjusting the form and parameters of the kernel function. A variety of kernel-based methods can be readily constructed by means of combining different linear algorithms with the kernel method.

3 CHUNKLET-BASED KERNEL FEATURE EXTRACTION

In this section, the description of proposed CKRCA method is given. The detailed derivation process is demonstrated below.

According to the kernel method, a nonlinear mapping from the input space to a high dimensional Hilbert space should be obtained firstly, i.e. $\Phi : R^M \to H, X \to \Phi(X)$. For learning the optimal nonlinear transformation, the total covariance matrix and within-chunklets covariance matrix in H-space are defined as follows:

$$\hat{C}_c^H = \frac{1}{N_c} \sum_{k=1}^{K} \sum_{i=1}^{n_k} \left(\Phi\left(x_i^{(k)}\right) - d_k^H \right) \left(\Phi\left(x_i^{(k)}\right) - d_k^H \right)^T, \tag{5}$$

$$\hat{C}_t^H = \frac{1}{N_c} \sum_{k=1}^{K} \sum_{i=1}^{n_k} \left(\Phi\left(x_i^{(k)}\right) - d^H \right) \left(\Phi\left(x_i^{(k)}\right) - d^H \right)^T \tag{6}$$

where d_k^H are the mean of the samples in k^{th} chunklet, and d^H is the mean of the samples in all chunklet in the high dimensional space.

The CKRCA can find an optimal nonlinear transformation matrix via maximizing the total covariance matrix and minimizing the within-chunklets covariance matrix simultaneously. The objective is to choose an optimal nonlinear transformation in the H-space via a chunklet constraints-based FDA criterion, and the following objective function is maximized:

$$J(W) = \arg\max \frac{\mid W^T \hat{C}_t^H W \mid}{\mid W^T \hat{C}_c^H W \mid} \tag{7}$$

where W is the optimal nonlinear transformation matrix which can be represented as $W = [w_1, w_2, \ldots, w_m]$.

The each of the M column vectors is a span of all data points in the feature space. Then the vector w_i can be rewritten as

$$w_i = \Phi(X)\alpha_i \tag{8}$$

where α_i is the coefficients, $\alpha_i = [\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{in}]^T$.

Therefore, the transformation matrix can be represented as

$$W = \Phi(X)\alpha \tag{9}$$

where α is the coefficients matrix, $\alpha = [\alpha_1, \alpha_2, \ldots, \alpha_m]$.

On the basis of (9), the objective function (7) can be rewritten as follows

$$J(\alpha) = \arg\max \frac{\mid \alpha^T \Phi(X)^T \hat{C}_t^H \Phi(X)\alpha \mid}{\mid \alpha^T \Phi(X)^T \hat{C}_c^H \Phi(X)\alpha \mid}. \tag{10}$$

Two new matrixes can be defined, such that

$$K_t = \Phi(X)^T \hat{C}_t^H \Phi(X), \tag{11}$$

$$K_c = \Phi(X)^T \hat{C}_c^H \Phi(X). \tag{12}$$

Before giving the concrete form of K_t and K_c, some intermediate quantities should be expressed as follows:

$$\begin{aligned} \upsilon_i^{(k)} &= \Phi(X)^T \Phi(x_i^{(k)}) \\ &= \left[\Phi(x_1), \Phi(x_i), \ldots, \Phi(x_N)\right]^T \Phi\left(x_i^{(k)}\right) \\ &= \left[\Phi(x_1)^T \Phi\left(x_i^{(k)}\right), \Phi(x_2)^N \Phi\left(x_i^{(k)}\right), \ldots, \Phi(x_N)^T \Phi\left(x_i^{(k)}\right)\right]^T \end{aligned} \tag{13}$$

where $\upsilon_i^{(k)}$ denotes the i^{th} data point of the k^{th} chunklet in inner product space, can be rewritten as

$$\upsilon_i^{(k)} = \left[k\left(x_1, x_i^{(k)}\right), k\left(x_2, x_i^{(k)}\right), \ldots, k\left(x_N, x_i^{(k)}\right)\right]^T, \tag{14}$$

μ denotes the mean vector of all chunklets in inner product space and can be defined as

$$\begin{aligned} \upsilon_i^{(k)} &= \Phi(X)^T d^H \\ &= \frac{1}{N_c} \sum_{k=1}^{K} \sum_{i=1}^{n_k} \Phi(X)^T \Phi\left(x_i^{(k)}\right) \\ &= \frac{1}{N_c} \sum_{k=1}^{K} \sum_{i=1}^{n_k} \upsilon_i^{(k)}, \end{aligned} \tag{15}$$

μ_k is the mean vector of the k^{th} chunklet in inner product space and can also be defined as

$$\begin{aligned} \upsilon^{(k)} &= \Phi(X)^T d_k^H \\ &= \frac{1}{N_k} \sum_{i=1}^{n_k} \Phi(X)^T \Phi\left(x_i^{(k)}\right) \\ &= \frac{1}{N_k} \sum_{i=1}^{n_k} \upsilon_i^{(k)}. \end{aligned} \tag{16}$$

Combining with (5), (6) and (11), (12), (13), K_t and K_c can be rewritten as follows:

$$K_t = \frac{1}{N_c} \sum_{k=1}^{K} \sum_{i=1}^{n_k} \left(\upsilon_i^{(k)} - \mu\right) \left(\upsilon_i^{(k)} - \mu\right)^T, \tag{17}$$

$$K_c = \frac{1}{N_c} \sum_{k=1}^{K} \sum_{i=1}^{n_k} \left(\upsilon_i^{(k)} - \mu_k\right) \left(\upsilon_i^{(k)} - \mu_k\right)^T. \tag{18}$$

Then, the objective function can be represented briefly as follows:

$$J(\alpha) = \arg\max \frac{\mid \alpha^T K_t \alpha \mid}{\mid \alpha^T K_c \alpha \mid}. \tag{19}$$

When the coefficients matrix α is confirmed, the transformation matrix W can be computed by (9). And based on the kernel method, a kernel matrix K_M can be computed as

$$\begin{aligned} K_M &= \Phi(X)^T \Phi(X) \\ &= k(X, X). \end{aligned} \tag{20}$$

Finally, a new feature subspace can be obtained

$$\begin{aligned} X_{new} &= W^T \Phi(X) \\ &= \alpha^T \Phi(X)^T \Phi(X) \\ &= \alpha^T K_M \end{aligned} \tag{21}$$

where $X_{new} \subset R^m$ $(m << M)$ represents the learned feature subspace, and has nonlinear separability compared with the original input data.

The procedure associated with the proposed CKRCA is illustrated in detail as follows.

4 EXPERIMENTS

4.1 Dataset Description and Experimental Settings

A public data set used in the experiment was acquired by the National Aeronautics and Space Administration's Airborne Visible/Infrared Imaging Spectrometer sensor (AVIRIS), i.e. Indian Pines 92AV3C [25]. The image was gathered in North-western Indiana in June 1992 with the size of 145 × 145 pixels. The number of original bands is 220. We selected 200 bands as our experimental data set by removing some Signal-to-noise (SNR) bands, from 104 to 108, 150 to 163 and 220. The number of land-cover classes were reduced to 9 from the initial 16 land-cover classes for the remaining has little label samples and cannot afford the reliable statistical image [26]. Figure 1 a) and b) shows the false color composite image and the corresponding ground truth areas of Indian Pines data set.

In order to demonstrate the performance of the proposed CKRCA method, we evaluated the accuracy of several classical linear methods, i.e. the PCA, FDA, and RCA, and kernel-based nonlinear methods, i.e. the KPCA and KFDA. We also compared the results with those obtained by using all original features. The classification process was performed using the support vector machine (SVM) classifier (LIBSVM library) with Gaussian kernel, and the kernel parameters were selected

Algorithm 1: The CKRCA Algorithm

Input:

– A data set $X = \{x_1, x_2, \ldots, x_N\} \subset R^{N \times M}$

– K chunklets C_k

Output:

– Feature subspace $X_{new} \subset R^{N \times m}$ $(m << M)$

Procedure:

1. A kernel matrix K_M can be computed by (20);
2. Compute K_t and K_c by (17) and (18);
3. Diagonalize K_t by eigenanalysis

3.1 Find U to satisfy $U^T K_t U = \Lambda_t$ and $U^T U = I$, here Λ_t is a diagonal matrix sorted in increasing order;

3.2 Form a matrix $\hat{U}$ by the last r column vectors of U with the nonzero eigenvalues;

3.3 Let $D_t = \hat{U}^T K_t \hat{U}$ be the $r \times r$ submatrix of Λ_t;

3.4 Let $\hat{K}_w = K_w + \lambda I$, here λ is a regularized parameter, which is a small and positive number, is the identity matrix with the same size of K_w;

3.5 Let $Z = \hat{U} D_t^{-\frac{1}{2}}$ and $K_z = Z^T \hat{K}_c Z$;

4. Diagonalize K_z by eigenanalysis

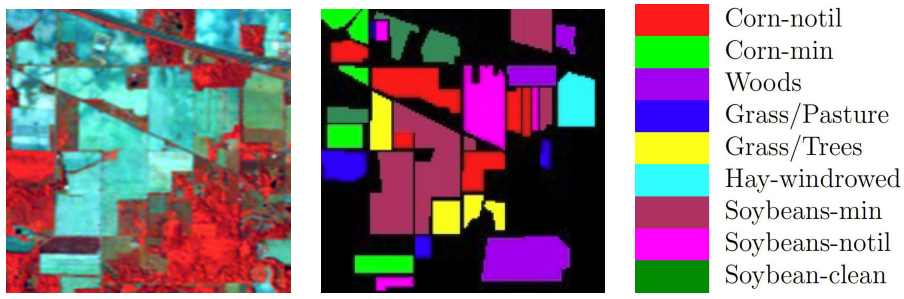
4.1 Find V to satisfy $V^T K_z V = \Lambda_c$ and $V^T V = I$, here Λ_c is a diagonal matrix sorted in descending order;

4.2 Assume the desired dimension is m, then form $\hat{V}$ by the first m column vectors of V with the smallest eigenvalues and let $D_c = \hat{V}^T K_z \hat{V}$;

5. Obtain optimal transformation matrix $\alpha = Z \hat{V} D_c^{-\frac{1}{2}}$;
6. Obtain the feature subspace using (21).

using the grid search strategy with 5-fold cross-validation, and the search range was set to $2^{-16} - 2^{16}$. The overall accuracy (OA) was used to measure the result of classification. The kernel function employed in the kernel-based feature extraction methods is Gaussian kernel function, and the kernel parameter was set to 10.

For comparing the performance of the above techniques accurately, we randomly selected five different percentage of samples (i.e. 5 %, 10 %, 15 %, 30 % and 40 %) as positive constraints/labeled pixels from the available ground truth data for the RCA and proposed CKPCA/the FDA and KFDA learning. These positive constraints have been divided into a number of several sets as chunklet in which any two samples may belong to the same class. According to the size of AVIRIS data set, the number of chunklets was set to 19. For each of the 9 classes, 20 % of the labeled pixels were chosen for SVM training, the classification performance is quantitatively estimated on the remaining 80 % ground truth pixels. The number of randomly selected five different percentage of positive constraints, the number of chunklets, and number of train/test pixels used in the experiments related to different land-cover classes are given in Table 1.



a) b)

Figure 1. a) False color composite image (using bands 73, 118, and 196); b) the available ground truth map with nine land cover classes of the Indian Pines data set

Class	Feature Subspace Learning						SVM Classifier	
	Positive Constraints					Chunklets	Train (pixels)	Test (pixels)
	5 %	10 %	15 %	30 %	40 %			
1 (Corn-notill)	72	135	210	420	630	3	260	1174
2 (corn-min)	48	90	140	280	420	2	156	678
3 (Grass/Pasture)	24	45	70	140	210	1	92	405
4 (Grass/Trees)	24	45	70	140	210	1	145	602
5 (Hay-windrowed)	24	45	70	140	210	1	81	408
6 (Soybeans-notill)	48	90	140	280	420	2	171	797
7 (Soybeans-min)	120	225	350	700	1 050	5	450	2 018
8 (Soybean-clean)	24	45	70	140	210	1	116	498
9 (Woods)	72	135	210	420	630	3	243	1 051

Table 1. The number of positive constraints (chunklets constraints) and total number of train/test pixels related to classes for the Indian Pines data set

4.2 Comparison with Linear Methods

The classification accuracy obtained by using all the features, the proposed CKRCA and the other linear methods, i.e. PCA, FDA, and RCA, are shown in Table 2. For the four feature extraction methods, the reduced feature size is experimentally set to the optimal feature number, which is corresponding to the best overall accuracy (OA).

From Table 2, one can see that the proposed CKRCA achieved the best classification accuracy compared with the PCA, FDA, and RCA in all different percentage of chunklet constraints, and had a better performance with extracted features varied from 12 to 17 than original all 200 features. For the RCA, the OA began to increase and obtained better performance than original features when the percentage of chunklet constraints exceeds 30 %. The classification accuracy got by the FDA increase was consistent with the percentage of labeled samples. But the classification accuracies of the FDA are lower than the PCA when the percentage of labeled samples

Original Feature	Feature size	200				
	OA (%)	82.02				
PCA	Feature size	13				
	OA (%)	75.95				
Percentage of chunklet constraints/labeled samples		5 %	10 %	15 %	30 %	40 %
FDA	Feature size	9				
	OA (%)	66.24	71.84	75.19	79.43	79.32
RCA	Feature size	13	12	11	10	12
	OA (%)	70.34	76.62	80.07	84.04	84.34
CKRCA	Feature size	17	12	16	14	17
	OA (%)	83.71	84.88	86.45	87.68	88.21

Table 2. Comparison of proposed CKRCA method with the PCA, FDA, and RCA, original features in different percentage of chunklet constraints/labeled samples for Indian Pines data set

was under 15 % and with original features in all percentage of chunklet constraints. The PCA yielded the lower classification accuracy than with original features.

4.3 Comparison with Kernel-Based Nonlinear Methods

In this section, the kernel-based nonlinear methods were used to compare with the proposed CKRCA. The classification results based on the KPCA and KFDA are shown in Table 3. In the KPCA and KFDA, the optimal feature size is experimentally set to seven and nine, while in the CKRCA, the desired feature size is set according to the eigenanalysis results. From Table 3, we can also observe that the proposed CKRCA has the best performance compared with the KPCA and KFDA in all different percentage of chunklet constraints. In details, the classification accuracy of the CKRCA can be increased as much as 7.05 % to 11.93 % with respect to the KFDA, and 7.32 % to 11.82 % with respect to the KPCA.

4.4 Feature Size of CKRCA

In order to obtain the optimal low-dimensional space, the parameter m is introduced as the desired feature size in the proposed CKRCA. From the procedure of the CKRCA, the optimal low-dimensional subspace can be obtained by the eigenvectors corresponding to the smallest m eigenvalues of K_z. With the purpose of determining the optimal m, all eigenvalues of K_z with 19 chunklets and different percentages of positive constraints and the 21 highlighted smallest eigenvalues sorted in increasing order are shown in Figure 2 a). It can be seen from Figure 2 a) that the eigenvalues of K_z tend to be horizontal when the index of eigenvalues is greater than 18. Therefore, we show the overall classification accuracy versus the feature size from 1 to 18 in Figure 2 b). By contrast, the OA shown in Figure 2 b) at first increased sharply

Original Feature	Feature size	200				
	OA (%)	82.02				
KPCA	Feature size	7				
	OA (%)	76.39				
Percentage of chunklet constraints/labeled samples		5 %	10 %	15 %	30 %	40 %
KFDA	Feature size	9				
	OA (%)	76.66	77.81	77.73	76.74	76.28
CKRCA	Feature size	17	12	16	14	17
	OA (%)	83.71	84.88	86.45	87.68	88.21

Table 3. Comparison of proposed CKRCA method with the KPCA and KFDA, original features in different percentage of chunklet constraints/labeled samples for Indian Pines data set

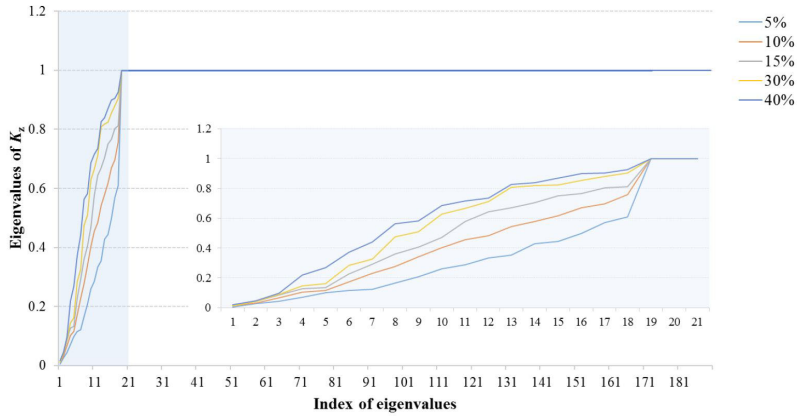
by increasing the number of the feature, then it reached the condition of minor fluctuation when more than 10 features were selected. Based on the aforementioned analysis, the feature size of the CKRCA can be derived experimentally by analyzing the eigenvalues of K_z.

4.5 Analysis of Execution Efficiency

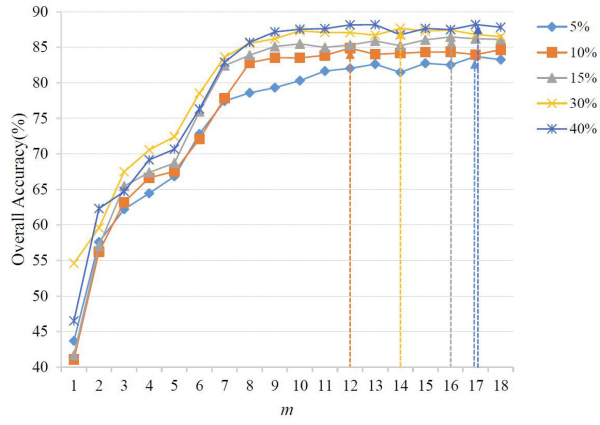
From the Sections 4.2 and 4.3, it can be seen that the proposed CKRCA method achieved the best classification accuracy compared with the three linear methods (i.e. PCA, FDA and RCA) and two kernel-based nonlinear methods (i.e. KPCA and KFDA). In general, kernel-based nonlinear methods would take more execution time than the corresponding linear methods for the computation of kernel matrix. In this section, we focus our attention to further analyse the execution efficiency for the kernel-based nonlinear methods (i.e. KPCA, KFDA and CKRCA) (see Table 4). Generally, eigenanalysis is the crucial and time-consuming process for the three kernel-based nonlinear methods. As expected, the computational time of the KFDA and proposed CKRCA is larger than the KPCA, for the execution of two eigenanalysis steps during the procedure. In greater detail, the CKRCA exhibited a smaller execution time than the KFDA with the same percentage of prior information. The execution time taken from KFDA significantly increased when increasing the percentage of labeled samples. While the proposed CKRCA method is not sensitive to the percentage of positive/chunklet constraints and has better efficiency combining with the classification results.

4.6 Number of Chunklets Impact on CKRCA

Furthermore, we compared the impact of different number of chunklets on the proposed CKRCA method in this section. We carried out the experiments with 19, 38, and 57 different number of chunklets randomly generated from the available ground



a)



b)

Figure 2. Feature size analysis of proposed CKRCA with 19 chunklets and different percentages of positive constraints. a) Value of all eigenvalues (and zoom on the 21 smallest one) of K_z sorted in increasing order. b) Classification OA versus the feature size from 1 to 18.

truth. Figure 3 presents the corresponding classification performance with different number of chunklet yielded by the CKRCA on Indian Pines data set. By analysis of Figure 3, we can conclude that the accuracy of the CKRCA increased by increasing the number of chunklets in all different percentages of chunklet constraints except 5 %. In practical application, we can increase the number of chunklets to improve the accuracy when the available proportion of the chunklet constraints is limited.

Exectution Time(s)	**KPCA**	127.14				
	Percentage of chunklet constraints/ labeled samples	5 %	10 %	15 %	30 %	40 %
	KFDA	526.31	667.45	837.39	1 318.2	1 797.6
	CKRCA	366.51	358.69	351.49	363.68	360.17

Table 4. Comparison of proposed CKRCA method with the KPCA, and KFDA, original features in different percentages of chunklet constraints/labeled samples for Indian Pines data set

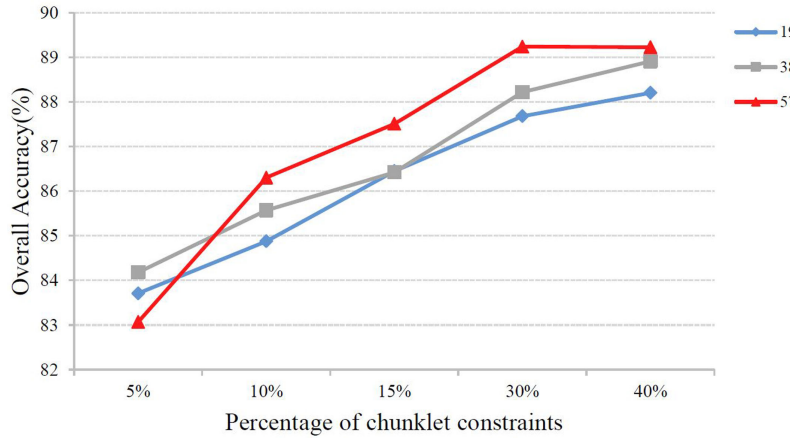


Figure 3. Classification performance (OA) of the proposed CKRCA with different percentages of chunklet constraints versus number of chunklets on Indian Pines data set

5 CONCLUSIONS

This paper presents a novel nonlinear feature extraction approach for dimensionality reduction of hyperspectral remote sensing image. The proposed approach is devised for addressing problems where the intrinsic nonlinear properties cannot be exploited by the traditional method and the available number of training samples is limited. The combination of kernel trick and the RCA with chunklet constraints in a semi-supervised model can improve the representation of the data. Comparative experiments with three classic linear methods and two popular kernel-based methods have been executed. Good results obtained on real hyperspectral data set demonstrated that the proposed approach combining chunklet constraints and kernel method can extract more effective features than the other conventional linear and kernel-based feature extraction techniques. Future work will consider the utilization of different kernel functions and metric learning methods in semi-supervised image feature extraction.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (61572-228, 61373067, 61373050), the National Key Basic Research Program of China (2015-CB453000), and the Science Technology Development Project from Jilin Province (20160101247JC, 20140520070JH). Associate Professor Chen Yang was the corresponding author for this paper.

Author Contributions

Conceived and designed the experiments: Chen Yang. Performed the experiments: Haishi Zhao. Analyzed the data: Laijun Lu, Renchu Guan. Wrote the paper: Chen Yang, Haishi Zhao

REFERENCES

[1] Smith, R. B.: Introduction to Hyperspectral Imaging. Microimages. `http://www.microimages.com/documentation/Tutorials/hyprspec.pdf`. Retrieved in June, 2008.

[2] Hughes, G. P.: On the Mean Accuracy of Statistical Pattern Recognizers. IEEE Transactions on Information Theory, Vol. 14, 1968, No. 1, pp. 55–63.

[3] Plaza, A.—Benediktsson, J. A.—Boardman, J. W. et al.: Recent Advances in Techniques for Hyperspectral Image Processing. Remote Sensing of Environment, Vol. 113, 2009, pp. 110–122, doi: 10.1016/j.rse.2007.07.028.

[4] Landgrebe, D. A.: Signal Theory Methods in Multispectral Remote Sensing. John Wiley & Sons, 2005.

[5] Fong, M.: Dimension Reduction on Hyperspectral Images. University California, Los Angeles, CA, 2007.

[6] Fukunaga, K.: Introduction to Statistical Pattern Recognition. Academic Press, 2013.

[7] Luo, R.—Liao, W.—Philips, W.—Pi, Y.: An Improved Semi-Supervised Local Discriminant Analysis for Feature Extraction of Hyperspectral Image. Joint Urban Remote Sensing Event (JURSE 2015), 2015.

[8] Adebanjo, H. M.—Tapamo, J. R.: Semi-Supervised Local Feature Extraction of Hyperspectral Images over Urban Areas. IEEE International Conference on Adaptive Science and Technology (ICAST), 2013, pp. 1–5.

[9] Chen, S.—Zhang, D.: Semisupervised Dimensionality Reduction with Pairwise Constraints for Hyperspectral Image Classification. IEEE Geoscience and Remote Sensing Letters, Vol. 8, 2011, No. 2, pp. 369–373.

[10] Shental, N.—Hertz, T.—Weinshall, D.—Pavel, M.: Adjustment Learning and Relevant Component Analysis. Computer Vision – ECCV 2002, Lecture Notes in Computer Science, Vol. 2353, 2002, pp. 776–790, doi: 10.1007/3-540-47979-1_52.

[11] Bar-Hillel, A.—Hertz, T.—Shental, N.—Weinshall, D.: Learning Distance Functions Using Equivalence Relations. Proceedings of the 20th International Conference on Machine Learning (ICML-2003), Washington, DC, USA, August 21–24, 2003, pp. 11–18.

[12] Bar-Hillel, A.—Hertz, T.—Shental, N.—Weinshall, D.: Learning a Mahalanobis Metric from Equivalence Constraints. Journal of Machine Learning Research, Vol. 6, 2005, No. 6, pp. 937–965.

[13] Hoi, S. C. H.—Liu, W.—Lyu, M. R.—Ma, W.-Y.: Learning Distance Metrics with Contextual Constraints for Image Retrieval. Proceedings of 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, New York, USA, Vol. 2, June 2006, pp. 2072–2078.

[14] Yang, C.—Bruzzone, L.—Zhao, H.—Liang, Y.—Guan, R.: Decorrelation–Separability-Based Affinity Propagation for Semisupervised Clustering of Hyperspectral Images. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, Vol. 9, 2015, No. 2, pp. 568–582.

[15] Bachmann, C. M.—Ainsworth, T. L.: Exploiting Manifold Geometry in Hyperspectral Imagery. IEEE Transactions on Geoscience and Remote Sensing, Vol. 43, 2005, No. 3, pp. 441–454.

[16] Rozel, A.—Clenet, H.—Douté, S.—Quantin, C.: Hyperspectral Data Processing Using Neural Networks: Preliminary Results for Mafic Minerals in SNC's Meteorites. 44th Lunar and Planetary Science Conference, Abstracts, 2013.

[17] Shawe-Taylor, J.—Cristianini, N.: Kernel Methods for Pattern Analysis. Cambridge University Press, 2004, doi: 10.1017/CBO9780511809682.

[18] Izquierdo-Verdiguier, E.—Gomez-Chova, L.—Bruzzone, L.—Camps-Valls, G.: Semisupervised Kernel Feature Extraction for Remote Sensing Image Analysis. IEEE Transactions on Geoscience and Remote Sensing, Vol. 52, 2014, No. 9, pp. 5567–5578.

[19] Fauvel, M.—Chanussot, J.—Benediktsson, J. A.: Kernel Principal Component Analysis for the Classification of Hyperspectral Remote Sensing Data over Urban Areas. EURASIP Journal on Advances in Signal Processing, Vol. 11, 2009, Art. No. 783194.

[20] Mei, F.—Zhao, C.—Wang, L.—Huo, H.: Anomaly Detection in Hyperspectral Imagery Based on Kernel ICA Feature Extraction. Proceedings of the 2008 Second International Symposium on Intelligent Information Technology Application (IITA '08), Vol. 1, 2008, pp. 869–873, doi: 10.1109/IITA.2008.98.

[21] Mika, S.—Rätsch, G.—Weston, J.—Schölkopf, B.—Müller, K.-R.: Fisher Discriminant Analysis with Kernels. Neural Networks for Signal Processing IX, Vol. 9, 1999, pp. 41–48.

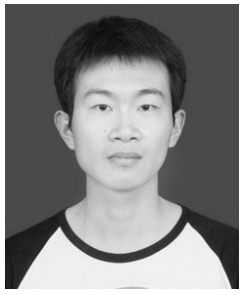
[22] Tsang, I. W.—Cheung, P. M.—Kwok, J. T.: Kernel Relevant Component Analysis for Distance Metric Learning. IEEE International Joint Conference on Neural Networks, Vol. 2, 2005, pp. 954–959.

[23] http://crsouza.com/2010/03/kernel-functions-for-machine-learning-applications.

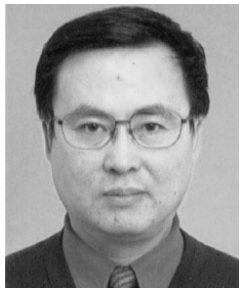
[24] Wang, H.—Yu, J.: Study on the Kernel-Based Methods and Its Model Selection. Journal of Southern Yangtze University (Natural Science Edition), Vol. 5, 2006, No. 4, pp. 500–504.

[25] AVIRIS NW Indiana's Indian Pines 1992 Data Set [Online]. Available on: `ftp://ftp.ecn.purdue.edu/biehl/MultiSpec/92AV3C(Originalfiles)` and `ftp://ftp.ecn.purdue.edu/biehl/PC_MultiSpec/ThyFiles.zip(groundtruth)`.

[26] Melgani, F.—Bruzzone, L.: Classification of Hyperspectral Remote Sensing Images with Support Vector Machines. IEEE Transactions on Geoscience and Remote Sensing, Vol. 42, 2004, No. 8, pp. 1778–1790.



Haishi Zhao received his B.Sc. degree in geography science from Jilin University, Changchun, China, in 2014. He is currently studying on his M.Sc. degree in digital geoscience at College of Earth Science, Jilin University, China. His research interests include hyperspectral image processing and machine learning.



Laijun Lu received his M.Sc. and Ph.D. degrees in mathematical geology from Jilin University, Changchun, China, in 1982 and 1992, respectively. He is currently Professor with the College of Earth Sciences and the Laboratory of Digital Geoscience, Jilin University. He was a Postdoctoral Fellow with the Mining and Petroleum of Exploration Postdoctoral Workstation, Northeastern University, Shenyang, China, from 1992 to 1994. He published six books and over 50 journal and conference papers. His research interests include digital geosciences and resources and environment information system, and application of spatial and time-series-related Markov process.

Chen Yang received her Ph.D. degree in digital geoscience from Jilin University, Changchun, China, in 2010. She was a visiting Ph.D. student at the Remote Sensing Laboratory, University of Trento, Italy in 2008. From December 2011 to June 2012, she worked as a Postdoctoral Fellow with the Remote Sensing Laboratory in the Department of Information Engineering and Computer Science, University of Trento, Italy. She is currently Associate Professor in the College of Earth Sciences, Jilin University, China. Her research interests include remote sensing image processing and machine learning.

Renchu Guan received his Ph.D. degree in bioinformatics from Jilin University, China, in 2010. From 2011 to 2012, he was a visiting scholar in University of Trento, Italy. He is currently Associate Professor in the College of Computer Science and Technology, Jilin University, China. He published over 30 papers. His research was featured in IEEE Transactions on Knowledge and Data Engineering, IEEE Transactions on Geoscience and Remote Sensing, IEEE Geoscience and Remote Sensing Letters, Engineering Applications of Artificial Intelligence, etc. He was the recipient of several grants from National Natural Science Foundation of China, Postdoctoral Science Foundation, etc. His research interests include machine learning, text mining and bioinformatics.