

Dimensionality Reduction Based on Determinantal Point Process and Singular Spectrum Analysis for Hyperspectral Images

Weizhao Chen¹, Zhijing Yang^{1*}, Faxian Cao¹, Yijun Yan², Meilin Wang¹, Chunmei Qing³, Yongqiang Cheng⁴

¹School of Information Engineering, Guangdong University of Technology, Guangzhou, 510006, China

²Department of Electronic and Electrical Engineering, University of Strathclyde, Glasgow, UK

³School of Electronic and Information Engineering, South China University of Technology, Guangzhou, China

⁴School of Engineering and Computer Science, University of Hull, Hull, HU6 7RX, UK

*yzhj@gdut.edu.cn

Abstract: Dimensionality reduction is of high importance in hyperspectral data processing, which can effectively reduce the data redundancy and computation time for improved classification accuracy. Band selection and feature extraction methods are two widely used dimensionality reduction techniques. By integrating the advantages of the band selection and feature extraction, we propose a new method for reducing the dimension of hyperspectral image data. First, a new and fast band selection algorithm is proposed for hyperspectral images based on an improved Determinantal point process (DPP). To reduce the amount of calculation, the Dual-DPP is used for fast sampling representative pixels, followed by kNN-based local processing to explore more spatial information. These representative pixel points are used to construct multiple adjacency matrices to describe correlation between bands based on mutual information. To further improve the classification accuracy, two-Dimensional Singular Spectrum Analysis (2D-SSA) is used for feature extraction from the selected bands. Experiments show that the proposed method can select a low-redundancy and representative band subset, where both data dimension and computation time can be reduced. Furthermore, it also shows that the proposed dimensionality reduction algorithm outperforms a number of state-of-the-art methods in terms of classification accuracy.

1. Introduction

Unlike traditional two-dimensional (2-D) images, hyperspectral data contains a 3-D structure, i.e. 2-D spatial information and 1-D spectral information. Abundant spectral information can describe the ground targets in more detail. Such advantages have facilitated the fast development of hyperspectral images in the field of remote sensing [1-2], where it has been successfully applied in urban mapping, environmental management, crop analysis, and mineral detection [3]. In addition to remote sensing, hyperspectral images has also been applied in lab-based applications such as forensics, pharmaceutical, medical, and food quality analysis [4-7].

Although more spectral bands can increase the representation of ground objects, not all the bands play equally important roles in hyperspectral data processing [8]. Excessive data dimensions will directly lead to increased computational complexity and computer memory resources. In addition, high dimensional dataset may cause the Hughes phenomenon, which will lead to a decline in classification accuracy [3].

To solve these problems, dimensionality reduction for hyperspectral images has attracted great attention in the past decade. Relevant techniques can be mainly divided into two categories, i.e. band selection and feature extraction.

Band selection can also be regarded as a feature selection process [9-14], which aims to select a subset of the spectral features from the original data whilst keep the dominant information and maintain the performance. In [15], Yao et al proposed a new filter-based feature extraction method for band selection. The advantage of band selection

is that it retains the physical information, characteristics and interpretable ability of the original data [9]. According to whether the sample labels are known or not, the band selection methods for hyperspectral images can be classified into supervised band selection [16-17] and unsupervised band selection [18-20].

Under the condition that a part of the sample labels is known, the supervised band selection method can select bands with higher correlation with class labels. In [12], the correlation between features and class labels is considered in designing a mutual-information based band selection algorithm. In comparison to unsupervised band selection methods, supervised band selection in general has better performance [21]. However, unsupervised band selection is still needed especially when there is no (sufficient) labelled data for supervised learning, regardless the time-consuming process of data labelling.

For most unsupervised band selection algorithms, they aim to preserve the information of the original data with lowest possible dimensional dataset. Therefore, the generic unsupervised band selection algorithm contains two steps. The first is to establish an effective criterion to evaluate the spectral band performance or the redundancy of the candidate band subset. The second is to find a search strategy to determine a suitable band subset.

Unlike band selection methods, feature extraction is to project the original data from the high-dimensional space to the low-dimensional space according to the criteria and obtain a low-dimensional data representation [22-25]. A number of feature extraction methods have been proposed for hyperspectral images dimensionality reduction. Based on how they deal with the spectral bands, these can be further

categorized into local-based methods and global-based methods.

Local-based methods map the data from high dimension to low dimension whilst maintaining the local structure. Examples include Locality Preserving Projection (LPP) [26] and Locally Linear Embedding (LLE) [27]. In [28], Guo et al proposed a novel sparse hashing method to embed the high-dimension features into a low-dimension Hamming space.

The global-based approaches aim to obtain a low dimensional representation of the original data whilst maintain the global structure of the original data. Typical examples include Principal Component Analysis (PCA) [29] and Isometric Feature Mapping (ISOMAP) [30]. In addition, some feature extraction methods can greatly remove the image noise, such as two-dimensional empirical mode decomposition (2D-EMD) [31] and two-dimensional singular spectrum analysis (2D-SSA) [22]. 2D-EMD can reconstruct each image using extracted low-order IMFs which can present the spatial structure of the image. 2D-SSA can decompose each image into varying trends, oscillations, and noise. The image can be reconstructed using the trend and selected oscillations whilst remove the noise. In general, feature extraction methods can achieve high accuracy in classification of hyperspectral images.

In summary, band selection and feature extraction methods have their own advantages and disadvantages for dimensionality reduction. Band selection methods can effectively maintain the original spectral band information and the correlation between spectral bands. Besides, band selection can directly remove the redundant information from the hyperspectral datasets. However, band selection normally is difficult to greatly improve the classification accuracy. On the contrary, feature extraction can effectively use the potential features of the hyperspectral image and achieve higher classification accuracy. However, feature extraction is usually applied to the original hyperspectral images and does not take into account the redundancy of the spectral bands. This will also result in the redundancy after feature extraction. Moreover, it will take more computation time for feature extraction because of the redundant information. In order to solve these problems, we combine band selection and feature extraction and propose a new dimension reduction method for hyperspectral image as detailed below.

First, we designed a fast band selection method using the spatial information. In [8], Yuan et al proposed a multiple graph to describe the complex structure between spectral bands by using spectral clustering. This method can consider latent structure between spectral bands in the high dimensional space thus it helps to make full use of the spatial information for more comprehensive measurement of the spectral redundancy. However, constructing a multiple graph structure by spectral clustering requires extremely high computation time and memory. To tackle this problem, we sample representative pixels to measure the correlation between bands. This sampling method consists of two steps: we first use Dual Determinantal Point Process (Dual-DPP) [32] to fast sample representative pixels, followed by kNN based strategy to determine neighbouring pixels to supplement the number of samples.

Afterwards, we use the mutual information to construct an adjacency matrix that can describe the

correlation of the spectral bands with representative pixels in each group. In this way, we can obtain multiple adjacency matrices from representative pixels in different groups. These adjacency matrices can measure redundancy of spectral bands using different spatial information. In addition, we also improved the original k-Determinantal point process (k-DPP) [33] to enable it to select a diversity and low-redundancy band subset from different adjacency matrices. Finally, to improve the classification accuracy, we use 2D-SSA to extract the spatial features from a low-redundancy and low-dimension band subset.

The main contributions of this paper can be summarized as follows:

1. A fast Dual-DPP and kNN based algorithm is proposed to sample sufficient representative pixels from each spectral band.
2. From the perspective of data distribution, spatial information from different regions is used to measure the redundancy of the spectral bands more comprehensively. Moreover, we have improved the original k-DPP algorithm so that it can select low-redundancy spectral band subsets on multiple similarity matrices.
3. We apply 2D-SSA to selected subset of bands to better extract spatial features to improve the classification accuracy under much lower feature dimensions whilst effectively reducing the computational time.

2. Related background

2.1 k-Determinantal Point Process

Determinantal point process (DPP) is an elegant probabilistic model of repulsion that arises in quantum physics and random matrix theory [32]. DPP is an effective and accurate sampling method, which can select a diversity and low-redundancy subset. DPP has many applications in real-world such as summarization, image search and news threading [32]. First proposed by Macchi [34], the DPP cannot set the cardinality k of the subset bands in advance. To solve this problem, Kulesza et al [33] proposed an improved DPP algorithm called k-DPP to model the sets of cardinality k as follows.

Given a discrete finite set $\gamma = \{1, 2, 3, \dots, N\}$, any candidate subset $Y \subseteq \gamma$ with cardinality $|Y| = k$ is selected according to the probability

$$P_L^k(Y) = \frac{\det(L_Y)}{\sum_{|Y'|=k} \det(L_{Y'})} \quad (1)$$

where Y' represents all the subsets that their cardinality, $|Y'|$, is equal to k . L_Y is a symmetric positive semidefinite matrix and its entries are indexed by the corresponding elements of Y . Similarly, $L_{Y'}$ is a symmetric positive semidefinite matrix whose entries are indexed by the corresponding elements of Y' . Each element of matrix L_Y represents the correlation of the corresponding two elements in Y .

For example, if the subset Y contains two elements $Y = \{i, j\}$, L_Y can be written as

$$L_Y = \begin{bmatrix} L_{ii} & L_{ij} \\ L_{ji} & L_{jj} \end{bmatrix} \quad (2)$$

The greater the correlation between elements i and j will result in the larger value of L_{ij} and L_{ji} . Then the probability of selecting candidate subset Y can be expressed as follows:

$$P_L^k(Y) \propto \det(L_Y) \propto \begin{vmatrix} L_{ii} & L_{ij} \\ L_{ji} & L_{jj} \end{vmatrix} \quad (3)$$

Obviously, the larger values of L_{ii} and L_{jj} mean that the elements i and j are more likely to belong to the subset Y . The large value of L_{ij} means that i and j seem unlikely to co-occur. The more diverse and lower-redundancy subset has larger $P_L^k(Y)$.

However, selecting Y from γ according to the equation (1) is an NP-hard problem. In [33], an approximate solution is given. The k-DPP mainly contains two steps [33]. In the first step, the eigenvector subset V with k eigenvectors are selected depending on the probability which is calculated by the corresponding eigenvalues. In the second step, we select the band subset according to eigenvectors in subset V . An example is given below to detail the selection process.

Given a positive semidefinite matrix L , its entries are indexed by the corresponding elements of γ . L can be decomposed as

$$L = \sum_{n=1}^N \lambda_n v_n v_n^T \quad (4)$$

λ_n is the eigenvalue of L and v_n is the corresponding eigenvector. Then we want to find a subset Y from γ whose cardinality is $|Y| = k$. First, we select k eigenvalues $S = \{\lambda_1, \lambda_2, \dots, \lambda_{k-1}, \lambda_k\}$ and the corresponding eigenvectors $V = \{v_1, v_2, \dots, v_{k-1}, v_k\}$. Define $S' \cup \lambda_k = S$, then the cardinality $|S'| = k - 1$. The probability of selecting the k -th eigenvalue, λ_k , corresponding to the eigenvectors v_k of L to add to S and V can be expressed as:

$$P(\lambda_k \in S) = \lambda_n \frac{e_k^{n-1}}{e_k^n} \quad (5)$$

where $e_k^n = \sum_{|Y|=k} \prod_{n \in Y} \lambda_n$. In the second step, denote the low-redundancy band subset as Y , whose cardinality is k , that is, $|Y| = k$ and $Y' \cup i = Y$. The probability of selecting the i -th band to add to Y can be written as:

$$P_r(i) = \frac{1}{|V|} \sum_{v_n \in V} (v_n^T e_i)^2 \quad (6)$$

where e_i is a column vector that the i element is one and the other elements are zero.

2.2 2D-Singular Spectrum Analysis

The 2D-SSA can extract the trend and select oscillations from hyperspectral images as features and remove the noise [22]. It contains four steps: Embedding 2-D Signal, SVD, Grouping, and Diagonal Averaging.

Embedding 2-D Signal: we set each spectral image P^{2D} with a size $n_x \times n_y$. Each spectral image can be represented as a matrix:

$$P^{2D} = \begin{bmatrix} p_{1,1} & p_{1,2} & \dots & p_{1,n_y} \\ p_{2,1} & p_{2,2} & \dots & p_{2,n_y} \\ \vdots & \vdots & \ddots & \vdots \\ p_{n_x,1} & p_{n_x,2} & \dots & p_{n_x,n_y} \end{bmatrix} \quad (7)$$

P^{2D} can be written as a trajectory matrix X^{2D} with Hankel-block-Hankel structure:

$$X^{2D} = \begin{bmatrix} H_1 & H_2 & \dots & H_{N_x-L_x+1} \\ H_2 & H_3 & \dots & H_{N_x-L_x+2} \\ \vdots & \vdots & \ddots & \vdots \\ H_{L_x} & H_{L_x+1} & \dots & H_{N_x} \end{bmatrix}_{L_x \times (n_x-L_x+1)} \quad (8)$$

where each submatrix H_r is Hankel type defined by

$$H_r = \begin{bmatrix} p_{r,1} & p_{r,2} & \dots & p_{r,n_y-L_y+1} \\ p_{r,2} & p_{r,3} & \dots & p_{r,n_y-L_y+2} \\ \vdots & \vdots & \ddots & \vdots \\ p_{r,L_y} & p_{r,L_y+1} & \dots & p_{r,n_y} \end{bmatrix}_{L_y \times (n_y-L_y+1)} \quad (9)$$

where L_x and L_y are the sizes of 2D-windows.

SVD: Matrix S is defined by the trajectory matrix X^{2D} as follows:

$$S = X^{2D} X^{2D^T} \quad (10)$$

The eigenvalues of S are denoted as $\{\lambda_1 \geq \lambda_2 \geq \dots \lambda_L\}$ and the corresponding eigenvectors are $\{\mu_1, \mu_2, \dots, \mu_L\}$, where $L = L_x \times L_y$.

Grouping: The total set of L components is divided into two disjoint sets according to their contribution. The contribution of each component is related to its eigenvalue which can be represented as

$$\eta_t = \frac{\sum_{l \in t} \lambda_l}{\sum_{l=1}^L \lambda_l} \quad (11)$$

t is the set of eigenvalues with large values. Then we select the set with higher η_t to reconstruct the matrix X_M^{2D} .

Diagonal Averaging: This process contains two-step Hankelization process. Sequential Hankelization is first applied within each block H_r and then applied between the blocks X^{2D} . Then 2D diagonal projection from group X_M^{2D} can be written as follows

$$Z_M^{2D} = \begin{bmatrix} z_{m_1,1} & z_{m_1,2} & \dots & z_{m_1,n_y} \\ z_{m_2,1} & z_{m_2,2} & \dots & z_{m_2,n_y} \\ \vdots & \vdots & \ddots & \vdots \\ z_{m_{n_x},1} & z_{m_{n_x},2} & \dots & z_{m_{n_x},n_y} \end{bmatrix} \quad (12)$$

Finally, the spectral image P^{2D} can be described as

$$P^{2D} = Z_1^{2D} + Z_2^{2D} + \dots + Z_M^{2D} = \sum_{m=1}^M Z_m^{2D} \quad (13)$$

3. The proposed band selection framework

In this section, we will give a detailed description of the proposed dimensionality reduction method. The flowchart of our proposed algorithm is showed in Fig. 1.

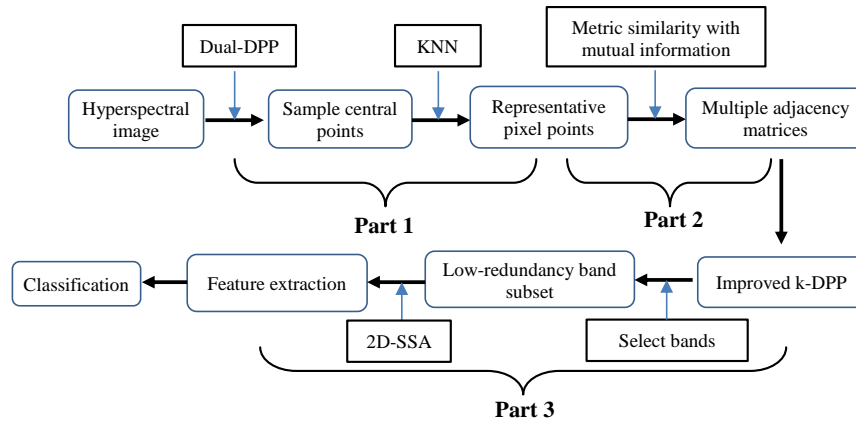


Fig.1. The flowchart of the proposed dimensionality reduction method

The proposed approach consists of three parts. In Part 1, a fast yet effective sampling algorithm based on Dual-DPP and kNN is introduced to sample the most representative pixels from the spectral datasets. In Part 2, these representative pixels are used to construct multiple adjacency matrices to describe the relationships of band pairs measured by the mutual information. In Part 3, an improved k-DPP algorithm is presented to select a low-redundancy subset of bands from multiple adjacencies. In addition, 2D-SSA is employed to extract the spatial features from the selected bands to further improve the classification accuracy. Details will be discussed in the following subsections.

In section 3.1, we introduce the proposed method to fast sample representative pixel points from hyperspectral images. The method of constructing multiple adjacency matrices is proposed in section 3.2. In section 3.3, we introduce the process of selecting bands with the proposed method and use 2D-SSA to extract spatial feature from the selected band subset.

3.1 Sample representative pixels

A hyperspectral image can be written as $B = \{b_1, b_2, b_3, \dots, b_l\} \in R^{n \times l}$, where n represents the number of pixels in each spectral band and l is the total number of spectral bands. b_i ($1 \leq i \leq l$) is the i -th spectral band. We use p_i^x to represent the pixel points in i -th spectral band.

In this section, we aim to select some representative pixel points from the hyperspectral spatial-domain for subsequent band selection. These pixel points must be diverse and widely distributed in the spatial domain. The algorithm k-DPP mentioned above can be applied for this purpose. However, the k-DPP inevitably needs to compute $L = BB^T \in R^{n \times n}$ and perform the eigen-decomposition on L . As shown in equation (4), if the dimension n of L is large, it will require a huge amount of computer memory and a huge amount of computation time. Therefore, it is impractical to apply this method directly.

To tackle this problem, the Dual-DPP [32] is applied here for fast sampling tasks. Different from $L = BB^T \in R^{n \times n}$, we construct another adjacency matrix $H = B^T B \in$

$R^{l \times l}$, where l is the number of bands and n is the number of pixels. Since l is much smaller than n , it's obvious that the dimension of H is much smaller than the dimension of L . Therefore, Dual-DPP requires less memory and less time than k-DPP in the decomposition process. H can be decomposed as

$$H = \sum_{n=1}^l \hat{\lambda}_n \hat{v}_n \hat{v}_n^T \quad (14)$$

where $\hat{\lambda}_n$ is eigenvalue of H and \hat{v}_n is the corresponding eigenvector. It can be simply proved that

$$v_n = B^T \hat{v}_n \quad (15)$$

Therefore, the sampling process by the Dual-DPP method also includes two steps. The first step is the same as the k-DPP method. According to equation (5), eigenvalues subset \hat{V} with \hat{k} eigenvalues \hat{v}_n are selected. In the second step, different from equation (6), we select i -th bands by the following probability

$$P_r(i) = \frac{1}{|\hat{V}|} \sum_{\hat{v}_n \in \hat{V}} ((B^T \hat{v}_n)^T e_i)^2 \quad (16)$$

We repeat the selection process for multiple times until the required number \hat{k} ($0 < \hat{k} \leq l$) is met.

Due to the dimensionality limit of matrix H , the value of \hat{k} is normally smaller than l . It will inevitably lead to insufficient samples. In order to solve this problem, we use KNN [35] to increase the number of pixel points.

First, the pixel points collected by Dual-DPP [32] are used as the central points, which can be recorded as a central point set $P_c = \{P_c^1, P_c^2, \dots, P_c^{\hat{k}}\} \in R^{l \times \hat{k}}$, where l is the number of bands and \hat{k} is the number of central points.

After that, around each central point P_c^α ($1 \leq \alpha \leq \hat{k}$), we use the kNN method [35] to find m pixel points that are closest to the central points. In this way, we can collect enough diversity pixel points. We represent the pixel points set as $\hat{B} = \{\hat{b}_1, \hat{b}_2, \hat{b}_3, \dots, \hat{b}_l\} \in R^{r \times l}$ which has \hat{k} clusters in each band and each cluster has m pixel points. $r = \hat{k} \times m$ represents the number of pixels in each spectral band and l is the total number of all spectral bands.

3.2 Construct multiple adjacency matrices with mutual information

Graph model can effectively capture the relationship between vertices. In [8], the author proposed a multi-graph structure by spectral clustering method that can effectively represent the relationship between bands. However, the spectral clustering methods are too complicated and require lots of computer memory.

To tackle this problem, considering the useful information in spatial domain, we use representative pixels to construct multiple adjacency matrices with mutual information metrics. With this model, we can consider the redundancy of spectral bands from the data distribution of each band and capture the complex relationships between the pairwise bands. Both aspects make our model more comprehensive in judging the redundancy between spectral bands.

We use a fast algorithm [36] to estimate the mutual information between two bands. Assume that pixels $(p_{i,\alpha}^x, p_{j,\alpha}^x)$, $x = 1, \dots, m$, in the α -th cluster from any pairwise bands \hat{b}_i and \hat{b}_j , and the joint probability of pixel points in α -th cluster from any pairwise bands \hat{b}_i and \hat{b}_j can be estimated as

$$f(p_{i,\alpha}, p_{j,\alpha}) = \frac{1}{m} \sum_{x=1}^m \frac{1}{2\pi h^2} e^{-\frac{1}{2h^2}((p_{i,\alpha}^x - p_{i,\alpha}^x)^2 + (p_{j,\alpha}^x - p_{j,\alpha}^x)^2)} \quad (17)$$

The marginal probability of pixels in the α -th cluster from any band \hat{b}_i can be estimated by

$$f(p_{i,\alpha}) = \frac{1}{m} \sum_{x=1}^m \frac{1}{\sqrt{2\pi h^2}} e^{-\frac{1}{2h^2}(p_{i,\alpha}^x - p_{i,\alpha}^x)^2} \quad (18)$$

Combining equations (17) and (18), we can get the mutual information of α -th cluster from any pairwise bands \hat{b}_i and \hat{b}_j as

$$I(\hat{b}_{i,\alpha}, \hat{b}_{j,\alpha}) = \frac{1}{m} \sum_{x=1}^m \log \frac{m \sum_{y=1}^m e^{-\frac{1}{2h^2}((p_{i,\alpha}^x - p_{i,\alpha}^y)^2 + (p_{j,\alpha}^x - p_{j,\alpha}^y)^2)}}{\sum_{y=1}^m e^{-\frac{1}{2h^2}(p_{i,\alpha}^x - p_{i,\alpha}^y)^2} \sum_{y=1}^m e^{-\frac{1}{2h^2}(p_{j,\alpha}^x - p_{j,\alpha}^y)^2}} \quad (19)$$

where $I(\hat{b}_{i,\alpha}, \hat{b}_{j,\alpha})$ indicates the mutual information between i -th and j -th bands calculated by the pixel points of α -th cluster. According to equation (19), we can calculate the mutual information between all bands and construct adjacency matrices L_{MI}^α . Through the above method, we can build multiple adjacency matrices $L_{MI} = \{L_{MI}^1, L_{MI}^2, L_{MI}^3, \dots, L_{MI}^\alpha\} (1 \leq \alpha \leq \hat{k})$, where $L_{MI}^\alpha \in R^{l \times l}$ is a symmetric matrix used to measure the correlation between spectral bands with pixel information from the α -th cluster.

3.3 Band selection and feature extraction

Taking into account the spatial information of the hyperspectral image, we have construct multiple adjacency matrices with mutual information mentioned in Section 3.2.

In this section, we propose an improved k-DPP to sample low-redundancy set of bands from these multiple adjacency matrices. According to Ref. [8], our algorithm can be written as follows:

$$P_{L_{MI}}^k(Y) = \frac{1}{k} \sum_{\alpha=1}^{\hat{k}} \frac{1}{e_{\alpha,k}^N} \sum_{|Y|=k} P^{V_Y^B}(Y') \prod_{n \in Y} \lambda_n^\alpha \quad (20)$$

where $e_{\alpha,k}^N = \sum_{|Y|=k} \prod_{n \in Y} \lambda_n^\alpha$ are the eigenvalue polynomials of adjacency matrices L_{MI} , and $P_{L_{MI}}^k(Y)$ is the probability of sampling a band subset Y from the all band set γ . λ_n^α is eigenvalue of L_{MI}^α and v_n^α is corresponding eigenvector. The probability of selecting i -th band to add to Y can be written as:

$$P_r(i) = \frac{1}{\hat{k}} \sum_{\alpha=1}^{\hat{k}} \frac{1}{|V^\alpha|} \sum_{v_n^\alpha \in V^\alpha} (v_n^{\alpha T} e_i)^2 \quad (21)$$

Our algorithm consists of two steps. The first step is to select the k feature vectors in the same way as stated in Section 2.1. The main difference is that we have to select different feature vectors on different adjacent matrixes. In the second step, we select k bands, according to the probability calculated in equation (21). Finally, we input the selected spectral band subset into the 2D-SSA algorithm for feature extraction. Details of the final proposed method are summarized in **Algorithm 1**.

4. Experimental result

In this section, we use two well-known hyperspectral image datasets to test the performance of the proposed method. Several representative unsupervised dimensionality reduction algorithms are used for comparisons, including PCA [29], LPP [26] and EMAP [37]. All the experiments were performed on Matlab and run on Intel CPU E5-1620@3.5GHz with 32GB RAM.

4.1. Hyperspectral Datasets

AVRIS sensor: Indian Pines dataset

The India Pines dataset was obtained by the Airborne Visible/Infrared Imaging Spectrometer sensor (AVIRIS) in 1992 [38]. The image covers Indian Pines test site in North-western Indiana. It has 220 spectral bands covering the spectrum range of 0.2-2.4 μm . And 20 water absorption spectral bands [104-108], [150-163] are removed and remaining 200 spectral bands form the dataset. It has 145×145 spatial pixels, including 16 classes ground truth.

ROSIS sensor: Pavia University dataset

The Pavia University was collected by the Reflective Optics System Imaging Spectrometer optical sensor covering the Pavia University, Italy [39]. The dataset has 115 spectral bands covering the spectrum range of 0.43-0.86 μm . 12 noisy spectral bands are removed to remain 103 bands for the dataset. It has 610×340 spatial pixels, including ground truth in 9 classes.

Algorithm 1 Dimensionality reduction based on DPP and SSA

Input:
 B : a hyperspectral image
 k : number of selected bands

Output:
 Y^* : The feature map of B

- 1) Sample representative pixel points by Dual-DPP and kNN
- 2) **for** pixel points of each cluster **do**
 calculate multiple adjacency matrices L_{MI}^α by equations (19)
- 3) **end for**
- 4) **for** each adjacency matrices L_{MI}^α **do**
- 5) Characteristic decomposition: $\{v_n^\alpha, \lambda_n^\alpha\}_{n=1}^l \leftarrow L_{MI}^\alpha$
- 6) Select k eigenvectors with probability $\lambda_n^\alpha \frac{e^{\frac{n-1}{B,k-1}}}{e_{B,k}^\alpha}$
- 7) **end for**
- 8) Select band subset Y with equations (21)
- 9) Extract spatial features Y^* from Y

Output: Y^*

4.2. Experimental parameter setup

In this section, we will introduce the parameter settings. We collected 20 pixel points as central points, and each central point selects the closest 50 and 30 pixel points in Indian Pines image and Pavia University image, respectively. The parameter, h , i.e. the width of the kernel in equation (18), is set to be 0.00032 and 0.0024 in Indian Pines image and Pavia University image, respectively. We use the SVM algorithm [40] to test the performance of the selected bands. **The parameters C and RBF kernel of SVM are determined by five-fold cross validation.** In all experiments, 5% of the samples are randomly selected for training, and the rest of the data are used for testing **and all experiments are repeated 10 times.** The number of training samples and test samples for Indian Pines image and Pavia University image are shown in Tables 1 and 2, respectively. The results of the experiment are evaluated with three authoritative indexes: overall accuracy (OA), average accuracy (AA) and kappa coefficient [41].

Table 1 The numbers of training and test samples of each class in the Indian Pines image

Class	Training	Test	Samples
1. Alfalfa	3	51	54
2. Corn-notill	72	1362	1434
3. Corn- mintill	42	792	834
4. Corn	12	222	234
5. Grass- pasture	25	472	497
6. Grass-trees	37	710	747
7. Grass-pas- turemowed	2	24	26
8. Hay- windrowed	24	465	489
9. Oats	2	18	20
10. Soybean- notill	48	920	968
11. Soybean- mintill	123	2345	2468

12. Soybean- clean	31	583	614
13. Wheat	11	201	212
14. Woods	65	1229	1294
15. Building- grass-trees	19	361	380
16. Stone- steel-towers	5	90	95
total	521	9845	10366

Table 2 The numbers of training and test samples of each class in Pavia university image

Class	Training	Test	Samples
1. Asphalt	332	6299	6631
2. Meadows	933	17716	18649
3. Gravel	105	1994	2099
4. Trees	153	2911	3064
5. Painted metal sheets	67	1278	1345
6. Bare soil	252	4777	5029
7. Bitumen	67	1263	1330
8. Self-blocking bricks	184	3498	3682
9. Shadows	48	899	947
Total	2141	40635	42776

4.3. Classification performance

In this section, we choose different dimensionality reduction methods to reduce the dimension of the data and compare the performance of the reduced dimension data on the classification task. All experiments are repeated 10 times, and the relevant mean and standard of classification accuracy deviation are shown in Table 3 and Table 4, respectively.

From Table 3, it can be seen that except the PCA algorithm, the classification accuracies of the other three data dimensionality reduction methods are all higher than that of the original data in Indian Pines dataset. Compared to the original dataset, our algorithm has 11% increase in the indexes of OA and AA, and 12% increase in the indexes of Kappa. Compared to EMAP, our algorithm has 1.7-2% increase in the indexes of OA and Kappa, and 0.3% slightly lower in the indexes of AA. Our method can achieve good results, mainly because our algorithm can effectively select a wide distribution of light bands, which can effectively maintain the original data information. Moreover, our algorithm can effectively extract spatial feature information.

As can be seen from Table 4, the four dimensionality reduction methods can achieve good results. Compared with other methods, the PCA method has the worst result for dimensionality reduction. Our proposed method achieves the best results. Compared to PCA and LPP, our method has a 7-8% improvement in the three indicators. Although EMAP is a powerful method, the performance of our method also has a 0.6-1% improvement in classification accuracy.

Table 3 Classification accuracy in the Indian Pines with PCA, LPP, EMAP and the proposed method (Best result of each row is marked in bold type)

Class	Baseline	PCA	LPP	EMAP	Proposed
Dimensionality	200	36	36	36	36
1. Alfalfa	58.04±18.05	61.18±13.95	62.75±15.52	90.20±5.06	80.78±16.74
2. Corn-notill	79.21±2.55	78.98±2.28	82.72±1.57	87.25±2.43	92.72±3.31
3. Corn- mintill	70.88±5.47	71.28±5.50	71.19±5.53	92.50±4.56	89.17±2.65
4. Corn	60.09±10.18	59.55±8.30	71.26±6.85	84.23±4.53	91.04±7.55
5. Grass- pasture	91.50±3.79	90.68±3.64	93.35±1.81	90.68±3.48	91.33±5.43
6. Grass-trees	95.06±2.00	94.99±2.00	94.59±1.95	96.10±1.27	96.85±2.32
7. Grass-pas- turemowed	72.50±11.32	77.08±10.80	77.50±7.91	93.75±2.20	78.33±23.64
8. Hay- windrowed	97.76±1.13	97.31±1.88	97.72±1.44	99.57±0.10	96.02±2.53
9. Oats	69.44±21.32	77.78±15.71	81.67±13.11	67.78±18.29	72.22±24.00
10. Soybean- notill	72.32±6.11	72.50±3.53	72.70±4.06	83.96±3.33	89.53±4.52
11. Soybean- mintill	82.74±2.75	81.69±2.91	84.11±2.92	92.67±2.89	95.03±1.30
12. Soybean- clean	74.89±3.66	73.07±3.28	78.95±5.43	79.04±4.79	86.04±4.15
13. Wheat	98.46±1.29	98.06±1.25	98.96±0.83	99.05±1.09	96.62±4.86
14. Woods	94.59±2.03	94.30±2.36	94.87±1.06	97.70±2.75	96.64±1.69
15. Building- grass-trees	50.08±5.37	49.42±5.86	51.72±6.33	89.17±4.83	89.11±4.92
16. Stone- steel-towers	89.22±5.59	89.11±5.73	89.44±5.52	98.00±0.88	92.78±5.84
OA (%)	81.85±0.83	81.40±1.11	83.43±0.74	91.20±0.48	92.94±0.87
AA (%)	78.55±2.70	79.19±1.72	81.47±1.86	90.10±0.95	89.64±2.33
Kappa (%)	79.28±0.96	78.78±1.25	81.06±0.87	89.96±0.54	91.95±1.00

Table 4 Classification accuracy in the Pavia University with PCA, LPP, EMAP and the proposed method (Best result of each row is marked in bold type)

Class	Baseline	PCA	LPP	EMAP	Proposed
Dimensionality	200	36	36	36	36
1. Asphalt	89.50±1.42	89.50±1.42	90.23±1.48	98.70±0.29	98.15±0.56
2. Meadows	96.48±0.65	96.48±0.65	96.74±0.30	99.11±0.29	99.89±0.06
3. Gravel	69.88±2.52	69.88±2.52	70.11±1.29	93.12±2.48	90.18±2.52
4. Trees	91.91±1.47	91.91±1.47	91.83±2.00	97.75±1.03	97.09±1.18
5. Painted metal sheets	99.37±0.34	99.37±0.34	99.40±0.30	99.69±0.15	99.87±0.16
6. Bare soil	73.97±2.61	73.97±2.61	75.00±1.27	94.51±1.39	99.33±0.36
7. Bitumen	80.19±2.80	80.19±2.80	80.97±3.00	92.94±1.46	96.12±1.96
8. Self-blocking bricks	85.26±2.54	85.26±2.54	86.56±1.71	96.73±1.21	94.75±0.81
9. Shadows	97.93±1.21	97.93±1.21	98.06±1.40	99.83±0.19	96.52±1.06
OA (%)	89.77±0.43	89.77±0.43	90.26±0.34	97.75±0.16	98.24±0.31
AA (%)	87.17±0.69	87.17±0.69	87.66±0.58	96.93±0.22	96.88±0.60
Kappa (%)	86.33±0.58	86.33±0.58	86.99±0.46	97.01±0.21	97.67±0.41

4.4. Comparison with 2D-SSA using all bands

In this section, we analyse the proposed dimensionality reduction algorithm and the original 2D-SSA from the two perspectives of data dimension reduction and the computation time. The original 2D-SSA is used for feature extraction for hyperspectral images using all bands. In our dimensionality reduction method, the 2D-SSA is applied for feature extraction on a part of bands selected by the proposed band selection method. The results are given as follows.

In Table 5 and Table 6, we analyse the differences in the classification accuracy of the feature extraction with entire bands and the feature extraction with selected bands. The numbers in parentheses indicate the numbers of the selected hyperspectral bands. The numbers of bands for the

Indian Pines and Pavia University datasets are 200 and 103, respectively. To compare the dimensionality reduction effects of different dimensions, 40 to 80 bands are tested and the dimensionality reduction performances are given with the increment of 10 dimensions each time.

In Fig. 2 and Fig. 3, we illustrate the classification accuracy and computation time of 2D-SSA with all bands and DPP+SSA with the increasing number of bands. The results of feature extraction with all bands are taken as baseline. For example, the line labelled 2D-SSA (200) in Fig.2 (a) and the labelled 2D-SSA (103) are taken as baseline. The time in the table contains all times for band selection, feature extraction and classification. It can be seen that, in Indian Pines, when the dimension exceeds 60, the

classification accuracy of our proposed algorithm is very close to or even slightly higher than the original 2D-SSA. On Pavia University, the experiment results of the two algorithms are similar in terms of classification accuracy, which means that the performance is very close. However, the dimension of dataset extracted by our method is lower and takes much less time. This shows that the features extracted by the original 2D-SSA are partially redundant, and our method can effectively remove the redundancy. The reasons why it takes less time are analysed as follows.

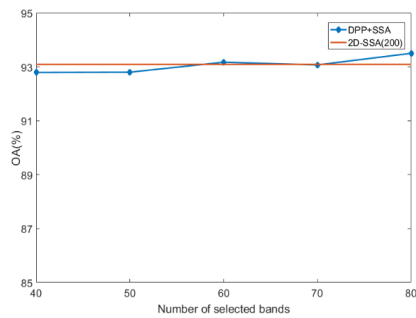
In band selection, the most time consuming is eigen-decomposition of similarity kernel matrix and the construction of adjacency matrix with mutual information. Therefore, we use the Dual-DPP to avoid the decomposition of high-dimensional matrices and reduce the computation time. Besides, we design a method for collecting the spatial representative pixel points with diversity and good coverage. This method avoids the estimation of mutual information on the entire hyperspectral dataset and effectively reduces the calculation time.

Table 5 Classification accuracy and computation time in the Indian Pines image of 2D-SSA and DPP+SSA (Best result of each row is marked in bold type)

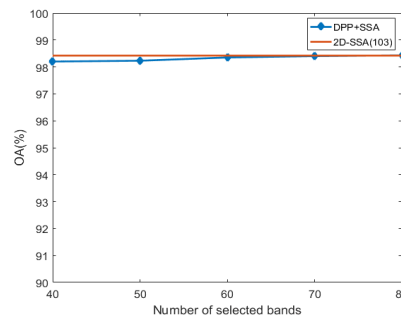
	2D-SSA (200)	DPP+SSA (40)	DPP+SSA (50)	DPP+SSA (60)	DPP+SSA (70)	DPP+SSA (80)
OA (%)	93.09±0.99	92.79±0.85	92.80±0.81	93.17±0.62	93.07±0.94	93.50±0.53
AA (%)	89.25±1.92	89.22±2.01	89.29±2.45	89.56±1.75	89.30±2.24	89.52±1.75
Kappa (%)	92.12±1.13	91.78±0.98	91.79±0.92	92.21±0.70	92.03±1.08	92.59±0.60
Time(s)	23.25	8.47	10.81	13.96	17.02	21.10

Table 6 Classification accuracy and computation time in the Pavia University image of 2D-SSA and DPP+SSA (Best result of each row is marked in bold type)

	2D-SSA (103)	DPP+SSA (40)	DPP+SSA (50)	DPP+SSA (60)	DPP+SSA (70)	DPP+SSA (80)
OA (%)	98.42±0.23	98.20±0.18	98.23±0.20	98.35±0.25	98.40±0.20	98.43±0.20
AA (%)	97.19±0.38	96.98±0.30	96.92±0.33	97.09±0.46	97.16±0.38	97.20±0.44
Kappa (%)	97.90±0.30	97.62±0.24	97.66±0.27	97.81±0.33	97.87±0.26	97.91±0.27
Time(s)	80.76	35.70	44.45	52.43	64.00	74.02



(a)



(b)

Fig.2. Classification accuracy in the Indian Pines image (a) and Pavia University image (b) of 2D-SSA and DPP+SSA

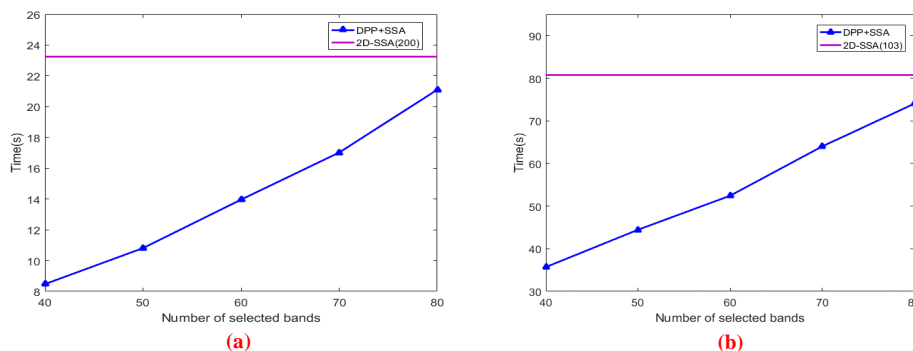


Fig.3. Computation time in the Indian Pines image (a) and Pavia University image (b) of 2D-SSA and DPP+SSA

5. Conclusion

In this paper, a new unsupervised dimensionality reduction algorithm is proposed for band selection of hyperspectral image. We make full use of the spatial information of hyperspectral image and construct multiple adjacency matrices to describe the correlation of spectral bands. In addition, we designed a fast spatial sampling method for hyperspectral, collecting some representative sample points, which can effectively reduce the amount of data calculation. An improved k-DPP is presented to sample a diversity and low-redundancy band subset from the multi adjacency matrices. Finally, duo to its ability of extracting features, 2D-SSA is employed and integrated with the improved k-DPP to build our unsupervised dimensionality reduction algorithm which can efficiently extract the spatial features of spectral images. The experimental results on two well-known hyperspectral datasets show that the proposed algorithm achieves higher classification accuracy with effectively compressing the data dimension.

Acknowledgements

This work is supported by the National Nature Science Foundation of China (no. 61471132) and the Training program for outstanding young teachers in higher education institutions of Guangdong Province (no. YQ2015057).

References

[1] W. Li, C. Chen, H. Su, and Q. Du.: ‘Local binary patterns and extreme learning machine for hyperspectral imagery classification,’ *IEEE Trans. Geosci. Remote Sens.*, 2015, 53, (7), pp. 3681-3693
 [2] M. Uzair, A. Mahmood, and A. Mian.: ‘Hyperspectral face recognition with spatiospectral information fusion and PLS regression,’ *IEEE Trans. Image Process.*, 2015, 24, (3), pp. 1127-1137
 [3] W. Zhao, S. Du.: ‘Spectral-Spatial Feature Extraction for Hyperspectral Image Classification: A Dimension Reduction and Deep Learning Approach,’ *IEEE Trans. Geosci. Remote Sens.*, 2016, 54, (8), pp.4544-4554

[4] G. Reed, K. Savage, D. Edwards, and N. N. Daeidet.: ‘Hyperspectral imaging of gel pen inks: An emerging tool in document analysis,’ *Sci. Justice*, 2014, 54, (1), pp. 71-80
 [5] P. Y. Sacre et al.: ‘Data processing of vibrational chemical imaging for pharmaceutical applications,’ *J. Pharmaceutical Biomed. Analy.*, 2014, 101, pp. 123-140
 [6] N. Neittaanmäki-Perttu et al.: ‘Detecting field cancerization using a hyperspectral imaging system,’ *Lasers Surgery Med.*, 2013, 45, (17), pp. 410-417
 [7] T. Kelman, J. Ren, and S. Marshall: ‘Effective classification of Chinese tea samples in hyperspectral imaging,’ *Artif. Intell. Res.*, 2013, 2, (4), pp. 87-96
 [8] Y. Yuan, X. Zheng, X. Lu.: ‘Discovering Diverse Subset for Unsupervised Hyperspectral Band Selection,’ *IEEE Trans. Image Process.*, 2017, 26, (1), pp. 51-64
 [9] J. Feng, L. Jiao, F. Liu, T. Sun, X. Zhang.: ‘Unsupervised feature selection based on maximum information and minimum redundancy for hyperspectral images,’ *Pattern Recognit*, 2016, 51, pp. 295-309
 [10] J. Feng, L. Jiao, F. Liu, T. Sun, and X. Zhang.: ‘Mutual-information based semi-supervise hyperspectral band selection with high discrimination, high information, and low redundancy’, *IEEE Trans. Geosci. Remote Sens.*, 2015, 53, (5), pp. 2956-2969
 [11] X. Geng, K. Sun, L. Ji, and Y. Zhao, ‘A fast volume-gradient-based band selection method for hyperspectral image’, *IEEE Trans. Geosci. Remote Sens.*, 2014, 52, (11), pp. 7111-7119
 [12] J. Feng, L. C. Jiao, X. Zhang, and T. Sun. ‘Hyperspectral band selection based on trivariate mutual information and clonal selection’, *IEEE Trans. Geosci. Remote Sens.*, 2014, 52, (7), pp. 4092-4105
 [13] X.H. Cao, C.C. Wei, J.G. Han, L.C. Jiao.: ‘Hyperspectral Band Selection Using Improved Classification Map,’ *IEEE Geosci. Remote Sens. Lett.*, 2017, 14(11), pp.2147-2151
 [14] X.H. Cao, J.G. Han, S.Y. Yang, D.C. Tao, L.C. Jiao.: ‘Band selection and evaluation with spatial information,’ *International Journal of Remote Sensing.*, 2016, 37(19), pp.4501-4520

- [15] C. Yao, Y.F. Liu, B. Jiang, J.G. Han, J.W. Han.: 'LLE Score: A New Filter-Based Unsupervised Feature Selection Method Based on Nonlinear Manifold Embedding and Its Application to Image Recognition,' *IEEE Trans. Image Process.*, 2017, 26(11), pp.5257-5269
- [16] J. M. Sotoca, F. Pla.: 'Supervised feature selection by clustering using conditional mutual information-based distances', *Pattern Recognit.* 2010, 43, (6), pp. 2068-2081
- [17] H. Yang, Q. Du, H. Su, and Y. Sheng.: 'An efficient method for supervised hyperspectral band selection', *IEEE Geosci. Remote Sens. Lett.*, 2011, 8, (1), pp. 138-142
- [18] C. Sui, Y. Tian, Y. Xu, and Y. Xie.: 'Unsupervised band selection by integrating the overall accuracy and redundancy'. *IEEE Geosci. Remote Sens. Lett.*, 2015, 2, (1), pp. 185-189
- [19] C. Wang, M. Gong, M. Zhang, and Y. Chan.: 'unsupervised hyperspectral image band selection via column subset selection.' *IEEE Geosci. Remote Sens. Lett.*, 2015, 12, (7), pp. 1411-1415
- [20] M. Zhang, J. Ma, and M. Gong.: 'Unsupervised Hyperspectral Band Selection by Fuzzy Clustering With Particle Swarm Optimization', *IEEE Geosci. Remote Sens. Lett.*, 2017, 14, (5), pp.773-777
- [21] C. Gong, D. Tao, S. J. Maybank, W. Liu, G. Kang, and J. Yang.: 'Multimodal curriculum learning for semi-supervised image classification', *IEEE Trans. Image Process.*, 2016, 25, (7), pp. 3249-3260
- [22] J. Zabalza, J. Ren, J. Zheng, J. Han, H. Zhao, S. Li, S. Marshall.: 'Novel Two-Dimensional Singular Spectrum Analysis for Effective Feature Extraction and Data Classification in Hyperspectral Imaging,' *IEEE Trans. Geosci. Remote Sens.*, 2015, 53, (8), pp. 4418-4433
- [23] J. Zabalza, J. Ren, J. Zheng, H. Zhao, C. Qing, Z. Yang, P. Du, S. Marshall.: 'Novel segmented stacked autoencoder for effective dimensionality reduction and feature extraction in hyperspectral imaging,' *Neurocomputing.*, 2016, 185, pp:1-10
- [24] L. Lei, S. Prasad, J.E. Fowler, L.M. Bruce.: 'Locality-preserving dimensionality reduction and classification for hyperspectral image analysis,' *IEEE Trans. Geosci. Remote Sens.*, 2012, 50, (4), pp. 1185-1198
- [25] B.C. Kuo, C.H. Li, J.M. Yang.: 'Kernel nonparametric weighted feature extraction for hyperspectral image classification', *IEEE Trans. Geosci. Remote Sens.*, 2009, 47, (4), pp. 1139-1155
- [26] Xiaofei He and Partha Niyogi.: 'Locality Preserving Projections', in *NIPS*, 2003.
- [27] M. Belkin and P. Niyogi.: 'Laplacian eigenmaps and spectral techniques for embedding and clustering,' in *Proc. Av. Neural Inf. Process. Syst.*, 2001, pp. 585-591.
- [28] Y.C. Guo, G.G. Ding, L. Li, J.G. Han, L. Shao.: 'Learning to Hash With Optimized Anchor Embedding for Scalable Retrieval,' *IEEE Trans. Image Process.*, 2017, 26(3), pp.1344-1354
- [29] G. Licciardi, P. R. Marpu, J. Chanussot, and J. A. Benediktsson.: 'Linear versus nonlinear pca for the classification of hyperspectral data based on the extended morphological profiles,' *IEEE Geosci. Remote Sens. Lett.*, 2012, 9, (3), pp.447-451
- [30] C. M. Bachmann, T. L. Ainsworth, and R. A. Fusina.: 'Exploiting manifold geometry in hyperspectral imagery,' *IEEE Trans. Geosci. Remote Sens.*, 2005, (43), (3), pp. 441-454
- [31] B. Demir and S. Ertürk.: 'Empirical mode decomposition of hyperspectral images for support vector machine classification,' *IEEE Trans. Geosci. Remote Sens.*, 2010, 48, (11), pp. 4071-4084
- [32] A. Kulesza and B. Taskar.: 'Determinantal Point Processes for Machine Learning", *Foundations and Trends in Machine Learning.* 2012, 5, (2-3), pp.123-286
- [33] A. Kulesza and B. Taskar: 'k-DPPs: Fixed-size determinantal point Processes'. in *Proc. Int. Conf. on Machine Learning.*, 2011, pp. 1193-1200
- [34] O. Macchi.: 'The coincidence approach to stochastic point processes,' *Advances in Applied Probability.*, 1975, 7, (1), pp. 83-122, 1975
- [35] T. Cover and P. Hart.: 'Nearest neighbor pattern classification.' *IEEE Trans. Inf. Theory.*, 1967,13, (1), pp. 21-27
- [36] P. Qiu, A. J. Gentles, S. K. Plevritis.: 'Fast calculation of pairwise mutual information for gene regulatory network reconstruction', *Comput. Methods Programs Biomed.* 2009, 94, (2), pp.177-180
- [37] P. R. Marpu, M. Pedergrana, M. D. Mura, J. A. Benediktsson, and L. Bruzzone.: 'Automatic generation of standard deviation attribute profiles for spectral-spatial classification of remote sensing data,' *IEEE Geosci. Remote Sens. Lett.*, 2013.10,(2),pp. 293-297,
- [38] R. O. Green, M. L. Eastwood, C. M. Sarture, T. G. Chrien, M. Aronsson, B. J. Chipendale, J. A. Faust, B. E. Pavri, C. J. Chovit, M. Solis, M. R. Olah, O. Williams.: 'Imaging spectroscopy and the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS).' *Remote. Sens. Environ.* 1998, 65, (3), pp. 227-248
- [39] S. Holzwarth, A. Müller, M. Habermeyer, R. Richter, A. Hausold, S. Thiemann, P. Strobl, Sens - DAIS 7915 / ROSIS Imaging Spectrometers at DLR, in: 3rd EARSeL Workshop on Imaging Spectroscopy, Herrsching, 13-16 May 2003, p. 12
- [40] Y. Bazi, F. Melgani.: 'Toward an optimal SVM classification system for hyper-spectral remote sensing images,' *IEEE Trans. Geosci. Remote Sens.*, 2006, 44, (11) pp. 3374-3385
- [41] G.M. Foody.: 'Status of land cover classification accuracy assessment,' *Remote Sens. Environ.*, 2002, 80, pp.185-201