A Novel Geometric Mean Feature Space Discriminant Analysis Method for Hyperspectral Image Feature Extraction

Li Li[1,2] · Hongwei Ge[1,2] · Jianqiang Gao[3] · Yixin Zhang[4] · Yubing Tong[5] · Jun Sun[1,2]

Abstract

Hyperspectral image contains abundant spectral information with hundreds of spectral continuous bands that allow us to distinguish different classes with more details. However, the number of available training samples is limited and the high dimensionality of hyperspectral data increases the computational complexity and even also may degrade the classification accuracy. In addition, the bottom line is that only original spectral is difficult to well represent or reveal intrinsic geometry structure of the hyperspectral image. Thus, feature extraction is an important step before classification of high dimensional data. In this paper, we proposed a novel supervised feature extraction method that uses a new geometric mean vector to construct geometric between-class scatter matrix (S_b^G) and geometric within-class scatter matrix (S_w^G) instead of traditional mean vector of state-of-the-art methods. The geometric mean vector not only can reveal intrinsic geometry structure of the hyperspectral image, but also can improve the ability of learning nonlinear correlation features by maximum likelihood classification (MLC). The proposed method is called geometric mean feature space discriminant analysis (GmFSDA) that uses three measures to produce the extracted features. GmFSDA, at first, maximizes the geometric between-spectral scatter matrix to increase the difference between extracted features. In the second step of GmFSDA, maximizes the between-class scatter and minimizes the within-class scatter simultaneously. The experimental results on three real-world hyperspectral image datasets show the better performance of GmFSDA in comparison with other feature extraction methods in small sample size situation by using MLC.

Keywords Hyperspectral image · Feature extraction · Geometric mean vector · Feature space discriminant analysis · Classification

Mathematics Subject Classification 68T10 · 68U10

1 Introduction

Hyperspectral image is a cube data containing abundant spectral information with hundreds of contiguous spectral bands. The spectral response curve of each pixel is the plot of bands intensity values versus band numbers in the hyperspectral image and allow us to better

---

distinguish many subtle objects and materials. Therefore, in the practical application field of land cover classification, hyperspectral imagery technology has been widely applied. However, there are many challenges for hyperspectral image analysis, such as, the number of available training samples is limited and the high dimensionality of hyperspectral data increases the computational complexity and also may degrade the classification accuracy. In fact, hyperspectral imaging can provide superior capability to discriminate land-cover types than multispectral does. But, with increasing the number of spectral bands of hyperspectral image, the classification accuracy will change greatly [1]. Firstly, the classification accuracy increases slowly and reaches the maximum. And then, it decreases dramatically when a small number of training samples are available. Generally speaking, feature reduction method can relieve the Hughes phenomenon and achieve higher classification accuracy in image classification tasks. Feature reduction method is used to find a transformation that maps the original data to a lower dimensional space where the essential discriminative information can be mainly preserved. The feature reduction method can be broadly divided into two groups: feature selection and feature extraction [2–11]. For feature selection methods, there is no new features will be generated and only a subset of original features is selected. Inspired by the clonal selection theory in an artificial immune system, Zhang et al. [11] gave a stochastic search skill. In this skill, the dimensionality reduction method is formulated as an optimization problem that searches an optimum with less features in a feature space. Feature extraction mainly aims at reducing the dimensionality of original data while keeping as much intrinsic information as possible, and also exploring the inherent low-dimensional structure, reducing the computational complexity, and improving the performance of data analysis. Generally speaking, features selected by feature selection techniques maintain the physical meaning of original data while features obtained by feature extraction methods are discriminative than those selected by feature selection techniques. The authors [7] have been pointed out that the feature extraction techniques can provide more effective features than by using feature selection techniques. Therefore, feature extraction is a very important preprocessing step for hyperspectral image classification.

There are three feature extraction techniques in hyperspectral image processing field, such as unsupervised, supervised and semi-supervised. Unsupervised feature extraction methods mainly focus on another representation of the original data in a lower dimensional subspace, satisfying some criterions, and usually do not concern class discrimination, and without the need for labeled samples. Principle component analysis (PCA) [12] is one of the best known unsupervised feature extraction method and widely used for hyperspectral image [1,13,14]. The main idea of PCA is to project the original data into a new subspace by minimizing the reconstruction error in the mean squared sense. The nonlinear version of PCA has shown more effective than PCA in hyperspectral image analysis in addition to the computational burden [15]. Villa et al. [16] proposed an unsupervised classification method for hyperspectral image with low spatial resolution. Due to the complicated and high-dimensional data observation, the hyperspectral remote sensing image unsupervised classification still leave huge challenges, Zhang et al. [42] proposed a hyperspectral image unsupervised classification framework based on robust manifold matrix factorization. To solve the high dimensionality of the hyperspectral image, Zhang et al. gave a unified low-rank matrix factorization to jointly perform the dimensionality reduction and data clustering.

Supervised feature extraction methods main rely on the existence of labeled samples to infer class separability. Linear discriminant analysis (LDA) [17] and nonparametric weighted feature extraction (NWFE) [18] are the most popular supervised feature methods and the two methods have been widely used in hyperspectral image classification. As we all know, in LDA method, the between-class scatter matrix (S_b) is maximized and the within-class scatter

matrix (S_w) is minimized simultaneously. But, there are three inherent drawbacks with LDA. (i) The performance of LDA will be very poor when the S_w is singular, especially in high-dimensional data. (ii) LDA can extract maximum c − 1 features (where c is the number of classes), which may not be sufficient to represent essential information of the original data in hyperspectral image classification. (iii) LDA works very poor when the classes have non normal-like or multi-modal mixture distributions. NWFE uses the weighted mean for calculation of nonparametric scatter matrices for extraction of more than c − 1 features.

Kernel principal component analysis (KPCA) [19] and generalized discriminant analysis (GDA) [20] are the nonlinear forms of PCA and LDA by using kernel trick (mapping the input data from the original space to a convenient feature space where inner products in the feature space can be computed by a kernel function). However, the computational burden of nonlinear representations (KPCA and GDA) is very heavy.

Semi-supervised feature extraction methods need labeled samples and unlabeled samples [21], and try to find a good projection. This method can preserve certain potential properties of the original data by adding a regularization term. In addition, there are some local feature extraction methods have been proposed to preserve the properties of local neighborhoods of hyperspectral images, such as, locality preserving projection (LPP) [22], linear local tangent space alignment (LLTSA) [23], maximum margin projection (MMP) [24] and monogenic binary coding (MBC) [25] with face representation and recognition. Zhou et al. [26] extracted the discriminant information from the filtered image by using a spectral-domain regularized local discriminant embedding and a spatial-domain local pixel neighborhood preserving embedding. Then the spatial-spectral discriminant information is incorporated to produce multi-scale spatial-spectral features. Of course, in real-world applications, the precise assessment of the global or local structures of the original data manifold is very difficult. Hence, all kinds of hybrid criterions are used to process the hyperspectral image such as [27–32]. In addition, there are some research results based on multiple features, such as [43–47]. In these papers, the authors gave a unified low-dimensional representation of these multiple features for subsequent classification.

Recently, from a curve fitting point of view, the geometric aspects of spectral response curve and the rich feature information of original data has been considered in hyperspectral image classification by Hosseini et al. [2,3,33,34] and Li et al. [35]. In these methods, the reflection coefficients vector from a hyperspectral image have been used in classification process. Hosseini considered the elements of the feature vector are the points of a curve by using rational function curve fitting (RFCF). Li et al. extracted the feature of hyperspectral image by using Maclaurin series function curve fitting (MFCF). In [35], Li et al. pointed out the RFCF feature extraction has three main shortcomings. The RFCF and MFCF two feature extraction methods do not need to transform the data to the projection subspace. But, the two feature extraction methods not whole reveal inherent geometric structure of hyperspectral image, because the S_b and S_w (from the two methods) are formed by using sample average. In other words, the two methods do not reveal the inherent geometric structure in essence, especially the existence of outliers in hyperspectral image. Imani et al. [36] proposed a feature space discriminant analysis (FSDA) to produce the extracted features for hyperspectral image. The main idea of FSDA method is reduce the redundant spectral information in the extracted features by introducing the between-spectral scatter based on class sample average vector of hyperspectral image. In fact, for a few of given samples with non-ideal conditions, the assessment result and the inherent geometric structure in essence are very weak by using the class sample average vector. Therefore, the classification performance of FSDA method with MLC will decline significantly. In order to further reveal the inherent geometric structure and obtain the more robust features from hyperspectral image, we proposed a supervised feature

Table 1 The information of each class for IPS image

No	Class name	Number of pixels	No	Class name	Number of pixels
C1	Alfalfa	46	C9	Oats	20
C2	Corn-no till	1428	C10	Soybeans-no till	972
C3	Corn-min till	830	C11	Soybeans-min till	2455
C4	Corn	237	C12	Soybeans-clean till	593
C5	Hay-windowed	483	C13	Wheat	205
C6	Grass/trees	730	C14	Woods	1265
C7	Grass/pasture-mowed	28	C15	Building-grass-tree-drives	386
C8	Grass/pasture	478	C16	Stone-steel towers	93

extraction method in this paper that has good performance using small training set. The proposed method uses a geometric mean vector to construct geometric between-class scatter matrix (S_b^G) and geometric within-class scatter matrix (S_w^G) instead of traditional mean vector of state-of-the-art methods. Hence, the proposed method is called geometric mean feature space discriminant analysis (GmFSDA).

Let $\mathcal{X} = \{x_i : x_i \in \mathbb{R}^d\}_{i=1}^N$ be the training dataset with geometric mean **m** and belonging to c classes. Each class has a geometric mean m_i of n_i data points where $N = \sum_{i=1}^c n_i$. For extraction of p features from $d \times 1$ original feature vector (**x**), a transformation matrix **A** is used. The extracted feature vector will be: $\mathbf{y}_{p \times 1} = \mathbf{A}_{p \times d} \mathbf{x}_{d \times 1}$. In other words, GmFSDA introduce the geometric between-spectral scatter and maximizes it to reduce the redundant spectral information in the extracted features. Similarly to FSDA method, GmFSDA is a two steps method, at first, uses a primary projection to make features as different from each other as possible. After that, maximizes the between-class scatter and minimizes the within-class scatter by using another projection matrix simultaneously.

The rest of this paper is organized as follows. In Sect. 2, we introduce a new feature extraction method: geometric mean feature space discriminant analysis (GmFSDA). Experimental results and discussion are shown in Sect. 3. Finally the conclusion of this paper is listed in Sect. 4.

2 Geometric Mean Feature Space Discriminant Analysis (GmFSDA)

2.1 The Basic Idea of GmFSDA

In this subsection, extracting discriminative spectral features and building powerful representation framework for hyperspectral image are very important with a few of given samples. There are some popular feature extraction methods just only use the class discrimination for feature extraction, such as PCA, KPCA, MMP, NWFE, RFCF and MFCF. In our proposed GmFSDA method, we further consider the difference between spectral bands in the transformed fe ature space in addition to separability between classes. In other words, there are three main purposes of GmFSDA. (i) The intrinsic geometry structure of the hyperspectral image can be revealed exactly and the GmFSDA algorithm can learn nonlinear correlation features with well discriminating power by MLC. (ii) The produced features are as different from each other as possible. (iii) Separability between classes is increased. The geometric mean
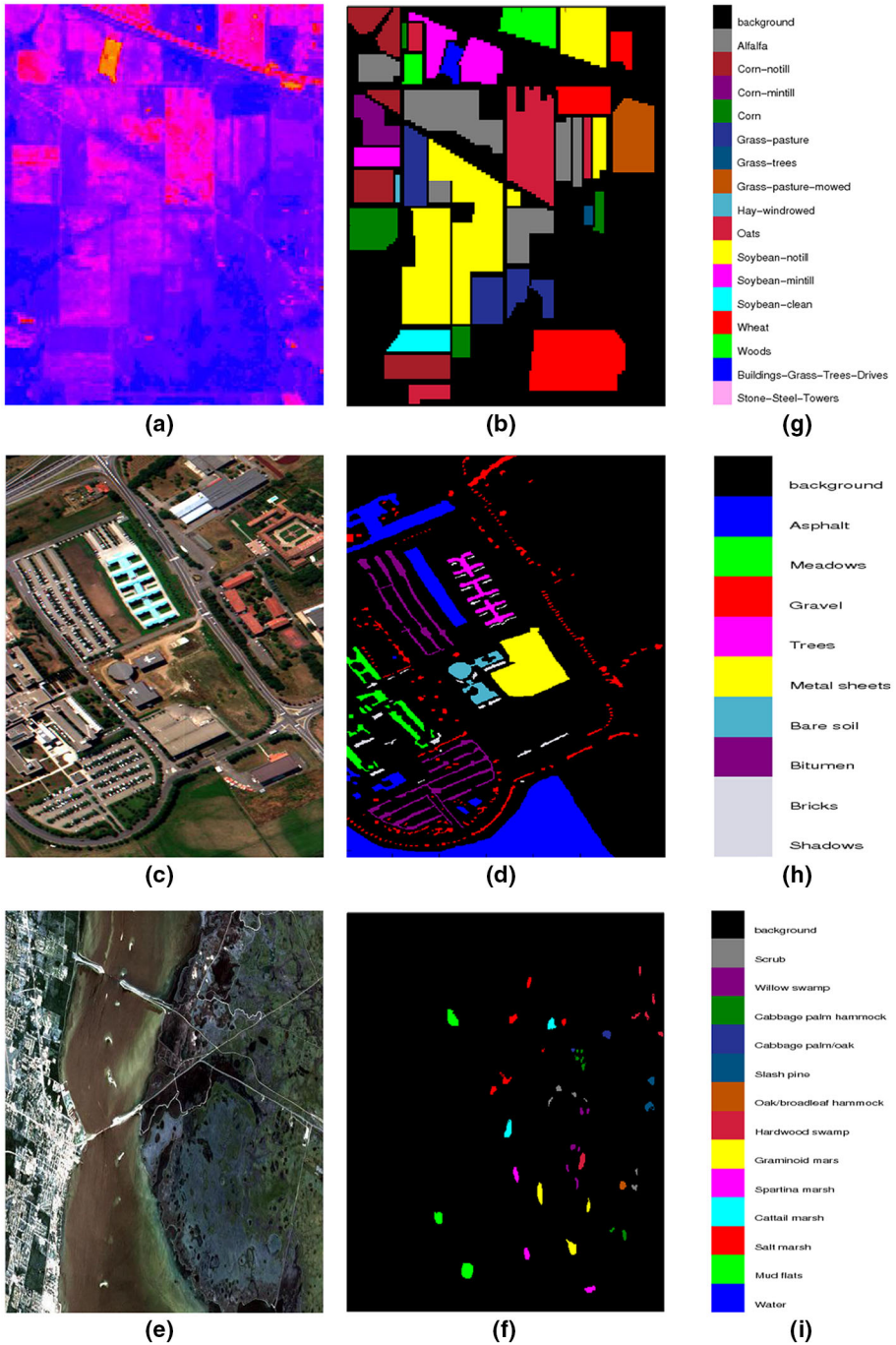
Fig. 1 **a** False color image of IPS, **b** ground truth map (GTM) of IPS, **c** false color image of Pavia University, **d** GTM of Pavia University, **e** false color image of KSC, **f** GTM of KSC, **g** class name of IPS, **h** class name of Pavia University, **i** class name of KSC

Table 2 The information of each class for Pavia University image

No	Class name	Number of pixels	No	Class name	Number of pixels
C1	Asphalt	6852	C6	Bare soil	5104
C2	Meadow	18,686	C7	Bitumen	1356
C3	Gravel	2207	C8	Bricks	3878
C4	Tree	3436	C9	Shadow	1026
C5	Metal sheet	1378			

Table 3 The information of each class for KSC image

No	Class name	Number of pixels	No	Class name	Number of pixels
C1	Scrub	761	C8	Graminoid marsh	431
C2	Willow	243	C9	Spartina marsh	520
C3	CP Hammock	256	C10	Cattail marsh	404
C4	CP/oak	252	C11	Salt marsh	419
C5	Slash pine	161	C12	Mud flats	503
C6	Oak/broadleaf	229	C13	water	927
C7	Hardwood swamp	105			

Table 4 Parameters of feature extraction methods in the experiments

Method	Parameters	Method	Parameters
PCA	–	RFCF	$L = 0, 1, \ldots, D - 1, M = D - L - 1$
KPCA	–	MFCF	$L = 1, 2, \ldots, D$
MMP	G	FSDA	–
NWFE	$p = 0.5$	GmFSDA	–

of n positive numbers $x_1, x_2, \ldots, x_n$ is defined as positive n-th root of their product [37]:

$$x_G = \sqrt[n]{x_1 \cdot x_2 \cdot \cdots \cdot x_n} = \left(\prod_{i=1}^{n} x_i \right)^{\frac{1}{n}} \tag{1}$$

It is important to emphasize that the calculation of geometric mean is either insoluble or meaning less if a data set contains negative numbers. For instance, the geometric mean of two numbers 4 and 9, is the square root of their product, i.e., $x_G = \sqrt[2]{4 \cdot 9} = 6$.

Let m_{ik}^{G} be the geometric mean of class k and spectral feature i, the number of classes and the number of features (spectral bands) are c and d, respectively. The geometric mean of classes in d dimensions is as follows:

$$\begin{array}{ccccc|c}
m_{11}^G & m_{12}^G & m_{13}^G & \cdots & m_{1c}^G & \mathbf{h}_1 \\
m_{21}^G & m_{22}^G & m_{23}^G & \cdots & m_{2c}^G & \mathbf{h}_2 \\
\vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\
m_{d1}^G & m_{d2}^G & m_{d3}^G & \cdots & m_{dc}^G & \mathbf{h}_d \\
\hline
\mathbf{M}_1 & \mathbf{M}_2 & \mathbf{M}_3 & \cdots & \mathbf{M}_c &
\end{array} \tag{2}$$

where the geometric mean m_{ik}^G as following Eq. (3), n_k is the number of samples per class, x_{ikj} is the j-th training sample of class k in i-th band.

$$m_{ik}^G = \left(\prod_{j=1}^{n_k} x_{ikj} \right)^{\frac{1}{n_k}}, \quad (i = 1, 2, \ldots, d; k = 1, 2, \ldots, c) \tag{3}$$

we can observe the difference between classes from the above matrix vertically, and meanwhile we can observe the difference between spectral features from the above matrix horizontally [formula (2)]. In the first step of GmFSDA, we maximize the difference between spectral features and in the second step, we maximize the difference between classes. Therefore, two types of vectors can be obtained from the above matrix.

$$\mathbf{h}_i = \left[m_{i1}^G, m_{i2}^G, \ldots, m_{ic}^G \right]^T, \quad (i = 1, 2, \ldots, d) \tag{4}$$

$$\mathbf{M}_k = \left[m_{1k}^G, m_{2k}^G, \ldots, m_{dk}^G \right]^T, \quad (k = 1, 2, \ldots, c) \tag{5}$$

Hence, $\mathbf{h}_i$ is the geometric mean of c classes in ith dimension and $\mathbf{M}_k$ is the geometric mean of class k in d dimensions. That is to say, $\mathbf{h}_i$ is the representative of feature (band) i and $\mathbf{M}_k$ is the representative of class k. In order to produce features as different from each other as possible and increase the class discrimination in GmFSDA method, we must maximize the geometric between-spectral scatter and maximize the between-class scatter and then obtain the projection matrix. The geometric between-spectral scatter matrix (S_f^G) can be calculated by following Eq. (6).

$$S_f^G = \sum_{i=1}^d (\mathbf{h}_i - \bar{\mathbf{h}})(\mathbf{h}_i - \bar{\mathbf{h}})^T \tag{6}$$

where

$$\bar{\mathbf{h}} = \frac{1}{d} \sum_{i=1}^d \mathbf{h}_i \tag{7}$$

$(\mathbf{h}_i)_{c \times 1}$ is a column vector and S_f^G is a $c \times c$ matrix. In the proposed GmFSDA method, the original feature space can be transformed to a new feature space by using the obtained a primary projection matrix $W \in \mathbb{R}^{c \times c}$ such that features have more difference from each other in the transformed space. We maximize the $tr(S_f^G)$ and sort the eigenvalues of S_f^G descending, then the projection matrix W can be obtained by using the eigenvectors associated with theses eigenvalues. Hence, in the transformed space, we can obtain the transformed vectors $\mathbf{g}_i \in \mathbb{R}^{c \times 1}$ $(i = 1, 2, \ldots, d)$ as follows:

$$\mathbf{g}_i = W \mathbf{h}_i, \quad (i = 1, 2, \ldots, d) \tag{8}$$

The class geometric mean matrix in formula (2) is transformed to formula (9) with the above transformations.

$$
\begin{array}{c}
\left.\begin{array}{ccccc}
r_{11}^G & r_{12}^G & r_{13}^G & \cdots & r_{1c}^G \\
r_{21}^G & r_{22}^G & r_{23}^G & \cdots & r_{2c}^G \\
\vdots & \vdots & \vdots & \cdots & \vdots \\
r_{d1}^G & r_{d2}^G & r_{d3}^G & \cdots & r_{dc}^G \\
\end{array}\right| \begin{array}{c} \mathbf{g}_1 \\ \mathbf{g}_2 \\ \vdots \\ \mathbf{g}_d \end{array} \\
\hline
\begin{array}{ccccc} \mathbf{R}_1 & \mathbf{R}_2 & \mathbf{R}_3 & \cdots & \mathbf{R}_c \end{array}
\end{array}
\tag{9}
$$

The vectors $\mathbf{h}_i$ and $\mathbf{M}_k$ are transformed into vectors $\mathbf{g}_i$ and $\mathbf{R}_k$ in terms of formula (9).

$$
\mathbf{g}_i = [r_{i1}^G, r_{i2}^G, \ldots, r_{ic}^G]^T, \quad (i = 1, 2, \ldots, d)
\tag{10}
$$

$$
\mathbf{R}_k = [r_{1k}^G, r_{2k}^G, \ldots, r_{dk}^G]^T, \quad (k = 1, 2, \ldots, c)
\tag{11}
$$

In fact, we know that the difference between spectral bands is increased in the obtained feature space from formula (9). Therefore, in the second step of GmFSDA, we maximize the between-class scatter matrix (S_b) and increase the class discriminant in the obtained feature space. The S_b can be calculated as follows:

$$
S_b = \sum_{k=1}^{c} (\mathbf{R}_k - \overline{\mathbf{R}})(\mathbf{R}_k - \overline{\mathbf{R}})^T
\tag{12}
$$

where

$$
\overline{\mathbf{R}} = \frac{1}{c} \sum_{k=1}^{c} \mathbf{R}_k
\tag{13}
$$

Similar to above analysis, we maximize $tr(S_b)$. For extraction of p features from the original $d \times 1$ feature vector ($\mathbf{x}$), the projection matrix A can be obtained by using p eigenvectors of S_b associated with p largest eigenvalues.

$$
\mathbf{y}_{p \times 1} = A_{p \times d} \mathbf{x}_{d \times 1}
\tag{14}
$$

2.2 The Effective of Geometric Mean

In probability theory and statistics, the geometric mean of a set of n positive numbers is defined as Eq. (1). As sample average, the geometric mean [37] can also be used to accurately estimate the central tendency. Generally speaking, the geometric mean is more resistant to outliers (or skewed data). We can see the effective of geometric mean by following example. Suppose we have Data $= [4.3, 4.0, 20, 4.1, 1, 4.2, 4.4]$ with the outliers "1" and "20" then the average value $x_m = 6$ and the geometric mean $x_G \approx 4.274$. So, we observe that 4.274 is more closer to the central tendency ($\frac{4+4.1+4.2+4.3+4.4}{5} = 4.2$) than 6. Therefore, the geometric mean is more effective than average value.

2.3 Development and Improvement of GmFSDA

In order to improve the classification accuracy, we proposed to minimize the within-class scatter matrix in addition to maximize the between-class scatter matrix simultaneously in the second step of GmFSDA for class discrimination. But, we only can extract maximum
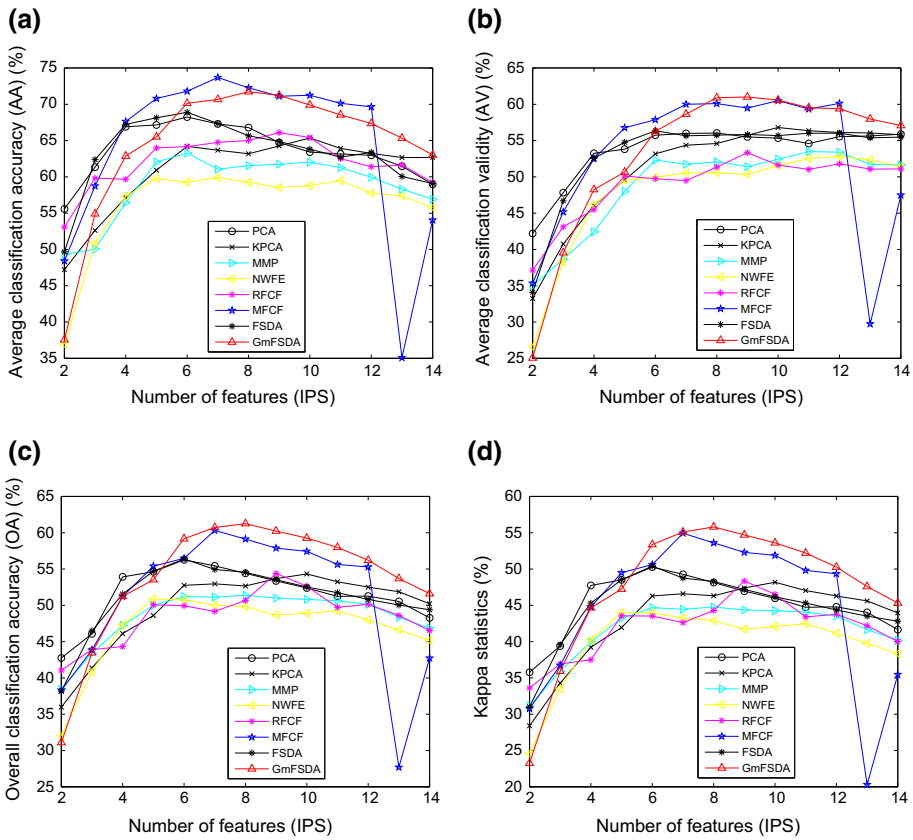
Fig. 2 Classification accuracy measures for IPS dataset using different feature extraction methods and various number of features. **a** AA, **b** AV, **c** OA, **d** kappa coefficient

$c - 1$ features from data, because the bank of S_b of Eq. (12) is limited. So, we also gave a nonparametric geometric mean between-class scatter matrix that allows us to extract more than $c - 1$ features. In hyperspectral image process, we use the training samples of classes instead of the geometric mean of classes in the improved GmFSDA. So, the training sample matrix is as follows:

$$X_j = \begin{bmatrix} x_{11j} & x_{12j} & \dots & x_{1cj} \\ x_{21j} & x_{22j} & \dots & x_{2cj} \\ \vdots & \vdots & \dots & \vdots \\ x_{d1j} & x_{d2j} & \dots & x_{dcj} \end{bmatrix}, \quad (j = 1, 2, \dots, n_k) \tag{15}$$

where x_{ikj} ($i = 1, 2, \dots, d; k = 1, 2, \dots, c; j = 1, 2, \dots, n_k$) is the j-th training sample of class k in i-th band. n_k is the number of training samples each class. So, in this case, the vector $\mathbf{h}_{ij}$ is defined as follows:

$$\mathbf{h}_{ij} = [x_{i1j}, x_{i2j}, \dots, x_{icj}]^T, \quad (i = 1, 2, \dots, d; j = 1, 2, \dots, n_k) \tag{16}$$

According to aforementioned of Sect. 2.1, $\mathbf{h}_{ij}$ must contain a representative from each class and the representative of each class is the geometric mean of it. Of course, when we want to
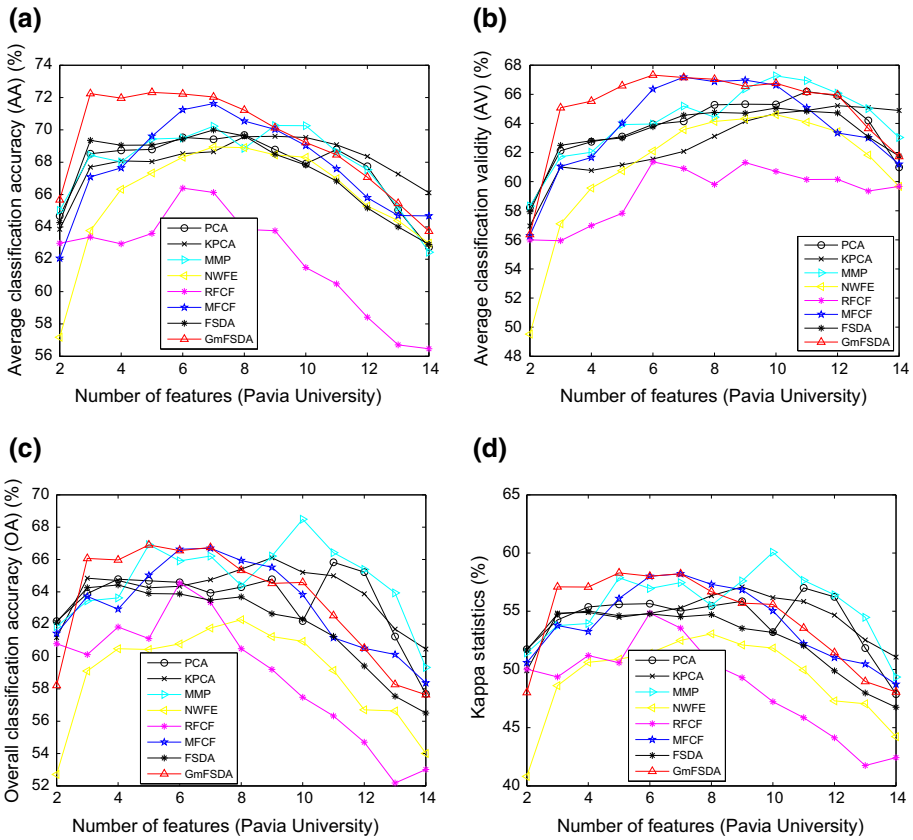
(a)

(b)

(c)

(d)

Fig. 3 Classification accuracy measures for Pavia University dataset using different feature extraction methods and various number of features. **a** AA, **b** AV, **c** OA, **d** kappa coefficient

use the training samples of classes instead of the geometric mean of them, the representative of each class is a training sample from that class. That is to say, all classes must be involved in the composing of $\mathbf{h}_{ij}$. In our paper, the same number of training samples must be used in all classes. Of course, the minimize number of training samples per class can be obtained by using $n_t = min\{n_k\}_{k=1}^c$. According to Sect. 2.1, in the first step of GmFSDA, the geometric between-spectral scatter matrix (S_f^G) is calculated as follows:

$$S_f^G = \sum_{j=1}^{n_t} \sum_{i=1}^{d} (\mathbf{h}_{ij} - \overline{\mathbf{h}}_j)(\mathbf{h}_{ij} - \overline{\mathbf{h}}_j)^T \tag{17}$$

where $\overline{\mathbf{h}}_j = \frac{1}{d} \sum_{i=1}^{d} \mathbf{h}_{ij}$. The projection matrix $W \in \mathbb{R}^{c \times c}$ can be obtained by maximizing $tr(S_f^G)$. The original feature space of hyperspectral image is transformed into a new feature space by using obtained projection matrix W such that the spectral bands have more difference from each other in it. In other words, the spectral scatter is increased in the transformed feature space with using following transformation:

$$(\mathbf{g}_{ij})_{c \times 1} = W\mathbf{h}_{ij}, \quad (i = 1, 2, \ldots, d; \ j = 1, 2, \ldots, n_t) \tag{18}$$
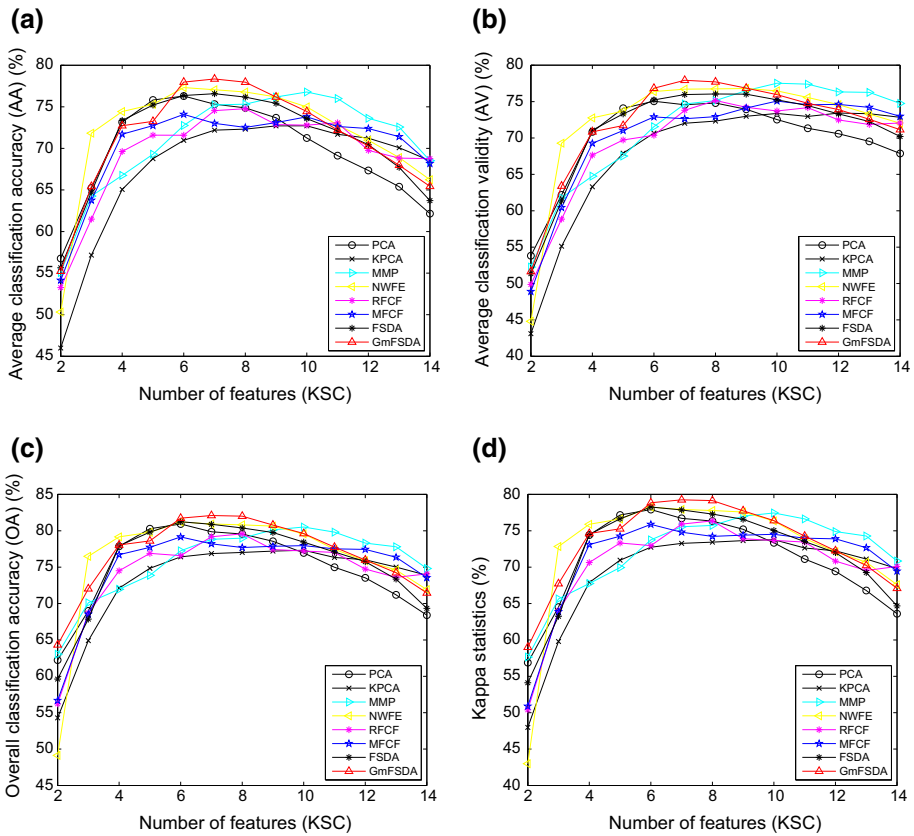
Fig. 4 Classification accuracy measures for KSC dataset using different feature extraction methods and various number of features. **a** AA, **b** AV, **c** OA, **d** kappa coefficient

The sample matrix X_j ($j = 1, 2, \ldots, n_t$) in (15) is transformed to (19) as follows by using the above transformations.

$$
R_j = \begin{bmatrix} r_{11j}^G & r_{12j}^G & \cdots & r_{1cj}^G \\ r_{21j}^G & r_{22j}^G & \cdots & r_{2cj}^G \\ \vdots & \vdots & \cdots & \vdots \\ r_{d1j}^G & r_{d2j}^G & \cdots & r_{dcj}^G \end{bmatrix}, \quad (j = 1, 2, \ldots, n_k)
\tag{19}
$$

The vectors $\mathbf{g}_{ij}$ and $\mathbf{R}_{kj}$ are defined as follows:

$$
\mathbf{g}_{ij} = \begin{bmatrix} r_{i1j}^G, r_{i2j}^G, \ldots, r_{icj}^G \end{bmatrix}^T, \quad (i = 1, 2, \ldots, d; \ j = 1, 2, \ldots, n_t)
\tag{20}
$$

$$
\mathbf{R}_{kj} = \begin{bmatrix} r_{1kj}^G, r_{2kj}^G, \ldots, r_{dkj}^G \end{bmatrix}^T, \quad (k = 1, 2, \ldots, c; \ j = 1, 2, \ldots, n_t)
\tag{21}
$$

Therefore, in the second step of GmFSDA, we maximize the class discrimination and the scatter matrices S_b and S_w can be obtained in the transformed feature space as follows:

$$S_b = \sum_{j=1}^{n_t} \sum_{k=1}^{c} (\mathbf{R}_{kj} - \overline{\mathbf{R}})(\mathbf{R}_{kj} - \overline{\mathbf{R}})^T \tag{22}$$

$$S_w = \sum_{k=1}^{c} \sum_{j=1}^{n_t} \sum_{i=1}^{n_t} (\mathbf{R}_{ki} - \mathbf{R}_{kj})(\mathbf{R}_{ki} - \mathbf{R}_{kj})^T \tag{23}$$

where

$$\overline{\mathbf{R}} = \frac{1}{c \times n_t} \sum_{j=1}^{n_t} \sum_{k=1}^{c} \mathbf{R}_{kj} \tag{24}$$

We know that the matrix S_w may be singular in high-dimensional small-sized data sets, and thus, the solution cannot be obtained. Therefore, we use the regularization skill to cope with it as follows [36]:

$$S_w = \frac{1}{2} S_w + \frac{1}{2} diag(S_w) \tag{25}$$

The projection matrix A is usually transformed into the following eigenvalue problem.

$$S_b A = \lambda S_w A \tag{26}$$

Therefore, the projection matrix A can be obtained by maximizing $tr(S_w^{-1} S_b)$ with above Eq. (26) and the Fisher criterion [38,39]. For extracting p features from $d \times 1$ original feature vector (x), the p eigenvectors of $S_w^{-1} S_b$ associated with the largest p eigenvalues of $S_w^{-1} S_b$ compose the projection matrix A. Thus, we have: $\mathbf{y}_{p \times 1} = A_{p \times d} \mathbf{x}_{d \times 1}$.

From above analysis, if the matrix S_w is singular, we also will use the Fisher criterion and then we will also obtain the projection matrix A, that is because the following theorems 1 and 2 from [39]. In [39], the Fisher's criterion function $F(a)$ as follows. First of all, theorem 1 shows the subspace where we can derive the discriminant vectors based on maximizing the Fisher's criterion. But, we can extract maximum $r = rank(S_w)$ features from original data, i.e., $A \in \mathbb{R}^{r \times d}$.

$$F(a) = \frac{a^T S_b a}{a^T S_w a + a^T S_b a} \tag{27}$$

Theorem 1 *Let $V_0 = span\{\alpha_i | S_w \alpha_i = 0, \alpha_i \in \mathbb{R}^n, i = 1, 2, \ldots, n - r\}$, where n is the dimensionality of samples, S_w is the within-class scatter matrix of the samples, and r is the rank of S_w, S_b denote the between-class scatter matrix of the samples. For each $\tilde{a} \in v_0$ which satisfies $\tilde{a}^T S_b a \neq 0$, it will maximize the function $F(a) = a^T S_b a / (a^T S_w a + a^T S_b a)$.*

Proof (i) Since both S_b and S_w are real symmetric, $a^T S_b a \geq 0$ and $a^T S_w a \geq 0$, for all $a \in \mathbb{R}^n$, it follows that

$$0 \leq a^T S_b a \leq a^T S_b a + a^T S_w a \Rightarrow 0 \leq F(a) = \frac{a^T S_b a}{a^T S_w a + a^T S_b a} \leq 1 \tag{28}$$

It is obvious that $F(a) = 1$ if and only if $a^T S_b a \neq 0$ and $a^T S_w a = 0$.
(ii) For each $\hat{a} \in V_0$, $\tilde{a}$ can be represented as a linear combination of the set $\{\alpha_i\}$, i.e., $\tilde{a} = \sum_{i=1}^{n-r} b_i \alpha_i$, where b_i is the projection coefficient of $\tilde{a}$ with respect to α_i. Therefore, we

Table 5 The accuracy of classes obtained by MLC, 19 training samples and 8 extracted features for IPS dataset

Class	GmFSDA		FSDA		MFCF		RFCF		NWFE		MMP		KPCA		PCA	
	PA	UA	PA	UA	PA	UA	PA	UA	PA	UA	PA	UA	PA	UA	PA	UA
C1	38.71	88.89	53.85	77.78	39.06	92.59	40.0	88.89	18.28	62.96	28.95	81.48	33.8	88.89	41.51	81.48
C2	59.44	52.73	40.06	30.02	63.21	51.95	44.04	31.72	34.48	37.05	52.01	32.15	50.0	29.24	50.89	28.53
C3	53.62	44.76	36.34	58.08	36.95	59.68	30.91	56.1	45.45	29.59	36.78	59.19	29.55	75.46	28.69	64.24
C4	37.45	82.11	22.87	61.47	42.46	76.15	30.11	65.6	23.8	56.88	21.35	56.42	24.28	58.26	20.82	60.55
C5	65.66	65.52	84.41	67.67	76.76	54.09	72.22	47.63	73.88	64.01	92.35	65.09	74.02	20.26	89.33	68.53
C6	87.43	83.12	94.37	49.51	94.66	64.84	88.43	69.9	91.9	46.27	82.71	55.84	91.08	54.57	91.35	57.95
C7	69.23	100.0	37.5	100.0	15.52	100.0	34.62	100.0	15.25	100.0	33.33	100.0	27.59	88.89	29.03	100.0
C8	97.02	92.16	96.74	96.95	100.0	84.53	99.26	87.58	97.47	83.88	96.2	88.24	96.8	92.37	93.87	93.46
C9	2.632	100.0	2.326	100.0	3.333	100.0	8.333	100.0	2.0	100.0	2.128	100.0	1.587	100.0	2.86	100.0
C10	58.71	64.01	42.45	48.37	48.61	76.92	38.57	48.16	38.81	48.06	37.56	55.61	56.27	59.81	40.88	44.7
C11	71.52	61.12	60.73	39.16	67.4	34.11	62.17	33.46	57.77	41.05	64.42	41.17	65.36	39.04	62.41	34.89
C12	39.53	64.81	27.23	42.51	44.6	65.51	23.49	41.29	21.66	37.28	28.31	37.63	24.34	27.35	27.02	39.55
C13	97.04	88.17	98.24	89.78	99.25	71.51	98.14	84.95	87.17	87.63	96.18	81.18	95.38	88.71	97.21	93.55
C14	88.37	78.09	90.03	78.25	85.33	74.72	81.25	85.87	88.45	81.14	89.74	84.27	77.42	86.68	90.75	85.87
C15	36.65	44.14	28.64	64.31	31.33	71.12	39.28	50.41	26.05	50.68	26.05	47.14	29.2	50.68	32.77	58.58
C16	84.62	89.19	95.45	85.14	78.89	95.95	43.79	90.54	76.83	85.14	67.74	85.14	96.36	71.62	96.88	83.78
$\overline{PA}$ and $\overline{UA}$	61.73	74.93	56.95	68.06	57.96	73.35	52.16	67.63	49.95	63.23	53.49	66.91	54.56	64.49	56.02	68.48

Table 6 The accuracy of classes obtained by MLC, 19 training samples and 5 extracted features for Pavia University dataset

Class	GmFSDA		FSDA		MFCF		RFCF		NWFE		MMP		KPCA		PCA	
	PA	UA	PA	UA	PA	UA	PA	UA	PA	UA	PA	UA	PA	UA	PA	UA
C1	76.93	67.98	75.89	55.88	77.01	61.69	79.96	52.54	67.0	53.56	79.93	52.64	83.02	47.94	76.49	51.76
C2	86.15	61.2	85.31	53.89	86.0	56.11	81.64	65.32	86.72	57.95	88.35	63.82	82.59	65.12	87.11	56.48
C3	51.09	48.26	30.72	35.24	39.71	47.17	29.07	53.24	38.37	37.93	30.18	37.39	29.64	35.1	28.46	35.51
C4	61.07	86.22	57.56	92.36	56.07	91.86	50.69	80.19	54.83	87.21	61.07	91.86	55.32	92.8	57.7	92.04
C5	99.77	96.62	99.77	95.88	99.77	95.44	97.38	93.08	99.77	95.14	99.85	96.84	99.77	93.97	99.77	95.88
C6	34.6	58.23	30.81	56.58	31.21	54.4	30.39	40.61	35.48	56.91	37.9	60.94	31.35	37.86	33.16	60.92
C7	49.48	74.87	32.6	67.61	37.24	73.0	34.77	75.39	31.37	67.31	30.62	74.64	32.04	74.35	30.97	67.39
C8	58.61	65.9	54.61	57.58	59.38	57.01	43.39	28.92	51.68	59.26	55.68	53.95	55.31	66.18	56.9	55.58
C9	73.08	96.23	67.42	97.22	68.88	96.72	66.59	89.08	66.01	96.43	68.16	96.52	60.43	97.52	64.71	97.42
$\overline{PA}$ and $\overline{UA}$	65.64	72.83	59.41	68.03	61.7	70.38	57.1	64.26	59.03	67.97	61.31	69.85	58.83	67.87	59.47	68.11

Table 7 The accuracy of classes obtained by MLC, 19 training samples and 7 extracted features for KSC dataset

| Class	GmFSDA		FSDA		MFCF		RFCF		NWFE		MMP		KPCA		PCA	
	PA	UA	PA	UA	PA	UA	PA	UA	PA	UA	PA	UA	PA	UA	PA	UA
C1	95.69	44.88	96.68	50.94	93.95	50.27	97.08	62.67	95.49	34.23	92.33	37.33	93.28	57.95	95.72	42.18
C2	87.27	85.71	84.0	84.38	89.69	77.68	86.61	86.61	87.26	82.59	88.35	81.25	83.41	78.57	89.23	77.68
C3	85.71	75.95	81.25	76.79	64.84	59.92	73.89	48.95	82.96	78.06	79.66	59.49	51.68	32.49	74.04	64.98
C4	57.77	51.07	34.67	40.77	21.29	22.75	29.33	28.33	23.06	40.77	41.74	39.06	33.03	30.9	29.59	24.89
C5	31.18	74.65	39.84	70.42	33.57	65.49	30.03	78.87	37.91	73.94	30.09	73.94	27.23	77.46	31.69	72.54
C6	25.65	47.14	27.3	48.1	26.12	44.29	32.91	49.05	26.45	43.33	19.42	41.43	32.95	40.48	25.27	43.81
C7	74.49	84.88	78.65	81.4	74.04	89.53	78.72	86.05	74.74	82.56	65.81	89.53	65.83	91.86	66.67	86.05
C8	74.0	84.95	72.14	84.22	62.9	86.41	64.56	81.8	70.16	82.77	64.52	85.19	57.14	77.67	58.63	87.38
C9	90.15	95.01	90.15	93.21	87.4	85.83	88.25	88.42	87.59	93.01	92.51	86.23	86.96	87.82	91.11	92.02
C10	92.19	91.95	95.42	91.95	98.48	84.42	97.78	91.69	99.44	91.69	99.72	92.73	80.46	82.34	98.49	84.94
C11	99.45	90.25	96.87	92.75	98.64	90.5	91.75	97.25	95.31	91.5	96.54	90.75	83.52	95.0	98.94	93.25
C12	82.47	88.43	83.77	86.36	77.71	84.3	86.65	73.76	84.05	84.92	81.55	86.78	72.93	60.12	81.38	87.6
C13	100.0	97.91	100.0	97.8	100.0	97.8	100.0	96.7	100.0	97.91	100.0	97.91	100.0	91.3	100.0	97.8
$\overline{PA}$ and $\overline{UA}$	76.62	77.91	75.44	76.85	71.43	72.24	73.66	74.63	74.19	75.18	73.25	73.97	66.8	69.54	72.37	73.47

(a)

(b)

(c)

(d)
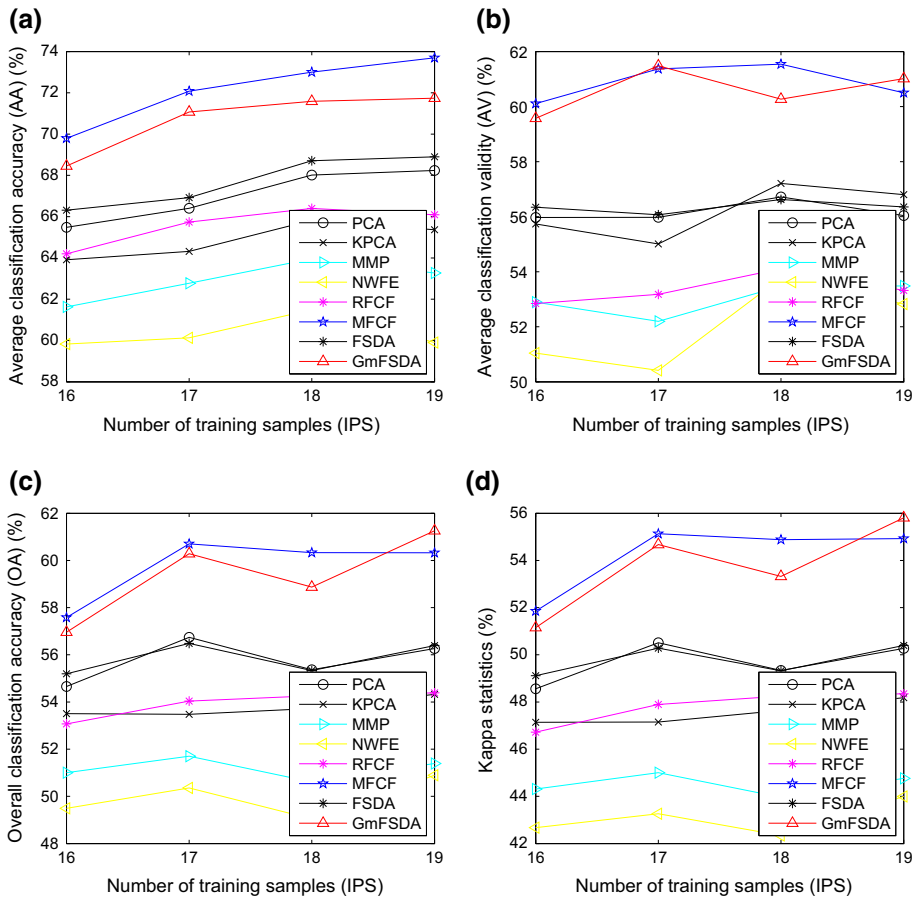
Fig. 5 Influence of the number of training samples per class on the PCA, KPCA, MMP, NWFE, RFCF, MFCF, FSDA and GmFSDA for the IPS dataset. **a** AA, **b** AV, **c** OA, **d** kappa coefficient

have $S_w \tilde{a} = S_w \sum_{i=1}^{n-r} b_i \alpha_i = \sum_{i=1}^{n-r} b_i S_w \alpha_i = 0 \Rightarrow \tilde{a}^T S_w \tilde{a} = 0$. We can conclude that for each $\tilde{a} \in V_0$ which satisfies $\tilde{a}^T S_b \tilde{a} \neq 0$, the function $F(a)$ will be maximized in terms of (i) and (ii).

An arbitrary vector $\tilde{a} \in V_0$ that maximizes $F(a)$ is not necessarily the optimal discriminant vector when the small sample size problem occurs. The main reason is that $\tilde{a}^T S_b \tilde{a}$ is not guaranteed to reach the maximal value. Theorem 2 will show that the within-class scatter matrix of all the transformed samples in V_0 is a complete zero matrix. Instead, the discriminant vector set can be derived directly from the between-class scatter matrix. □

Theorem 2 *Let QQ^T be a transformation which transforms the samples in V (original feature space) into a subspace V_0, where $Q = [\alpha_1, \ldots, \alpha_{n-r}] \in \mathbb{R}^{n \times (n-r)}$ and each α_i satisfies $S_w \alpha_i = 0$, for $i = 1, \ldots, n-r$, where the subspace V_0 is spanned by the orthonormal set of α_i's. If all the samples are transformed into the subspace V_0 through QQ^T, then the within-class scatter matrix $\tilde{S}_w$ of the transformed samples in V_0 is a complete zero matrix.*

Proof Suppose x_m^k is the feature vector extracted from the m-th sample of the k-th class, and that the database comprised c classes, where each class contains n_t samples. Let y_m^k denote the
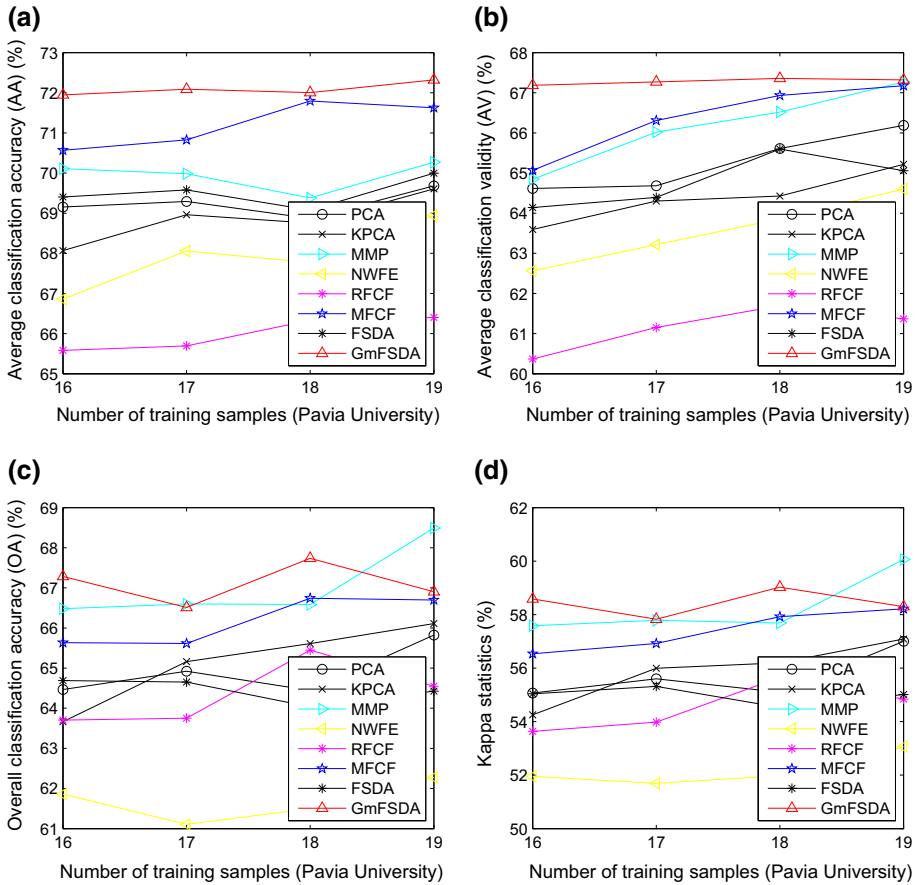
Fig. 6 Influence of the number of training samples per class on the PCA, KPCA, MMP, NWFE, RFCF, MFCF, FSDA and GmFSDA for the Pavia University dataset. **a** AA, **b** AV, **c** OA, **d** kappa coefficient

transformed feature vector of x_m^k through the transformation QQ^T. That is, $y_m^k = QQ^T x_m^k$, $\bar{y}^k = QQ^T \bar{x}^k$ and $\bar{y} = QQ^T \bar{x}$, where $\bar{x}^k = \frac{1}{n_t} \sum_{j=1}^{n_t} x_m^k$ and $\bar{x} = \frac{1}{c \times n_t} \sum_{k=1}^{c} \sum_{j=1}^{n_t} x_m^k$. Thus,

$$
\begin{aligned}
\tilde{S}_w &= \sum_{k=1}^{c} \sum_{j=1}^{n_t} \left(y_m^k - \bar{y}^k \right) \left(y_m^k - \bar{y}^k \right)^T \\
&= \sum_{k=1}^{c} \sum_{j=1}^{n_t} \left(QQ^T x_m^k - QQ^T \bar{x}^k \right) \left(QQ^T x_m^k - QQ^T \bar{x}^k \right)^T \\
&= QQ^T \sum_{k=1}^{c} \sum_{j=1}^{n_t} \left(x_m^k - \bar{x}^k \right) \left(x_m^k - \bar{x}^k \right) QQ^T \\
&= QQ^T S_w QQ^T = 0 (since S_w Q = 0)
\end{aligned}
\tag{29}
$$

From the above two theorems, we know that maximizing the between-class scatter in V_0 is equal to maximizing the total scatter in V_0. Therefore, in order to extract more than r features, we use the Eq. (25) for hyperspectral image.
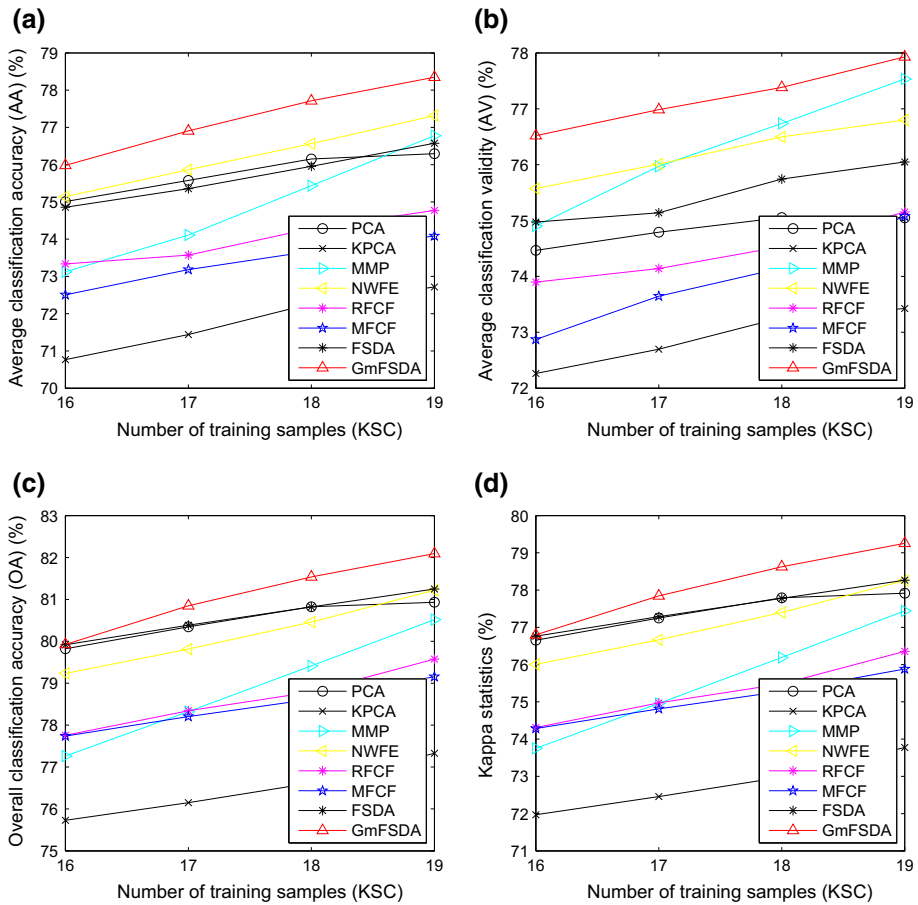
(a)

(b)

(c)

(d)

Fig. 7 Influence of the number of training samples per class on the PCA, KPCA, MMP, NWFE, RFCF, MFCF, FSDA and GmFSDA for the KSC dataset. **a** AA, **b** AV, **c** OA, **d** kappa coefficient

As mentioned above, the proposed GmFSDA feature extraction method not only can reveal intrinsic geometry structure of the hyperspectral image, but also it can learn nonlinear correlation features with well discriminating power by MLC.   □

3 Experiments and Analysis

In this experiments, the performances of all tested methods were quantitatively compared using the average classification accuracy (AA), average classification validity (AV), overall accuracy (OA) and kappa coefficient. The classification process was carried out using MATLAB 7.11 version on a computer equipped with an Intel Core i7 processor at 3.40-GHz. In order to evaluate the capability of the GmFSDA feature extraction in classification of hyperspectral image data, the proposed GmFSDA feature extraction method together with common feature extraction methods such as PCA, KPCA, MMP, NWFE, RFCF, MFCF and FSDA were tested on two hyperspectral datasets. The AA, AV, OA, and Kappa criteria [2,3,35] were
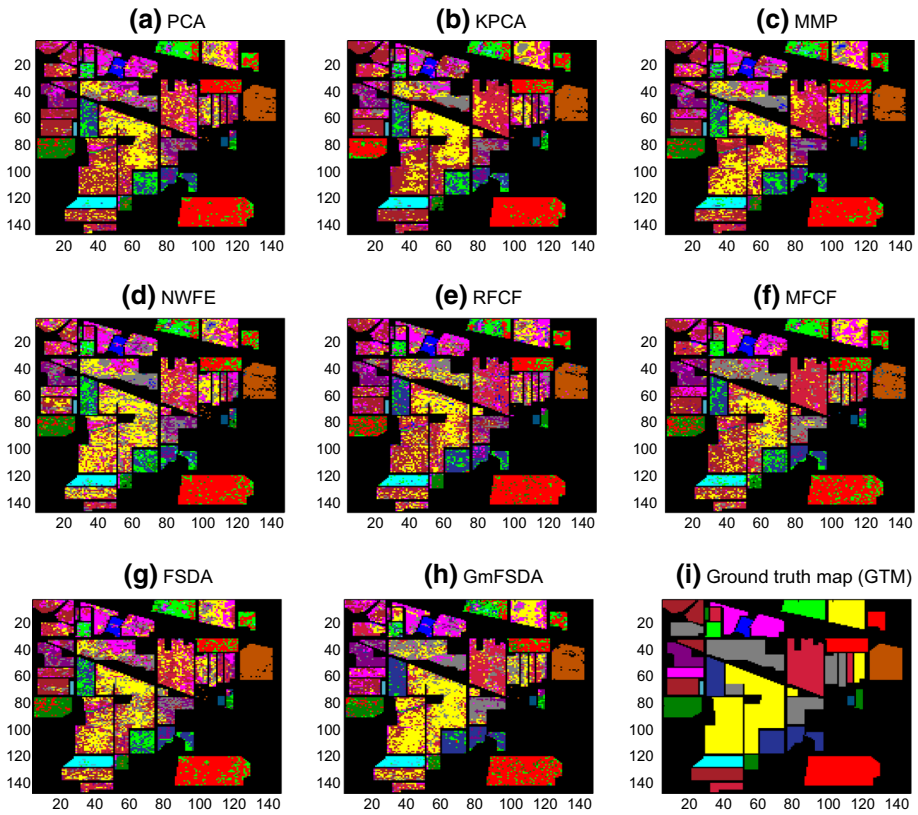
Fig. 8 Class maps obtained by MLC, 19 training samples and 8 extracted features for IPS dataset. **a** PCA class map, **b** KPCA class map, **c** MMP class map, **d** NWFE class map, **e** RFCF class map, **f** MFCF class map, **g** FSDA class map, **h** GmFSDA class map, **i** ground truth class map

used in this comparison, and these are defined as follows:

$$AA = \frac{1}{C} \sum_{c=1}^{C} ACC(c) \tag{30}$$

Where $ACC(c) = n_c / N_c$ denotes classification accuracy for each subject class, n_c denotes the number of pixels of class c correctly classified and N_c denotes the number of test pixels in that class.

$$AV = \frac{1}{C} \sum_{c=1}^{C} VAL(c) \tag{31}$$

Where $VAL(c) = n_c / m_c$ denotes classification validity for each subject class, m_c denotes the number of all the pixels labeled as class c in output class map.

$$OA = \frac{n}{N} \tag{32}$$

The measure is similar to AA, expect that classes are not considered subject but as a whole. n denotes the number of all the pixels correctly classified and N denotes the number of all
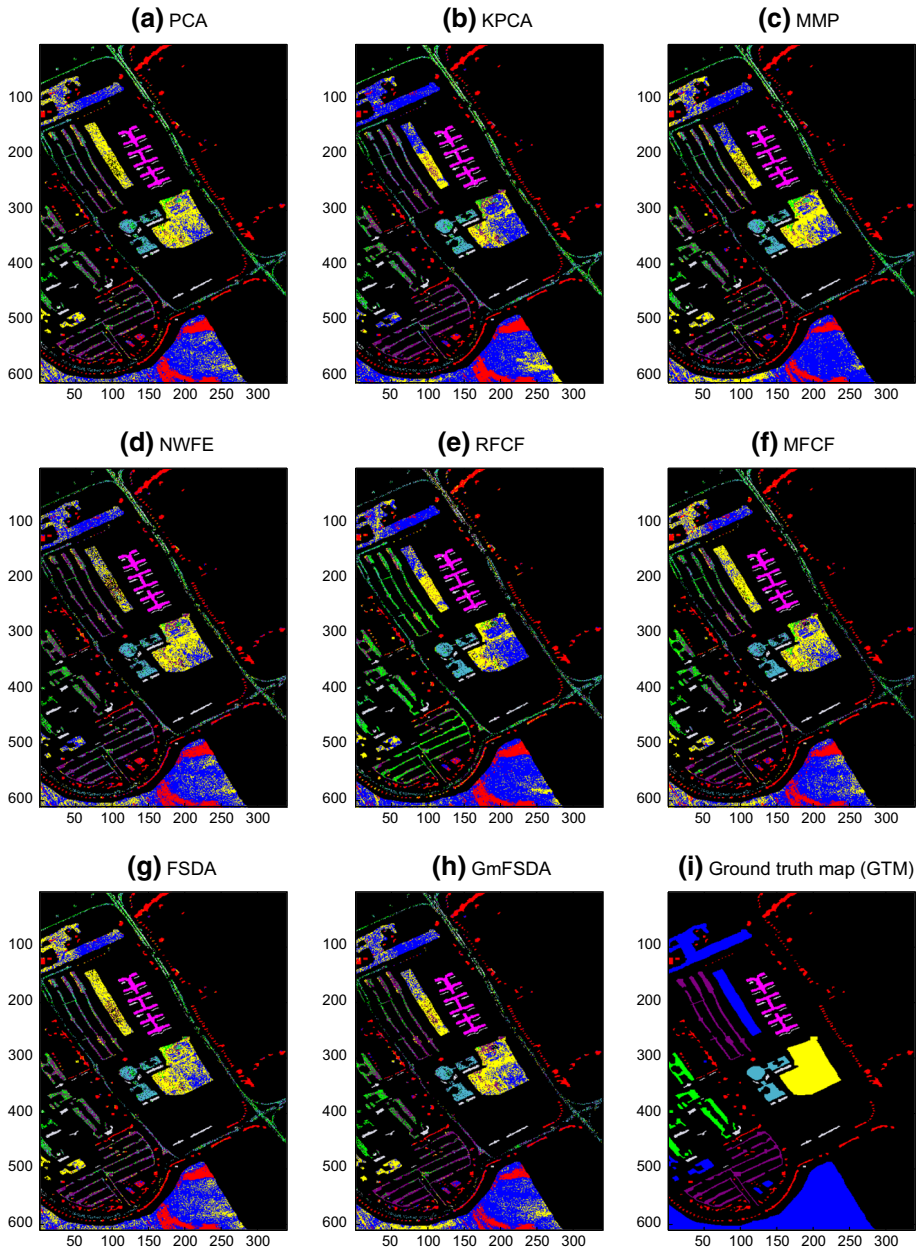
Fig. 9 Class maps obtained by MLC, 19 training samples and 5 extracted features for Pavia University dataset. **a** PCA class map, **b** KPCA class map, **c** MMP class map, **d** NWFE class map, **e** RFCF class map, **f** MFCF class map, **g** FSDA class map, **h** GmFSDA class map, **i** ground truth class map
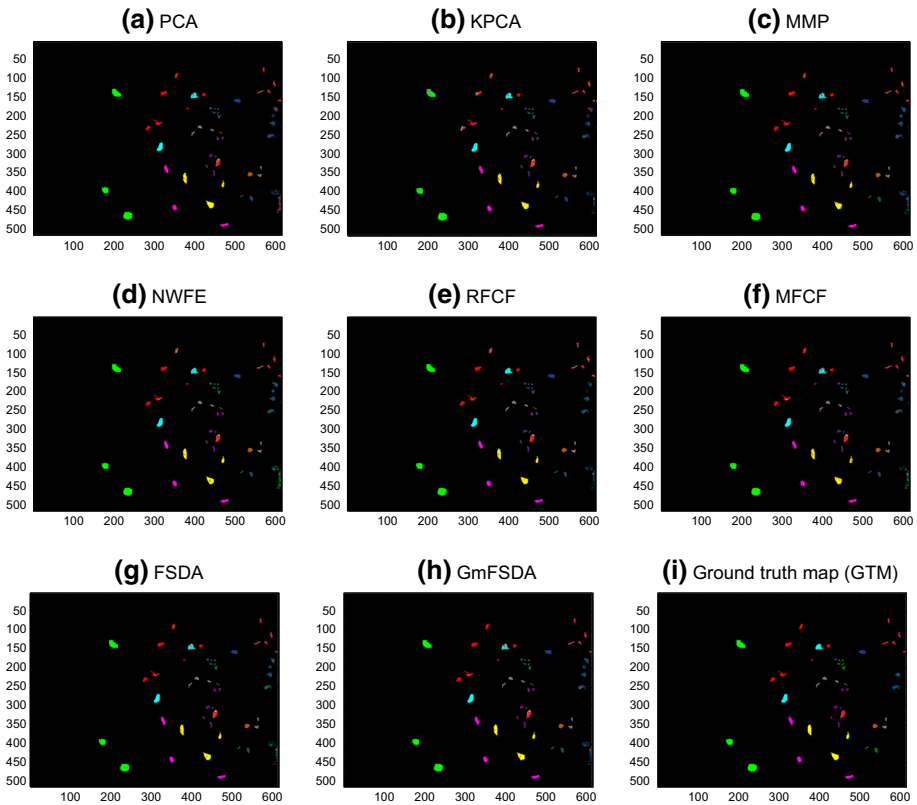
Fig. 10 Class maps obtained by MLC, 19 training samples and 7 extracted features for KSC dataset. **a** PCA class map, **b** KPCA class map, **c** MMP class map, **d** NWFE class map, **e** RFCF class map, **f** MFCF class map, **g** FSDA class map, **h** GmFSDA class map, **i** ground truth class map

pixels in the test set.

$$Kappa = \frac{OA - P_e}{1 - P_e} \tag{33}$$

Where $P_e = (\sum_{c=1}^{C} m_c \cdot N_c)/N^2$.

In our experiments, the product's accuracy (PA) and the user's accuracy (UA) are also used to evaluate the performance of the above mentioned eight methods in addition to AA, AV, OA, and kappa coefficient. The PA and UA are defined as Eqs. (34) and (35), respectively:

$$PA_i = \frac{x_{i,i}}{x_{+i}} \tag{34}$$

$$UA_i = \frac{x_{i,i}}{x_{i+}} \tag{35}$$

where $x_{i,i}$ is the value on the major diagonal of the i-th row in the confusion matrix, x_{i+} is the total number of the i-th row, and x_{+i} is the total number of the i-th column. The $\overline{PA}$ and $\overline{UA}$ are the average results of PA_i and UA_i (i = 1, 2, . . . , c) and are defined as Eqs. (36) and (37), respectively.

Table 8 The AA, AV, OA and kappa statistic, and standard deviation of accuracy measures, and the number of feature corresponding to the best result for each method with IPS datasetn, 19 training samples and MLC

Methods	AA (std)	#	AV (std)	#	OA (std)	#	Kappa (std)	#
PCA	68.2382 ± 1.1812	6	56.0409 ± 1.2497	8	56.2675 ± 2.0817	6	50.2710 ± 2.2480	6
KPCA	65.3626 ± 2.2538	10	56.8014 ± 2.7258	10	54.3258 ± 2.0091	10	48.1818 ± 2.0465	10
MMP	63.2736 ± 2.6279	6	53.4819 ± 2.4375	11	51.3947 ± 1.9891	8	44.7651 ± 2.1572	8
NWFE	59.9026 ± 3.9166	7	52.8222 ± 1.5037	12	50.8919 ± 1.9936	5	43.9978 ± 1.8740	5
RFCF	66.0930 ± 3.0307	9	53.3203 ± 2.7678	9	54.3901 ± 2.0469	9	48.3474 ± 2.2592	9
MFCF	73.6930 ± 1.2594	7	60.5006 ± 2.2567	10	60.3258 ± 2.0543	7	54.9269 ± 2.2621	7
FSDA	68.8992 ± 1.4278	6	56.3515 ± 1.7770	6	56.3922 ± 1.7436	6	50.3963 ± 1.9660	6
GmFSDA	71.7358 ± 1.7474	8	61.0145 ± 2.6996	9	61.2609 ± 3.0879	8	55.8064 ± 3.3379	8

Table 9 The AA, AV, OA and kappa statistic, and standard deviation of accuracy measures, and the number of feature corresponding to the best result for each method with Pavia University dataset, 19 training samples and MLC

Methods	AA (std)	#	AV (std)	#	OA (std)	#	Kappa (std)	#
PCA	69.6811 ± 1.7187	8	66.1885 ± 3.0652	11	65.8246 ± 4.6731	11	57.0023 ± 4.9869	11
KPCA	69.6053 ± 1.2558	9	65.2174 ± 1.5747	12	66.1094 ± 2.6022	9	57.0868 ± 2.7379	9
MMP	70.2710 ± 2.9336	9	67.2703 ± 2.3047	10	68.4913 ± 4.4261	10	60.0634 ± 4.8494	10
NWFE	68.9454 ± 1.5332	7	64.6024 ± 2.0973	10	62.2756 ± 2.8930	8	53.0646 ± 2.8456	8
RFCF	66.4014 ± 1.7329	6	61.3675 ± 2.4682	6	64.5470 ± 3.8054	6	54.8388 ± 3.9860	6
MFCF	71.6272 ± 1.6648	7	67.1766 ± 2.5207	7	66.6984 ± 3.8655	7	58.2162 ± 3.9257	7
FSDA	69.9977 ± 1.4827	7	65.0567 ± 1.9851	10	64.4254 ± 2.9795	4	55.0154 ± 2.9876	4
GmFSDA	72.3211 ± 1.2953	5	67.3188 ± 2.2130	6	66.9032 ± 2.8414	5	58.2957 ± 2.7849	5

Table 10 The AA, AV, OA and kappa statistic, and standard deviation of accuracy measures, and the number of feature corresponding to the best result for each method with KSC dataset, 19 training samples and MLC

Methods	AA (std)	#	AV (std)	#	OA (std)	#	Kappa (std)	#
PCA	77.9155 ± 1.2557	6	75.0478 ± 1.1952	6	80.9327 ± 1.1187	6	77.9155 ± 1.2557	6
KPCA	73.7732 ± 1.6517	10	73.4248 ± 1.5222	12	77.3288 ± 1.4679	10	73.7732 ± 1.6517	10
MMP	77.4447 ± 2.2146	10	77.5321 ± 1.0263	10	80.5177 ± 1.9714	10	77.4447 ± 2.2146	10
NWFE	78.2570 ± 2.8395	6	76.8010 ± 1.6949	9	81.2188 ± 2.5175	6	78.2570 ± 2.8395	6
RFCF	76.3562 ± 1.3986	8	75.1518 ± 1.3918	8	79.5770 ± 1.2559	8	76.3562 ± 1.3986	8
MFCF	75.8815 ± 1.4570	6	75.0780 ± 1.4301	10	79.1539 ± 1.3185	6	75.8815 ± 1.4570	6
FSDA	78.2674 ± 1.4869	6	76.0480 ± 1.3754	8	81.2490 ± 1.3273	6	78.2674 ± 1.4869	6
GmFSDA	79.2568 ± 2.9542	7	77.9291 ± 1.5900	7	82.0931 ± 2.6149	7	79.2568 ± 2.9542	7

$$\overline{\mathrm{PA}} = \frac{1}{c} \sum_{i=1}^{c} \mathrm{PA}_i \tag{36}$$

$$\overline{UA} = \frac{1}{c} \sum_{i=1}^{c} UA_i \qquad (37)$$

where c is the number of class in dataset.

3.1 Hyperspectral Data Description

We use three popular datasets for doing our experiments: Indian Pines Site (IPS), Pavia University and KSC. The well-known Indian Pines Site (IPS) dataset was obtained by the Airborne Visible Infrared Imaging Spectrometer (AVIRIS) sensor over the Indian Pines Site (IPS) in northwestern Indiana in 1992 [40,41]. The hyperspectral image contains 220 spectral reflectance bands and the size is 145×145 pixels. The spatial resolution is 20 m. The ground truth available is designated into 16 classes (see Table 1) and is not all mutually exclusive. 20 spectral bands were removed because of the noise and water absorption phenomena and $N = 200$ bands were left. The three-channel false-color composition and the reference land-cover of the AVIRIS data of IPS are shown in Fig. 1.

The Pavia University dataset was gathered by the Reflective Optics System Imaging Spectrometer (ROSIS) sensor over the engineering school at university of Pavia. It is 610×340 pixels, and the spatial resolution is 1.3 m per pixel. 12 spectral bands were removed due to the noise, and the remaining 103 spectral bands were processed. 9 classes of interest are considered (see Table 2): tree, asphalt, bitumen, gravel, metal sheets, shadows, bricks, meadows, and soil. The three bands synthetic false color image and the reference land-cover image are shown in Fig. 1.

The third dataset has been acquired over the Kennedy Space Center (KSC) Florida, in March, 1996 by AVIRIS sensor [41]. AVIRIS acquires data in 224 bands of 10 nm width with center wavelengths from 400–2500 nm. The KSC data acquired from an altitude of approximately 20 km, have a spatial resolution of 18 m. After removing water absorption and low SNR bands, $N = 176$ bands were used for the analysis. The image size is 614×512 pixels, and contains 13 different classes of land-cover (see Table 3). The three-channel false-color composition and the reference land-cover of the KSC data are also shown in Fig. 1.

3.2 Experiments and Results

In this subsection, the extracted features of PCA, KPCA, MMP, NWFE, RFCF, MFCF and proposed GmFSDA methods have been fed into an maximum likelihood classifier (MLC) and then the classification results have been compared for IPS, Pavia University and KSC three hyperspectral images data. Moreover, there are many parameters of all the above mentioned feature extraction methods needed to be manually set (see Table 4). In Table 4, the G denotes Gaussian kernel, p denotes regularization parameter, D denotes the number of features for all feature extraction methods and $D = 2, 3, \ldots, 14$. In RFCF feature extraction method, the parameter L is changed in the range of 0 to $D-1$, and M is selected regarding to the constraint: $M + L + 1 = D$. Of course, we have know that there are many uncertain factors for RFCF method [35]. The best classification results can be obtained with $L = 0, 1, D-2, D-1$ in terms of the research results of [2,3].

In our experiments, the parameter $L = 1$ has been used to extract feature for RFCF method. The training samples are selected randomly from entire scene and the remaining samples are used as testing samples. In order to demonstrate the capability of all above mentioned methods, the 19 samples each class is used for training the MLC. Because of random selection

of training samples, we do each experiment 10 times and the average results are reported. In the experiment, the number of features is set from 2–14 in all methods and each time. The classification accuracy measures for IPS, Pavia University and KSC datasets corresponding to each approach versus the number of features are shown in Figs. 2, 3 and 4, respectively.

In the three experiments, we evaluate the performance of proposed GmFSDA in comparison with PCA, KPCA, MMP, NWFE, RFCF, MFCF and FSDA in small sample size situation for IPS, Pavia University and KSC datasets. We obtain the average classification accuracy (AA), average classification validity (AV), overall accuracy (OA) and kappa coefficient versus the number of extracted features using 16 training samples (see Figs. 2, 3, 4). The MLC is used for classification of reduced data. From Figs. 2, 3 and 4, we can obtain the fact that the superiority of the proposed GmFSDA feature extraction method in comparison with the competing algorithms is apparent. The four measures index, such as AA, AV, OA, and kappa coefficient have been dramatically improved by GmFSDA feature extraction method in comparison to other methods.

As shown in Fig. 2, with the increase of the number of features, the AA, OA and kappa coefficient are increased dramatically and then dropped slowly. We also noticed that the robustness of the proposed GmFSDA method is better than MFCF method, especially when the number of extracted feature is 13. In terms of the research results of Ref. [35] and Fig. 2, for IPS dataset, the similarity of spectral response curves is quite remarkable for most of its classes. Hence, the most feature extraction techniques cannot discriminate its classes very well and the classification results are not much satisfactory. The AA of MFCF is higher than our proposed GmFSDA method.

From Figs. 3 and 4, we can observe that the proposed GmFSDA method can reach the highest accuracy in AA, AV, OA and kappa coefficient, especially the MFCF method cannot obtain the best results in KSC dataset. Of course, for Pavia University dataset, the fluctuation of these four indicators i.e., AA, AV, OA, and kappa coefficient is larger than other two datasets. However, for KSC dataset, the similarity of spectral response curves is not quite remarkable for most of its different classes. Hence, most feature extraction techniques can discriminate its classes satisfactorily. Of course, in the three datasets, the proposed GmFSDA feature extraction method outperforms the other methods.

In addition, the proposed GmFSDA feature extraction method outperforms the PCA, KPCA, MMP, NWFE, RFCF, MFCF and FSDA with small sample size in PA and UA. The accuracy results for 19 training samples cases are shown in Tables 5, 6 and 7.

In order to further evaluate the performance of our proposed GmFSDA method, the AA, AV, OA and kappa coefficient with different numbers of training samples are shown in Figs. 5, 6 and 7 for IPS, Pavia University and KSC datasets, respectively. Figures 5, 6 and 7 illustrate the classification accuracy and kappa coefficient of each method with different number of training samples. It can be observed that the performances of all the compared methods will improve as the number of training samples increases. Furthermore, the proposed GmFSDA method consistently presents excellent performances with different numbers of training and test samples in addition to the IPS dataset. As shown in Fig. 5, it can be observed that the classification accuracy of MFCF is higher than our proposed GmFSDA method in terms of AA, AV, OA and kappa coefficient. The main reason is that for IPS dataset, the similarity of spectral response curves by using MFCF can reveal the class discriminant. However, this robustness is poor with different numbers of extracted features in terms of Fig. 2.

Figures 8, 9 and 10 show the classification maps obtained by PCA, KPCA, MMP, NWFE, RFCF FSDA and proposed GmFSDA methods with MLC and 8, 5 and 7 extracted features and 19 training samples for IPS, Pavia University and KSC datasets, respectively. From the visual point of view, the capability of GmFSDA feature extraction method is also demonstrated in

terms of Figs. 8, 9 and 10. Hence, the results of GmFSDA feature extraction algorithm are better than its nearest competitors, such as FSDA, MFCF and RFCF.

The number of features can be determined in terms of the best overall accuracy of each method (see Tables 8, 9, 10). In order to further evaluate the performance of our proposed GmFSDA method, the averages and standard deviations of accuracy measures are shown in Tables 8, 9 and 10 for IPS, Pavia University and KSC datasets, respectively. The std denotes the standard deviation of accuracy measures, and # denotes the number of feature corresponding to the best result of each method.

From Tables 8, 9 and 10, the average accuracy of proposed GmFSDA method is higher than that of the other seven methods. However, its standard deviation is bigger than FSDA feature extraction method for the IPS and KSC different training sets in terms of Tables 8 and 10. Actually, in GmFSDA method, we transform the feature space two times. At first, the original feature space is transformed to a new feature space by using a primary projection and the features have more discrimination in the new space. Then, the second feature space can be obtained by using a secondary projection in the first obtained feature space, and the capability of class discrimination can be improved. The main difference of GmFSDA with other methods is in the primary projection. The primary projection can improve the performance of classification significantly when the number of available training samples is limited. On the whole, the proposed GmFSDA method has better robustness, and GmFSDA not only can reveal intrinsic geometry structure of the hyperspectral image, but also it can learn nonlinear correlation features with well discriminating power by MLC.

4 Conclusion

In this paper, an efficient GmFSDA supervised feature extraction method is proposed for hyperspectral image classification. The proposed method mainly consider the geometry structure and outlier data value of hyperspectral images pixels. Inspired by the research result of M. Imani and H. Ghassemian, the average feature space discriminant analysis (FSDA) can extract an efficient features for hyperspectral image. Hence, we use a more efficient geometric mean instead of average vales for hyperspectral image data feature extraction. The proposed method need to transform the original data to a new space. So, GmFSDA uses two projection matrices for feature extraction. The primary projection matrix is used to maximize the geometric between-spectral scatter. The secondary projection matrix which is applied to the feature space obtained in the first step, maximizes the separability between classes. The performance of GmFSDA is compared to other popular and state-of-the-art supervised/unsupervised feature extraction methods with IPS, Pavia University and KSC datasets. The GmFSDA can capture the intrinsic geometrical nature of hyperspectral images pixels. The experimental results show that the proposed method can achieve feature extraction and obtain better classification accuracy than the other seven methods.

References

1. David L (2002) Hyperspectral image data analysis as a high dimensional signal processing problem. IEEE Signal Process Mag 19(1):17–28
2. Hosseini SA, Ghassemian H (2016) Hyperspectral data feature extraction using rational function curve fitting. Int J Pattern Recognit Artif Intell 30(01):1650001. https://doi.org/10.1142/S0218001416500014
3. Hosseini SA, Ghassemian H (2016) Rational function approximation for feature reduction in hyperspectral data. Remote Sens Lett 7(2):101–110
4. Imani M, Ghassemian H (2017) High-dimensional image data feature extraction by double discriminant embedding. Pattern Anal Appl 20(2):473–484
5. Imani M, Ghassemian H (2016) Binary coding based feature extraction in remote sensing high dimensional data. Inf Sci 342:191–208
6. Jia X, Kuo BC, Crawford MM (2013) Feature mining for hyperspectral image classification. Proc IEEE 101(3):676–697
7. Maji P, Garai P (2013) Fuzzy-rough simultaneous attribute selection and feature extraction algorithm. IEEE Trans Cybern 43(4):1166–1177
8. Li S, Qiu J, Yang X et al (2014) A novel approach to hyperspectral band selection based on spectral shape similarity analysis and fast branch and bound search. Eng Appl Artif Intell 27:241–250
9. Esfandian N, Razzazi F, Behrad A (2012) A clustering based feature selection method in spectro-temporal domain for speech recognition. Eng Appl Artif Intell 25(6):1194–1202
10. Dernoncourt D, Hanczar B, Zucker JD (2014) Analysis of feature selection stability on high dimension and small sample data. Comput Stat Data Anal 71:681–693
11. Zhang L, Zhong Y, Huang B et al (2007) Dimensionality reduction based on clonal selection for hyperspectral imagery. IEEE Trans Geosci Remote Sens 45(12):4172–4186
12. Hotelling H (1933) Analysis of a complex of statistical variables into principal components. Educ Psychol 24(6):417–441
13. Liao W, Pizurica A, Scheunders P et al (2013) Semisupervised local discriminant analysis for feature extraction in hyperspectral images. IEEE Trans Geosci Remote Sens 51(1):184–198
14. Plaza A, Martinez P, Plaza J et al (2005) Dimensionality reduction and classification of hyperspectral image data using sequences of extended morphological transformations. IEEE Transn Geosci Remote Sens 43(3):466–479
15. Fauvel M, Chanussot J, Benediktsson JA (2009) Kernel principal component analysis for the classification of hyperspectral remote sensing data over urban areas. EURASIP J Adv Signal Process. Article ID 783194. https://doi.org/10.1155/2009/783194
16. Villa A, Chanussot J, Benediktsson JA et al (2013) Unsupervised methods for the classification of hyperspectral images with low spatial resolution. Pattern Recognit 46(6):1556–1568
17. Chang CI, Ren H (2000) An experiment-based quantitative and comparative analysis of target detection and image classification algorithms for hyperspectral imagery. IEEE Trans Geosci Remote Sens 38(2):1044–1063
18. Kuo BC, Landgrebe DA (2004) Nonparametric weighted feature extraction for classification. IEEE Trans Geosci Remote Sens 42(5):1096–1105
19. Scholkopf B, Smola A, Muller KR (1998) Nonlinear component analysis as a kernel eigenvalue problem. Neural Comput 10:1299–1319
20. Baudat G, Anouar F (2000) Generalized discriminant analysis using a kernel approach. Neural Comput 12(10):2385–2404
21. Chen S, Zhang D (2011) Semisupervised dimensionality reduction with pairwise constraints for hyperspectral image classification. IEEE Geosci Remote Sens Lett 8(2):369–373
22. He X, Niyogi P (2004) Locality preserving projections. In: Advances in neural information processing systems, pp 153–160
23. Zhang T, Yang J, Zhao D et al (2007) Linear local tangent space alignment and application to face recognition. Neurocomputing 70(7):1547–1553
24. He X, Cai D, Han J (2008) Learning a maximum margin subspace for image retrieval. IEEE Trans Knowl Data Eng 20(2):189–201
25. Yang M, Zhang L, Chi-Keung Shiu S, Zhang Z (2012) Monogenic binary coding: an efficient local feature extraction approach to face recognition. IEEE Trans Inf Forensics Secur 7:1738–1751
26. Zhou Y, Peng J, Chen CLP (2015) Dimension reduction using spatial and spectral regularized local discriminant embedding for hyperspectral image classification. IEEE Trans Geosci Remote Sens 53:1082–1095
27. Li L, Ge H, Gao J (2017) A spectral-spatial kernel-based method for hyperspectral imagery classification. Adv Sp Res 59(4):954–967

28. Gao J, Xu L (2016) A novel spatial analysis method for remote sensing image classification. Neural Process Lett 43(3):805–821
29. Gao J, Xu L, Shen J et al (2015) A novel information transferring approach for the classification of remote sensing images. EURASIP J Adv Signal Process 2015(1):38
30. Gao J, Xu L, Huang F (2016) A spectral-textural kernel-based classification method of remotely sensed images. Neural Comput Appl 27(2):431–446
31. Gao J, Xu L (2015) An efficient method to solve the classification problem for remote sensing image. AEU-Int J Electron Commun 69(1):198–205
32. Gao J, Xu L, Shi A et al (2014) A kernel-based block matrix decomposition approach for the classification of remotely sensed images. Appl Math Comput 228:531–545
33. Hosseini A, Ghassemian H (2012) Classification of hyperspectral and multispectral images by using fractal dimension of spectral response curve. In: 2012 20th Iranian conference on electrical engineering (ICEE). IEEE, pp 1452–1457
34. Hosseini SA, Ghassemian H (2013) A new hyperspectral image classification approach using fractal dimension of spectral response curve. In: 2013 21st Iranian conference on electrical engineering (ICEE). IEEE, pp 1–6
35. Li L, Ge H, Gao J (2018) Hyperspectral image feature extraction using maclaurin series function curve fitting. Neural Process Lett. https://doi.org/10.1007/s11063-018-9825-5
36. Imani M, Ghassemian H (2015) Feature space discriminant analysis for hyperspectral data feature reduction. ISPRS J Photogramm Remote Sens 102:1–13
37. Elton EJ, Gruber MJ, Brown SJ et al (2009) Modern portfolio theory and investment analysis. Wiley, Hoboken
38. Rietz HL (1916) The mathematical theory of probabilities. Science 43(1121):896–897
39. Chen LF, Liao HYM, Ko MT et al (2000) A new LDA-based face recognition system which can solve the small sample size problem. Pattern Recognit 33(10):1713–1726
40. Purdue Research Foundation, Hyperspectral images by multiSpec (2015). https://engineering.purdue.edu/~biehl/MultiSpec/
41. Universidad-del-Pais-Vasco, Hyperspectral remote sensing scenes (2014). http://www.ehu.eus/ccwintco/index.php?title=Hyperspectral_Remote_Sensing_Scenes
42. Zhang L, Zhang L, Du B et al (2019) Hyperspectral image unsupervised classification by robust manifold matrix factorization. Inf Sci 485:154–169
43. Zhang L, Zhang L, Tao D et al (2011) On combining multiple features for hyperspectral remote sensing image classification. IEEE Trans Geosci Remote Sens 50(3):879–893
44. Zhang L, Zhang L, Tao D et al (2013) A modified stochastic neighbor embedding for multi-feature dimension reduction of remote sensing images. ISPRS J Photogramm Remote Sens 83:30–39
45. Zhang Q, Zhang L, Zhang L et al (2015) Ensemble manifold regularized sparse low-rank approximation for multiview feature embedding. Pattern Recognit 48(10):3102–3112
46. Zhang L, Zhang Q, Du B et al (2016) Simultaneous spectral-spatial feature selection and extraction for hyperspectral images. IEEE Trans Cybern 48(1):16–28
47. Zhu X, Zhang L, Zhang L et al (2016) Multidomain subspace classification for hyperspectral images. IEEE Trans Geosci Remote Sens 54(10):6138–6150

Affiliations

Li Li[1,2] · Hongwei Ge[1,2] · Jianqiang Gao[3] · Yixin Zhang[4] · Yubing Tong[5] · Jun Sun[1,2]

✉ Hongwei Ge
  ghw8601@163.com

✉ Jianqiang Gao
  gaojianqiang@mail.jnmc.edu.cn; jianqianggaohh@126.com

  Li Li
  liliiot@163.com; li_iot@vip.jiangnan.edu.cn

Yixin Zhang
zyx@jiangnan.edu.cn

Yubing Tong
yubing@mail.med.upenn.edu

Jun Sun
sunjun_wx@hotmail.com

1   Jiangsu Provincial Engineering Laboratory of Pattern Recognition and Computational Intelligence,
    Jiangnan University, Wuxi 214122, China

2   School of Internet of Things, Jiangnan University, Wuxi 214122, Jiangsu, China

3   School of Medical Information Engineering, Jining Medical University, Rizhao 276826, Shandong,
    China

4   School of Science, Jiangnan University, Wuxi 214122, Jiangsu, China

5   Medical Image Processing Group, Department of Radiology, University of Pennsylvania,
    Philadelphia, PA 19104, USA