Unsupervised Band Selection Based on Evolutionary Multiobjective Optimization for Hyperspectral Images

Maoguo Gong, *Senior Member, IEEE*, Mingyang Zhang, and Yuan Yuan, *Senior Member, IEEE*

***Abstract*—Band selection is an important preprocessing step for hyperspectral image processing. Many valid criteria have been proposed for band selection, and these criteria model band selection as a single-objective optimization problem. In this paper, a novel multiobjective model is first built for band selection. In this model, two objective functions with a conflicting relationship are designed. One objective function is set as information entropy to represent the information contained in the selected band subsets, and the other one is set as the number of selected bands. Then, based on this model, a new unsupervised band selection method called multiobjective optimization band selection (MOBS) is proposed. In the MOBS method, these two objective functions are optimized simultaneously by a multiobjective evolutionary algorithm to find the best tradeoff solutions. The proposed method shows two unique characters. It can obtain a series of band subsets with different numbers of bands in a single run to offer more options for decision makers. Moreover, these band subsets with different numbers of bands can communicate with each other and have a coevolutionary relationship, which means that they can be optimized in a cooperative way. Since it is unsupervised, the proposed algorithm is compared with some related and recent unsupervised methods for hyperspectral image band selection to evaluate the quality of the obtained band subsets. Experimental results show that the proposed method can generate a set of band subsets with different numbers of bands in a single run and that these band subsets have a stable good performance on classification for different data sets.**

***Index Terms*—Band selection, evolutionary algorithm (EA), hyperspectral image, multiobjective optimization.**

I. Introduction

HYPERSPECTRAL images contain rich information on a wide range of spectra with a high spectral resolution [1]. Due to this character, hyperspectral images have been successfully introduced in many applications, such as target detection [2], medical imaging [3], earth monitoring [4], etc. However, the detailed spectra lead to an increase in the dimensions of the data, which brings the "Hughes phenomenon" for classification. The "Hughes phenomenon" is that an increase in dimensions of limited training samples will cause a decrease in classification accuracy [5], which brings a huge challenge for classification. Moreover, hyperspectral images consist of hundreds of spectra, which are highly correlated in adjacent bands. Heavy redundancy among these adjacent bands makes no contribution to classification and even causes misclassification, which brings another challenge for classification. Therefore, dimensionality reduction is required for the hyperspectral image processing [6], [7], which can reduce the dimensions of the data and select meaningful features to represent the original full feature set without losing crucial information.

For dimensionality reduction, feature extraction and feature selection are the two most popular ways. Feature extraction is used to find an appropriate mapping to transform a high-dimensionality feature space into a low-dimensionality feature space. In this process, the most informative content of the original high-dimensionality feature space is preserved. The related works are principal component analysis [8], independent component analysis [9], Fisher's linear discriminant analysis [10], Gaussian process latent variable model [11], and local linear embedding [12]. On the contrary, feature selection chooses the most representative feature subsets from the original set. Compared with feature extraction, feature selection can preserve the physical information of the original data [13], [14]. The selected feature subset is not distorted, and this characteristic is crucial to represent the significant information of the hyperspectral image. The bands of a hyperspectral image can be considered as features. Therefore, feature selection, i.e., band selection, is a hot issue in dimensionality reduction.

Band selection can be implemented in two ways: supervised and unsupervised. The supervised band selection often sets a criterion function to measure the similarity between the selected band and the labeled image and adopt some optimization strategies to find the best band subsets [15]–[19]. The unsupervised band selection tends to find a feature subset with the most representative information among the whole bands. For

Manuscript received June 12, 2015; revised July 19, 2015; accepted July 23, 2015. This work was supported in part by the National Natural Science Foundation of China under Grants 61273317, 61422209, and 61172143, the National Program for Support of Top-notch Young Professionals of China, the Specialized Research Fund for the Doctoral Program of Higher Education under Grant 20130203110011, and the Fundamental Research Fund for the Central Universities under Grant K5051202053.

M. Gong and M. Zhang are with the Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, International Research Center for Intelligent Perception and Computation, Xidian University, Xi'an 710071, China (e-mail: gong@ieee.org).

Y. Yuan is with the Center for OPTical IMagery Analysis and Learning (OPTIMAL), State Key Laboratory of Transient Optics and Photonics, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, China (e-mail: yuany@opt.ac.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TGRS.2015.2461653

the unsupervised band selection, some efficient criteria have been proposed, such as various variants of PCA [20]–[23] and band-clustering-based criteria [24]–[26]. Stochastic searching method, clustering method, matrix computation, and ranking method have been adopted to optimize these criteria [7], [20]–[29]. These supervised and unsupervised methods mostly transform the band selection into a single-objective optimization problem, in which they devote to maintaining significant information or reducing the redundancy. However, for band selection methods whether supervised or unsupervised, the appropriate number of bands required for band selection to preserve the significant information is still an open issue [7]. In other words, it is difficult to determine how many bands are required to be selected into band subsets.

In this paper, to address this open issue, a novel multiobjective model for band selection is proposed, based on which a multiobjective optimization band selection (MOBS) method is proposed. In this method, the band selection problem is modeled as a multiobjective optimization problem (MOP), and two objective functions with a conflicting relationship are designed. One is set as the inverse of the sum of the information entropy of the selected bands, and the other one is the number of selected bands. They are required to be minimized simultaneously with a conflicting relationship. To solve this MOP, a multiobjective evolutionary algorithm (MOEA) is proposed to optimize the two objective functions simultaneously. Meanwhile, a novel update operation is also designed to restrain the redundancy among bands in the proposed MOBS which is based on the Kullback–Leibler (KL) divergence [30] introduced from information theory. The KL divergence is set as a guideline for band subsets to reduce the redundancy among the selected bands. In the framework of the proposed MOBS, band subsets with different numbers of bands can be optimized in a cooperative way. They can communicate with each other by exchanging their bands. The relationship among band subsets with different numbers of bands is used to improve the quality of results. Compared with other methods, the proposed method can obtain a set of band subsets with different numbers of bands in a single run, and the communication among band subsets with different numbers of bands is used to promote the optimization process.

The rest of this paper is organized as follows. Section II introduces some background knowledge and the motivation. Section III describes the proposed algorithm. Section IV shows the experimental results on three data sets: Indian Pines, Pavia University, and Salinas. Finally, Section V gives the conclusion remarks of this paper.

II. Background and Motivation

A. Problem Statement and Literature Review

Due to the abundant spectral information contained in the hyperspectral image, unsupervised band selection is designed to select the most informative band subset to represent the original band set. For band selection, there is still an open issue on how many bands are required to be selected to represent the original full bands. In many papers, researchers have proposed different methods to handle this issue. In [7], Chang proposed to use a concept of virtual dimensionality (VD) [31] to estimate the appropriate number of selected bands which is also used in [32]. It can be interpreted that one signal source can be only located to a separate dimension, which means that the same size of dimensions as the VD is required to represent the original data. Meanwhile, it is also noted that the value of VD can only provide an estimate for the number of bands and not the exact one which is hardly possible to know for real data [7]. For band selection methods based on band sorting [23], [33], bands are ranked according to some criteria. Then, these ranked bands are selected from top to bottom. The number of selected bands is equal to the rank of the last selected band. Some works consider the band selection as a clustering problem. In [26] and [27], the number of selected bands is set as the number of clusters. Moreover, for different applications, different methods are proposed. For classification, some works list a series of band subsets with a range of number of dimensions to show classification performance, such as [29] and [34]–[37]. These methods always get these band subsets by implementing separate runs. For visualization, the number of selected bands is set to 3 to match the RGB three channels [37]. For spectral unmixing, Chang proposed that the number of selected bands is equal to twice the number of endmembers for endmember separation [38].

B. Brief Introduction to Evolutionary Multiobjective Optimization

Without loss of generality, the MOPs in this paper can be set for minimization. A MOP can be stated as the following:

$$\begin{aligned} &\text{minimize } F(\mathbf{x}) = (f_1(\mathbf{x}), \ldots, f_m(\mathbf{x}))^T \\ &\text{subject to } \mathbf{x} \in \Omega \end{aligned} \tag{1}$$

where Ω is the feasible space, $\mathbf{x}$ is a solution to the MOP, R^m is the objective space, and $F\colon \Omega \to R^m$ consists of m real-valued objective functions. In most instances, the objectives in a MOP are contradictory to each other, which means that no point in feasible space can minimize all of the objectives simultaneously. Hence, multiobjective optimization [39], [40] is designed to find the best tradeoff relationship among them simultaneously.

For minimization, a solution $\mathbf{x_u}$ is said to dominate another solution $\mathbf{x_v}$ if and only if

$$\begin{aligned} &\forall\, i = 1, 2, \ldots, m f_i(\mathbf{x_u}) \le f_i(\mathbf{x_v}) \\ &\exists\, j = 1, 2, \ldots, m f_j(\mathbf{x_u}) < f_j(\mathbf{x_v}). \end{aligned} \tag{2}$$

A point $\mathbf{x}^*$ in Ω is called a Pareto optimal solution to (1) on the condition that there is no such point $\mathbf{x}$ in Ω that makes $F(\mathbf{x})$ dominate $F(\mathbf{x}^*)$. Then, $F(\mathbf{x}^*)$ is termed as the Pareto optimal vector. The objectives in a Pareto optimal vector have such a relationship that a decrease in one objective causes an increase in the others. All of the Pareto optimal points constitute a set called Pareto optimal set [41], and their corresponding Pareto optimal objective vectors are termed as Pareto optimal front (PF) [41].

For multiobjective optimization, it has been recognized that evolutionary algorithms (EAs) are well suited because EAs can deal with a set of possible solutions simultaneously [39], [42]. Since [43], various EAs to deal with MOPs have been proposed,

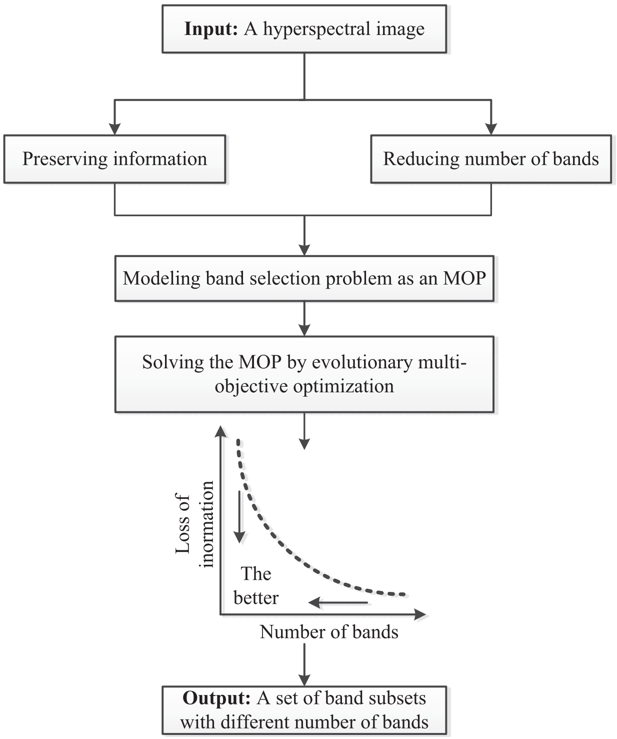

Input: A hyperspectral image
Preserving information
Reducing number of bands
Modeling band selection problem as an MOP
Solving the MOP by evolutionary multi-objective optimization
Loss of inormation
The better
Number of bands
Output: A set of band subsets with different number of bands

Fig. 1. Framework of MOBS.

such as [44]–[48], and these EAs are termed as MOEAs. MOEAs seek to obtain a set of Pareto optimal solutions for approximating the true PF in a single run. Due to the character of MOEAs, it offers a new way to solve the problem in hyperspectral image processing. There has been an instructive and efficient application of MOEAs in clustering of hyperspectral image [49], which receives a good result.

C. Motivation of Using Evolutionary Multiobjective Optimization

As described previously, it is difficult to figure out the appropriate number of the selected bands [7]. This can be taken as an ill-posed problem. Therefore, an appropriate way to address this issue is to show a series of band subsets with different numbers of bands. With more options of band subsets, decision makers can judge and select the most appropriate band subset according to the problem required. Most band selection methods fix the number of selected bands in each run. Consequently, if a series of band subsets with a range of proper numbers of bands are required, these algorithms have to run repeatedly, and the number of bands needs to be preset in each run. Therefore, as shown in Fig. 1, the band selection problem is formulated as a MOP to solve this limitation. The framework of evolutionary multiobjective optimization is adopted. As described in Section II-B, a MOEA has the ability to handle a series of solutions simultaneously. For band selection, these solutions can be set as band subsets with different numbers of bands, and they can be obtained in a single run.

There is another advantage in MOEAs that the communication among band subsets with different numbers of bands is used. The communication here refers to that band subsets with different numbers of bands exchange their bands and update themselves under some guidelines, which is crucial for the optimization process. In EAs, these band subsets with different numbers of bands can be set as different populations. The communication among different populations can be taken as the concept of coevolutionary mechanisms. In [50]–[52], it has been shown that coevolutionary mechanisms have the ability to improve the efficiency of the optimization process significantly. Therefore, a promising way for multiobjective optimization is to combine coevolutionary mechanisms with our optimization process. Therefore, to adopt coevolutionary mechanisms in our method, we propose a MOEA based on decomposition by dividing the MOP into a series of subproblems. Then, these subproblems coevolve for Pareto optimal solutions in a cooperative way. The proposed MOEA constructs a bridge for band subsets with different numbers of bands to communicate with each other. Moreover, MOEAs have the advantage that they have few restrictions on the mathematical model of objective functions [39], [40].

III. Proposed MOBS Method

A. Novel Multiobjective Model for Band Selection

Band selection is designed to reduce the computation complexity, maintain significant information, and reduce redundancy [23]. As discussed in Section I, it is hard to find out the appropriate number of selected bands, since this is an ill-posed problem. If the number of bands is too small, the computation complexity is low, but significant information may be missing and vice versa. To exploit this issue, we build a novel multiobjective model to deal with this ill-posed problem. In this model, the band selection problem is formulated as a MOP. The two objective functions of the MOP are designed as follows:

$$\min F(\mathbf{x}) = \begin{cases} f_1(\mathbf{x}) = \text{length}(\mathbf{x}) \\ f_2(\mathbf{x}) = 1 \Big/ \sum\limits_{i=1}^{f_1(\mathbf{x})} H(\mathbf{x_i}) \end{cases} \tag{3}$$

where $\mathbf{x}$ is a solution vector containing a selected band subset, length is the number of selected bands, and $H(x_i)$ is the information entropy of the ith selected band. These two objective functions need to be minimized simultaneously. From the definition of the Shannon entropy [30], the entropy of a continuous random variable Y with a probability density function $f(y)$ $y \in \Omega$ can be written as the following:

$$h(Y) = -\int\limits_{\Omega} f(y) \log f(y) dx. \tag{4}$$

For a discrete random variable Y, its entropy can be expressed as follows:

$$H(Y) = -\sum_{y \in \Omega} p(y) \log p(y)$$

$$\text{subject to } \sum_{y \in \Omega} p(y) = 1 \tag{5}$$

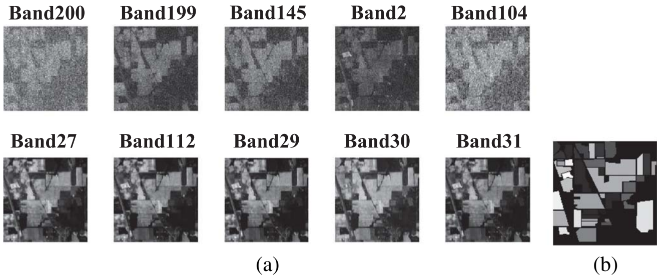


Fig. 2. (a) Examples of AVIRIS 200th, 199th, 145th, 2nd, and 104th band images with the least entropy and 27th, 112th, 29th, 30th, and 31st band images with the largest entropy. (b) Reference label image.

where $p(y)$ stands for the probability of which an event $y \in \Omega$ occurs. For hyperspectral images, each band is considered as a collection of outputs of a random variable Y, and the histogram of the band can be taken as the probability distribution of the band. Consequently, the estimation of the probability function $p(y)$ can be calculated as follows:

$$p(y) = \frac{h(y)}{mn} \tag{6}$$

where $h(y)$ represents the gray-level histogram of a band, and a band consists of $m \times n$ pixels. Hence, the entropy of a band can be considered as the amount of the information content of a band [53], [54]. According to the character of entropy, the larger the entropy is, the more image details the band contains. As shown in Fig. 2, from the data set Indian Pines described in Section IV, we select five bands: 200th, 199th, 145th, 2nd, and 104th, which have the least five entropies, and the other five bands: 27th, 112th, 29th, 30th, and 31st, which have the largest five entropies. It is obvious that the five bands with large entropy are more distinct than those five bands with less entropy. Then, it can be inferred that the bands with larger entropy are more informative and more similar to the reference label image which is crucial to classification. The bands with less entropy are more noisy and indistinct. Thus, the information entropy criterion is designed as the second objective function to preserve the image details and restrain the noisy bands. The form of the second objective function is designed to construct a conflicting relationship with the first objective function. In this way, the band selection problem is transformed into a MOP.

B. Tchebycheff Decomposition Strategy

First, a Pareto optimal solution to a MOP can be approximated by an optimal solution to a single scalar optimization problem where the objective is an aggregation of all of the objectives in the MOP. Thus, the approximation of a MOP can be decomposed into a series of single scalar objective optimization subproblems, which is a fundamental idea for many traditional mathematical methods to approximate the PF [41], [48], [55]. Second, in this paper, we design a novel band selection algorithm which can generate a series of band subsets with different numbers of bands in a single run. As described in Section II-C, coevolutionary mechanisms can improve the efficiency of the optimization process significantly. Therefore, a framework is required in which a series of band subsets with different numbers of bands can be optimized in a cooperative way. Based on the two considerations discussed previously, in the MOBS, we divide the MOP into a series of subproblems so that these band subsets with different band numbers can be considered as these single scalar optimization subproblems. In this way, it is convenient for traditional mathematical methods to solve a MOP, and it also constructs the required framework discussed previously. In the proposed method, the decomposition strategy is alternative.

Several methods for transforming the MOP into a series of single scalar optimization problems have been proposed, such as the weighted sum approach [41], [55], the Tchebycheff approach (TE) [41], [55], and the boundary intersection approach [56], [57]. Among them, the TE approach is widely used for its good performance. The derivation of the TE approach is as follows [41], [55].

A widely used method to aggregate the objective functions is the simple weighted sum approach, which is shown in

$$g = \sum_{i=1}^{m} \left(\lambda_i f_i(\mathbf{x})\right)^P \tag{7}$$

where $f_i(\mathbf{x})$ is the ith objective function in a MOP and m is the number of objective functions. λ_i is the weight set typically by the decision maker such that $\sum_{i=1}^{m} \lambda_i = 1$. However, in some cases, a function's independent optimal value may be unattainable. To approximate all of the functions' independent optimal values efficiently, in the function value space, a reference point is defined, which represents the optimal values that all of the functions can or cannot obtain. With the introduction of the reference point, the most common extension of (7) is as follows:

$$g = \left[\sum_{i=1}^{m} \lambda_i^P \left(f_i(\mathbf{x}) - z_i^*\right)^P\right]^{\frac{1}{P}} \tag{8}$$

where $\mathbf{z}^* = (z_1^*, \ldots, z_m^*)^T$ is the reference point. When $P \to \infty$, the limit of (8) is the form of TE approach, which is shown as follows:

$$\begin{aligned} g^{te}(\mathbf{x}|\boldsymbol{\lambda}, \mathbf{z}^*) &= \max_{1 \le i \le m} \left(\lambda_i \left|f_i(\mathbf{x}) - z_i^*\right|\right) \\ \text{subject to } &\mathbf{x} \in \Omega \end{aligned} \tag{9}$$

where $g^{te}(\mathbf{x}|\boldsymbol{\lambda}, \mathbf{z}^*)$ is a single scalar objective function. Generally, $\mathbf{z}_{\mathbf{i}}^* = \min(f_i(\mathbf{x})|\mathbf{x} \in \Omega)$ for every $i = 1, \ldots, m$.

For each subproblem, a weight vector $\boldsymbol{\lambda}^{\mathbf{i}}$ is allocated. $\boldsymbol{\lambda}^{\mathbf{i}}$ defines a direction of optimization for the ith subproblem. As shown in Fig. 3, a point $\mathbf{P}$ is on the direction determined by the $\boldsymbol{\lambda}^{\mathbf{i}}$. The ideal point is $\mathbf{Z}^*$. Obviously, the two contour lines represent the two distances of the two objectives between the point $\mathbf{P}$ and the ideal point $\mathbf{Z}^*$. In (9), the single scalar objective subproblem $g^{te}(\mathbf{x}|\boldsymbol{\lambda}, \mathbf{z}^*)$ represents the contour line with a larger value. Iteratively, when $g^{te}(\mathbf{x}|\boldsymbol{\lambda}, \mathbf{z}^*)$ is minimized, the point $\mathbf{P}$ moves toward the ideal point $\mathbf{Z}^*$ along the direction determined by the $\boldsymbol{\lambda}^{\mathbf{i}}$. Finally, the optimal solution to (9) is obtained, which is also a Pareto optimal solution to the MOP. In TE, a MOP is divided into a series of subproblems with a series of $\boldsymbol{\lambda}^{\mathbf{i}}$, which means that a series of directions of

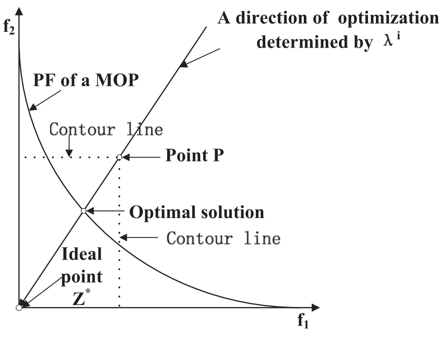


Fig. 3. Illustration of the Tchebycheff approach.

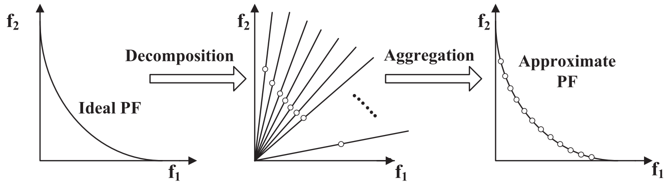


Fig. 4. Process of decomposition strategy.

optimization is determined. In each direction, an optimal solution to a single scalar subproblem is obtained. Then, these optimal solutions aggregate to approximate the Pareto optimal set of the MOP. This process is shown in Fig. 4.

C. Restriction in Local Search

Due to the redundancy among bands, a restriction strategy is designed for subproblems coevolving to reduce redundancy. In the information theory [30], two popular criteria are widely used to measure the dissimilarity: mutual information (MI) and KL divergence. In some related works [15], [16], [19], MI is well used to measure the dissimilarity between the selected band and the label. However, it is required to calculate joint distribution in MI [30], which is quite time-consuming for hyperspectral images. Considering the low complexity of computation, we select the KL as the dissimilarity criterion in the proposed method. The performances of these two criteria in the proposed framework will be shown in Section IV-E.

The definition of KL in the discrete domain can be stated as follows [30]:

$$D_{\mathrm{KL}}(Y_i||Y_j) = \sum_{y\in\Omega} p_i(y)\log\frac{p_i(y)}{p_j(y)} \tag{10}$$

where two random variables Y_i and Y_j are set in space Ω, and $p_i(y)$ and $p_j(y)$ are the probability distribution of these two random variables. As a dissimilarity measurement, the formula is required to be symmetrical for the two variables. Thus, symmetrical KL divergence is often adopted [26]. Here, we revise (10) as the following:

$$D_{\mathrm{KLS}}(Y_i||Y_j) = \sum_{y\in\Omega} p_i(y)\log\frac{p_i(y)}{p_j(y)} + \sum_{y\in\Omega} p_j(y)\log\frac{p_j(y)}{p_i(y)}. \tag{11}$$

As we can see from (11), the KL divergence is nonnegative, which is suitable for a dissimilarity measurement, and only when $p_i(y)$ is equal to $p_j(y)$ can the KL be zero. The KL divergence can detect the dissimilarity between two distributions [30]. The larger the KL is, the more distinct the two distributions are.

For a hyperspectral image, the two random variables Y_i and Y_j represent the ith and jth bands, and the probability distribution of these two random variables can be estimated by calculating the $p(y)$ as (6). For high-dimensionality data, the computation of the gray-level histogram is affordable. Therefore, the KL divergence is introduced as a restriction in the proposed algorithm. The solutions will learn and communicate collaboratively under the instruction of the KL-divergence-based criterion. Details will be described in Section III-D.

D. Implementation of the MOBS

Through the TE approach, the MOP model is divided into N subproblem model. Each subproblem has the following elements: weight vector, values of objective functions, neighborhood, reference point, and band subsets, i.e., solution. These subproblems are implemented with the newly designed update operation to update their elements and optimize their objective functions.

In the update operation, every solution is allocated a neighborhood, and a local search is implemented in the neighborhood. The band subsets in a neighborhood can exchange their elements with each other to generate new band subsets, and these new band subsets will coevolve under the instruction of the objective functions and the KL-based restriction. Under the instructive measurements, the update operation can focus on both maintaining image detail content and reducing the redundancy among the selected bands. In this way, the quality of these band subsets is ensured. In each generation, the quality of these band subsets is improved through the update operation. Iteratively, the best band subsets will be obtained. The detailed process of the MOBS algorithm is as follows.

1) Initialization:

a) Set the weight vectors, and generate initial solutions:

We set the weight vector $[\boldsymbol{\lambda}^{\mathbf{1}},\ldots,\boldsymbol{\lambda}^{\mathbf{N}}]^T$ as an even distribution, and through the TE approach, the MOP model is decomposed into N subproblem model. The subproblems are numbered as i, where $i = 1,\ldots,N$. For each subproblem i, a randomly generated solution $\mathbf{x}^{\mathbf{i}}$ is allocated. Here, the solution is a real number coding vector with a size M, where M represents the number of selected bands, and the real numbers of a solution represent the selected bands, e.g., a solution vector: $[10, 20, 30]^T$ means selecting the 10th, 20th, and 30th bands.

b) Set the reference point:

For the two objectives in the proposed method, we set a reference point vector $\mathbf{z}^* = [z_1, z_2]^T$ as the following:

$$z_i = \min\left(f_i^1(\mathbf{x}),\ldots,f_i^N(\mathbf{x})\right)\ i = 1,2 \tag{12}$$

where N is the number of subproblems and i is the serial index of the objectives. As (12) shows, the reference point is just the minimum value of the corresponding N objectives.

c) Calculate the initial subproblems and set the neighborhood: Calculate the single objective of the subproblems as (13), and allocate each single objective to the corresponding subproblem

$$g^{te}(\mathbf{x}|\boldsymbol{\lambda}, \mathbf{z}^*) = \max_{1 \le i \le 2} \left(\lambda_i |f_i(\mathbf{x}) - z_i^*|\right). \tag{13}$$

It is noticed that g^{te} is consecutive of $\boldsymbol{\lambda}$ and that the optimal solution of $g^{te}(\mathbf{x}|\boldsymbol{\lambda}^{\mathbf{i}}, \mathbf{z}^*)$ is closed to the ones of the subproblems which have a closed weight vector. The information contained in the neighbors of a subproblem can help in optimizing the subproblem. Therefore, we set a neighborhood to store the neighbors of each subproblem. In the proposed method, the neighborhood of the ith subproblem is defined as a set of weight vectors which are closest to the weight vector of the ith subproblem, where $i = 1, \ldots, N$. The distance metrics here is Euclidean distance, and we calculate the Euclidean distance between every two weight vectors to select T closest weight vectors for every weight vector. Then, every subproblem is allocated with a weight vector and a neighborhood. The neighborhood is stored as $NH(j) = (j_1, \ldots, j_T)$, where j represents the serial index of the subproblems. In addition, since the neighborhoods are decided by the weight vectors and the weight vectors are unchanged, the neighborhood of every subproblem is fixed. While with the update of generations, the solution of every subproblem is improved from generation to generation.

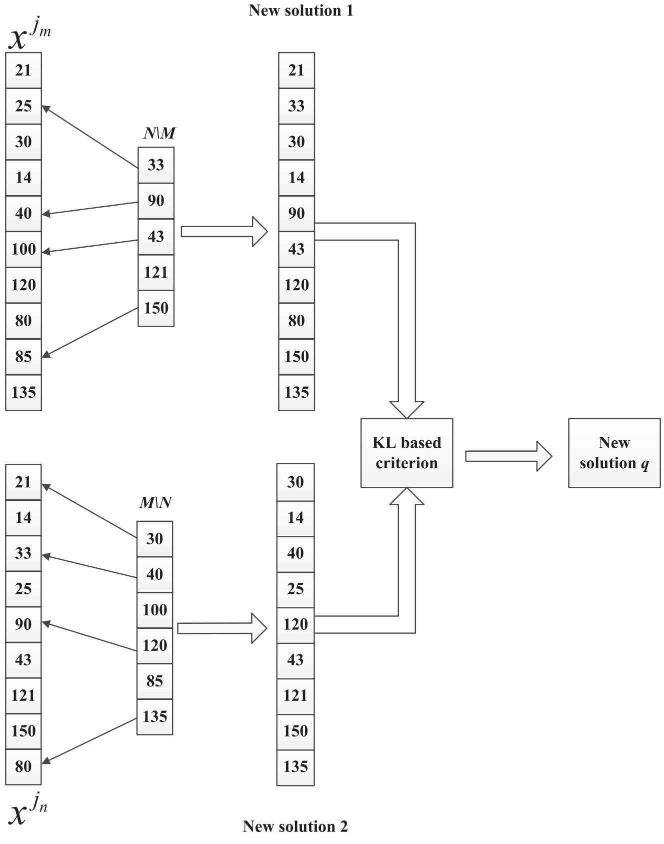


Fig. 5. Illustration of the genetic operation.

2) Update Operation: For each ith subproblem, $i = 1, \ldots, N$, do the following steps.

a) *Genetic operation for generating new solutions:* We select two indexes j_m, j_n randomly from the corresponding neighborhood $NH(j)$. Then, a new solution $\mathbf{q}$ is generated with two parent solutions $\mathbf{x}^{\mathbf{j_m}}$ and $\mathbf{x}^{\mathbf{j_n}}$ by using genetic operation. Traditional genetic operation needs the solutions with the same size, which is obviously impractical in this case. To overcome this problem, a new genetic operation is designed. Different from traditional binary genetic operation, in the proposed method, the real number coding is adopted. Therefore, the length of the solution vectors may not be the same. First, we detect the two parent $\mathbf{x}^{\mathbf{j_m}}$ and $\mathbf{x}^{\mathbf{j_n}}$ solutions to build two sets: $\mathbf{M}$ and $\mathbf{N}$ which represent the selected bands in $\mathbf{x}^{\mathbf{j_m}}$ and $\mathbf{x}^{\mathbf{j_n}}$, respectively. From the two sets $\mathbf{M}$ and $\mathbf{N}$, two difference sets $\mathbf{M} \setminus \mathbf{N}$ and $\mathbf{N} \setminus \mathbf{M}$ can be obtained. Then, we randomly select h elements in $\mathbf{x}^{\mathbf{j_n}}$ to be replaced by h elements randomly selected in $\mathbf{M} \setminus \mathbf{N}$. Meanwhile, another h elements in $\mathbf{x}^{\mathbf{j_m}}$ are selected to be replaced by h elements randomly selected in $\mathbf{N} \setminus \mathbf{M}$. The number h is supposed to be like this: $h \le \min(n_{j_m}, n_{j_n})$, where n_i represents the number of elements in $\mathbf{x}^{\mathbf{i}}$, $i = j_m, j_n$. Then, we obtain two new solutions: $\mathbf{q}^{\mathbf{j_m}}$ and $\mathbf{q}^{\mathbf{j_n}}$. Finally, the restriction strategy is introduced. We calculate the KL divergence of the two new solutions as (11) and compare their values. The solution with a larger value of KL divergence is chosen as the final new solution $\mathbf{q}$. In this process, different solutions exchange elements with each other in a neighborhood to generate two new solutions, and based on the KL divergence criterion, the two new solutions complete the competition strategy to generate a final new solution with a better property, which is supposed to be a good solution to the ith subproblem. For example, a solution $\mathbf{x}^{\mathbf{j_m}}$ with 11 bands and a solution $\mathbf{x}^{\mathbf{j_n}}$ with 10 bands are given as Fig. 5, and $\mathbf{M} \setminus \mathbf{N}$ and $\mathbf{N} \setminus \mathbf{M}$ can be obtained. There are five elements in $\mathbf{N} \setminus \mathbf{M}$ and six elements in $\mathbf{M} \setminus \mathbf{N}$. Obviously, $h \le 5$. Without loss of generality, let us set $h = 4$. As described previously, four positions are randomly selected in $\mathbf{x}^{\mathbf{j_m}}$, $\mathbf{x}^{\mathbf{j_n}}$, $\mathbf{M} \setminus \mathbf{N}$, and $\mathbf{N} \setminus \mathbf{M}$. Crossovers occur between $\mathbf{x}^{\mathbf{j_m}}$ and $\mathbf{N} \setminus \mathbf{M}$, and $\mathbf{x}^{\mathbf{j_n}}$ and $\mathbf{M} \setminus \mathbf{N}$. Then, calculate the KL divergence of the two new solutions, and compare their values. Finally, we get the final new solution $\mathbf{q}$ with a larger value of KL divergence. This process is shown in Fig. 5.

b) *Update of the reference points:* For each $i = 1, 2$, if reference point $z_i > f_i(\mathbf{q})$, then $z_i = f_i(\mathbf{q})$. In this process, the new solution $\mathbf{q}$ is put into calculation of the two objectives, and we obtain $f_i(\mathbf{q})$, $i = 1, 2$. Then, we compare the new objectives with the current reference points $\mathbf{z}^*$. As our method is for minimizing optimization, when the values of the new objectives are less than the values of the reference points, the reference points are replaced by the new objectives.

c) *Update of the current solutions:* For the current solution $\mathbf{x}^{\mathbf{i}}$ to the ith subproblem, if $g^{te}(\mathbf{q}|\boldsymbol{\lambda}^{\mathbf{i}}, \mathbf{z}) \le g^{te}(\mathbf{x}^{\mathbf{i}}|\boldsymbol{\lambda}^{\mathbf{i}}, \mathbf{z})$, then $\mathbf{x}^{\mathbf{i}} = \mathbf{q}$. In this process, the new solution $\mathbf{q}$ and the current solution $\mathbf{x}^{\mathbf{i}}$ are put into calculation of the ith subproblem objective, and $g^{te}(\mathbf{q}|\boldsymbol{\lambda}^{\mathbf{i}}, \mathbf{z})$ and $g^{te}(\mathbf{x}^{\mathbf{i}}|\boldsymbol{\lambda}^{\mathbf{i}}, \mathbf{z})$

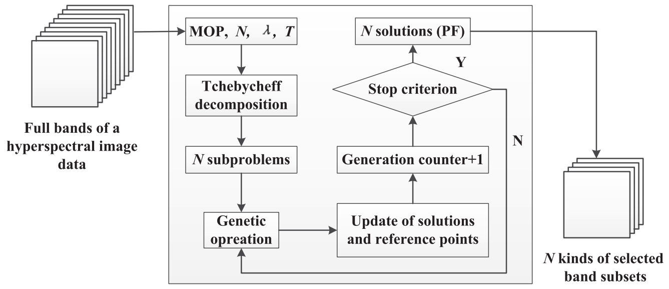

Full bands of a hyperspectral image data
MOP, N, λ, T
Tchebycheff decomposition
N subproblems
Genetic opreation
Update of solutions and reference points
Generation counter+1
Stop criterion
Y
N
N solutions (PF)
N kinds of selected band subsets

Fig. 6. Process of the MOBS.

are obtained. Then, compare their values, and update the current solutions.

3) Stopping Criterion: The number of generations has been prefixed. When the number of generation reaches the prefixed number, the algorithm stops.

The whole process of the MOBS is shown in Fig. 6. From the aforementioned description, the band subsets with different numbers of bands are involved in the optimization process, where the coevolutionary mechanism is combined with the optimization process. Moreover, the information-entropy-based objective function can restrain the noisy bands and exploit the bands with detailed information. Meanwhile, the KL-based restriction strategy can reduce the redundancy among the selected bands. Thus, it can be inferred that the MOBS can obtain band subsets with good quality.

IV. Experimental Studies

In this section, to test the performance of the proposed method, the following unsupervised algorithms based on different frameworks are adopted as the comparison algorithms. The first method is maximum-variance PCA (MVPCA) [23], which is a classic band selection method based on band sorting framework. MVPCA sorts bands by constructing a loading factor matrix based on PCA transformation of the original data and then sorts the variances from high to low. The second and third methods are based on clustering framework termed as Walumi and Waludi [26]. Information criteria are adopted into the process of clustering to reduce redundancy among bands. The fourth method is proposed recently in [34], termed as VGBS. VGBS selects the bands with the maximum determinant of the covariance matrix with a fast computation strategy. Moreover, another latest proposed method based on a multitask sparsity pursuit (MTSP) framework [37] is also adopted as a competitor, which transforms a band selection problem into a single-objective optimization problem and optimizes it with an immune clonal strategy. The last competitor is a fundamental band selection method termed as uniform band selection (UBS) [7], [23], which selects band subsets uniformly from original full bands. The experiments will be implemented on three benchmark hyperspectral image data sets to evaluate the performances of the proposed method and competitors.

On each hyperspectral image, the following comparisons are conducted:

1) classification performance;
2) analysis of the selected bands.

Two popular indexes, namely, overall accuracy (OA) and average accuracy (AA), will be involved in the experiments to evaluate the performance of the classification. The OA represents the proportion of the exactly classified samples to the total samples, which reflects the classification performance of the pixels in the whole image. The OA will be tested in band subsets with different numbers of bands to show the performance of the classification in various numbers of features. The AA represents the mean of the accuracy values of all of the ground truth classes, which reflects the performance of the classification for each class. Thus, the AA is tested in a fixed number of bands, where all of the competitors can obtain a good OA to ensure that the value of AA is meaningful.

To obtain classifier-robust classification results, three widely used supervised classifiers are adopted here, which are extreme learning machine (ELM) [58]–[60], support vector machine (SVM) [61], [62], and k-nearest neighborhood (KNN) [63]. Two parameters in ELM need to be preset here: the number of hidden neural nodes, which is set as 800, and the type of activation function, which is set as a sigmoidal function. In SVM, a one-against-all approach is adopted to deal with multiclass classification. The kernel function is set as a radial basis function, and the optimal two parameters are determined by fivefold cross validation. In KNN, the number of nearest neighbors is set as 3, and the distance metric is set as the Euclidean distance. In all of the experiments, 50% samples of each class are randomly selected to constitute the training sets, and the remaining samples constitute the testing sets.

For analyzing the selected bands, two aspects are considered: the contained information and the redundancy among the selected bands. As mentioned previously, the entropy is adopted to measure the informative bands. Meanwhile, we adopt the mean spectral divergence (MSD) [34], [64] and the mean spectral angle (MSA) [34], [64] as the criteria to evaluate the redundancy among the selected bands. The definitions of the MSD and MSA are as follows:

$$\mathrm{MSD} = \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} D_{\mathrm{KLS}(i,j)} \tag{14}$$

$$\mathrm{MSA} = \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha(i,j) \tag{15}$$

where the D_{KLS} in (14) is just the KL divergence in (11); the $\alpha(i,j)$ in (15) refers to the spectral angle between the ith band B_i and the jth band B_j which is calculated as $\alpha(i,j) = \arccos(B_i^T B_j / \|B_i\| \|B_j\|)$; and n is the number of selected bands. As we can see, the MSD is an information-theory-based criterion, and the MSA is a geometric-based vector-angle criterion [34], [64]. From the two aspects, the evaluation of the mean redundancy among selected bands can be quantitatively expressed. The larger the values of the MSD and MSA are, the less redundancy is contained among the selected bands.

If the number of selected bands is too large, the complexity of the computation is still high. Considering the practical value of the selected bands, we set the number of selected bands ranging from 1 to 30. In the following experiments, we set the number of subproblems N as 150 because there are 30 kinds of numbers of the selected bands, and each number contains five solutions.

TABLE I
GROUND TRUTH CLASSES FOR THE INDIAN PINES AND THEIR RESPECTIVE SAMPLE NUMBER

Class	Ground truth	Samples	Training
C1	Alfalfa	46	23
C2	Corn-notill	1428	714
C3	Corn-mintill	830	415
C4	Corn	237	118
C5	Grass-pasture	483	241
C6	Grass-trees	730	365
C7	Grass-pasture-mowed	28	14
C8	Hay-windrowed	478	239
C9	Oats	20	10
C10	Soybean-notill	972	486
C11	Soybean-mintill	2455	1227
C12	Soybean-clean	593	296
C13	Wheat	205	102
C14	Woods	1265	632
C15	Bldg-Grass-Trees-Drives	386	193
C16	Stone-Steel-Towers	93	46

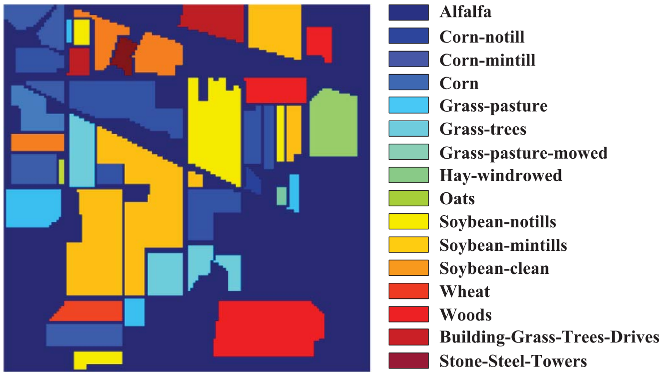


Fig. 7. Indian Pines label image.

The number T of the members in a neighborhood can be set ranging from 5 to 51, and the iteration generation is set as 50.

A. Results on Indian Pines Data Set

This hyperspectral image was gathered by the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) sensor over the Indian Pines in 1992, which is a classic data set widely used in the remote sensing research. It consists of 145 × 145 pixels and 224 spectral reflectance bands. Due to the water absorption, the 104th–108th, the 150th–163th, and the 220th bands are removed manually. Thus, in this experiment, the remaining 200 bands are used for band selection. There are 16 classes ground truth, and the numbers of training samples of each class are shown in Table I. Moreover the 16-class ground truth label image is shown in Fig. 7.

1) Classification Performance: The result of the MOBS is shown in Fig. 8. The horizontal axis represents the value of the first objective function, and the vertical axis represents the value of the second objective function. As we can see, the number of selected bands ranges from 1 to 30. Each number of selected bands has five solutions, and the solution which dominates the other four solutions is chosen as the representative one among the five solutions.

To obtain an average classification result, we implement this experiment 30 times and randomly select the training set and the test set in each time. The numbers of training samples for each class are shown in Table I. The mean OA values and mean AA values obtained under the three classifiers are shown in Fig. 9. Fig. 9(a)–(c) shows the OA values under ELM, SVM, and KNN, respectively. As we can see, MOBS achieves the best or comparable OA when the number of selected bands p is larger than 10 under the three classifiers. When p ranges from 5 to 10, MOBS has comparable results with Walumi and Waludi and better than other competitors under ELM and SVM. Under KNN, MOBS has the best results when p is over 5. When p ranges from 1 to 5, all of the competitors have the comparable results under the three classifiers. Fig. 9(d) shows the values of AA under the three classifiers. In Fig. 9(d), it is obvious that MOBS achieves better AA than other competitors, which reflects a good performance in ground truth classification. It is interesting to notice that the information-theory-based criteria: MOBS, Walumi, and Waludi have better classification performance on Indian Pines.

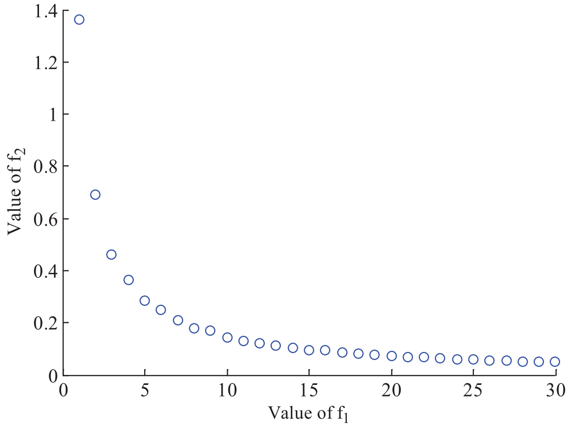


Fig. 8. MOBS results on Indian Pines: Each point represents a band subset.

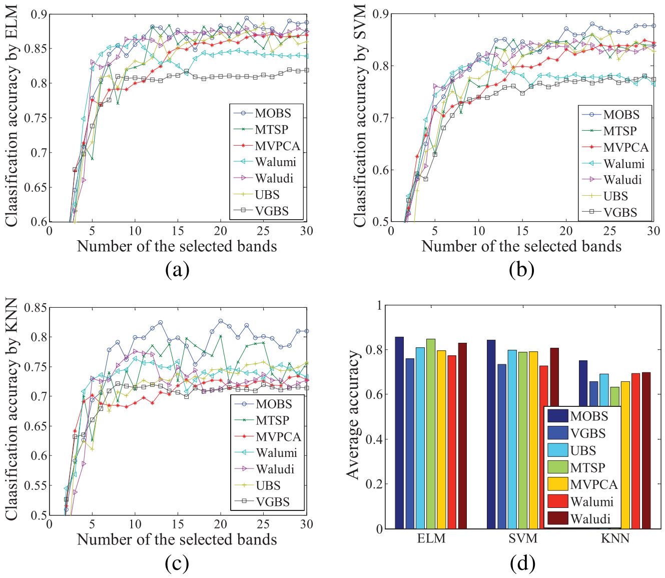


Fig. 9. (a) Overall accuracy by ELM on Indian Pines. (b) Overall accuracy by SVM on Indian Pines. (c) Overall accuracy by KNN on Indian Pines. (d) Average accuracy by ELM, SVM, and KNN with 21 selected bands on Indian Pines.

2) Analysis of the Selected Bands: The representative selected band subset by MOBS is $[9, 12, 13, 15, 21, 36, 65, 66, 69,$

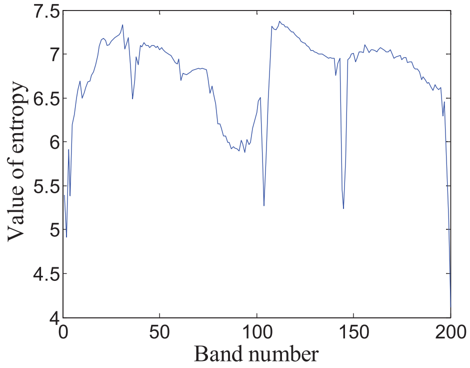
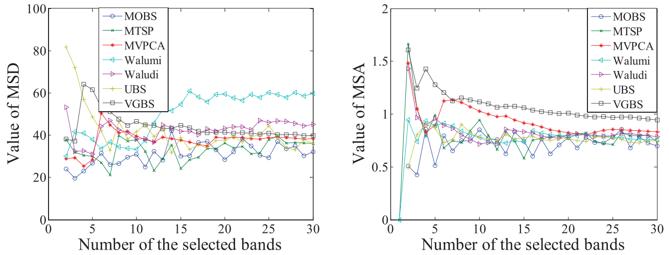


Fig. 10. Entropy of each band on Indian Pines.



(a) (b)

Fig. 11. (a) Value of MSD on Indian Pines. (b) Value of MSA on Indian Pines.

TABLE II
GROUND TRUTH CLASSES FOR THE PAVIA UNIVERSITY AND THEIR RESPECTIVE SAMPLE NUMBER

Class	Ground truth	Samples	Training
C1	Asphalt	6631	3315
C2	Meadows	18649	9324
C3	Gravel	2099	1049
C4	Trees	3064	1532
C5	Painted metal sheets	1345	672
C6	Bare Soil	5029	2514
C7	Bitumen	1330	665
C8	Self-Blocking Bricks	3682	1841
C9	Shadows	947	473

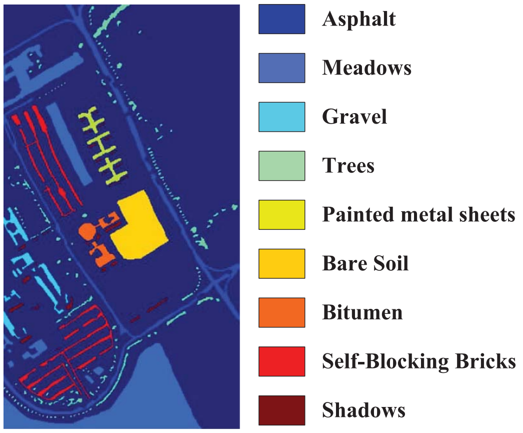


Fig. 12. University of Pavia label image.

$80, 83, 90, 109, 116, 122, 134, 138, 157, 178, 179, 188]^T$. It can be seen that the result obtained by MOBS has less continuous bands, which means less redundancy among adjacent bands. Fig. 10 shows the entropy of each band in Indian Pines. As we can see from it, there are some bands with extreme low entropy compared with their adjacent bands such as bands $[1, 2, 3]^T$, $[104, 105]^T$, $[144, 145]^T$, $[198, 199, 200]^T$, etc. These bands can be taken as noisy bands with little information. It can be found that the MOBS selects band subsets avoiding the regions with low entropy to restrain the noisy bands. Moreover, the selected bands are not centered at the same region with high entropy. Instead, they distribute over the regions with a relatively smooth regions with high entropy, which reduce the redundancy among the adjacent bands.

In Fig. 11(a) and (b), we find that MOBS has comparable values with MTSP and UBS in MSD and comparable values with MTSP, UBS, Walumi, and Waludi in MSA. MOBS achieves the best classification performance, while it does not achieve the best performance in reducing redundancy on Indian Pines. The reason of this phenomenon is that MOBS selects less noisy bands in its band subsets. For a hyperspectral image containing some noisy bands which have low entropy, these bands are always distinct from the other informative bands, which can cause large values of MSD and MSA with other bands. From (14) and (15), it can be inferred that, when more noisy bands are selected, the values of MSD and MSA will increase. However, these noisy bands make no contribution to the classification and cause misclassification. For example, the MSD and MSA of band subset $[1, 20]^T$ are 46.57 and 1.75, while the MSD and MSA of band subset $[20, 112]^T$ are 18.69 and 0.46. It is obvious that band subset $[1, 20]^T$ has a better MSD and MSA. However, since band 1 is more noisy than band 112, the band subset $[20, 112]^T$ has a better classification result than band subset $[1, 20]^T$ does.

B. Results on Pavia University

This hyperspectral image was captured by the Reflective Optics System Imaging Spectrometer optical sensor over the University of Pavia, which is an urban area. It consists of 115 spectral bands in the original data set, and 12 noisy bands are removed, so there are 103 bands remaining in the data set for the experiment. For each band image, there are 610 × 340 pixels in it, and 9 classes of the representative urban ground truth are labeled. The nine class labels and the number of each class training samples are listed in Table II, and the label image is shown in Fig. 12.

1) Classification Performance: The result of the MOBS on this data set is shown in Fig. 13. The OA values under the three classifiers with different numbers of selected bands and AA values are shown in Fig. 14(a)–(d), respectively. As shown in Fig. 14(a) and (b), under the classifiers ELM and SVM, MOBS achieves comparable classification performance with other competitors when the number of bands p ranges from 1 to 15. When p is over 15, MOBS achieves the best classification performance compared with the other competitors. In Fig. 14(c), under the classifier KNN, MOBS achieves comparable classification performance with the other competitors when p ranges from 1 to 10. When p is over 10, MOBS achieves the best classification performance compared with the other competitors. From Fig. 14(d), under the classifiers ELM and KNN, MOBS

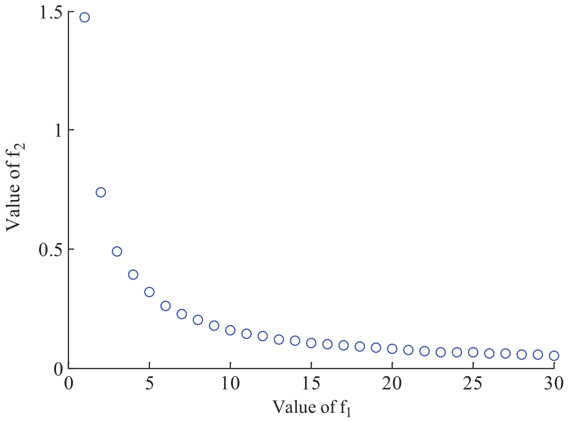
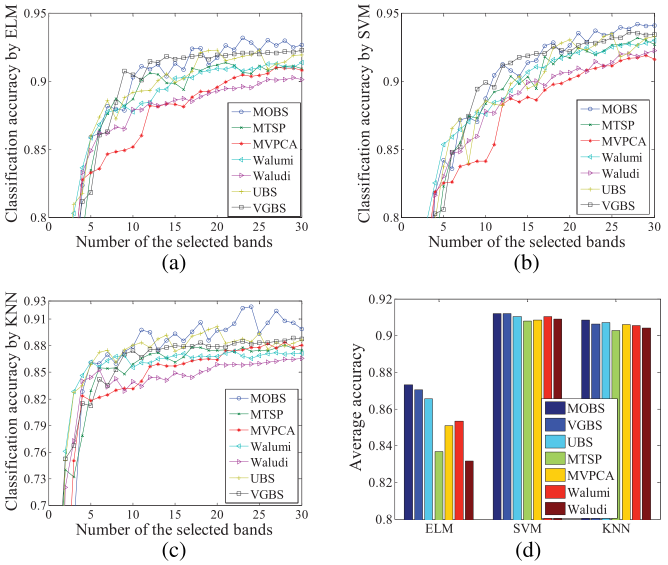


Fig. 13. MOBS results on Pavia University: Each point represents a band subset.



(a) (b) (c) (d)

Fig. 14. (a) Overall accuracy by ELM on Pavia University. (b) Overall accuracy by SVM on Pavia University. (c) Overall accuracy by KNN on Pavia University. (d) Average accuracy by ELM, SVM, and KNN with 21 selected bands on Pavia University.

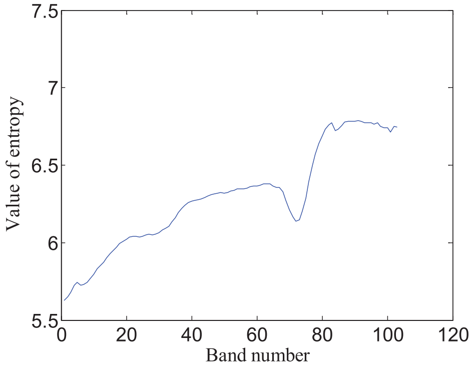
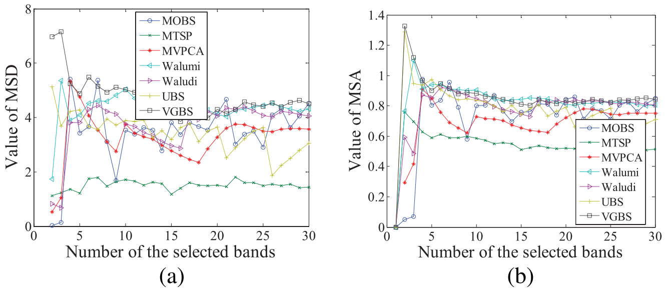


Fig. 15. Entropy of each band on Pavia University.



(a) (b)

Fig. 16. (a) Value of MSD on Pavia University. (b) Value of MSA on Pavia University.

obtains the best average accuracy, and under the classifier SVM, MOBS and VGBS obtain the same best average accuracy.

2) Analysis of the Selected Bands: The entropy of each band is shown in Fig. 15. As can be seen in Fig. 15, the curve of the entropy is relatively smooth, with a few sharp decreasing regions. The representative selected bands by MOBS are $[2, 4, 7, 10, 13, 17, 20, 31, 35, 52, 54, 59, 68, 77, 81, 82, 83, 85, 94, 95, 101]^T$. The selected bands distribute uniformly over the smooth regions and avoid the sharp decreasing regions. The values of MSD and MSA are shown in Fig. 16(a) and (b), respectively. VGBS and Walumi achieve the largest MSD. MOBS achieves comparable MSD with Waludi, UBS, and MVPCA and outperforms MTSP. For MSA, MOBS achieves comparable performance with VGBS, Waludi, Walumi, and UBS and outperforms MVPCA and MTSP.

C. Results on Salinas Data Set

The last hyperspectral image date set was gathered by the AVIRIS sensor over Salinas Valley, which is another classic remote sensing image. In the original data set, there are 224 bands in total. The 108th–112th, 154th–167th, and 224th bands are removed for the water absorption. Therefore, the data set remains 204 bands, and each band consists of 512 × 214 pixels. In the Salinas scene, there are 16 classes ground truth labeled. The 16-class labels and the number of each class training samples are listed in Table III, and the label image is shown in Fig. 17.

1) Classification Performance: The result of MOBS is shown in Fig. 18. The values of OA are shown in Fig. 19(a)–(c), and the values of AA are shown in Fig. 19(d). From Fig. 19(a) and (c), MOBS achieves the best classification performance when the number of bands p is over 5 under classifiers ELM and KNN. From Fig. 19(b), under classifier SVM, MOBS and UBS achieve comparable classification performance and outperform other competitors. Since the ground truth in Salinas is relatively easy for classification. The AA values of the seven competitors show similar results with each other. As shown in Fig. 19(d), under classifier ELM, MOBS achieves the highest AA of 0.9885, while UBS and MVPCA achieve comparable AAs of 0.9875 and 0.9876, respectively. Under classifier SVM, MOBS achieves the best AA of 0.9638, while VGBS, UBS, MVPCA, Walumi, and Waludi achieve nearly the same AA of 0.9635 with the one of MOBS. Under classifier KNN, MOBS achieves the same highest AA of 0.9633 with UBS, MVPCA, and Waludi and outperforms VGBS, MTSP, and Walumi.

2) Analysis of the Selected Bands: The representative band subset obtained by MOBS is $[9, 19, 20, 22, 34, 36, 38, 48, 59, 63, 95, 97, 105, 126, 138, 145, 152, 166, 172, 185, 200]^T$. The entropy of each band in Salinas is shown in Fig. 20. As shown

TABLE III
GROUND TRUTH CLASSES FOR THE SALINAS AND THEIR RESPECTIVE SAMPLE NUMBER

Class	Ground truth	Samples	Training
C1	Brocoli-green-weeds-1	2009	1004
C2	Brocoli-green-weeds-2	3726	1863
C3	Fallow	1976	988
C4	Fallow-rough-plow	1394	697
C5	Fallow-smooth	2678	1339
C6	Stubble	3959	1979
C7	Celery	3579	1789
C8	Grapes-untrained	11271	5635
C9	Soil-vinyard-develop	6203	3101
C10	Corn-senesced-green	3278	1639
C11	Lettuce-romaine-4wk	1068	534
C12	Lettuce-romaine-5wk	1972	986
C13	Lettuce-romaine-6wk	916	458
C14	Lettuce-romaine-7wk	1070	535
C15	Vinyard-untrained	7268	3634
C16	Vinyard-vertical	1807	903

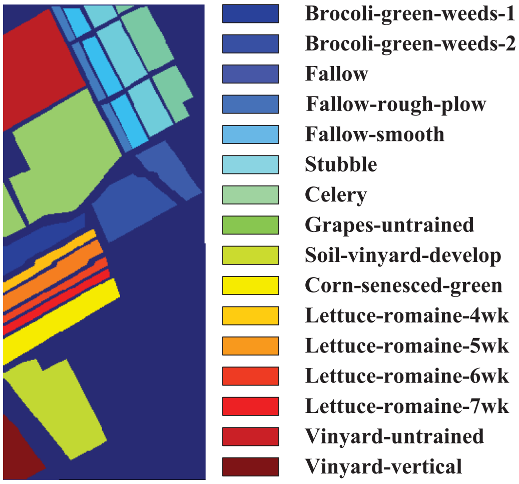


Fig. 17. Salinas label image.

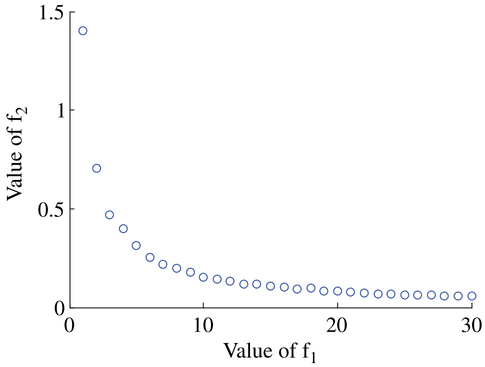
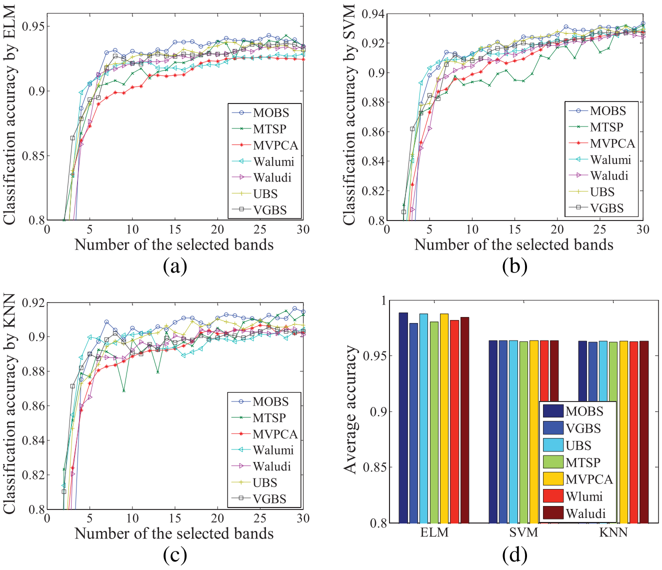


Fig. 18. MOBS results on Salinas: each point represents a band subset.



(a) (b) (c) (d)

Fig. 19. (a) Overall accuracy by ELM on Salinas. (b) Overall accuracy by SVM on Salinas. (c) Overall accuracy by KNN on Salinas. (d) Average accuracy by ELM, SVM, and KNN with 21 selected bands on Salinas.

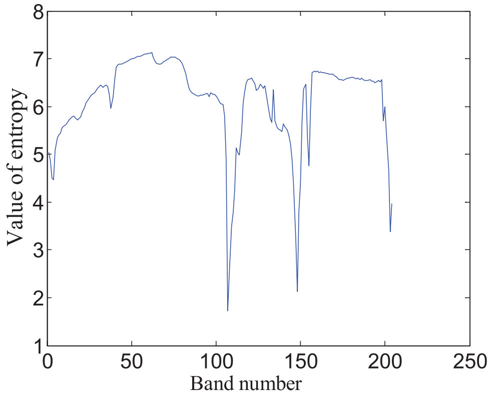
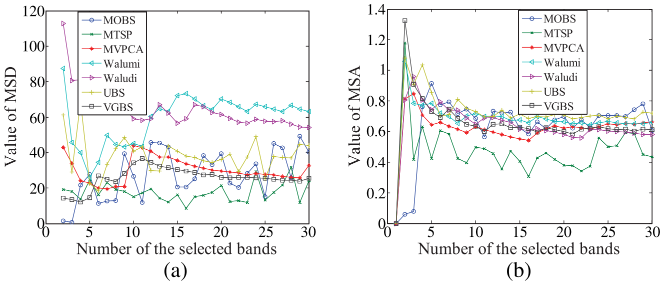


Fig. 20. Entropy of each band on Salinas.



(a) (b)

Fig. 21. (a) Value of MSD on Salinas. (b) Value of MSA on Salinas.

in Fig. 20, the selected bands distribute over relatively smooth regions with a high entropy and avoid sharp decreasing regions.

The values of MSD and MSA of the result are shown in Fig. 21(a) and (b), respectively. As we can see, for MSD, Walumi and Waludi have the largest values, and MOBS achieves comparable values with the rest competitors. For MSA, MOBS achieves the comparable values with Waludi, Walumi, and UBS and outperforms VGBS, MVPCA, and MTSP.

D. Computational Time Complexity Analysis

To record the execution time, experiments are implemented on a computer with an Intel Core i5-3470 3.2-GHZ CPU and 4-GB random access memory. The adopted Walumi and Waludi are implemented in Visual Studio C++, and the other methods are implemented in MATLAB. In Table IV, the execution times of these six algorithms are listed. Since UBS selects bands uniformly with *a priori* knowledge of the whole bands, we do not record UBS in Table IV. The matrix-computing-based methods

TABLE IV
COMPUTATIONAL TIME (IN SECONDS) FOR SELECTING TEN BANDS ON INDIAN PINES

Method	Indian Pines	Pavia University	Salinas
MOBS	8.81	37.85	22.86
MTSP	39.57	36.79	34.45
MVPCA	0.11	0.40	0.39
Walumi	20.00	14.00	33.00
Waludi	101.00	220.00	415.00
VGBS	0.34	0.57	0.74

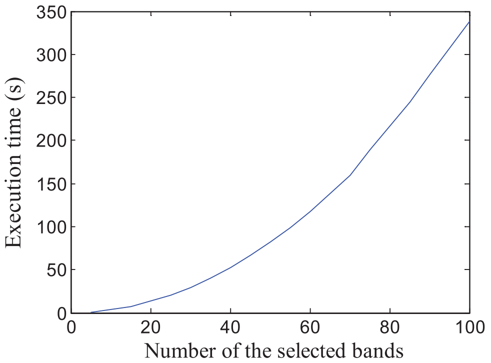


Fig. 22. Execution time of MOBS when the number of the selected bands ranges from 5 to 100 on Indian Pines.

MVPCA and VGBS outperform other methods in this aspect. However, they cannot offer a stable good classification performance on these three data sets. The execution time of MOBS is moderate among the competitive algorithms. While considering that MOBS can offer equal or best performance of classification in these three data sets and can obtain band subsets with different numbers of bands in a single run, the execution time is acceptable. Fig. 22 shows the execution time of MOBS when the number of selected bands p ranges from 5 to 100. As we can see, when p is small, especially ranging from 1 to 50, the implementation of MOBS is efficient. In addition, an enhancement in hardware or a parallel computing method can further improve the speed of implementing. Therefore, it can be claimed that MOBS possesses a value of practical applications.

E. Sensitivity Analysis of MOBS

1) Sensitivity in Relation to Similarity Criteria: In local search strategy, KL is introduced as a criterion to restrain redundancy. MI is also an alternative criterion to measure the similarity from information theory. The performances of the two criteria-based MOBSs which are termed as MOBS-KL and MOBS-MI are tested, and we compare their classification performances on Indian Pines. As shown in Fig. 23(a) and (b), under the ELM, MOBS-KL outperforms MOBS-MI slightly, and under the SVM, MOBS-KL and MOBS-MI have comparable results with each other. Therefore, the choice between KL and MI has little influence on classification performance. However, MOBS-MI is more time-consuming to implement due to its requirement to calculate joint entropy. When N is 150 and the numbers of bands are ranging from 1 to 30, MOBS-MI costs 134.91 s for each iteration, while MOBS-KL costs only 17.77 s. Therefore, KL is adopted as a criterion for redundancy in the proposed method.

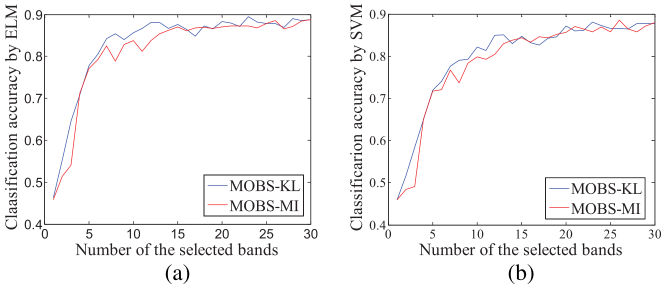


Fig. 23. (a) OA of the two criteria-based MOBSs by ELM on Indian Pines. (b) OA of the two criteria-based MOBSs by SVM on Indian Pines.

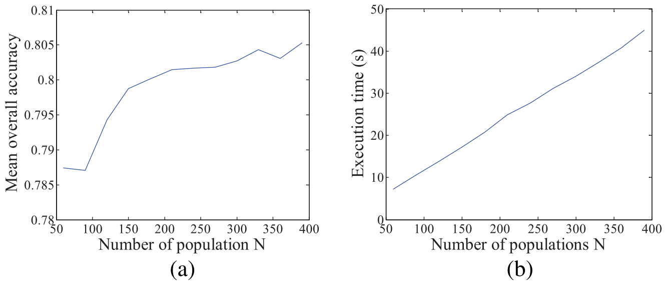


Fig. 24. (a) Mean OA when N ranges from 60 to 390 on Indian Pines. (b) Execution time of each iteration when N ranges from 60 to 390 on Indian Pines.

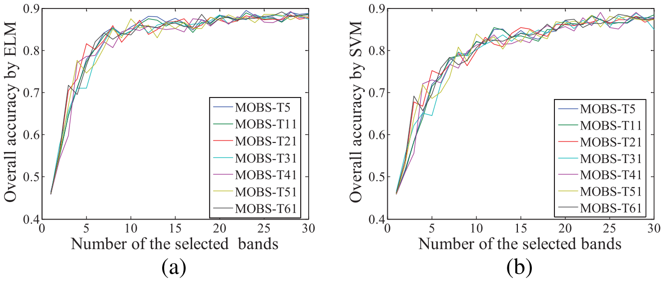


Fig. 25. (a) OA values via MOBS with different T values by ELM on Indian Pines. (b) OA values via MOBS with different T values by SVM on Indian Pines.

2) Sensitivity in Relation to the Number of Population N: In the proposed method, we select band subsets with the number of bands ranging from 1 to 30. Since each number of bands is allocated five solutions, the number of population N can be calculated by 30 × 5. To evaluate the influence of the numbers of population N, we still select band subsets with the number of bands ranging from 1 to 30 on Indian Pines, while N is ranging from 60 to 390 with a step of 30. The parameter T is kept the same proportion to N in each experiments. Fig. 24(a) shows the mean OA of the obtained band subsets with different N values. The mean OA increases with the N increasing. The reason is that, when N increases, the diversity of the population increases, which can avoid the local optimum and increase the probability of finding the global optimal solution. While in Fig. 24(b), there is an obvious increase in time consumption for each iteration when N is increasing. Therefore, considering both classification performance and efficiency of implementing, N typically ranges between 150 and 240.

3) Sensitivity in Relation to the Number of Neighborhood T: To evaluate the influence of T on classification performance, T is set to 5, 11, 21, 31, 41, 51, and 61, respectively, while N is maintained at 150. The results are shown in Fig. 25. It can be found that the value of T has little influence on classification performance. In the proposed method, the parameter T controls the range of neighborhood local searching. Since all of these neighborhoods have overlapping regions with adjacent ones, each neighborhood can communicate with the whole populations eventually. Therefore, the results can achieve global optimum with various range of neighborhood local searching, i.e., values of T. In our experiments, the proportion between T and N typically ranges from 0.05 to 0.5.

F. Discussion

Experimental results show that MOBS has equal or best classification performance among comparison algorithms in the three data sets. From the analysis of similarity criteria in local search, it can be claimed that the similarity criterion here is alternative. The framework of evolutionary multiobjective optimization plays a crucial role in selecting good band subsets. Meanwhile, it is noticed that MOBS is not the best method in reducing the redundancy. From this, we can see that less redundancy does not mean a good performance in classification directly when hyperspectral images contain some severely noisy bands. The redundancy has to be maintained a certain extent, and the significant informative content in a band subsets also plays a crucial role in classification. Due to focusing on both aspects, MOBS can obtain a competitive classification performance.

V. Conclusion

A novel band selection model based on multiobjective optimization has been established in this paper. Based on this model, a new unsupervised band selection algorithm called MOBS has been proposed for hyperspectral images. This paper has proposed a new method for the issue of handling the appropriate number of selected bands. A MOEA is designed to solve the MOP for band selection. Moreover, the coevolutionary mechanism is adopted to improve the efficiency of the optimization process in MOBS. The proposed algorithm can obtain band subsets with various numbers of the selected bands in a single run, and these band subsets possess a high quality of classification performance. The experimental results show that the proposed algorithm can obtain band subsets with a stable competitive performance in classification on the three benchmark data sets: Indian Pines, Pavia University, and Salinas.

In the future, our work will explore more objective functions to be suitable for band selection. A further study on the spatial information will be explored to improve the performance. Moreover, we will also pay interest in optimizing the search strategy to reduce the time complexity.

References

[1] M. Borengasser, W. Hungate, and R. Watkins, *Hyperspectral Remote Sensing Principles and Applications*. Boca Raton, FL, USA: CRC Press, 2008.

[2] Y. Yuan, Q. Wang, and G. Zhu, “Fast hyperspectral anomaly detection via high-order 2-D crossing filter,” *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 4, pp. 620–630, Feb. 2015.

[3] H. Akbari, Y. Kosugi, K. Kojima, and N. Tanaka, “Detection and analysis of the intestinal ischemia using visible and invisible hyperspectral imaging,” *IEEE Trans. Biomed. Eng.*, vol. 57, no. 8, pp. 2011–2017, Aug. 2010.

[4] S. Liu, L. Bruzzone, F. Bovolo, and P. Du, “Hierarchical change detection in multitemporal hyperspectral images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 1, pp. 244–260, Jan. 2015.

[5] G. F. Hughes, “On the mean accuracy of statistical pattern recognizers,” *IEEE Trans. Inf. Theroy*, vol. 14, no. 1, pp. 55–63, Jan. 1968.

[6] A. Plaza, P. Martinez, J. Plaza, and R. Perez, “Dimensionality reduction and classification of hyperspectral image data using sequences of extended morphological transformations,” *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 3, pp. 466–479, Mar. 2005.

[7] C.-I. Chang and S. Wang, “Constrained band selection for hyperspectral imagery,” *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 6, pp. 1575–1585, Jun. 2006.

[8] A. Agarwal, T. EI-Ghazawi, H. EI-Askary, and J. Le-Moigne, “Efficient hierarchical-PCA dimension reduction for hyperspectral image,” in *Proc. IEEE Int. Symp. Signal Process. Inf. Technol.*, Giza, Egypt, 2007, pp. 353–356.

[9] J. Wang and C.-I. Chang, “Independent component analysis-based dimensionality reduction with applications in hyperspectral image analysis,” *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 6, pp. 1586–1600, Jun. 2006.

[10] W. Lei, S. Prasad, J. E. Fowler, and L. M. Bruce, “Locality-preserving dimensionality reduction and classification for hyperspectral image analysis,” *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 4, pp. 1185–1198, Apr. 2012.

[11] X. Gao, X. Wang, D. Tao, and X. Li, “Supervised Gaussian process latent variable model for dimensionality reduction,” *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 41, no. 2, pp. 425–434, Apr. 2011.

[12] S. T. Roweis and L. K. Saul, “Nonlinear dimensionality reduction by locally linear embedding,” *Science*, vol. 90, no. 5500, pp. 2323–2326, Dec. 2000.

[13] C. Hou, F. Nie, X. Li, D. Yi, and Y. Wu, “Joint embedding learning and sparse regression: A framework for unsupervised feature selection,” *IEEE Trans. Cybern.*, vol. 44, no. 6, pp. 793–804, Jun. 2014.

[14] X. Fang *et al.*, “Locality and similarity preserving embedding for feature selection,” *Neurocomputing*, vol. 128, pp. 304–315, Mar. 2014.

[15] B. Guo, S. R. Gunn, R. I. Damper, and J. D. B. Nelson, “Band selection for hyperspectral image classification using mutual information,” *IEEE Geosci. Remote Sens. Lett.*, vol. 3, no. 4, pp. 522–526, Oct. 2006.

[16] L. Zhang, Y. Zhong, B. Huang, J. Gong, and P. Li, “Dimensionality reduction based on clonal selection for hyperspectral imagery,” *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 12, pp. 4172–4186, Dec. 2007.

[17] L. Bruzzone, F. Roli, and S. B. Serpico, “An extension of the Jeffreys–Matusita distance to multiclass cases for feature selection,” *IEEE Trans. Geosci. Remote Sens.*, vol. 33, no. 6, pp. 1318–1321, Nov. 1995.

[18] W. Siedlecki and J. Sklansky, “A note on genetic algorithms for large scale feature selection,” *Pattern Recognit. Lett.*, vol. 10, no. 5, pp. 335–347, Nov. 1989.

[19] 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.*, vol. 52, no. 7, pp. 4092–4105, Jul. 2014.

[20] A. A. Green, M. Berman, P. Switzer, and M. D. Craig, “A transformation for ordering multispectral data in terms of image quality with implications for noise removal,” *IEEE Trans. Geosci. Remote Sens.*, vol. 26, no. 1, pp. 65–74, Jan. 1988.

[21] J. B. Lee, A. S. Woodyatt, and M. Berman, “Enhancement of high spectral resolution remote sensing data by a noise-adjusted principal components transform,” *IEEE Trans. Geosci. Remote Sens.*, vol. 28, no. 3, pp. 295–304, May 1990.

[22] R. E. Roger, “A fast way to compute the noise-adjusted principal components transform matrix,” *IEEE Trans. Geosci. Remote Sens.*, vol. 32, no. 6, pp. 1194–1196, Nov. 1994.

[23] C.-I. Chang, Q. Du, T. S. Sun, and M. L. G. Althouse, “A joint band prioritization and band decorrelation approach to band selection for hyperspectral image classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 37, no. 6, pp. 2631–2641, Nov. 1999.

[24] H. J. Su, H. Yang, Q. Du, and Y. H. Sheng, “Semisupervised band clustering for dimensionality reduction of hyperspectral imagery,” *IEEE Geosci. Remote Sens. Lett.*, vol. 8, no. 6, pp. 1135–1139, Nov. 2011.

[25] G. C. T. Jee and C. Wu, “Unsupervised cluster-based band selection for hyperspectral image classification,” in *Proc. ICACSEI*, Beijing, China, Jul. 2013 pp. 562–565.

[26] A. M. Usó, F. Pla, J. M. Sotoca, and P. García-Sevilla, “Clustering based hyperspectral band selection using information measures,” *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 12, pp. 4158–4171, Dec. 2007.

[27] S. Jia, Z. Ji, Y. Qian, and L. Shen, "Unsupervised band selection for hyperspectral imagery classification without manual band removal," *IEEE J. Sel. Topics Appl. Earth Observ.*, vol. 5, no. 2, pp. 531–543, Apr. 2012.

[28] K. Sun, X. Geng, and L. Ji, "A new sparsity-based band selection method for target detection of hyperspectral image," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 2, pp. 329–333, Feb. 2015.

[29] K. Sun, X. Geng, L. Ji, and Y. Lu, "A new band selection method for hyperspectral image based on data quality," *IEEE J. Sel. Topics Appl. Earth Observ.*, vol. 7, no. 6, pp. 2697–2703, Jun. 2014.

[30] E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, Jul.–Oct. 1948.

[31] C.-I. Chang, *Hyperspectral Imaging: Techniques for Spectral Detection and Classification*. New York, NY, USA: Plenum, 2003.

[32] M. Ghamary Asl, M. R. Mobasheri, and B. Mojaradi, "Unsupervised feature selection using geometrical measures in prototype space for hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 7, pp. 3774–3787, Jul. 2014.

[33] S. D. Stearns, B. E. Wilson, and J. R. Peterson, "Dimensionality reduction by optimal band selection for pixel classification of hyperspectral imagery," in *Proc. 16th SPIE, Appl. Digit. Image Process.*, Oct. 1993, vol. 2028, pp. 118–127.

[34] 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.*, vol. 52, no. 11, pp. 7111–7119, Nov. 2014.

[35] C. Sui, Y. Tian, Y. Xu, and Y. Xie, "Unsupervised band selection by integrating the overall accuracy and redundancy," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 1, pp. 185–189, Jan. 2015.

[36] Q. Du, and H. Yang, "Similarity-based unsupervised band selection for hyperspectral image analysis," *IEEE Geosci. Remote Sens. Lett.*, vol. 5, no. 4, pp. 564–568, Oct. 2008.

[37] Y. Yuan, G. Zhu, and Q. Wang, "Hyperspectral band selection by multitask sparsity pursuit," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 2, pp. 631–644, Feb. 2015.

[38] C.-I. Chang and K. Liu, "Progressive band selection of spectral unmixing for hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 4, pp. 2002–2017, Apr. 2014.

[39] K. Deb, *Multi-Objective Optimization Using Evolutionary Algorithms*. Chichester, U.K.: Wiley, 2001.

[40] C. A. Coello Coello, D. A. Van Veldhuizen, and G. B. Lamont, *Evolutionary Algorithms for Solving Multi-Objective Problems*. Norwell, MA, USA: Kluwer, 2002.

[41] K. Miettinen, *Nonlinear Multiobjective Optimization*. Norwell, MA, USA: Kluwer, 1999.

[42] C. M. Fonseca and P. J. Fleming, "An overview of evolutionary algorithms in multiobjective optimization," *Evol. Comput.*, vol. 3, no. 1, pp. 1–16, 1995.

[43] J. D. Schaffer, "Multiple objective optimization with vector evaluated genetic algorithms," in *Proc. 1st Int. Conf. Genetic Algorithms*, J. J. Grefensttete, Ed. Hillsdale, NJ, USA: Lawrence Erlbaum, 1987, pp. 93–100.

[44] K. Deb, S. Agrawal, A. Pratap, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II," *IEEE Trans. Evol. Comput.*, vol. 6, no. 2, pp. 182–197, Apr. 2002.

[45] C. A. C. Coello, G. T. Pulido, and M. S. Lechuga, "Handling multiple objectives with particle swarm optimization," *IEEE Trans. Evol. Comput.*, vol. 8, no. 3, pp. 256–279, Jun. 2004.

[46] J. Knowles and D. Corne, "The Pareto archived evolution strategy: A new baseline algorithm for multiobjective optimisation," in *Proc. Congr. Evol. Comput.*, Washington, DC, USA, Jul. 1999, pp. 98–105.

[47] E. Zitzler, M. Laumanns, and L. Thiele, "SPEA2: Improving the strength Pareto evolutionary algorithm for multiobjective optimization," in *Proc. Evol. Methods Des. Optim. Control Appl. Ind. Problems*, K. C. Giannakoglou, D. T. Tsahalis, J. Périaux, K. D. Papailiou, and T. Fogarty, Eds., Athens, Greece, 2001, pp. 95–100.

[48] Q. Zhang and H. Li, "MOEA/D: A multi-objective evolutionary algorithm based on decomposition," *IEEE Trans. Evol. Comput.*, vol. 11, no. 6, pp. 712–731, Dec. 2007.

[49] A. Paoli, F. Melgani, and E. Pasolli, "Clustering of hyperspectral images based on multiobjective particle swarm optimization," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 12, pp. 4175–4188, Dec. 2009.

[50] K. C. Tan, Y. J. Yang, and C. K. Goh, "A distributed cooperative coevolutionary algorithm for multiobjective optimization," *IEEE Trans. Evol. Comput.*, vol. 10, no. 5, pp. 527–549, Oct. 2006.

[51] C. K. Goh and K. C. Tan, "A competitive–cooperative coevolutionary paradigm for dynamic multiobjective optimization," *IEEE Trans. Evol. Comput.*, vol. 13, no. 1, pp. 103–127, Feb. 2009.

[52] M. A. Potter and K. A. De Jong, "Cooperative coevolution: An architecture for evolving coadapted subcomponents," *Evol. Comput.*, vol. 8, no. 1, pp. 1–29, 2000.

[53] P. Groves and P. Bajcsy, "Methodology for hyperspectral band and classification model selection," in *Proc. IEEE Workshop Adv. Tech. Anal. Remotely Sens. Data*, Greenbelt, MD, USA, 2003, pp. 120–128.

[54] P. Bajcsy and P. Groves, "Methodology for hyperspectral band selection," *Photogramm. Eng. Remote Sens. J.*, vol. 70, no. 7, pp. 793–802, Jul. 2004.

[55] R. T. Marler and J. S. Arora, "Survey of multi-objective optimization methods for engineering," *Struct. Multidiscipl. Optim.*, vol. 26, no. 6, pp. 369–395, Apr. 2004.

[56] I. Das and J. E. Dennis, "Normal-boundary intersection: A new method for generating Pareto optimal points in multicriteria optimization problems," *SIAM J. Optim.*, vol. 8, no. 3, pp. 631–657, Aug. 1998.

[57] A. Messac, A. Ismail-Yahaya, and C. Mattson, "The normalized normal constraint method for generating the Pareto frontier," *Struct Multidisc. Optim.*, vol. 25, pp. 86–98, Jul. 2003.

[58] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: Theory and applications," *Neurocomputing*, vol. 70, no. 1, pp. 489–501, Dec. 2006.

[59] X. Liu, S. Lin, J. Fang, and Z. Xu, "Is extreme learning machine feasible? A theoretical assessment (part I)," *IEEE Trans. IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 1, pp. 7–20, Jan. 2015.

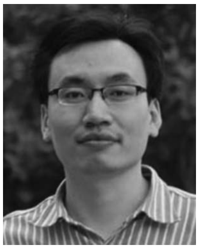
[60] S. Lin, X. Liu, J. Fang, and Z. Xu, "Is extreme learning machine feasible? A theoretical assessment (part II)," *IEEE Trans. IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 1, pp. 21–34, Jan. 2015.

[61] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge, U.K.: Cambridge Univ. Press, 2000.

[62] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 1–39, Apr. 2001. [Online]. Available: www.csie.ntu.edu.tw/~cjlin/libsvm

[63] T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory*, vol. IT-13, no. 1, pp. 21–27, Jan. 1967.

[64] J. Yin, Y. Wang, and J. Hu, "A new dimensionality reduction algorithm for hyperspectral image using evolutionary strategy," *IEEE Trans. Ind. Informat.*, vol. 8, no. 4, pp. 935–943, Nov. 2012.



Maoguo Gong (M'07–SM'14) received the B.S. degree in electronic engineering and the Ph.D. degree from Xidian University, Xian, China, in 2003 and 2009, respectively.

He has been a Teacher with Xidian University since 2006, where he was promoted to Associate Professor and Full Professor in 2008 and 2010, respectively, both with exceptive admission. He has published over 50 papers in journals and conferences. He is the holder 15 granted patents as the first inventor. His current research interests include computational intelligence with applications to optimization, learning, data mining, and image understanding.

Mingyang Zhang received the B.S. degree in automation from Xidian University, Xian, China, in 2012, where he is currently working toward the Ph.D. degree.

His current research interests include computational intelligence and image understanding.

Yuan Yuan (M'05–SM'09) is currently a Full Professor with the Chinese Academy of Sciences, Beijing, China. She is the author or coauthor of over 150 papers, including about 100 in reputable journals such as IEEE Transactions and *Pattern Recognition*, as well as conference papers in CVPR, BMVC, ICIP, and ICASSP. Her current research interests include visual information processing and image/video content analysis.