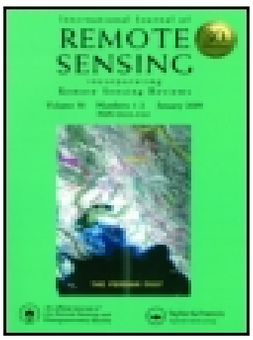

International Journal of Remote Sensing

ISSN: 0143-1161 (Print) 1366-5901 (Online) Journal homepage: http://www.tandfonline.com/loi/tres20

Unsupervised hyperspectral feature selection based on fuzzy *c*-means and grey wolf optimizer

Fuding Xie, Cunkuan Lei, Fangfei Li, Dan Huang & Jun Yang

To cite this article: Fuding Xie, Cunkuan Lei, Fangfei Li, Dan Huang & Jun Yang (2018): Unsupervised hyperspectral feature selection based on fuzzy *c*-means and grey wolf optimizer, International Journal of Remote Sensing

To link to this article: https://doi.org/10.1080/01431161.2018.1541366

Published online: 13 Nov 2018.

高光谱影像的光谱信息和二维空间信息共同构成了高光谱数据立方体。在光谱空间中，每个像素点反映为一条连续光谱响应曲线，不同的目标地物在高光谱影像中表现为不同的辐射强度。在影像空间中，每个波段则对应着一幅二维影像，通常被视为一个特征.相比波段提取，波段选择能更好地保留原始波段的物理意义及光谱特征，同时又达到了降维的目的，因而深受广大高光谱领域工作者的青睐。

INTERNATIONAL JOURNAL OF REMOTE SENSING
https://doi.org/10.1080/01431161.2018.1541366

Unsupervised hyperspectral feature selection based on fuzzy *c*-means and grey wolf optimizer

Fuding Xie[a], Cunkuan Lei[a], Fangfei Li[a], Dan Huang[b] and Jun Yang [a]

[a]College of Urban and Environment, Liaoning Normal University, Dalian, Liaoning, P.R.China.; [b]College of Computer Science, Liaoning Normal University, Dalian, Liaoning, P.R.China.

ABSTRACT

Hyperspectral image (HSI) with hundreds of narrow and consecutive spectral bands provides substantial information to discriminate various land-covers. However, the existence of redundant features/bands not only gives rise to increasing of computation time but also interferes the classification result of hyperspectral images. Obviously, it is a very challenging problem how to select an effective feature subset from original bands to reduce the dimensionality of the hyperspectral dataset. In this study, a novel unsupervised feature selection method is suggested to remove the redundant features of HSI by feature subspace decomposition and optimization of feature combination. Feature subset decomposition is achieved by the fuzzy *c*-means (FCM) algorithm. The optimal feature selection is based on the optimization process of grey wolf optimizer (GWO) algorithm and maximum entropy (ME) principle. To evaluate the effectiveness of the proposed method, experiments are conducted on three well-known hyperspectral datasets, Indian Pines, Pavia University, and Salinas. Six state-of-the-art feature selection methods are used to compare with the proposed method. Experimental results successfully confirm the superior performance of our proposal with respect to three classification accuracy indices overall accuracy (OA), average accuracy (AA) and kappa coefficient (κ).

ARTICLE HISTORY
Received 26 March 2018
Accepted 22 October 2018

KEYWORDS
hyperspectral image; feature selection; FCM algorithm; GWO algorithm

1. Introduction

Hyperspectral remote sensing provides us spectral signatures with hundreds of consecutive and narrow spectral channels that dramatically characterized the instinct physical and chemical meanings of ground objects. A wealth of spectral information makes it sensitive to discriminate various land covers, which has widely drawn some applications in remote sensing and other research fields, including urban planning, mine exploration, environment monitoring, and machine learning (Slavkovikj et al. 2016; Yue et al. 2015; Liu et al. 2015; Li et al. 2016; Veen 2009).

Broad endeavours have been dedicated to the dimension reduction of hyperspectral images since 2000. Generally speaking, conventional dimension reduction (DR) methods

CONTACT Jun Yang yangjun@lnnu.edu.cn College of Urban and Environment, Liaoning Normal University, 850 Huanghe Road, Dalian, Liaoning, P.R.China.

can be separated into two parts: feature extraction (Chiang, Chang, and Ginsberg 2000; Agarwal et al. 2008; Liu et al. 2017) and feature selection.

Unlike feature extraction, feature selection achieves the purpose of dimensionality reduction, while inheriting the original data of the physical interpretability of the pixel's spectral characteristics. This is the preferred method for researchers. Liu et al. (2017) inspired by the scale selection, spatial and spectral information based approach was discussed. Based on the notion of saliency, scale, and image description that are intrinsically linked, the saliency band selection turns into a problem of scale selection. Unlike most traditional band selection methods of sequentially selecting one band at a time, Yu, Song, and Chang (2018) developed a novel technique to hyperspectral band subset selection that simultaneously selects multiple bands as a band subset. By way of integrating the strong points of both feature selection and feature extraction, a newly designed multiple bands selection method was proposed, which could discern the significance of original band dataset and ease the drawback of band extraction that cannot maintain the physical interpretability of the original band dataset (Zhang et al. 2017).

According to whether there is a label or not, feature selection methods can be further separated into three categories: supervised (Martínez Sotoca and Pla 2010; Yang et al. 2011; Cao, Xiong, and Jiao 2016; Liu and Wang et al. 2017), semi-supervised (Camps-Valls, Bandos Marsheva, and Zhou 2007; Zhang et al. 2013; Feng et al. 2015; Cao et al. 2017) and unsupervised (Mitra, Murthy, and Pal 2002; Du and Yang 2008; Breaban and Luchian 2011; Feng et al. 2016). Supervised feature selection measures the correlation of features with the help of class labels to search for the most discriminant feature subset. Likewise, semi-supervised feature selection seeks the most discriminative feature subset by using both of the limited labelled samples and the unlabelled samples. As a matter of fact, the acquisition of class labels for hyperspectral image (HSI) requires expert's help, which is time-consuming and expensive due to the large amount of labour consumed. Hence, unsupervised band selection (UBS) method (with class label unknown) is given priority to the research of dimensionality reduction.

Unsupervised feature selection can be actualized by band ranking (Peng, Long, and Ding 2005; Zhang et al. 2007; Tesmer, Perez, and Zurada 2009; Kim and Rattakorn 2011; Peng, Shen et al. 2013) and band clustering (Zhang et al. 1999; Guo et al. 2006; Martínez-Usómartinez-Uso et al. 2007; Jia et al. 2015; Bevilacqua and Berthoumieu 2017) methods. Clustering-based feature selection methods are performed in a defined similarity space by dividing the features into disjoint clusters, so that the features in the same group are more similar than that in different groups. Accordingly, representative bands could be selected from each cluster. A feature selection approach based on fuzzy *c*-means clustering and genetic algorithm (FCM-GA) was defined in Wu (2012). The author used fuzzy *c*-means (FCM) algorithm to classify the band data and took a genetic algorithm (GA) as a search strategy. Guo et al. (2006) proposed a fast and effective feature selection algorithm. The algorithm initially considered all spectral bands as potential clustering centres and determines the representative point for each cluster through the message passing between data points. Bevilacqua and Berthoumieu (2017) thought that bands are clustered in hyperspectral cube space, then, a representative image was emerging from each cluster as the member of the selected band set.

The representative feature combination can be found without calculating all the possible options in the simulated evolutionary based methods. It is common to choose a measure as fitness function rather than the second order statistics calculation, for instance, particle swarm optimization (PSO) (Ghamisi and Benediktsson 2014; Ghamisi, Couceiro, and Benediktsson 2015), genetic algorithm (GA) (Bazi and Melgani 2006; Wu 2012; Ghamisi and Benediktsson 2014), artificial bee colony (ABC) (Karaboga and Basturk 2008; Karaboga and Ozturk 2011) and grey wolf optimizer (GWO) (Mirjalili, Mirjalili, and Lewis 2014; Medjahed et al. 2016; MirjaliMirjalili et al. 2016; Mirjalili 2015; Rodríguez et al. 2017; Khairuzzaman, Kayom, and Chaudhury 2017; Zhang et al. 2018). The PSO method has the advantage of fast convergence and lower computational complexity. However, this method is easy to fall into a local minimum. Despite GA has extensive search capability, it is an algorithm without the memory of previous behaviour. In order to combine the advantages of these two algorithms, a new band selection approach based on the hybridization of GA and PSO was depicted in Ghamisi, Couceiro, and Benediktsson (2015). Karaboga et al. proposed a novel optimization algorithm called ABC algorithm, which is inspired by the natural foraging behaviour of bees in the searching process. It is used initially for optimizing multivariable numerical problems in Karaboga and Basturk (2008). A novel band selection method called chaotic binary coded gravitational search algorithm (CBGSA) was put forward for reducing the dimensionality of hyperspectral datasets, which has been proved to be superior to the GA methods owing to the preferable ability of searching optimal solution (Wang et al. 2017). The memetic algorithm as a newly designed method was designed to obtain stable and superior optimal band subsets. For sake of optimizing the objective function, a novel crossover operator and a local search were developed by Zhang et al. (2017b). Medjahed et al. (2016) first applied directly GWO algorithm to hyperspectral feature selection, aiming at reducing the dimension of hyperspectral images. It can be seen from their work that the results of dimensionality reduction are not always ideal.

Grey wolf optimizer, introduced by Mirjalili, Mirjalili, and Lewis (2014), is capable of providing very competitive results compared to other optimization algorithms. To optimize problems with multiple objectives, Mirjalili et al. further proposed a multi-objective grey wolf optimizer (MirjaliMirjalili et al. 2016). By comparing the experimental results of GWO and other five optimization algorithms on eight standard datasets, the effectiveness of the grey wolf optimizer in training multi-layer perceptrons was analysed (Mirjalili 2015). GWO algorithm was enhanced by introducing a new hierarchical operator in the simulation of the hunting process (Rodríguez et al. 2017). Khairuzzaman, Kayom, and Chaudhury (2017) modified GWO method by using multilevel thresholding technique and successfully applied it to image segmentation. A novel hybrid algorithm based on biogeography-based optimization and grey wolf optimizer is presented in Zhang et al. (2018). Because GWO has the advantages of simple operation, high availability and fast convergence, and is beneficial to approximate the global optimal solution, it is chosen as the optimization strategy for feature selection in this study.

Feature selection techniques typically attempt to seek a feature combination which is capable of achieving maximum inter-class separability and minimal redundancy between features. What is more, to randomly select m features out of the n original features, there are $n!/(n-m)!m!$ possible results. In the case of high-dimensional space, the workload for selecting the optimal feature combination from all feature combinations is

of significant amount. This fact suggests us to develop new methods to cope with this problem.

In this study, a novel feature selection method is introduced based on FCM and GWO algorithms. The original features are first divided into different feature subsets by using the FCM algorithm. The use of clustering strategy enables similar features to be partitioned into a feature subset. Then, *K* features are randomly selected from each feature subset to compose a new feature combination, in which the redundancy between features is small because the features in different subsets are not similar. Finally, GWO algorithm is used to optimize the obtained feature to make it with perfect classification ability. Unlike the use of overall accuracy (OA) as a fitness function in the process of optimization, the utilization of maximum entropy (ME) principle effectively reduces the computation time of feature selection. The feasibility of the proposed framework is demonstrated by carrying out on three typical hyperspectral datasets. Experimental results and comparison with six state-of-the-art methods show that the proposed approach can achieve superior performance.

2. FCM and GWO algorithms

2.1. *FCM algorithm*

Clustering is to partition the dataset into distinct clusters or groups, such that data belonged to the same cluster are similar to each other, and data from different clusters are dissimilar (Winkler, Klawonn, and Kruse 2011; Zhang et al. 2014). Formally, dataset (DS) is decomposed into *p* clusters as follows:

$$\mathrm{DS} = \mathrm{C}_1 \cup \mathrm{C}_2 \cup \cdots \cup \mathrm{C}_\mathrm{p}, \mathrm{C}_\mathrm{i} \cap \mathrm{C}_\mathrm{j} = \emptyset, \mathrm{i} \neq \mathrm{j}, \mathrm{i}, \mathrm{j} = 1, 2, \cdots, \mathrm{p} \tag{1}$$

where C_i indicates cluster *i*.

FCM proposed by Dunn (1973) is one of the most popular and prominent clustering algorithms in machine learning. Different from hard clustering approaches, FCM is a soft clustering method and permits samples to be in two or three clusters at the same time, with different membership degrees.

Suppose DS = {x_1, x_2, …, x_n}to be a dataset with *n* samples in *q*-dimensional space R^q. FCM clustering algorithm may divide DS into $1 < C < n$ clusters by an iterative process. The objective function *J* in FCM is defined as follows:

$$\mathrm{Min}: \begin{cases} \mathrm{J}(\mathrm{U}, \mathrm{V}) = \sum_{\mathrm{j}=1}^{\mathrm{C}} \sum_{\mathrm{i}=1}^{\mathrm{n}} \mu_{\mathrm{i,j}}^{\mathrm{m}} (\mathrm{x_i} - \mathrm{v_j})^2 \\ \mu_{\mathrm{i,j}} \in [0, 1], \sum_{\mathrm{j}=1}^{\mathrm{C}} \mu_{\mathrm{i,j}} = 1 \end{cases} \tag{2}$$

where $\mathbf{U} = (\mu_{i,j})_{n \times p}$ is membership degree matrix; $\mathbf{V} = (v_1, v_2, \ldots, v_C)$ is centroid matrix; $v_j \in R^q$ is the centroid vector of cluster *j*; $\mu_{i,j}$ is the membership degree of sample x_i belonging to cluster *j*. The parameter $m > 1$ in Equation (2) is a fuzzy factor which controls the membership of each sample.

Obviously, this is an extreme optimization problem with a constrained condition. Solving it by the Lagrange multiplier method, membership degree $\mu_{i,j}$ and centroid v_j of can be updated constantly by Equation (3) and Equation (4) respectively,

$$\mu_{i,j} = \left[\sum_{k=1}^{C} \left(\frac{(x_i - v_j)^2}{(x_i - v_k)^2} \right)^{\frac{1}{m-1}} \right]^{-1} \tag{3}$$

$$v_j = \frac{\sum_{i=1}^{n} \mu_{i,j}^m x_i}{\sum_{i=1}^{n} \mu_{i,j}^m} \tag{4}$$

FCM clustering algorithm can be simply described by the following steps.

Input: Data set DS, threshold ε, pre-assigned cluster number C.

Output: Membership degree matrix $\mathbf{U}$.

Step1: Choose randomly C samples as the initial centroids.

Step2: Update membership degree $\mu_{i,j}$ by Equation (3).

Step 3: Determine $\left| J^{(k+1)}(\mathbf{U}, \mathbf{V}) - J^{(k)}(\mathbf{U}, \mathbf{V}) \right| < \varepsilon$. If it is true, then stop.

Step 4: Update cluster centroid v_j by Equation (4), go to step 2.

In this study, we set fuzzy factor $m = 2$ in all experiments. For hyperspectral data, one can consider each feature as a sample. Then, all initial features are grouped into different clusters.

2.2. *Maximum entropy (ME)*

To select the most informative feature subset for improving the classification accuracy, it is an important and crucial problem to choose/define fitness function in an optimization process. In this work, the ME will be adopted as the measurement to evaluate the band image information in GWO algorithm. The ME principle points out that the model with the ME is the best in all probability distributions.

In information theory, Shannon entropy is thought as the basic information unit. Let $\boldsymbol{Y}$ be a random vector, its quantity of information can be calculated in discrete form as follows:

$$E(Y) = -\sum_i p(y_i) \log_2 p(y_i) \tag{5}$$

where $E(\boldsymbol{Y})$ is the entropy of $\boldsymbol{Y}$, $p(y_i$ denotes the probability and y_i is the i^{th} component of $\boldsymbol{Y}$.

The histogram-based method is commonly used to estimate the probability $p(y_i)$. To find the most distinctive feature subset which the information of raw features is maximally preserved, a simple method is to directly collect those features with high entropy. Provided that there are s features in this feature subset, the problem can be defined mathematically

$$\frac{1}{s} \sum_{i=1}^{s} E(Y_i) \tag{6}$$

where $E(Y_i)$ is the entropy of feature i obtained by histogram method and Equation (5).

Another key problem in the feature selection process is to properly choose the optimization strategy. The desired optimization algorithm should not fall into local optimum and converge to the global optimum. Luckily, grey wolf optimizer has such characteristics.

2.3. *Grey wolf optimizer (GWO)*

Mirjalili, Mirjalili, and Lewis (2014) proposed a more powerful grey wolf optimizer which simulated the hunting and living behaviour of a group of grey wolves. From the point of life and management, the authors divided hierarchically all grey wolves into four classes denoted by alpha wolf, beta wolf, delta wolf, and omega wolf successively. In the first class, there is only a grey wolf, alpha wolf, which is the leader and controls the whole wolf pack. The beta wolf in the second class is only obedient to alpha wolf and helps it to make some decisions. Meanwhile, beta wolf commands the rest two lower-level wolves. Beta wolf will be alpha wolf if the alpha wolf passes away. Grey wolves in the lowest class are called omega which are responsible for collecting useful information to submit to the alpha wolf and the beta wolf. The rest is known as delta wolves.

The hunting process of grey wolves is carried out in three stages (Mirjalili, Mirjalili, and Lewis 2014):

- Tracking, chasing and approaching the prey.
- Pursuing, encircling and harassing the prey.
- Attack toward the prey.

The mathematical model of social hierarchy and hunting process can be described as follows:

- α: the fittest solution
- β: the second best solution
- δ: the third best solution.
- ω: the rest of the candidate solutions.

The grey wolves encircle prey can be modelled:

$$D = \left| C \cdot X_{\mathrm{p}}(t) - X(t) \right| \tag{7}$$

$$X(t+1) = X_{\mathrm{p}}(t) - D \times A \tag{8}$$

where t denotes the current iteration;D represents the distance between the grey wolf and the prey; A and C are coefficient vectors; X_{p} is the position vector of the prey andX is the position vector of a grey wolf. Here, the symbol represents the corresponding multiplication of each component of two vectors.

The coefficient vectors A and C can be obtained by:

$$A = 2a \cdot \mathrm{r}_1 - a \tag{9}$$

$$C = 2 \times \mathrm{r}_2 \tag{10}$$

where a is a constant vector ($a, a, \ldots, a$) that the constant a is decreased linearly from 2 to 0 in the iteration process; r_1 and r_2 are random vectors.

The mathematical model of encircling prey can be described: The mathematical model of hunting are:

$$D_\alpha = |C_1 \cdot X_\alpha(t) - X| D_\beta = |C_2 \cdot X_\beta(t) - X| D_\delta = |C_3 \cdot X_\delta(t) - X| \tag{11}$$

$$X_1 = X_\alpha - A_1 \times D_\alpha X_2 = X_\beta - A_2 \times D_\beta X_3 = X_\delta - A_3 \times D_\delta \tag{12}$$

$$X(t+1) = (X_1 + X_2 + X_3)/3 \tag{13}$$

Repeating (7)-(13) constantly, one can lastly find the best solution X_α. This optimization algorithm can be used to select the informative feature if we slightly revise Equation (13) by rounding operation.

The GWO algorithm is summarized in the following steps.

Step1: Initialize the grey wolf population $\boldsymbol{X}_i$ (i = 1, 2, …, N); and parameters a; maximum iterations.

Step2: Choose α, β and δ wolves from the population by Equation (5).

Step3: Update the position of each wolf by revised Equation (13).

Step4: Compute Equation (7) – (12) and go to step2.

Step5: Output a until max number of iterations or algorithm convergence is satisfied.

3. The proposed feature selection method

In what follows, we will describe the proposal in details.

It is well known that there exist a large number of redundant features in hyperspectral data. From the feature point of view, features with analogous radiance value should have similar classification capability. It implies that one can pick up several features from these spectra to represent them so as to reduce the dimensionality of hyperspectral data. The premise of doing so is that the features with analogous radiance value have been put into the same cluster. These clusters are obviously obtained by using FCM algorithm for all initial features. It is sometimes difficult to properly preassign cluster number C in FCM algorithm. Fortunately, this problem can be solved by the visual result of the spectral curve of hyperspectral data. This process is also known as feature subspace decomposition, aiming at reducing the redundant information.

On the basis of the feature cluster, K different features randomly selected from each cluster constitute a new feature vector (grey wolf) as shown in Figure 1. The component $b_{i,j}$ represents the j^{th} feature selected from the i^{th} cluster. This selection method effectively reduces the redundancy between features. However, we cannot guarantee that the obtained grey wolf has a good classification ability. So it is necessary to use grey wolf algorithm to make it with good classification ability while preserving the redundancy as small as possible.

Suppose that there are N grey wolves in the initial population. To calculate the entropy of each grey wolf in the initial population by using of Equation (5) and Equation (6) and sort them down with the biggest in front, the wolves corresponded to the first three maximum entropies are named α, β and δ wolf successively. The rest grey wolves are considered as ω wolves. Under the absence of location of the prey, we

$b_{1,1}$	….	$b_{1,K}$	……	$b_{C,1}$	…	$b_{C,K}$

Figure 1. Grey wolf consisted of $s = k \times C$ features.

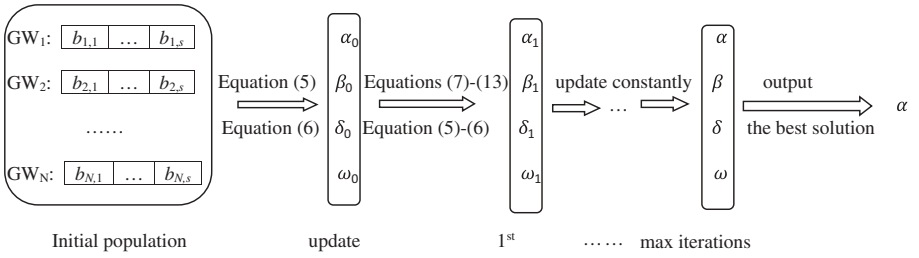

Figure 2. The work frame of grey wolf optimizer.

may replace approximately the position vector of the prey with the position of the α wolf. Each of the grey wolves has the ability to know the location of prey and encircles them.

Encircling prey during the hunt can be carried out by Equation (7) – (10) and setting $a = 2$.

After encircling prey, each wolf should update its position timely in terms of the locations of α, β and δ wolf to enclose the prey. Hunting process is accomplished by computing Equation (11) – (13).

Repeating constantly the process of encircling and hunting, one can lastly obtain the best solution α. The process of encircling and hunting the prey can be designed by a flow chart as shown in Figure 2.

The proposed feature selection algorithm FCM-GWO can be summarized as follows:

Input: Hyperspectral dataset; parameters and maximum iterations in GWO.

Output: the most informative feature subset.

Step1: Partition the initial features into C clusters by FCM algorithm introduced in subsection 2.1.

Step2: Select randomly K different features from each cluster to make up a grey wolf.

Step3: Initialize population in GWO by repeatedly executing step2 N times.

Step4: Call the GWO algorithm mentioned in subsection 2.3.

The complexity of FCM algorithm is $O(CBn)$; C is the class number, B is the number of features, and n denotes the number of pixels in the hyperspectral dataset. The complexity of Step2 and Step3 is constant, $O(1)$. GWO algorithm can be completed in time $O(tBn)$; t is the number of iterations. Therefore, the complexity of the proposal is $O(Bn)$.

4. Experimental results

In this study, the goal of hyperspectral feature selection is to classify the hyperspectral data. Usually, the different classifiers will result in different classification results on the same hyperspectral data set. As a popular and attractive classifier, support vector machine (SVM) has been widely applied in remote sensing field (Bazi and Melgani 2006; Feng et al. 2014; Ghamisi, Couceiro, and Benediktsson 2015). In this study, the support vector machine with five-fold cross-validation (SVM-5) is employed to accomplish the last classification for HSI.

To evaluate the performance of the proposed FCM-GWO feature selection method, the test experiments have been carried out on three typical hyperspectral images, Indian Pines, Pavia University, and Salinas. A detailed description of these three datasets can be found on webpage http://www.ehu.es/ccwintco/index.php/Hyperspectral_Remote_Sensing_Scenes. The data used in this study are also downloaded from this website. The key information of three hyperspectral datasets is listed in Table 1. The false-colour images and the ground truth image of three datasets are shown in Figure 3-Figure 5.

In all experiments, 5%, 10%, 15%, and 20% of pixels from each class are randomly labelled to form the training sets. The rest is to be classified directly. The initial population in GWO algorithm is made up of $N = 30$ grey wolves. GWO algorithm will cease by

Table 1. The description of three hyperspectral datasets. (Note: Airborne Visible Infrared Imaging Spectrometer (AVIRIS); Reflective Optical System Imaging Spectrometer (ROSIS)).

	Indian Pines	Pavia University	Salinas
class number	16	9	16
bands/features	200	103	204
size	145 × 145	610 × 340	512 × 217
sensor	AVIRIS	ROSIS	AVIRIS
resolution	20m	1.3m	3.7m
samples	10,249	42,776	54,129

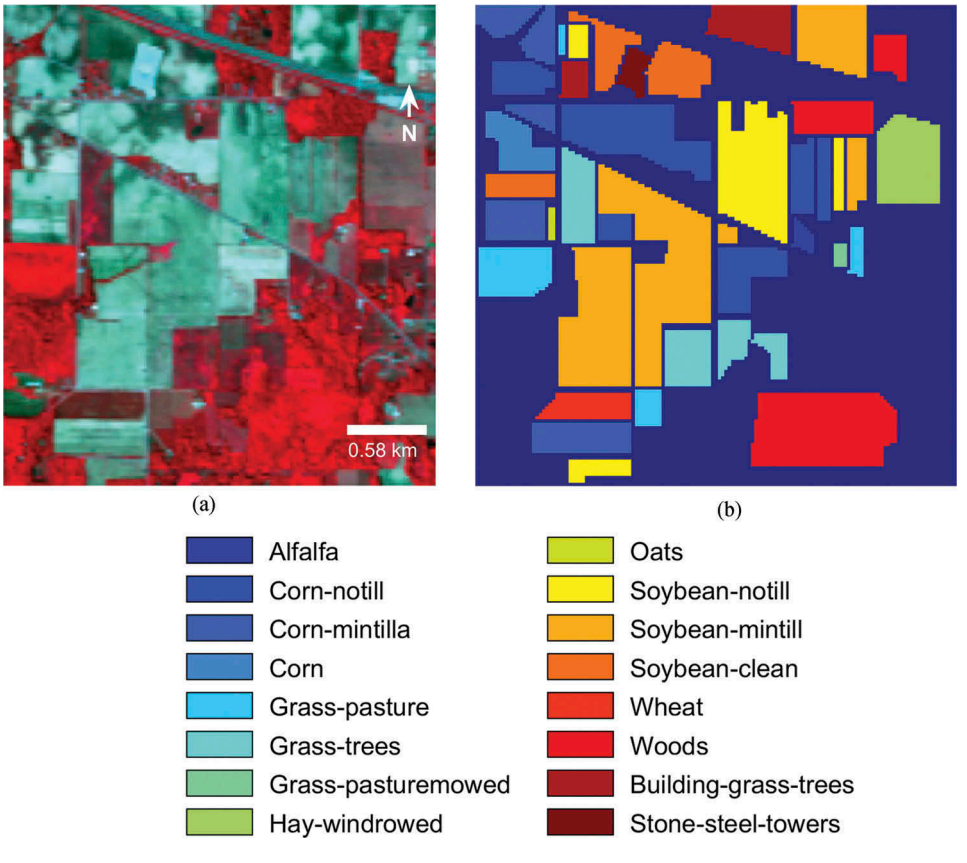

Figure 3. Indian Pines dataset. (a) False-colour image and (b) Ground truth.

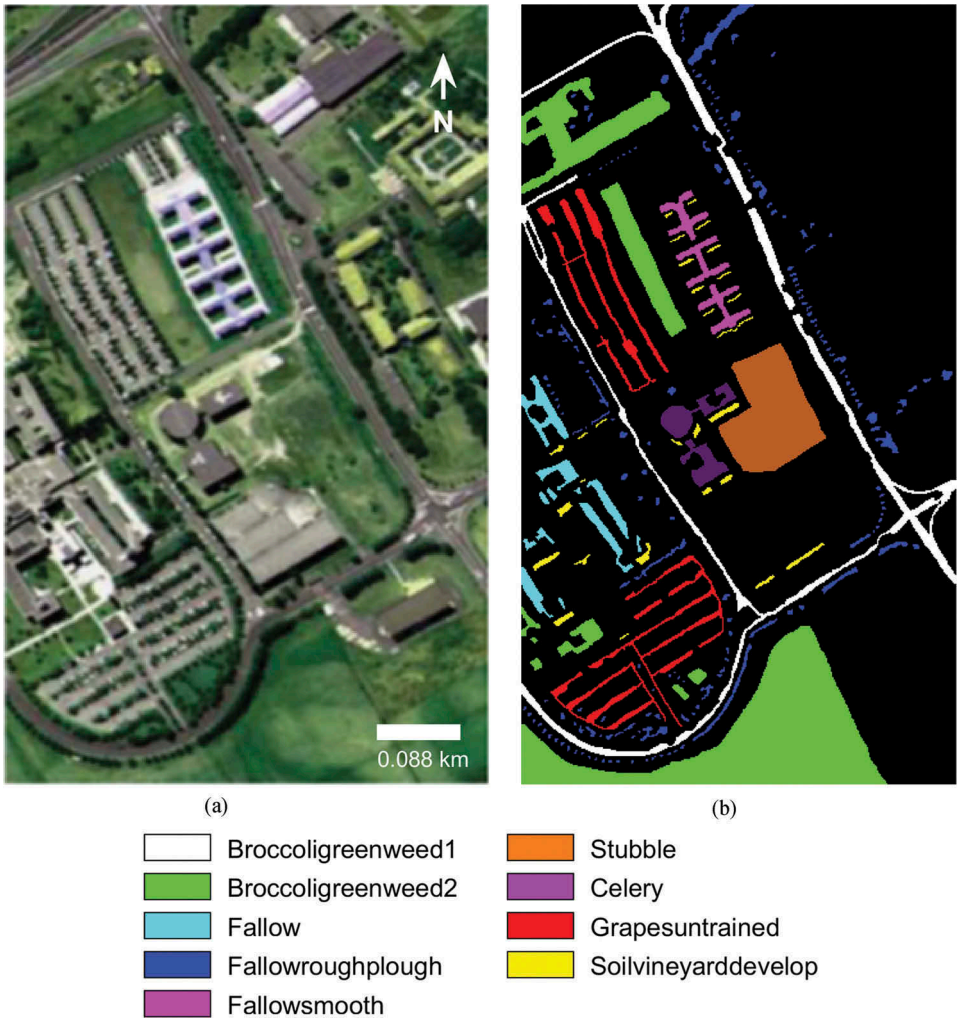

Figure 4. Pavia University dataset. (a) False-colour image and (b) Ground truth.

setting maximum iterations 50. The experiments are carried out over 30 independent runs with a random choice of training and test sets. Two representative supervised feature selection methods mRMR (Peng, Long, and Ding 2005), TMI-CSA (Feng et al. 2014), two typical semi-supervised feature selection methods SDIR (Feng et al. 2015), SSFC (Quinzán, Sotoca, and Pla 2009) and two competitive unsupervised feature selection methods FCM-GA (Wu 2012), MIMR-CSA (Feng et al. 2016) are chosen as the comparison methods.

Three popular indices, OA, AA, and κ, are employed to evaluate the performance of the proposal FCM-GWO quantitatively.

4.1. *Classification results on three hyperspectral datasets*

By setting C = 5, 3 and 6 for Indian Pines, Pavia University and Salinas data sets respectively, clustering results of features have been obtained by FCM. After grouping

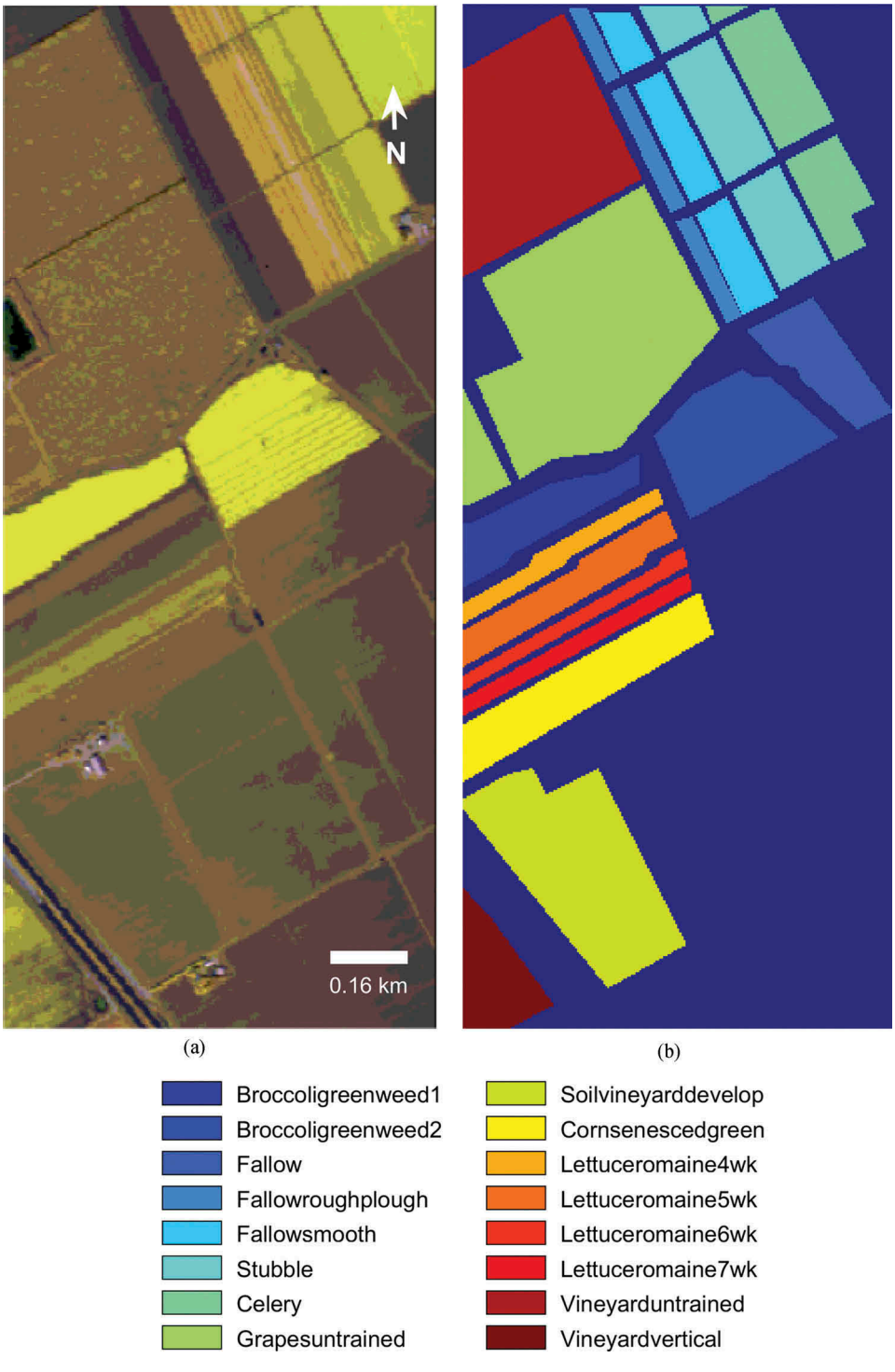

Figure 5. Salinas dataset. (a) False-colour image and (b) Ground truth.

all features into clusters, *K* different features are randomly selected from each cluster to

compose a grey wolf, $K = 1, 2, \ldots, 9$ for Indian Pines, $K = 2, 3, 5, 7, 9, 11$ for Pavia University dataset and $K = 1, 2, \ldots, 6$ for Salinas dataset. The most informative feature combination has been returned by GWO algorithm.

Table 2 reports the average classification accuracy over OA, AA, and κ with its standard deviation on Indian Pines dataset. The first column indicates the sampling ratios, and the first row denotes the band numbers of the most informative feature subsets obtained by WGO algorithm. The best classification result over OA is 86.64% ± 0.8% while sampling 20% and selecting 45 features. It can be seen that, as a whole, there is a growth tendency for each index from top-left to lower-right in Table 2. Classification results become to be better and better as the increasing of the number of selected features. There is a little variety from 30 features to 45 features for three indices.

In Table 3, the classification accuracies over OA are greater than 90% starting from the number 9 of selected features for different labelling proportions, in particular for 5%. Because the labelled samples are divided equally into 5 parts in SVM with five-fold cross-validation, that is to say, that only 1% of samples are as training samples and 4% as testing samples while 5% of samples are randomly labelled. In this case, the classification results on Pavia University dataset demonstrate powerfully the effectiveness of the proposed method. This conclusion has also been proved by classification results on Salinas dataset reported in Table 4. When the number of the selected band is equal to 21, 27 and 33, respectively, there is little difference in classification accuracy.

From Table 4, one can surprisingly see that all classification accuracies are above 90% for different sampling ratios and different numbers of the selected band. Especially, the fact that AA values (except one) in Table 4 all exceed 95% shows the outstanding performance of the GWO algorithm on Salinas dataset. The most distinctive feature subsets with different band numbers have been discovered by the introduced method. As the increasing of band numbers (from 12 to 36), classification result over three indices has a slight variation while sampling 10%, 15% and 20%. This phenomenon indicates that the bands appended later have no distinguished classification capability. It is worthwhile to consider how to determine the number of optimal bands.

4.2. *The comparison of classification results*

In order to impartially compare our FCM-GWO method with other six feature selection algorithms, the number of selected features is taken as 30 for the Indian Pines and Salinas datasets, and 20 for the Pavia University dataset as the other six algorithms. The same strategy of randomly sampling 20% is adopted in all experiments. For Pavia University dataset, the number 20 of selected feature cannot be achieved since we group all bands into three clusters and take k bands from each cluster. Therefore, we use the classification results of 15 and 21 features. All comparison results of seven algorithms on three datasets are listed in Tables 5–7 respectively.

From Table 5 one can see that the classification accuracies for each index increase orderly from top to down. This interesting fact shows that two supervised feature selection methods have gained bad classification results and three unsupervised feature selection methods have achieved the good ones. The result obtained by MIMR-CSA approach is closed to the one got by the proposed method. The differences between these two methods are only

Table 2. The classification results over OA, AA, and κ on Indian Pines dataset. The numbers marked in bold in each line represent the best classification results.

		5	10	15	20	25	30	35	40	45
5%	OA	70.62 ± 2.0	74.58 ± 2.3	75.80 ± 1.7	76.04 ± 1.7	75.97 ± 3.9	75.92 ± 1.4	76.51 ± 2.4	76.37 ± 1.1	**76.92 ± 2.5**
	AA	61.26 ± 13.4	64.49 ± 15.5	63.76 ± 9.5	65.91 ± 12.9	64.55 ± 14.3	66.95 ± 16.4	66.36 ± 13.3	65.11 ± 9.5	**67.63 ± 13.7**
	κ	66.10 ± 2.6	70.75 ± 2.8	72.21 ± 2.1	72.48 ± 2.0	72.36 ± 4.6	72.39 ± 1.6	73.07 ± 2.9	72.93 ± 1.2	**73.54 ± 2.7**
10%	OA	75.91 ± 0.9	79.00 ± 1.0	80.78 ± 1.4	81.08 ± 1.0	80.89 ± 1.2	80.98 ± 1.7	81.75 ± 1.6	81.11 ± 1.4	**82.24 ± 1.2**
	AA	69.09 ± 10.8	71.92 ± 12.7	74.51 ± 10.5	75.43 ± 11.7	74.83 ± 11.1	74.79 ± 11.1	75.96 ± 12.0	73.27 ± 10.1	**76.55 ± 10.7**
	κ	72.30 ± 0.9	75.96 ± 1.1	77.99 ± 1.7	78.35 ± 1.1	78.11 ± 1.3	78.23 ± 2.1	79.11 ± 1.8	78.38 ± 1.5	**79.69 ± 1.4**
15%	OA	77.41 ± 1.8	81.15 ± 1.6	82.94 ± 1.1	83.58 ± 0.9	83.61 ± 1.2	84.27 ± 0.9	84.14 ± 1.0	83.82 ± 1.6	**84.96 ± 0.6**
	AA	72.50 ± 10.2	76.73 ± 8.7	78.61 ± 9.0	79.83 ± 6.7	80.32 ± 7.1	80.06 ± 8.1	78.97 ± 8.4	78.75 ± 10.3	**81.12 ± 8.5**
	κ	74.02 ± 2.3	78.41 ± 2.0	80.48 ± 1.3	81.24 ± 1.1	81.26 ± 1.4	82.01 ± 1.0	81.87 ± 1.1	81.51 ± 1.8	**82.80 ± 0.7**
20%	OA	78.59 ± 1.0	82.90 ± 1.1	84.86 ± 0.7	84.91 ± 1.3	84.82 ± 0.7	85.92 ± 0.8	85.91 ± 0.6	86.67 ± 0.4	**86.64 ± 0.8**
	AA	73.78 ± 9.4	80.24 ± 6.5	81.94 ± 6.6	82.88 ± 7.1	82.12 ± 6.8	83.49 ± 5.5	82.53 ± 7.8	82.36 ± 6.5	**83.94 ± 7.3**
	κ	75.42 ± 1.1	80.44 ± 1.2	82.70 ± 0.8	82.75 ± 1.5	82.66 ± 0.8	83.92 ± 1.0	83.90 ± 0.7	83.63 ± 0.4	**84.73 ± 0.9**

Table 3. The classification results over OA, AA, and κ on Pavia University dataset. The numbers marked in bold each line represent the best classification results.

		6	9	15	21	27	33
5%	OA	86.92 ± 0.7	90.75 ± 0.3	91.01 ± 0.5	92.80 ± 0.4	92.77 ± 0.4	**92.82 ± 0.3**
	AA	83.73 ± 3.0	88.27 ± 2.0	88.31 ± 2.8	90.27 ± 1.9	90.63 ± 2.1	**90.67 ± 1.8**
	κ	82.33 ± 1.2	87.64 ± 0.4	87.98 ± 0.7	90.41 ± 0.5	90.37 ± 0.6	**90.44 ± 0.3**
10%	OA	87.73 ± 0.3	91.65 ± 0.2	92.13 ± 0.5	93.70 ± 0.3	93.78 ± 0.2	**93.80 ± 0.4**
	AA	84.90 ± 2.1	89.14 ± 1.6	89.96 ± 1.6	91.30 ± 1.9	91.71 ± 1.4	**91.87 ± 1.3**
	κ	83.45 ± 0.4	88.86 ± 0.3	89.50 ± 0.7	91.61 ± 0.5	91.72 ± 0.3	**91.75 ± 0.6**
15%	OA	88.06 ± 0.5	91.98 ± 0.3	92.71 ± 0.6	94.10 ± 0.2	**94.26 ± 0.2**	94.19 ± 0.3
	AA	85.34 ± 2.5	89.52 ± 1.3	90.50 ± 1.3	91.85 ± 1.2	**92.25 ± 1.4**	92.25 ± 1.3
	κ	83.92 ± 0.7	89.29 ± 0.4	90.27 ± 0.8	92.15 ± 0.3	**92.36 ± 0.2**	92.27 ± 0.4
20%	OA	88.43 ± 0.2	92.30 ± 0.3	92.96 ± 0.2	94.42 ± 0.2	94.48 ± 0.2	**94.59 ± 0.3**
	AA	85.77 ± 1.5	89.96 ± 1.4	90.88 ± 1.0	92.30 ± 0.9	92.54 ± 0.9	**92.74 ± 1.0**
	κ	84.41 ± 0.3	89.73 ± 0.4	90.6 ± 0.2	92.58 ± 0.3	92.66 ± 0.3	**92.80 ± 0.3**

Table 4. The classification results over OA, AA, and κ on Salinas dataset. The numbers marked in bold each line represent the best classification results.

		6	12	18	24	30	36
5%	OA	90.71 ± 0.7	92.02 ± 0.6	91.84 ± 0.4	92.32 ± 0.6	92.16 ± 0.4	**92.69 ± 0.2**
	AA	94.37 ± 2.0	95.27 ± 1.7	95.37 ± 1.2	95.61 ± 1.4	95.44 ± 2.0	**95.85 ± 1.4**
	κ	90.13 ± 0.8	91.10 ± 0.7	90.90 ± 0.5	91.44 ± 0.7	91.26 ± 0.4	**91.85 ± 0.3**
10%	OA	91.26 ± 0.4	92.91 ± 0.2	92.87 ± 0.2	93.11 ± 0.1	93.17 ± 0.2	**93.55 ± 0.1**
	AA	95.02 ± 1.2	96.09 ± 1.1	96.28 ± 1.0	96.31 ± 0.9	96.36 ± 1.1	**96.56 ± 1.0**
	κ	90.25 ± 0.5	92.09 ± 0.2	92.06 ± 0.2	92.31 ± 0.2	92.39 ± 0.2	**92.81 ± 0.2**
15%	OA	91.61 ± 0.3	93.32 ± 0.2	93.25 ± 0.2	93.47 ± 0.0	93.56 ± 0.3	**93.95 ± 0.2**
	AA	95.43 ± 0.8	96.52 ± 0.6	96.63 ± 0.9	96.69 ± 0.6	96.71 ± 1.0	**96.89 ± 0.7**
	κ	90.64 ± 0.3	92.55 ± 0.2	92.47 ± 0.3	92.72 ± 0.0	92.82 ± 0.4	**93.25 ± 0.2**
20%	OA	91.79 ± 0.1	93.47 ± 0.3	93.57 ± 0.2	93.68 ± 0.3	93.85 ± 0.2	**94.21 ± 0.3**
	AA	95.61 ± 0.6	96.62 ± 0.6	96.84 ± 0.6	96.82 ± 0.7	96.97 ± 0.8	**97.11 ± 0.7**
	κ	90.84 ± 0.1	92.72 ± 0.4	92.83 ± 0.2	92.96 ± 0.3	93.14 ± 0.2	**93.54 ± 0.3**

Table 5. The comparison results of seven methods over OA, AA, and κ on Indian Pines dataset.

	OA	AA	κ
mRMR	71.7 ± 2.6	54.8 ± 2.9	67.2 ± 3.2
TMI-CSA	72.9 ± 2.0	55.5 ± 2.3	68.7 ± 2.3
SSFC	73.3 ± 2.4	56.2 ± 2.9	69.1 ± 2.8
SDIR	78.6 ± 1.0	61.9 ± 1.4	75.4 ± 1.1
FCM-GA	79.2 ± 0.9	67.5 ± 1.5	76.2 ± 1.0
MIMR-CSA	84.5 ± 0.9	79.4 ± 1.8	82.3 ± 1.0
FCM-GWO	**85.9 ± 0.8**	**83.5 ± 5.5**	**83.9 ± 1.0**

1.4% for OA and 1.6% for κ. The distinct difference existed in mRMR and FCM-GWO is about 14% and 16% for OA and κ respectively. From the point of average accuracy, the advantage of our results is obvious. The excellent performance of our proposal is also shown in Figure 6. Our classification accuracies of each class over OA a higher or equal to the other six algorithms (except the 4^{th} and 15^{th} classes), especially for the 1^{st}, 7^{th}, 9^{th} 10^{th}, 12^{th} and 16^{th} classes. Five algorithms, mRMR, TMI-CSA, SSFC, SDIR and FCM-GA, display a poor classification performance for the 1^{st}, 4^{th}, 7^{th}, and 9^{th} classes. This may be directly leads to low classification accuracy on this dataset.

The comparison result of these methods on Pavia University image is reported in Table 6. The last two rows list classification results obtained by the proposal while

Table 6. The comparison results of seven methods on Pavia University dataset.

	OA	AA	κ
mRMR	85.6 ± 2.3	79.7 ± 3.2	80.6 ± 3.3
TMI-CSA	86.2 ± 1.6	79.4 ± 2.7	81.4 ± 2.2
SSFC	86.4 ± 1.2	78.8 ± 2.5	81.8 ± 1.6
SDIR	90.5 ± 0.7	85.5 ± 1.7	87.3 ± 1.0
FCM-GA	89.3 ± 0.6	86.4 ± 1.0	85.7 ± 0.9
MIMR-CSA	92.5 ± 0.3	89.1 ± 0.6	90.1 ± 0.3
FCM-GWO (15 bands)	**93.0 ± 0.2**	**90.9 ± 1.0**	**90.6 ± 0.2**
FCM-GWO (21 bands)	**94.4 ± 0.2**	**92.3 ± 0.9**	**92.6 ± 0.3**

Table 7. The comparison results of seven methods over OA, AA and κ on Salinas dataset.

	OA	AA	κ
mRMR	90.4 ± 0.5	94.3 ± 0.4	89.3 ± 0.5
TMI-CSA	90.8 ± 0.2	94.3 ± 0.2	89.8 ± 0.2
SSFC	91.0 ± 0.4	94.7 ± 0.3	90.0 ± 0.4
SDIR	92.7 ± 0.2	96.0 ± 0.3	91.9 ± 0.2
FCM-GA	92.7 ± 0.3	96.1 ± 0.3	91.9 ± 0.4
MIMR-CSA	93.5 ± 0.2	96.8 ± 0.1	92.7 ± 0.2
FCM-GWO	**93.9 ± 0.2**	**97.0 ± 0.6**	**93.1 ± 0.2**

sampling 20%. The performance of the proposal defeats all the other six algorithms while selecting 15 bands. The MIMR-CSA is still the most competitive method with other five approaches. Figure 10 exhibits the classification accuracy over OA of theses seven algorithms for each class on Pavia University dataset. From Figure 7, we can see that they acquire much the same accuracies on the 2nd, 5th and 9th classes. Their classification results present a notable difference for the 3rd and 7th classes, in particular, TMI-CSA for the 3rd class and SSFC for the 7th class. Maybe, these two methods have not selected the features which may distinguish effectively these two classes from the others. However, the proposed FCM-GWO method shows as usual perfect classification ability on the 3rd, 6th and 7th classes.

Satisfactory classification results of these seven methods on Salinas dataset are listed in Table 7. The classification accuracies over three indices are nearly outnumbered 90%. These seven algorithms all exhibit their outstanding classification ability. There is a little difference among three unsupervised feature selection methods, FCM-GA, MIMR-CSA, and FCM-GWO. The classification results reported in Table 7 show that the most informative features have been selected by these seven algorithms. The superiority of the proposed FCM-GWO algorithm is mainly reflected by the classification results on the 11th and 15th classes. For samples in the 1st,7th–9th and 12th classes, seven methods can nearly recognize correctly them. From Figure 8, it is easily can be seen that the better classification results for the 8th class and the 15th class have not been obtained by seven algorithms. As pointed out in Gualtieri and Chettri (1999), these two classes were, in fact, the same. Maybe, this is the radical reason why they cannot be classified correctly by nearly all existed methods. As a whole, the proposed unsupervised feature selection algorithm FCM-GWO has gained satisfying classification results on these three hyperspectral datasets. Compared with the proposal, the other unsupervised MIMR-CSA method is the most competitive approach. However, our method is more easily

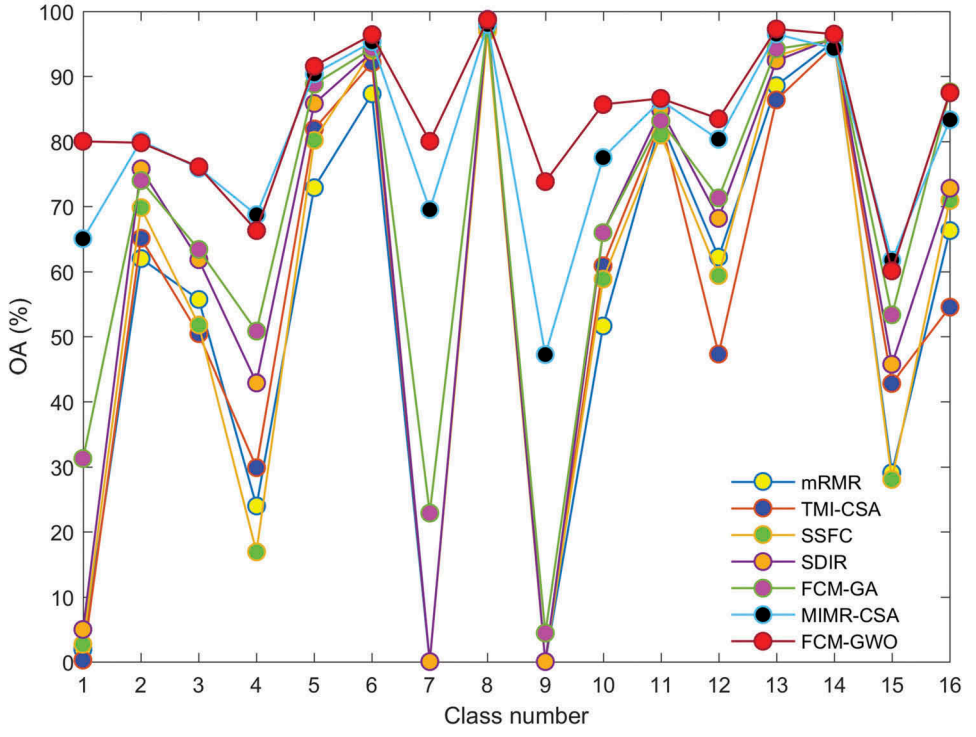

Figure 6. Classification accuracies of each class over OA on Indian Pines dataset.

understandable than MIMR-CSA technique. Experimental results also show that unsupervised feature selection methods may select the more informative features on the absence of *a priori* knowledge. It implies that the unsupervised feature selection method has a bright future in the study of dimension reduction in hyperspectral remote sensing.

Medjahed et al. (2016) applied straightly GWO method to select candidate band subsets and defined five objective functions as a fitness function to evaluate them. Based on the selected band subset, they used *k*-nearest neighbour (*k*-NN) classifier to perform classification. The best accuracy over OA they provided on three datasets is 73.67% (60 bands) for Indian Pines, 88.17% (9 bands) for Pavia University and 95.38% (62 bands) for Salinas dataset, respectively. We here have not compared our results with theirs since the different classifiers and numbers of the selected feature are utilized. On Pavia University dataset, the classification accuracy over OA 91.65% (9 Bands) obtained by our method is greater than their 88.17% while sampling 10%. In a sense, the superiority of the proposal is demonstrated.

In the case of randomly labelling 10% of samples in each class, we also compare our experimental results with those obtained by OMPBS (Liu et al. 2018) and MABS (Zhang et al. 2017b). Two data sets, Indian Pines and Pavia University, were used to test the performance of OMPBS. The highest accuracy acquired by OMPBS over OA is 90.7% (16 selected bands) on Pavia University data set, and 73.5% (34 selected bands) on Indian Pines data set, respectively (Liu et al. 2018). Accordingly, the average classification accuracy provided by the proposed method over OA on Pavia University and Indian

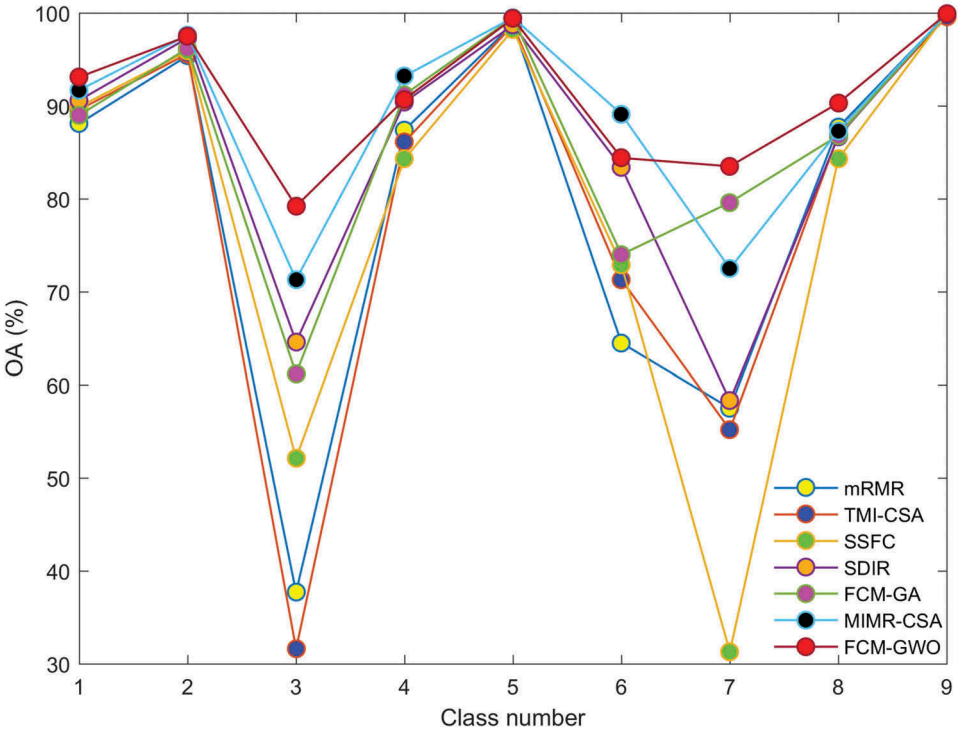

Figure 7. Classification accuracies of each class over OA on Pavia University dataset.

Pines data sets is 92.13% (15 features) and 80.98% (30 features) respectively. Although the authors reported their experimental results on three hyperspectral data sets in a broken line diagram, not the specific accuracy values, the visual results show that our classification results are similar to those reached by MABS method on Indian Pines and Salinas data sets, but obviously better than their on Pavia University data set (Zhang et al. 2017b). The comparison results indicate that feature combination selected by our method is more distinctive than those chosen by OMPBS and MABS methods.

Experimental results on three data sets show that the proposed feature selection outperforms the other six algorithms. The reasons behind this are: on the one hand, the use of clustering and selection strategies minimizes the redundancy of the resulting grey wolf and maintains crucial information. The grey wolf formed by selecting the same number of representative features from each feature subset makes each feature subset have the same contribution to the classification. Doing so clearly contributes to the improvement of classification capacity. On the other hand, different from the strategy of selecting crossover mutation or clone used in GA and CSA optimization algorithm, GWO adopts the method of global progressive optimization. Finally, it is the effective combination of clustering, selection, and optimization that makes our algorithm have perfect classification ability.

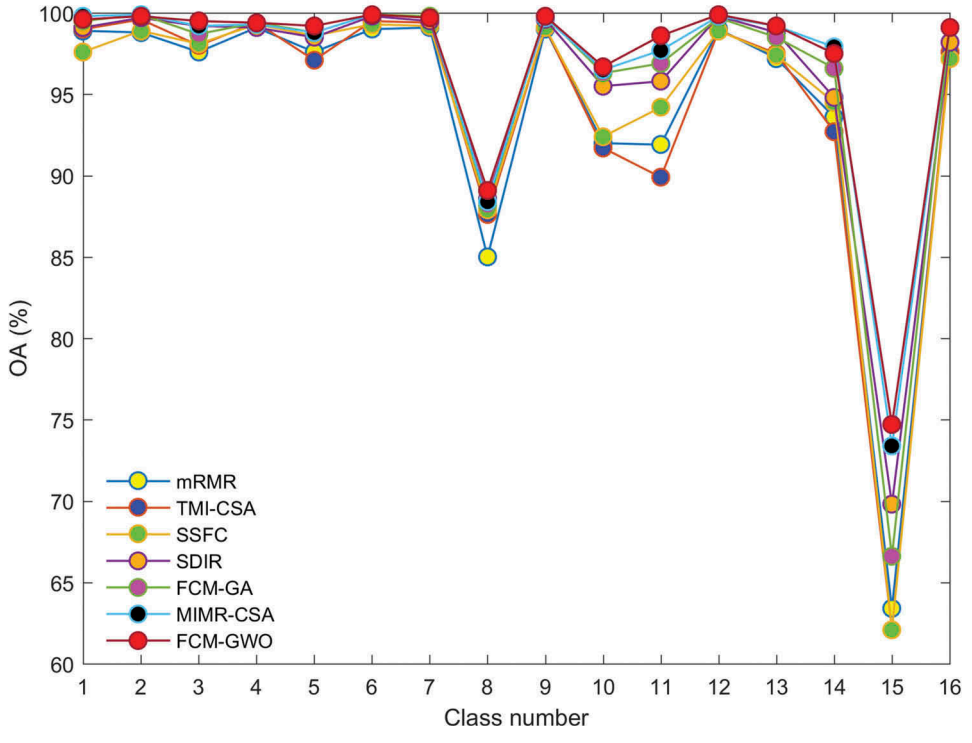

Figure 8. Classification accuracies of each class over OA on Salinas dataset.

4.3. *The analysis of parameters and computational time*

In what follows, taking Indian Pines dataset as an example, we analyse the effect of parameters on the average classification accuracy over OA and the computing time of several algorithms. Figure 9 shows how the average classification accuracy provided by seven algorithms varies with the number of selected features (NSF) from 10 to 100. The OA curves have a significant rise when NSF increases from 10 to 40, and then present a steady trend after NSF exceeds 80. The reason behind this is that the selected features are informative and complementary each other for classification while NSF is small. The redundancy between features increases evidently with the increasing of NSF and most of the selected features fail to provide useful information for classification. Compared with the other six methods, the most distinctive features can be selected by the proposed algorithm when NSF is less than 40. It manifests the effectiveness of our proposed strategy of feature subset decomposition, selection, and optimization.

Figure 10 records the computing time of the seven algorithms on the Indian Pines hyperspectral dataset (sampling 20%, 30 features). Compared to the other six algorithms, mRMR method has the shortest computing time due to its fast incremental search scheme. The change of cluster with differently selected features gives rise to recalculate the affinity function in each iteration; therefore, FCM-GA method is time-consuming. The use of fast hierarchical clustering technique not only makes SSFC algorithm take less time but also reduces the computing time

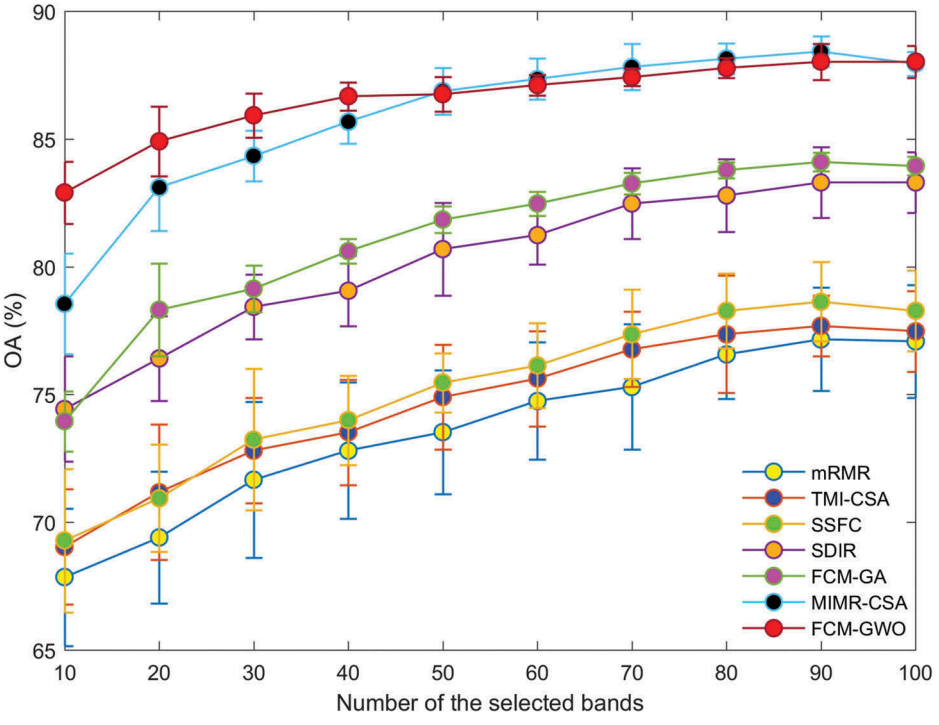

Figure 9. Mean overall accuracy of seven algorithms on the Indian Pines image as the number of selected bands increases from 10 to 100.

with the increase of feature number. Based on the pre-calculated entropy of each band, FCM-GWO method can complete the selection process in 10 s and is only slower than the mRMR algorithm.

The parameter *a* used in GWO algorithm controls the optimization range of the grey wolf.

The change of mean overall accuracy on the Indian Pines dataset with different parameter *a* is displayed in Figure 11. The classification accuracy shows a slight downward trend with the increase of *a* value from 1 to 5. When *a* is less than 2.5, the variation of *a* has little effect on classification accuracy. The reason why the classification accuracy decreases with the increase of *a* value may be that a very few features enter another feature subset during the optimization process. This further illustrates the necessity of feature subset decomposition and selection strategy in our proposal. The difference 1.4% of average classification accuracy proves to be no strong relationship between the classification result and the change of parameters.

5. Conclusion

In this study, an unsupervised feature selection method is proposed on the basis of fuzzy *c*-means and grey wolf optimizer algorithm. Experimental results on three

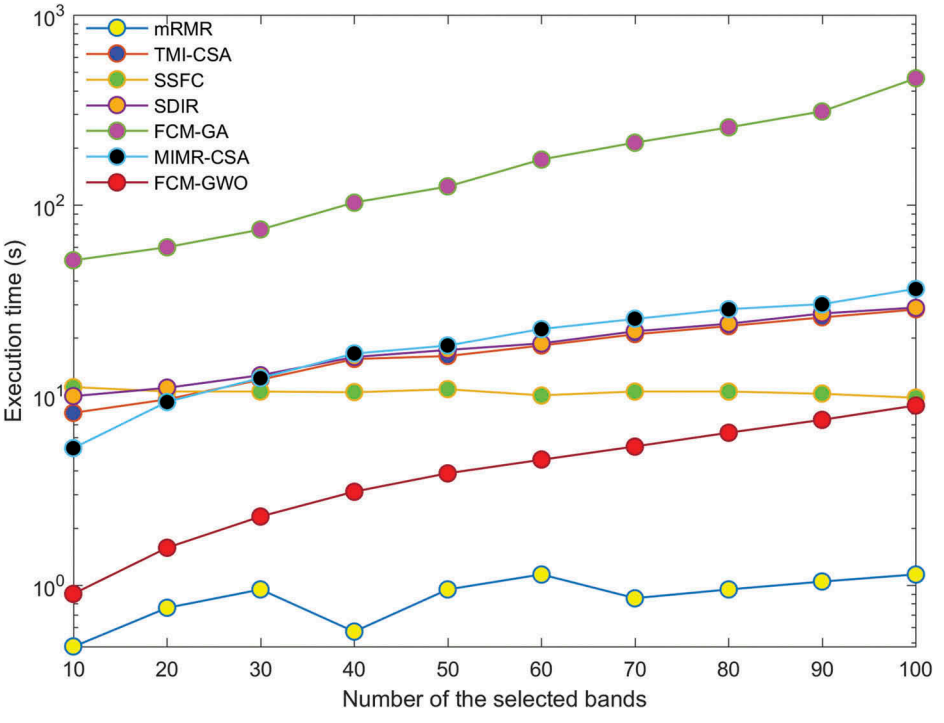

Figure 10. Execution times of seven methods on the Indian Pines image as the number of selected bands increases from 10 to 100.

typical hyperspectral datasets confirm successfully the validity of the suggested method. The comparison results of seven methods demonstrate that the unsupervised feature selection method may possess good feature selection ability. The relevant spectral bands are removed availably by feature subspace decomposition and optimization-based process. The semi-supervised SVM classifier is lastly used to test the effectiveness of the selected features.

The main contributions of the proposal are:

- A new unsupervised feature selection strategy is developed based on FCM and GWO.
- The introduced method is straightly and easily understandable since the computation is simple and readily comprehensible.
- The GWO method is slightly improved so as to apply it to the discrete case.
- The assignation of parameters, class number c, the number of grey wolf population N and max number of iterations, is trivial and unrelated to experience.

As one can see in Tables 2–4, the classification accuracy has not changed surprisingly as the number of selected features increases multiplicatively. It means that there still exist redundant bands in the features added later. Therefore, it is a

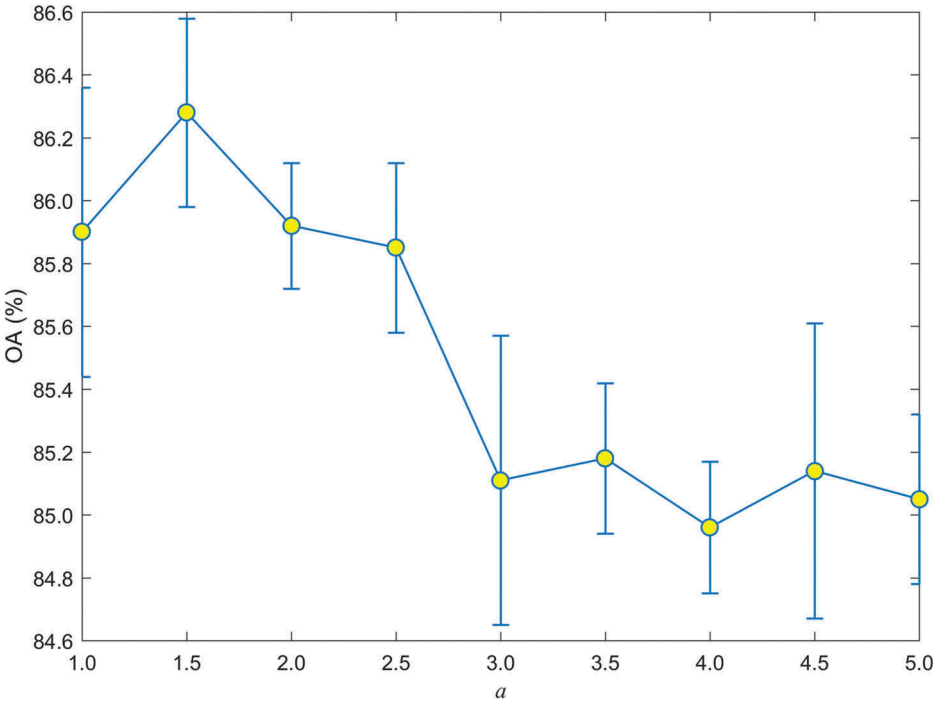

Figure 11. The curves of mean overall accuracy on the Indian Pines dataset with different parameter *a*.

challenging and valuable problem how to exactly or properly find the needed features for hyperspectral data classification.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the National Natural Science Foundation of China [61772252,41771178];

ORCID

Jun Yang http://orcid.org/0000-0001-6740-4358

References

Agarwal, A., T. El-Ghazawi, H. El-Askary, and J. Le-Moigne. 2008. "Efficient Hierarchical-PCA Dimension Reduction for Hyperspectral Imagery." *IEEE International Symposium on Signal Processing and Information Technology*, Giza, Egypt, 353–356. doi: 10.1007/s11517-007-0279-6

Bazi, Y., and F. Melgani. 2006. "Toward an Optimal SVM Classification System for Hyperspectral Remote Sensing Images." *IEEE Transactions on Geoscience & Remote Sensing* 44 (11): 3374–3385. doi:10.1109/TGRS.2006.880628.

Bevilacqua, M., and Y. Berthoumieu. 2017. "Unsupervised Hyperspectral Band Selection via Multi-Feature Information-Maximization Clustering". *IEEE International Conference on Image Processing*, Beijing, China, 540–544. doi: 10.4070/kcj.2016.0445

Breaban, M., and H. Luchian. 2011. "A Unifying Criterion for Unsupervised Clustering and Feature Selection." *Pattern Recognition* 44: 854–865. doi:10.1016/j.patcog.2010.10.006.

Camps-Valls, G., T. V. Bandos Marsheva, and D. Zhou. 2007. "Semi-Supervised Graph-Based Hyperspectral Image Classification." *IEEE Transactions on Geoscience & Remote Sensing* 45 (10): 3044–3054. doi:10.1109/TGRS.2007.895416.

Cao, X., C. Wei, J. Han, and L. Jiao. 2017. "Hyperspectral Band Selection Using Improved Classification Map." *IEEE Geoscience & Remote Sensing Letters* 99: 1–5.

Cao, X., T. Xiong, and L. Jiao. 2016. "Supervised Band Selection Using Local Spatial Information for Hyperspectral Image." *IEEE Geoscience & Remote Sensing Letters* 13 (3): 329–333.

Chiang, S., C. I. Chang, and I. W. Ginsberg 2000. "Unsupervised Hyperspectral Image Analysis Using Independent Component Analysis." *IEEE Geoscience and Remote Sensing Symposium*, Honolulu Hawaii, USA, 2000. *Proceedings. IGARSS* 2000. 7: 3136–3138.

Du, Q., and H. Yang. 2008. "Similarity-Based Unsupervised Band Selection for Hyperspectral Image Analysis." *IEEE Geoscience & Remote Sensing Letters* 5 (4): 564–568. doi:10.1109/LGRS.2008.2000619.

Dunn, J. C. 1973. "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters." *Journal of Cybernetics* 3 (3): 32–57. doi:10.1080/01969727308546046.

Feng, J., L. Jiao, F. Liu, T. Sun, and X. Zhang. 2015. "Mutual-Information-Based Semi-Supervised Hyperspectral Band Selection with High Discrimination, High Information, and Low Redundancy." *IEEE Transactions on Geoscience & Remote Sensing* 53 (5): 2956–2969. doi:10.1109/TGRS.2014.2367022.

Feng, J., L. Jiao, F. Liu, T. Sun, and X. Zhang. 2016. "Unsupervised Feature Selection Based on Maximum Information and Minimum Redundancy for Hyperspectral Images." *Pattern Recognition* 51 (C): 295–309. doi:10.1016/j.patcog.2015.08.018.

Feng, J., L. Jiao, X. Zhang, and T. Sun. 2014. "Hyperspectral Band Selection Based on Trivariate Mutual Information and Clonal Selection." *IEEE Transactions on Geoscience & Remote Sensing* 52 (7): 4092–4105. doi:10.1109/TGRS.2013.2279591.

Ghamisi, P., and J. A. Benediktsson. 2014. "Feature Selection Based on Hybridization of Genetic Algorithm and Particle Swarm Optimization." *IEEE Geoscience & Remote Sensing Letters* 12 (2): 309–313. doi:10.1109/LGRS.2014.2337320.

Ghamisi, P., M. S. Couceiro, and J. A. Benediktsson. 2015. "A Novel Feature Selection Approach Based on FODPSO and SVM." *IEEE Transactions on Geoscience & Remote Sensing* 53 (5): 2935–2947. doi:10.1109/TGRS.2014.2367010.

Gualtieri, J. A., and S. R. Chettri. 1999. "Support Vector Machine Classifiers as Applied to AVIRIS Data." *Summaries of the Eighth JPL Airborne Earth Science Workshop*, Pasadena, California, USA, February 8–11.

Guo, B., S. R. Gunn, R. I. Damper, and J. D. B. Nelson. 2006. "Band Selection for Hyperspectral Image Classification Using Mutual Information." *IEEE Geoscience & Remote Sensing Letters* 3 (4): 522–526. doi:10.1109/LGRS.2006.878240.

Jia, S., G. Tang, J. Zhu, and Q. Li. 2015. "A Novel Ranking-Based Clustering Approach for Hyperspectral Band Selection." *IEEE Transactions on Geoscience & Remote Sensing* 54 (1): 88–102. doi:10.1109/TGRS.2015.2450759.

Karaboga, D., and B. Basturk. 2008. "On the Performance of Artificial Bee Colony (ABC) Algorithm." *Applied Soft Computing* 8 (1): 687–697. doi:10.1016/j.asoc.2007.05.007.

Karaboga, D., and C. Ozturk. 2011. "A Novel Clustering Approach: Artificial Bee Colony (ABC) Algorithm." *Applied Soft Computing* 11 (1): 652–657. doi:10.1016/j.asoc.2009.12.025.

Khairuzzaman, A., M. Kayom, and S. Chaudhury. 2017. "Multilevel Thresholding Using Grey Wolf Optimizer for Image Segmentation." *Expert Systems with Applications* 86: 64–76. doi:10.1016/j.eswa.2017.04.029.

Kim, S. B., and P. Rattakorn. 2011. "Unsupervised Feature Selection Using Weighted Principal Components." *Expert Systems with Applications* 38 (5): 5704–5710. doi:10.1016/j.eswa.2010.10.063.

Li, H., Y. Wang, S. Xiang, J. Duan, F. Zhu, and P. Chunhong. 2016. "A Label Propagation Method Using Spatial-Spectral Consistency for Hyperspectral Image Classification." *International Journal of Remote Sensing* 37 (1): 191–211. doi:10.1080/01431161.2015.1125547.

Liu, K. H., S. Y. Chen, H. C. Chien, and M. H. Lu. 2018. "Progressive Sample Processing of Band Selection for Hyperspectral Image Transmission." *Remote Sensing* 10 (3): 64.

Liu, L., C. F. Li, Y. M. Lei, J. Y. Yin, and J. J. Zhao. 2017. "Feature Extraction for Hyperspectral Remote Sensing Image Using Weighted PCA-ICA." *Arabian Journal of Geosciences* 10 (14): 307. doi:10.1007/s12517-017-3090-1.

Liu, Y., C. Wang, Y. Wang, X. Bai, J. Zhou, and L. Bai. 2017. "Differential Weights-Based Band Selection for Hyperspectral Image Classification." *International Journal of Wavelets Multiresolution & Information Processing* 15 (6): 1–16. doi:10.1142/S0219691317500655.

Liu, Y., G. Cao, Q. Sun, and M. Siegel. 2015. "Hyperspectral Classification via Deep Networks and Superpixel Segmentation." *International Journal of Remote Sensing* 36 (13): 3459–3482. doi:10.1080/01431161.2015.1055607.

Martínez Sotoca, J., and F. Pla. 2010. "Supervised Feature Selection by Clustering Using Conditional Mutual Information-Based Distances." *Pattern Recognition* 43 (6): 2068–2081. doi:10.1016/j.patcog.2009.12.013.

Martínez-Usómartinez-Uso, A., F. Pla, J. M. Sotoca, and G.-S. Pedro. 2007. "Clustering-Based Hyperspectral Band Selection Using Information Measures." *IEEE Transactions on Geoscience & Remote Sensing* 45 (12): 4158–4171. doi:10.1109/TGRS.2007.904951.

Medjahed, S. A., T. Ait Saadi, A. Benyettou, and M. Ouali. 2016. "Gray Wolf Optimizer for Hyperspectral Band Selection." *Applied Soft Computing* 40 (C): 178–186. doi:10.1016/j.asoc.2015.09.045.

Mirjalili, S. 2015. "*How Effective Is the Grey Wolf Optimizer in Training Multi-Layer Perceptrons*." *Applied Intelligence* 43 (1): 150–161.

Mirjalili, S., S. Saremi, S. M. Mirjalili, and L. D. S. Coelho. 2016. "Multi-Objective Grey Wolf Optimizer: A Novel Algorithm for Multi-Criterion Optimization." *Expert Systems with Applications* 47: 106–119. doi:10.1016/j.eswa.2015.10.039.

Mirjalili, S., S. M. Mirjalili, and A. Lewis. 2014. "Grey Wolf Optimizer." *Advances in Engineering Software* 69 (3): 46–61. doi:10.1016/j.advengsoft.2013.12.007.

Mitra, P., C. A. Murthy, and S. K. Pal. 2002. "Unsupervised Feature Selection Using Feature Similarity." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24 (3): 301–312. doi:10.1109/34.990133.

Peng, H., F. Long, and C. Ding. 2005. "Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27 (8): 1226–1238. doi:10.1109/TPAMI.2005.159.

Quinzán, I., J. M. Sotoca, and F. Pla. 2009. "Clustering-Based Feature Selection in Semi-Supervised Problems." International Conference on Intelligent Systems Design and Applications. Pisa, Italy, 535–540.

Rodríguez, L., O. Castillo, J. Soria, P. Melin, F. Valdez, C. I. Gonzalez, G. E. Martinez, and J. Soto. 2017. "A Fuzzy Hierarchical Operator in the Grey Wolf Optimizer Algorithm." *Applied Soft Computing* 57: 315–328. doi:10.1016/j.asoc.2017.03.048.

Shen, L., Z. Zhu, S. Jia, J. Zhu, and Y. Sun. 2013. "Discriminative Gabor Feature Selection for Hyperspectral Image Classification." *IEEE Geoscience & Remote Sensing Letters* 10 (1): 29–33. doi:10.1109/LGRS.2012.2191761.

Slavkovikj, V., S. Verstockt, W. De Neve, S. Van Hoecke, and R. Van De Walle. 2016. "Unsupervised Spectral Sub-Feature Learning for Hyperspectral Image Classification." *International Journal of Remote Sensing* 37 (2): 309–326. doi:10.1080/01431161.2015.1125554.

Tesmer, M., C. A. Perez, and J. M. Zurada. 2009. "Normalized Mutual Information Feature Selection." *IEEE Transactions on Neural Networks* 20 (2): 189–201. doi:10.1109/TNN.2008.2005601.

Veen, L. 2009. "Interactive Exploration of Uncertainty in Fuzzy Classifications by Isosurface Visualization of Class Clusters." *International Journal of Remote Sensing* 30 (18): 4685–4705. doi:10.1080/01431160802651942.

Wang, M., Y. Wan, Z. Ye, X. Gao, and X. Lai. 2017. "A Band Selection Method for Airborne Hyperspectral Image Based on Chaotic Binary Coded Gravitational Search Algorithm." *Neurocomputing* 273: 57–67. doi:10.1016/j.neucom.2017.07.059.

Winkler, R., F. Klawonn, and R. Kruse. 2011. "Fuzzy C-Means in High Dimensional Spaces." *International Journal of Fuzzy System Applications* 1 (1): 1–16. doi:10.4018/ijfsa.2011010101.

Wu, J. 2012. "Unsupervised Intrusion Feature Selection based on Genetic Algorithm and FCM." *Lect. Notes Electr. Eng*. 154: 1005–1012.

Yang, H., D. Qian, S. Hongjun, and Y. Sheng. 2011. "An Efficient Method for Supervised Hyperspectral Band Selection." *IEEE Geoscience & Remote Sensing Letters* 8 (1): 138–142. doi:10.1109/LGRS.2010.2053516.

Yu, C., M. Song, and C. I. Chang. 2018. "Band Subset Selection for Hyperspectral Image Classification." *Remote Sensing* 10 (1): 113. doi:10.3390/rs10010113.

Yue, J., W. Zhao, S. Mao, and H. Liu. 2015. "Spectral–Spatial Classification of Hyperspectral Images Using Deep Convolutional Neural Networks." *Remote Sensing Letters* 6 (6): 468–477. doi:10.1080/2150704X.2015.1047045.

Zhang, D., M. Ji, J. Yang, Y. Zhang, and F. Xie. 2014. "A Novel Cluster Validity Index for Fuzzy Clustering Based on Bipartite Modularity." *Fuzzy Sets & Systems* 253 (9): 122–137. doi:10.1016/j.fss.2013.12.013.

Zhang, L., Y. Zhong, B. Huang, J. Gong, and P. Li. 2007. "Dimensionality Reduction Based on Clonal Selection for Hyperspectral Imagery." *IEEE Transactions on Geoscience & Remote Sensing* 45 (12): 4172–4186. doi:10.1109/TGRS.2007.905311.

Zhang, L., Zhang, Q., Du, B., Huang, X., et al. 2017. "Simultaneous Spectral-Spatial Feature Selection and Extraction for Hyperspectral Images." *IEEE Transactions on Cybernetics* 48 (1): 16–28. DOI:10.1109/TCYB.2016.2605044.

Zhang, M., J. Ma, M. Gong, H. Li, and J. Liu. 2017b. "Memetic Algorithm Based Feature Selection for Hyperspectral Images Classification." *IEEE Congress on Evolutionary Computation, Donostia, San Sebastián, Spain* 495–502.

Zhang, X., Q. Kang, J. Cheng, and X. Wang. 2018. "A Novel Hybrid Algorithm Based on Biogeography-Based Optimization and Grey Wolf Optimizer." *Applied Soft Computing* 67: 197–214. doi:10.1016/j.asoc.2018.02.049.

Zhang, X., Y. He, N. Zhou, and Y. Zheng. 2013. "Semisupervised Dimensionality Reduction of Hyperspectral Images via Local Scaling Cut Criterion." *IEEE Geoscience & Remote Sensing Letters* 10 (6): 1547–1551. doi:10.1109/LGRS.2013.2261797.

Zhang, Y., M. D. Desai, J. Zhang, and M. Jin 1999. "Adaptive Subspace Decomposition for Hyperspectral Data Dimensionality Reduction." *International Conference on Image Processing, 1999. ICIP 99. Proceedings*. IEEE 1999, Kobe, Japan, 2:326–329.