Knowledge-Based Systems 168 (2019) 39–48

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

A feature selection approach for hyperspectral image based on modified ant lion optimizer

Mingwei Wang [a,*], Chunming Wu [a], Lizhe Wang [b], Daxiang Xiang [c], Xiaohui Huang [b]

[a] Institute of Geological Survey, China University of Geosciences, Wuhan, 430074, PR China
[b] School of Computer Science, China University of Geosciences, Wuhan, 430074, PR China
[c] Changjiang River Scientific Research Institute, Changjiang Water Resources Commission, Wuhan, 430010, PR China

HIGHLIGHTS

- A novel feature selection approach for HSI based on MALO algorithm is proposed.
- A novel MALO algorithm based on the standard ALO algorithm and Lévy flight is proposed.
- Harmonic wavelet kernel is utilized to construct WSVM classifier for classification.
- A new evaluation criterion is formulated to estimate the performance of feature selection.

ARTICLE INFO

Article history:
Received 14 August 2018
Received in revised form 17 December 2018
Accepted 27 December 2018
Available online 4 January 2019

Keywords:
Hyperspectral image
Feature selection
Wavelet support vector machine
Ant lion optimizer
Lévy flight

ABSTRACT

Feature selection is one of the most important issues in hyperspectral image (HSI) classification to achieve high correlation between the adjacent bands. The main concern is selecting fewer bands with the highest accuracy as possible. Generally, it is a combinatorial optimization problem and cannot be fully solved by swarm intelligence algorithms. Ant lion optimizer (ALO) is a newly proposed swarm intelligence algorithm that mimics the swarming behaviour of antlions. In addition, wavelet support vector machine (WSVM) is able to enhance the stability of the classification result, and Lévy flight helps swarm intelligence algorithms jump out of the local optimum. Therefore, in this paper, a novel feature selection method based on a modified ALO (MALO) and WSVM is proposed to reduce the dimensionality of HSIs. The proposed method is compared with some state-of-the-art algorithms on some HSI datasets. Moreover, a new evaluating criteria is formulated to estimate the performance of feature selection, and the classification accuracy and selected number of bands are balanced as much as possible. Experimental results demonstrate that the proposed method outperforms other approaches, finds the optimal solution with a reasonable convergence orientation, and its classification accuracy is satisfied with fewer bands, it is robust, adaptive and might be applied for practical work of feature selection.

1. Introduction

Recent advances in remote sensing technology have made hyperspectral image (HSI) more widely available with dense sampling of several narrow continuous bands, and each band indicates a one-dimensional feature [1]. The high spectral resolution provides the potential for better discrimination of different physical objects, while also yielding large volumes of data. To reduce the data redundancy, feature selection has been one of the hot topics in the community of remote sensing [2]. Specifically, it is necessary to develop efficient and effective feature selection techniques to improve the accuracy and efficiency of classification.

Feature selection is a fundamental task for pattern recognition and data mining applications, especially for high-dimensional datasets. It is the process of selecting a subset of independent features for use in a model's construction [3]. Ideally, objects may be described more completely with more features and each feature should supplement an independent set of information [4]. From the evaluation perspectives, feature selection can be divided into two categories based on the searching strategy. In filter methods, the subset is constructed independently from the pattern recognition algorithm, and the merits of a subset are estimated in terms of the specific characteristics of the information [5]. However, the above feature selection methods are evaluated according to the correlation of each band, which will cost a lot of computational time, and it is difficult to satisfy the real-time processing. In the wrapper methods, the quality of a subset is evaluated based on a classifier model [6], and the commonly used classifiers can be

* Corresponding author.
E-mail address: wmwscola@sina.com (M. Wang).

https://doi.org/10.1016/j.knosys.2018.12.031
0950-7051/

divided into two categories: unsupervised and supervised. The supervised models need to know the class label of each training sample in advance, which results in a better classification result than unsupervised models in most cases. As a commonly used supervised classifier, support vector machine (SVM) is a powerful machine learning approach for data classification and regression problems with small sample sizes and high dimensionality [7], and it has been widely used in many fields [8–10]. In particular, many studies addressed the problem of HSI classification by using SVM, and they obtained higher classification accuracy than traditional approaches [11–14]. Moreover, kernel function plays a very important role in the performance of the SVM method. The basic idea of SVM is that the kernel function is used to map the input data into a higher dimensional feature space so that the classification problem becomes linearly separable [15]. In recent years, the combination of wavelet theories and SVM, namely, wavelet support vector machine (WSVM) has drawn considerable attention due to the powerful nonlinear mapping ability of SVM and the ability of wavelet kernel functions to locally analyse and extract the characteristic features from time series.

Subset generation is considered as a search process that selects the subset of items from the original dataset using complete, random or heuristic search. The complete search generates all possible subsets to select the optimal one. If the dataset totally includes N features, then $O(2^N)$ subsets will be generated and assessed, which is a non-polynomial hard problem for the larger size datasets. Random search selects the attributes and searches for the next subset randomly [16], which may be performed as a complete search in the worst case. As a combinatorial optimization problem, swarm intelligence algorithms with heuristic search can be utilized as guiding strategies in designing underlying heuristics to solve the feature selection problem [17], such as genetic algorithm [18], particle swarm optimization (PSO) algorithm [19], differential evolution (DE) algorithm [20], gravitational search algorithm (GSA) [21], cuckoo search (CS) algorithm [22], gray wolf optimizer (GWO) [23], grasshopper optimization algorithm (GOA) [24], salp swarm algorithm (SSA) [25,26] and dragonfly algorithm (DA) [27], which may demonstrate superior efficacy in tackling feature selection problems when compared to the exact methods.

For HSI feature selection, Wang [28] proposed a hybrid feature selection strategy based on artificial bee colony algorithm and SVM, which formed a wrapper to search for the optimal combination of bands with high classification accuracy. Su [29] proposed a PSO-based optimization system to simultaneously determine the optimal number of bands and select the independent bands for hyperspectral dimensionality reduction, which obviously outperformed the popular sequential forward selection (SFS) method. Ghosh [30] presented a novel feature selection technique by using DE algorithm for the subset generation in HSIs, which could result in a significant improvement over the existing algorithms with respect to the overall classification accuracy and Kappa coefficient. Wang [31] proposed a novel feature selection method based on GSA to reduce the dimensionality of airborne HSIs, and the overall classification result was satisfactory for the ground object targets or classes. Medjahed [32] proposed a new procedure for feature selection by using CS algorithm to optimize the objective function, which obtained satisfactory results compared to other approaches and classifier systems that used all of bands. Xie [33] presented a novel wrapper feature selection method based on GWO and SVM with five-fold cross-validation to ensure that the selected band subset had a good classification ability. However, the above algorithms have some parameters that need to be set by user, which may lead to the stagnation problem, and failure to find the optimal solution.

Ant lion optimizer (ALO) [34] is a newly proposed stochastic global search algorithm. Currently, ALO has been widely used in diverse applications, e.g. Raju [35] utilized ALO for the simultaneous optimization of the controller, which could find better results when facing a random load pattern as disturbance. Kamboj [36] applied ALO to solve the non-convex and dynamic economic load dispatch problem of electric power system, and experimental results shown that ALO had a good balance between exploration and exploitation that resulted in high local optima avoidance. Yao [37] used ALO for the route planning of unmanned aerial vehicle, which was superior to other swarm intelligence algorithms in terms of its robustness, convergence speed, accuracy and local minima avoidance. The performance of ALO does not depend on any parameters, which makes it not easy to trap into the local optimal, and it has stable converge to the global best solution. In the paper, feature selection is considered as a discrete optimization problem, which cannot be solved by using the standard ALO. Mafarja [38] proposed a binary coded ALO (BALO) to conduct feature selection for the datasets from the UCI machine learning repository, and experimental results demonstrated that BALO acquired the feature space for the optimal band combination regardless of the initialization of the stochastic operators. Moreover, Emary [39] presented a novel ALO that used Lévy flight (LF), which was called Lévy ALO (LALO) to improve the local optimization ability of the basic algorithm. However, LF distribution for each ant will increase the CPU time for the iteration process, which will decrease the efficiency of the algorithm.

Hence, a WSVM classifier and modified ALO (MALO) with LF distribution for the local disturbance of part of optimal ants and binary coding for the discrete optimization problem is considered here, and its utilization for feature selection of HSIs is investigated. In this work, we have made the following contributions:

- A MALO based on the standard ALO and LF distribution is proposed, and LF distribution has a strong exploration ability that is utilized to search for the optimal ants and adapt to the discrete optimization problem.
- Harmonic wavelet kernel is utilized to construct WSVM for classification, and it is tested on three HSI datasets.
- A new evaluation criteria is formulated to estimate the performance of feature selection, and the classification accuracy and selected number of bands is balanced as much as possible.
- A feature selection approach for HSI datasets based on MALO is proposed, and the classification accuracy is the optimal with fewer bands.

The rest of the paper is structured as below. Section 2 illustrates the basic principle of MALO. Section 3 describes the relative work of SVM and wavelet kernel function. The fundamental viewpoint of the proposed approach for feature selection of HSI is detailed in Section 4. Section 5 displays the experimental results and discussion. Finally, the paper is concluded in Section 6.

2. Overview of MALO

2.1. The mathematical model of ALO

In 2015, Mirjalili proposed a novel heuristic search algorithm called ant lion optimizer (ALO) that mimics the hunting behaviour of antlions in nature and has no parameters to adjust [34]. Thus, ALO has plenty of potential for avoiding the local optima stagnation, because of the use of random walk and roulette wheel. Exploration of the search space in ALO is guaranteed by the random selection of antlions and the random walk of ants around them, and exploitation of the search space is guaranteed by the adaptive shrinking boundaries of antlions' traps. The mathematical model of ALO can be explained with the help of the following steps.

As ants move stochastically in nature when search for food, the random walk of ants can be described as follows:

$$X^t = [0, cumsum(2r(t_1) - 1), cumsum(2r(t_2) - 1), \ldots, cumsum(2r(t_n) - 1)] \quad (1)$$

Here X^t is the random walk of ants, n is the maximum number of iterations, *cumsum* represents the cumulative sum, t is the step of random walk (iteration), and $r(t)$ is a stochastic function that is defined as follows:

$$r(t) = \begin{cases} 1 & \text{if } rand > 0.5 \\ 0 & \text{if } rand \leq 0.5 \end{cases} \quad (2)$$

where *rand* is a random number produced by a uniform distribution in the interval of [0,1].

To keep the random walk of ants inside the search space, the position of each ant is normalized by using the following min–max normalization equation:

$$X_i^t = \frac{(X_i^t - a_i)(b_i - c_i^t)}{d_t^t - a_i} + c_i^t \quad (3)$$

where a_i is the minimum random walk of ith variable, b_i is the maximum random walk of ith variable, c_i^t is the minimum of ith variable at iteration t, and d_i^t is the maximum of ith variable at iteration t.

With the mechanisms proposed so far, antlions are able to build traps that are proportional to their fitness and ants are required to move randomly. However, antlions shoot sands outwards the centre of the pit once they realize that an ant is in the trap. The behaviour slides down the trapped ant, which is trying to escape. To mathematically model the behaviour, the radius of an ant's random walk is decreased adaptively, which can be described as follows:

$$c^t = \frac{c^t}{I} \quad (4)$$

$$d^t = \frac{d^t}{I} \quad (5)$$

where $I = 10^{\omega \frac{t}{T}}$, T is the maximum number of iterations, and ω is a constant that is defined based on the current iteration ($\omega = 2$ when $t > 0.1T$, $\omega = 3$ when $t > 0.5T$, $\omega = 4$ when $t > 0.75T$, $\omega = 5$ when $t > 0.9T$, and $\omega = 6$ when $t > 0.95T$). Basically, the constant ω can adjust the accuracy level of exploitation.

2.2. LF distribution with random walk

LF distribution is a hypothesis in the field of biology that can optimize the searching efficiency. It is a random walk strategy in which the step-lengths have a probability distribution that is heavy-tailed [40]. Due to the ergodic and dynamic properties of random walk, LF distribution has been widely utilized in the field of evolutionary computation for solving complex optimization problems [41,42]. Supposed that the position of an ant is represented by X_i, and LF distribution transforms it to a new state LX_i. Therefore, LF distribution is adopted to construct MALO in the paper, and it is defined as Eq. (6):

$$LX_i = X_i + \alpha \oplus Levy(\lambda) \quad (6)$$

where LX_i denotes the new state of the ant, α is the step size that is related to the scales of the problem, and α is set as $\alpha = 1$ here.

To improve the performance of standard ALO in terms of its optimization ability, the local search is incorporated into ALO by means of utilizing the LF distribution, which is carried out for the current global best ants X_g, and the range around X_g is the most promising area for finding the optimal solution. The basic procedure of the local search is as follows. First, initialize the state with LF distribution according to X_g. Then, determine the value mapping to the solution space for the current iteration and increase the number of ants by using Eq. (6). Finally, compute the fitness value of each ant, and choose the better quality ants for the next iteration.

2.3. Binary coding for discrete optimization problem

On the other hand, feature selection is considered as a discrete optimization problem, which cannot be directly solved by using decimal coding. As for the binary coding form, it converts the population into a probability value for each individual of the binary vector, which forces the elements to take a value of 0 or 1. Thus, each ant in ALO adopts the binary 0–1 coding technique.

In the discrete binary environment, each dimension can be only characterized by 0 or 1. Moving through a dimension means that the corresponding variable changes from 0 to 1 or vice versa. In order to introduce a binary mode for ALO, the updating procedure of each ant may be considered as similar to the continuous algorithm. The main difference between the standard and binary coded ALO is that the updating of ants means switching between "0" and "1" in the binary algorithm. In other words, the coding considers the new ants' position to be either 0 or 1 with the given probability, which is updated using a condition as shown in Eq. (7).

$$R_i^t = \begin{cases} 1 & \text{if } rand < |\tanh(LX_i^t)| \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

where tanh() is the hyperbolic tangent function, and R_i^t is the binary coding form of ants' position.

2.4. Elitism with crossover operation

Elitism is an important characteristic of swarm intelligence algorithms, which allows them to maintain the optimal solutions obtained at any stage of the optimization process, but the operation is based on an addition operation that is not adapted to the binary coding form. Crossover takes more than one parent solution and produces a child solution from the whole population. It is an operation between the two binary solutions obtained from random walk [43]. The optimal antlion in each iteration is saved as the elite. Since the elite is considered to be the fittest antlion, it should be able to affect the movements of all the ants during the iterations. Therefore, it is assumed that the ants simultaneously walk around the elite and the antlion according to the roulette wheel as follows:

$$Ant_i^t = Crossover(R_A^t, R_E^t) \quad (8)$$

where R_A^t is the random walk around the antlion selected by the roulette wheel at iteration t, and R_E^t is the random walk around the elite at iteration t.

To summarize, a novel MALO is proposed in this paper by combining LF distribution, binary coding, and crossover operation to solve the problem of feature selection and obtain the optimal band combination for HSI classification. The current global best ants are randomly distributed in the local space by using Eq. (6), each ant is binary coded and the position is updated by using Eq. (7), and elitism with the crossover operation occurs by using Eq. (8). The specific operation process of the proposed technique will be introduced in the following.

3. The basic principle of WSVM

SVM is a machine learning approach based on statistical theory, which finds the optimal solution of the classification results with limited information about a small sample dataset. Different from other machine learning approaches, SVM uses the kernel function to transform a non-linear problem into a linear problem,

and reduce the complexity of the mapping [44]. Any function that satisfies Mercer's theorem can be used as a kernel function to compute the inner product in feature space. In this study, an admissible wavelet kernel is constructed, and then the architecture of wavelet SVM (WSVM) is proposed by combining the wavelet technique with SVM.

3.1. The basic theory of SVM

Consider a classification problem consisting of n instance-label pairs, $S = (x_i, y_i), (i = 1, 2, \ldots, n), x_i \in R$ is an instance vector and $y_i \in \{-1, +1\}$ is a class label. Classifier training finds a hyper-plane that separates the positive $(+1)$ samples from the negative (-1) samples. The training process involves the optimization of the following expression:

$$\phi(\omega, \varepsilon) = \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{n}\varepsilon_i \tag{9}$$

This expression is subject to the following constraints:

$$y_i[(\omega \cdot x_i) + b] \geq 1 - \varepsilon_i, i = 1, 2, \ldots, n \tag{10}$$

where ω is the normal vector of the hyper-plane, $\varepsilon_i \geq 0$ is the slack variable for measuring the classification errors, C is a positive constant or penalty factor for the error term $\sum_{i=1}^{n}\varepsilon_i$, and ϕ is a function that maps the input space to a higher dimensional feature space. The transformation of space is actually a transformation of the linearly non-separable problem to the easier linearly-separable problem in higher dimensions, and the transformation relies on the kernel function for SVM, which is defined as Eq. (11):

$$K(x, x') = \phi(x)^T \cdot \phi(x') \tag{11}$$

In the practical application of SVM, the most widely adopted kernel function is the radial basis function (RBF) kernel function, which is defined as:

$$K(x, x') = exp(\frac{-\|x - x'\|^2}{\sigma^2}) \tag{12}$$

However, for RBF kernel function, it is difficult to decompose the shift-invariant kernel into the dot product form of two identical functions, which makes it difficult to construct a linearly-separable space especially for high dimensional dataset.

3.2. Wavelet kernel function

The kernel that satisfies the Mercer condition is called an admissible support vector (SV) kernel, which ensures the global optimality of the solutions. The SV kernel can either be in the dot product form $K(x, x') = K\langle x, x'\rangle$ or the shift-invariant form $K(x, x') = K(x - x')$ in the feature space. In particular, it is difficult to decompose the shift invariant kernel into the dot product form of two identical functions. The following theorem can be sued in place of Mercer's theorem, since it provides the necessary and sufficient conditions to judge whether the shift-invariant kernel function is an allowed SV kernel [45].

Theorem 1. *The shift-invariant kernel function $K(x, x') = K(x - x')$ is an allowed SV kernel, if and only if the Fourier transform of $K(x)$ satisfies the following:*

$$F[K(\omega)] = (2\pi)^{-N/2}\int_{R^N} exp(-j\langle\omega \cdot x\rangle)K(x)dx \geq 0 \tag{13}$$

where N is the dimensionality of the wavelet kernel.

The basic principle of wavelet transform is to use a linear combination of wavelets to represent an arbitrary function $f(x)$. Supposing that $\phi(x)$ is a one dimensional mother wavelet and according to tensor product theory, the separable dth-dimensional wavelet can be written as follows:

$$\phi_d(x) = \prod_{i=1}^{d}\phi(x_i) \tag{14}$$

The shift-invariant wavelet kernel function can be constructed as follows:

$$\phi_d(x, x') = \prod_{i=1}^{d}\phi(\frac{x_i - x_i'}{\sigma_i}) \tag{15}$$

where $\sigma_i > 0$ is the wavelet scale factor.

The shift-invariant kernel function strictly satisfies the conditions of Theorem 1. However, there are few forms of wavelet function that can satisfy the above conditions. The mother wavelet is given as follows [46]:

$$\phi(x) = \frac{e^{i4\pi x} - e^{i2\pi x}}{i2\pi x} \tag{16}$$

It can be proved that the mother wavelet satisfies the allowed conditions of the shift-invariant kernel. The following shift-invariant kernel function named the Harmonic wavelet kernel, which is constructed from the mother wavelet as an allowed SV kernel and can be written as follows:

$$K(x, x') = \prod_{i=1}^{d}\frac{e^{i4\pi\frac{x_i - x_i'}{\sigma_i}} - e^{i2\pi\frac{x_i - x_i'}{\sigma_i}}}{i2\pi(\frac{x_i - x_i'}{\sigma_i})} \tag{17}$$

Harmonic wavelet kernel has translation orthogonality and also approximates an arbitrary function in the square integral space, such as the classification function. Since the wavelet kernel has the non-linear mapping ability, WSVM classifier has good adaptive and stable classification decision-making abilities.

4. The proposed method

4.1. Spectral entropy quality index

As for the data acquisition process of HSI, there are parts of high correlation bands because of the transformation of testing environment, which do not make any contribution to image classification and will actually reduce the efficiency. Spectral entropy quality (SEQ) index is a no-reference image quality assessment criteria that uses SVM to train an image distortion and quality prediction engine, and map the image to a different local space according to spectral entropy. It is capable of assessing the quality of band information across multiple distortion categories [47].

Due to the strong relationship between spectral entropy values and the degree and type of distortion, the block discrete cosine transform (DCT) coefficient matrix D is also computed on 8×8 blocks. The use of DCT rather than the discrete Fourier transform reduces blocked edge energy in the transform. Therefore, the spectral probability map based on DCT coefficient is expressed as follows:

$$p(i, j) = \frac{D(i, j)^2}{\sum_i\sum_j D(i, j)^2} \tag{18}$$

where $D(i, j)$ is the DCT coefficient matrix for each feature.

Then, the local spectral entropy is defined as follows:

$$E_f = -\sum_i\sum_j p(i, j)log_2 p(i, j) \tag{19}$$

Moreover, the correlation coefficient can intuitively evaluate the correlation with each band, and the main goal for feature selection is to select a subset of independent bands for use in the model's construction. Therefore, an independent band quality criteria is defined as follows:

$$Q = \frac{E_f}{C_r} \tag{20}$$

where E_f is the value of local spectral entropy, and C_r is the correlation coefficient between the current band with the band obtaining the maximum spectral entropy value.

Spectral entropy is an accurate descriptor of bands' energy spectrum, and the correlation coefficient emphasizes the main frequency and main orientations within a local patch. So, it is able to distinguish the quality of each band more clearly.

4.2. The coding scheme of the proposed method

The key issue to applying MALO is the presentation of the problem to be handled, that is how to make a suitable mapping between the problem solution and MALO. As for the problem of feature selection, each band has two candidate states as selected or deselected, which is easy to map using binary code. Here, the coding length is equal to the number of bands. Each bit of MALO is represented by "0" or "1", where "1" indicates that this band will be chosen for classification, and "0" indicates that this band will not be chosen. Supposing that the whole dataset has 10 bands, the coding of MALO is "0100101010". That is, the 2nd, 5th, 7th and 9th bands will be chosen to complete the classification task using WSVM, and other bands will be abandoned. The entire code can simultaneously indicate the solution for the optimal band combination.

4.3. The definition of objective function

In this section, the proposed feature selection technique for HSI datasets is developed to maximize the classification accuracy and minimize the number of redundant bands by using MALO. As one of the optimal classifiers, WSVM is utilized to complete the classification task here. The main procedure of the proposed method will be explained as follows.

In the proposed method, the classification accuracy using SVM is considered as the main factor of the fitness value.

$$Acc = \frac{T_P + T_N}{T_P + T_N + F_P + F_N} \tag{21}$$

In Eq. (21), the meaning of parameters is as below: T_P (True Positive): in case of test sample is positive and it is identified as positive, it is considered as a true positive; T_N (True Negative): in case of test sample is negative and it is identified as negative, it is considered as a true negative. F_P (False Positive), in case of test sample is negative and it is identified as positive, it is considered as a false positive. F_N (False Negative) in case of test sample is positive and it is identified as negative, it is considered as a false negative.

For feature selection based on MALO and WSVM, the classification accuracy is just one significant goal, and how to decrease the number of redundant and independent bands is another concerned goal. The ultimate goal of feature selection is to obtain a higher classification result that uses the fewer number of bands as possible. Thus, the objective function is defined as follows:

$$F(i) = \lambda \cdot Acc(i) + (1 - \lambda) \cdot log_{10} \frac{n_c}{n_s(i)} \tag{22}$$

where $F(i)$ is the fitness value of ith ant, n_c and $n_s(i)$ are respectively the total and selected number of bands, and $Acc(i)$ is the classification accuracy for each ant. λ is a weighting parameter, which is set as $\lambda = 0.9$ here.

4.4. Implementation of the proposed method

The proposed method is easy to be implemented. The main process of the proposed MALO to make feature selection and classification for HSIs is as follows:

Algorithm: Feature selection for HSIs optimized by MALO

Input: Construct the training and testing samples based on the band information of original HSIs, and the iteration number of MALO is t = 0.

Output: The classification accuracy based on the optimal band combination.

Build WSVM classifier model, remove part of high correlation bands by using spectral entropy and correlation coefficient based on Eq. (20);
Generate initial population of MALO and transform it as binary form, which express the band information of HSIs;
while The algorithm does not reach the termination condition **do**
    t = t + 1;
    Make classification by using WSVM classifier, and compute the fitness value of each ant by Eq. (22);
    Trapping in antlion's pits by using Eqs. (4) and (5);
    Random walk of ants by using Eq. (3);
    Determine the new state by the current global best ants according to Eq. (6);
    Change the population to the binary coding form by Eq. (7);
    Elitism with crossover operation by Eq. (8);
    if Current ant becomes fitter than corresponding antlion **then**
        The antlion updates its position to the latest position of the ant;
    end if
end while
return The optimal band combination, and make comparison with other feature selection approaches via the classification accuracy.

5. Experimental results and discussion

The proposed technique is implemented by the language of Matlab 2014b on a personal computer with a 2.30 GHz CPU, 8.00 G RAM under Windows 8 operating system.

5.1. Datasets description

To evaluate the performance of WSVM classifier and feature selection based on MALO, 3 HSI datasets were used in the experiment.

The first dataset was acquired by the NASA EO-1 satellite over the Okavango Delta, Botswana, in 2001. The size of the image is 1476 × 256 pixels with a spatial resolution of 30 m per pixel resolution over a 7.7 km strip in 242 bands covering the 0.4 to 2.5 μm portion of the spectrum in 10 nm windows. The noise bands that cover water absorption features were removed, and the remaining 145 bands are used for experiment. Fig. 1 shows a false colour composition of the image. The class names and corresponding numbers of ground truth observations that were used in the experiments are listed in Table 1.

The second dataset shown in Fig. 2 was acquired on March 23, 1996 at the Kennedy Space Center (KSC), Merritt Island, Florida, USA. The image is formed by 512 × 614 pixels and 224 bands with a spatial resolution of 18 m. The number of bands is reduced to 176

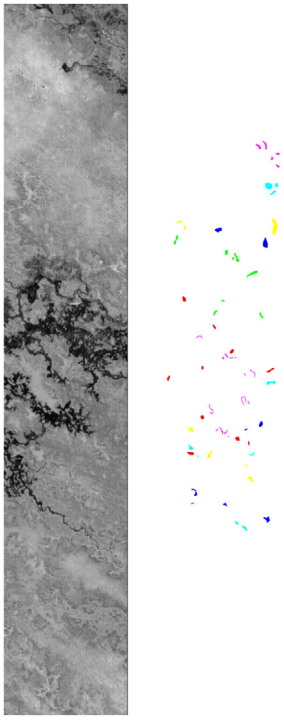

Fig. 1. Botswana HSI and its reference map.

Table 1
The number of available labelled samples and land-cover classes in Botswana dataset.

Class number	Class name	No. of labelled samples
1	Water	270
2	Hippo grass	101
3	Floodplain grasses 1	251
4	Floodplain grasses 2	215
5	Reeds	269
6	Riparian	269
7	Firescar	259
8	Island interior	203
9	Acacia woodlands	314
10	Acacia shrublands	248
11	Acacia grasslands	305
12	Short mopane	181
13	Mixed mopane	268
14	Exposed soils	95
	Total	3248

by removing water absorption and low signal-to-noise bands. The labelled samples were collected using land-cover maps derived from colour infrared photography provided by Landsat thematic mapper imagery. The class names and corresponding numbers of ground truth observations that were used in the experiments are listed in Table 2.

The third dataset was acquired by the AVIRIS sensor over the agricultural land of Indian Pines, Indiana, in the early growing season of 1992. These data were acquired in the spectral range 0.4 to 2.5 μm with spectral resolution of about 10 nm. The image consists of 145 $\times$ 145 pixels and 220 spectral bands with a spatial resolution of 20 m. Twenty water absorption and 15 noisy bands were removed and the remaining 185 bands were included as candidate features. Fig. 3 shows a false colour composition of the AVIRIS Indian Pines scene. The class names and corresponding numbers of ground truth observations that were used in the experiments are listed in Table 3.

To make an impartial comparison, the algorithms will end when the number of function evaluations reaches 300. For MALO, the

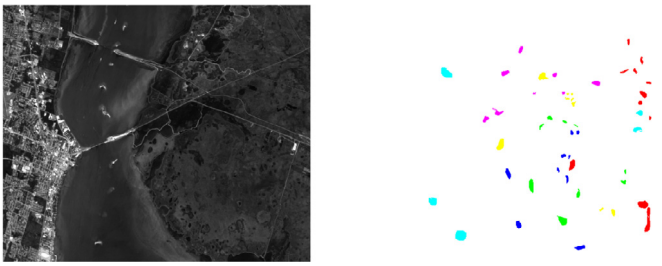

Fig. 2. KSC HSI and its reference map. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2
The number of available labelled samples and land-cover classes in KSC dataset.

Class number	Class name	No. of labelled samples
1	Scrub	761
2	Willow swamp	243
3	Cabbage palm hammock	256
4	Cabbage palm/Oak hammock	252
5	Slash pine	161
6	Oak/Broadleaaf hammock	229
7	Hardwood swamp	105
8	Graminoid marsh	431
9	Spartina marsh	520
10	Cattaial marsh	404
11	Salt marsh	419
12	Mud flats	503
13	Water	27
	Total	5211

Table 3
The number of available labelled samples and land-cover classes in Indian Pines dataset.

Class number	Class name	No. of labelled samples
1	Alfalfa	46
2	Corn-notill	1428
3	Corn-min	830
4	Corn	237
5	Grass/Pasture	483
6	Grass/Trees	730
7	Grass/Pasture-mowed	28
8	Way-windrowed	478
9	Oats	20
10	Soybeans-notill	972
11	Soybeans-min	2455
12	Soybean-clean	593
13	Wheat	205
14	Woods	1265
15	Bldg-Grass-Tree-Drives	386
16	Stone-steel towers	93
	Total	10,249

objective function will run 200 times for the standard ALO and 100 times for LF distribution. In addition, all of the algorithms conduct 30 independent operations. Although the computational complexity is O($nlogn$) for the algorithms above [48], LF distribution will decrease the random number generation and multiplication, which will cost less CPU time than that for MALO. For all HSI datasets, we randomly choose 10% samples of each category as training data, and remaining 90% are selected as testing data. In the paper, the proposed Harmonic wavelet kernel function is utilized as the kernel function of WSVM classifier. Moreover, some other feature selection techniques such as maximum Relevance Minimum Redundancy (mRMR) [49], Conditional Mutual Information Maximization (CMIM) [50], Joint Mutual Information (JMI) methods [51] and Relief algorithm [52] are also used to make comparisons here.

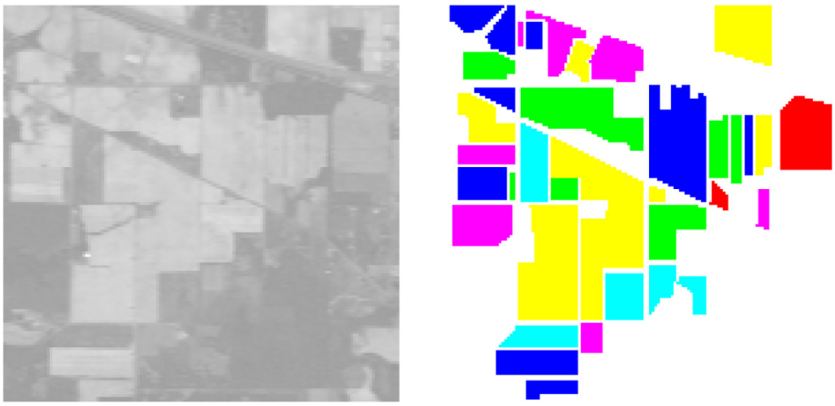

Fig. 3. Indian Pines HSI and its reference map. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 4
Parameter settings of different algorithms.

Parameters	Value
Population size	10
Dimension	Number of bands
Number of runs for each algorithm	30
f_m Mutation factor in DE algorithm	0.6
C_R Crossover rate in DE algorithm	0.9
G_0 Initial of gravitational variable in GSA	100
α User specified constant in GSA	10
p_a Detecting probability in CS algorithm	0.25
β Parameter in MALO and CS algorithm	1.5
a Correlation coefficient in GWO	[2, 0]

Meanwhile, we present some contrastive experimental results, including illustrative examples and performance evaluation tables, which clearly demonstrate the merits of the proposed method. Our primary interest is the optimal band combination, which is reflected by the fitness value of objective function that is defined as Eq. (22), and the Kappa coefficient, where a higher fitness value of objective function indicates a better optimization ability.

5.2. Parameter settings for different algorithms

As the operation process of MALO, the computational results do not depend on any parameter settings, and it avoids trapping into the local optimal to some extent. In addition, some of the commonly used swarm intelligence algorithms based feature selection methods are also assessed in the paper. As is illustrated in Section 2, MALO is utilized here. To make intuitive comparisons, all of the algorithms adopt the binary coding form, and DE algorithm [30], GSA [31], CS algorithm [32], GWO [33] and standard ALO are utilized here. Table 4 displays the parameter settings of these algorithms.

5.3. Experiments for removing high correlation bands

Three original HSI datasets (Botswana, KSC, and Indian Pines) are used in this section to validate the performance when removing high correlation bands. The change curve of the classification accuracy for three datasets is shown in Fig. 4.

As it is shown in Fig. 4, the classification accuracy gradually tends to be stable as the number of bands increases under WSVM, and it is boosted to some extent by using the original datasets. Moreover, the classification accuracy reflects the Hughes phenomenon by first increasing and then decreasing. The main goal of this section is to find the highest classification accuracy, use these bands to build new datasets, and then remove other high correlation bands. For the new datasets, the number of bands is 125, 141 and 153 for Botswana, KSC and Indian Pines datasets respectively, which reduces the data dimension compared with the original datasets. This is especially true for KSC dataset where the number of bands is decreased by 20% with a rather good classification accuracy, which is an effective operation to improve the classification efficiency for HSI datasets.

5.4. Experiments for different swarm intelligence algorithms

Three HSI datasets (Botswana, KSC, Indian Pines) with high correlation bands removed are used in this section to prove the performance of feature selection based on MALO and WSVM. Tables 5–6 show the fitness value, classification accuracy, and selected number of bands handled by different algorithms. In Table 5, Fiv and Acc respectively denote the average fitness value and the classification accuracy after 30 independent operations, and p-value is the correlation between fitness value and classification accuracy. In Table 6, Fn and Time represent the average selected number of bands and CPU time respectively for 30 independent operations in seconds.

According to the data in Tables 5–6, the optimization abilities of GWO and ALO are better than DE, GSA and CS, and the average fitness value exceeded 0.84 for Indian Pines dataset. In addition, GWO is based on the search for prey via the Top 3 solutions, but it requires a number of parameters to be set, and the exploration ability is limited by the initial setting. In ALO, ants simultaneously walk around the antlion and the elite according to the roulette wheel, and they do not depend on any parameters; LF distribution is more probability to enhance the exploitation ability of the current global best ants in the local search. Thus, MALO has the highest fitness value for all datasets, which illustrates that the optimization ability of MALO is the optimal compared with other 5 algorithms. As for the stability, the standard deviation of fitness value is lower than 0.007 for all algorithms, which proves that the swarm intelligence algorithms have a slight fluctuation for HSI datasets. In particular, ALO has the minimum standard deviation of fitness value for Indian Pines dataset, and the difference is only 0.0002, which is negligible for independent operations. More importantly, MALO has the optimal classification accuracy for all algorithms, which is over 0.8% higher than DE algorithm, and the number of properly classified samples is more than 70 for Indian Pines dataset. In addition, the average classification accuracy exceeded 93% for Botswana and KSC datasets, the p-values are lower than 0.05 and even 0.003 for MALO, which proves the significant improvements in the classification accuracy, and wavelet kernel function retains a stable classification accuracy for independent operations. The selected number of bands using MALO is less than those of all other algorithms, the average selected number of bands with DE algorithm is higher than 35 for all datasets, and the proposed

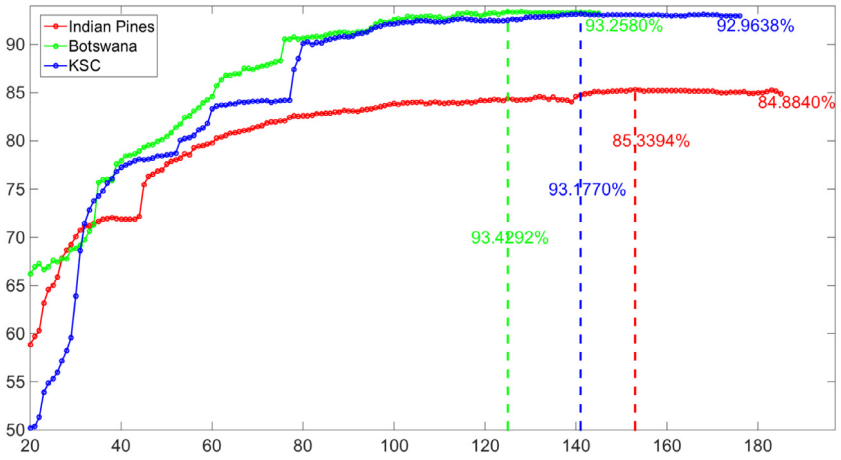

Fig. 4. The change curve of classification accuracy for 3 HSI datasets.

Table 5
The fitness value and classification accuracy of different algorithms.

Dataset	Meas.	DE	GSA	CS	GWO	ALO	MALO
Botswana	Fiv	0.9093 ± 0.0068	0.9115 ± 0.0057	0.9138 ± 0.0049	0.9148 ± 0.0041	0.9167 ± 0.0034	0.9191 ± 0.0025
	Acc(%)	92.72 ± 0.57	92.96 ± 0.48	93.24 ± 0.42	93.28 ± 0.39	93.39 ± 0.38	93.62 ± 0.36
	p-value	0.2205	0.1418	0.0736	0.0251	0.0097	0.0011
KSC	Fiv	0.9053 ± 0.0055	0.9074 ± 0.0039	0.9101 ± 0.0043	0.9119 ± 0.0037	0.9154 ± 0.0032	0.9177 ± 0.0024
	Acc(%)	92.30 ± 0.41	92.43 ± 0.36	92.62 ± 0.39	92.77 ± 0.39	93.00 ± 0.35	93.13 ± 0.32
	p-value	0.1595	0.1076	0.0552	0.0117	0.0076	0.0024
Indian Pines	Fiv	0.8337 ± 0.0029	0.8366 ± 0.0027	0.8399 ± 0.0023	0.8420 ± 0.0021	0.8456 ± 0.0020	0.8479 ± 0.0022
	Acc(%)	84.80 ± 0.37	85.02 ± 0.34	85.19 ± 0.29	85.31 ± 0.31	85.52 ± 0.26	85.65 ± 0.25
	p-value	0.0842	0.0455	0.0273	0.0081	0.0040	0.0005

Table 6
The average selected number of bands and CPU time of different algorithms.

Dataset	Meas.	DE	GSA	CS	GWO	ALO	MALO
Botswana	Fn	35.4333	34.1333	32.7667	31.3333	30.2667	28.5000
	Time	16.3971	14.9724	15.4466	15.5711	14.1462	12.3808
KSC	Fn	40.5333	38.9000	36.8667	35.7000	33.6000	31.7333
	Time	30.7889	28.6607	29.4297	29.7876	27.7137	24.3290
Indian Pines	Fn	54.6667	51.3000	48.3333	45.9667	40.7000	38.3333
	Time	250.1132	235.1242	239.2797	245.8634	229.1429	205.6103

method eliminates more than 70% bands from the whole datasets by removing high correlation bands. For KSC dataset, it only selects 26 bands from the 138 total bands with a rather good classification accuracy, and there are more than 30 bands to be selected by using other commonly used objective functions [31,53]. With regard to the operating efficiency, ALO has a faster convergence speed compared with DE, GSA, CS and GWO, and LF distribution decreases 2 multiplications compared with ALO. Therefore, the CPU time of MALO is 12% less than that of ALO, which generally agrees with the experiment results, and it only needs 11.38 s to select the optimal band combination for Botswana dataset. On the whole, we can conclude that MALO has the optimal optimization ability and operating efficiency, which makes it preferred choice adaptability to solve the feature selection problem.

5.5. Experiments for commonly used feature selection techniques

In this section, some commonly used feature selection approaches such as mRMR, CMIM, JMI methods and Relief algorithm are utilized to validate the effectiveness of the proposed method. Tables 7–9 show the Kappa coefficient and classification accuracy for three HSI datasets using some commonly used feature selection methods. In the tables, OA and Kappa are respectively the overall classification accuracy and Kappa coefficient for each category using different feature selection techniques.

Table 7
The overall classification accuracy and Kappa coefficient for Botswana dataset.

Class number	Full spectral (145)	mRMR (22)	CMIM (22)	JMI (22)	Relief (22)	Proposed (22)
1	100	100	100	100	100	100
2	96.70	79.12	87.91	96.70	96.70	98.90
3	100	96.90	95.13	95.58	97.79	100
4	92.75	48.19	87.05	95.85	93.78	93.78
5	88.43	82.23	85.54	85.95	89.26	89.67
6	78.10	59.92	73.97	73.14	68.60	81.40
7	98.28	96.57	97.85	98.28	98.28	98.28
8	98.91	91.26	97.81	90.16	90.16	98.36
9	91.87	73.14	80.92	80.57	87.28	91.52
10	87.00	72.20	78.48	76.23	79.37	91.03
11	95.26	91.61	91.24	93.07	91.97	93.07
12	91.41	87.73	93.87	92.64	91.41	92.64
13	95.02	80.08	89.21	92.12	93.36	96.27
14	97.65	95.29	97.65	98.82	98.82	97.65
OA(%)	93.26	82.10	89.12	89.70	90.45	93.98
Kappa	0.9270	0.8060	0.8821	0.8884	0.8965	0.9348

According to the data in Tables 7–9, it is observed that the classification accuracy by using the proposed feature selection technique is better than that by using the full spectral information, which is higher than 1% for Indian Pines dataset. In addition, the Kappa coefficient has exceeded 0.93 for Botswana dataset, which

Table 8
The overall classification accuracy and Kappa coefficient for KSC dataset.

Class number	Full spectral (176)	mRMR (26)	CMIM (26)	JMI (26)	Relief (26)	Proposed (26)
1	94.31	96.35	95.77	93.87	96.35	96.50
2	90.87	86.76	86.76	87.21	84.47	89.95
3	89.13	90.87	86.09	85.22	92.17	90.43
4	84.14	41.85	69.60	53.74	41.85	82.82
5	64.83	15.86	37.93	34.48	19.31	60.69
6	66.02	56.31	51.94	45.63	50.97	64.56
7	85.11	89.36	87.23	89.36	87.23	86.17
8	97.42	88.92	93.04	85.57	85.31	97.16
9	95.94	82.91	89.53	86.54	77.14	95.51
10	96.15	89.56	97.25	93.13	78.57	97.53
11	97.61	97.08	97.88	95.76	97.08	97.88
12	95.14	94.48	88.52	89.40	96.47	98.23
13	99.88	100	100	99.88	100	100
OA(%)	92.96	86.65	89.21	86.46	84.90	93.45
Kappa	0.9216	0.8512	0.8798	0.8490	0.8317	0.9271

Table 9
The overall classification accuracy and Kappa coefficient for Indian Pines dataset.

Class number	Full spectral (185)	mRMR (32)	CMIM (32)	JMI (32)	Relief (32)	Proposed (32)
1	29.27	0.00	9.76	31.71	17.07	58.54
2	83.42	60.70	44.36	69.88	69.18	83.74
3	72.02	56.22	30.66	9.91	29.72	69.75
4	71.83	51.17	20.19	31.46	30.99	70.89
5	91.49	19.54	84.60	75.40	83.91	91.03
6	96.50	91.78	94.37	93.00	91.17	97.72
7	68.00	4.00	0.00	4.00	0.00	68.00
8	98.84	96.98	96.05	97.44	99.07	99.53
9	55.56	0.00	0.00	5.56	16.67	55.56
10	78.17	53.26	17.26	0.34	52.69	80.00
11	85.60	75.10	89.54	88.37	81.08	86.33
12	85.96	37.45	19.29	14.04	38.95	89.89
13	94.57	90.76	92.93	90.22	77.17	95.11
14	97.19	97.36	97.36	97.19	96.84	97.72
15	52.16	14.99	37.18	35.73	40.63	60.23
16	88.10	84.52	84.52	96.43	94.05	88.10
OA(%)	84.88	66.56	64.61	64.18	70.49	85.90
Kappa	0.8272	0.6132	0.5812	0.5768	0.6583	0.8388

demonstrates that the precision is applicable to meeting practical demands. More importantly, the dimension or the selected number of bands is greatly reduced, and the number of bands is only 18% compared with the original datasets. As for the performance of feature selection, the classification accuracy of the proposed method provides better results than other approaches involved in this paper. Although the selected number of bands is the same, the overall classification accuracy is lower than 70% for Indian Pines dataset when using mRMR, CMIM and JMI methods; it is only 70.4945% when using Relief algorithm, and there are fewer or even no samples to be correctly classified for the Grass/Pasture-mowed and Oats categories. For Botswana and KSC datasets, the classification accuracy is lower than 90% when using mRMR, CMIM and JMI approaches; it is just 90.4517% when using Relief algorithm, and it is respectively 93.98% and 93.45% when using the proposed method. In short, it is proved that the proposed method is a robust, reliable and efficient feature selection technique for HSI datasets.

6. Conclusion and future directions

In the paper, a feature selection technique for HSIs is proposed. First, parts of high correlation bands are removed, and then a novel MALO is utilized to obtain the optimal band combination. Results are compared with some other feature selection techniques optimized by DE, GSA, CS, GWO and standard ALO. In general, it is observed that the classification accuracy is improved by removing the high correlation bands, and swarm intelligence algorithms can be well used to solve the problem of feature selection. Among these algorithms, MALO has a better performance, and it is able to find the optimal solution quickly and fast enough to meet some real-time applications. In terms of the standard deviation of classification accuracy, the experimental results always remain in a stable interval with wavelet kernel function and have almost no fluctuation. That is, MALO and the newly proposed feature selection evaluation criteria are more appropriate to be employed to reduce the data dimension based on WSVM classifier for HSI datasets. The classification accuracy is obviously higher than the original datasets and some traditional feature selection approaches such as mRMR, CMIM, JMI methods and Relief algorithm. In summation, WSVM has a satisfactory and stable performance for classification in most cases, and the disadvantage of heavy computational burden can be conquered to a great extent when it is combined with MALO. The proposed method is able to keep a good balance on the efficiency and classification accuracy, which makes it more appropriate for practical applications. In the future, it will be interesting to collect some airborne HSIs on a larger scale and conduct classification for each pixel of them. Moreover, as the data dimension increases, it is necessary to combine it with other quality assessment criteria for feature selection.

Acknowledgements

This work is funded the National Key Research & Development Program of China under Grant No. 2017YFC1404700, 2017YFC1502406-03, the National Natural Science Foundation of China under Grant No. U1711266, and the Fundamental Research Funds for the Central Universities under Grant No. CUG2017JM06.

References

[1] L. Wang, C. Zhao, Hyperspectral Image Processing, Springer, 2016.
[2] S.D. Backer, P. Kempeneers, W. Debruyn, P. Scheunders, A band selection technique for spectral classification, IEEE Geosci. Remote Sens. Lett. 2 (3) (2005) 319–323.
[3] C. Tang, X. Liu, M. Li, P. Wang, J. Chen, L. Wang, W. Li, Robust unsupervised feature selection via dual self-representation and manifold regularization, Knowl.-Based Syst. 145 (2018) 109–120.
[4] G. Chandrashekar, F. Sahin, A Survey on Feature Selection Methods, Pergamon Press, Inc., 2014.
[5] V. Bolon-Canedo, N. Scnchez-Marono, A. Alonso-Betanzos, Recent advances and emerging challenges of feature selection in the context of big data, Knowl.-Based Syst. 86 (2015) 33–45.
[6] M.M. Kabir, M.M. Islam, K. Murase, A new wrapper feature selection approach using neural network, Neurocomputing 73 (16) (2008) 3273–3283.
[7] A.P. Castaño, Support Vector Machines, Springer Science & Business Media, 2018.
[8] Z. Ye, M. Wang, C. Wang, H. Xu, P2P traffic identification using support vector machine and cuckoo search algorithm combined with particle swarm optimization algorithm, in: Frontiers in Internet Technologies, Springer, 2014, pp. 118–132.
[9] R. Sali, H. Shavandi, M. Sadeghi, A clinical decision support system based on support vector machine and binary particle swarm optimisation for cardiovascular disease diagnosis. international, Int. J. Data Min. Bioinform. 15 (4) (2016) 312–327.
[10] F. Kaytez, M.C. Taplamacioglu, E. Cam, F. Hardalac, Forecasting electricity consumption: A comparison of regression analysis, neural networks and least squares support vector machines, Int. J. Electr. Power Energy Syst. 67 (2015) 431–438.
[11] F. Melgani, L. Bruzzone, Classification of hyperspectral remote sensing images with support vector machines, IEEE Trans. Geosci. Remote Sens. 42 (8) (2004) 1778–1790.
[12] B.C. Kuo, H.H. Ho, C.H. Li, C.C. Hung, J.S. Taur, A kernel-based feature selection method for SVM With RBF kernel for hyperspectral image classification, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 7 (1) (2014) 317–326.
[13] C. Bo, H. Lu, D. Wang, Hyperspectral image classification via JCR and SVM models with decision fusion, IEEE Geosci. Remote Sens. Lett. 13 (2) (2016) 177–181.
[14] L. Yang, S. Yang, S. Li, R. Zhang, F. Liu, L. Jiao, Coupled compressed sensing inspired sparse spatial-spectral LSSVM for hyperspectral image classification, Knowl.-Based Syst. 79 (2015) 80–89.

[15] B. Scholkopf, A. Smola, Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond, MIT press, 2001.
[16] C. Lai, M.J.T. Reinders, L. Wessels, Random subspace method for multivariate feature selection, Pattern Recognit. Lett. 27 (10) (2006) 1067–1076.
[17] E.G. Talbi, Metaheuristics: From Design to Implementation, Wiley Online Library, 2009.
[18] H. Faris, A.-Z. AlaM, A.A. Heidari, I. Aljarah, M. Mafarja, M.A. Hassonah, H. Fujita, An intelligent system for spam detection and identification of the most relevant features based on evolutionary random weight networks, Inf. Fusion 48 (2019) 67–83.
[19] Y. Zhang, S. Wang, P. Phillips, G. Ji, Binary PSO with mutation operator for feature selection using decision tree applied to spam detection, Knowl.-Based Syst. 64 (1) (2014) 22–31.
[20] J. Li, A combination of de and svm with feature selection for road icing forecast, in: International Asia Conference on Informatics in Control, Automation and Robotics, Remote Sensing, 2010, pp. 509–512.
[21] J. Xiang, X.H. Han, F. Duan, Y. Qiang, X.Y. Xiong, Y. Lan, H. Chai, A novel hybrid system for feature selection based on an improved gravitational search algorithm and k-NN method, Appl. Soft Comput. 31 (2015) 293–307.
[22] M.A.E. Aziz, A.E. Hassanien, Modified cuckoo search algorithm with rough sets for feature selection, Neural Comput. Appl. 29 (4) (2018) 925–934.
[23] E. Emary, H.M. Zawbaa, A.E. Hassanien, Binary grey wolf optimization approaches for feature selection, Neurocomputing 172 (2016) 371–381.
[24] M. Mafarja, I. Aljarah, A.A. Heidari, A.I. Hammouri, H. Faris, A. Al-Zoubi, S. Mirjalili, Evolutionary population dynamics and grasshopper optimization approaches for feature selection problems, Knowl.-Based Syst. 145 (2018) 25–45.
[25] H. Faris, M.M. Mafarja, A.A. Heidari, I. Aljarah, A. Al-Zoubi, S. Mirjalili, H. Fujita, An efficient binary salp swarm algorithm with crossover scheme for feature selection problems, Knowl.-Based Syst. 154 (2018) 43–67.
[26] I. Aljarah, M. Mafarja, A.A. Heidari, H. Faris, Y. Zhang, S. Mirjalili, Asynchronous accelerating multi-leader salp chains for feature selection, Appl. Soft Comput. 71 (2018) 964–979.
[27] M. Mafarja, I. Aljarah, A.A. Heidari, H. Faris, P. Fournier-Viger, X. Li, S. Mirjalili, Binary dragonfly optimization for feature selection using time-varying transfer functions, Knowl.-Based Syst. 161 (2018) 185–204.
[28] L. Wang, Z. Liang, D. Liu, Artificial bee colony algorithm-based band selection for hyperspectral imagery, J. Harbin Inst. Tech. 47 (11) (2015) 82–88.
[29] H. Su, Q. Du, G. Chen, P. Du, Optimized hyperspectral band selection using particle swarm optimization, IEEE J. Sel. Top. Appl. Earth Obs Remote Sens. 7 (6) (2014) 2659–2670.
[30] A. Datta, S. Ghosh, A. Ghosh, Self-adaptive differential evolution for feature selection in hyperspectral image data, Appl. Soft Comput. 13 (4) (2013) 1969–1977.
[31] M. Wang, Y. Wan, Z. Ye, X. Gao, X. Lai, A band selection method for airborne hyperspectral image based on chaotic binary coded gravitational search algorithm, Neurocomputing 273 (2018) 57–67.
[32] S.A. Medjahed, T.A. Saadi, A. Benyettou, M. Ouali, Binary cuckoo search algorithm for band selection in hyperspectral image classification, IAENG Int. J. Comput. Sci. 42 (3) (2015) 183–191.
[33] F. Xie, F. Li, C. Lei, L. Ke, Representative band selection for hyperspectral image classification, ISPRS Int. J. Geo-Inf. 7 (2018) 338.
[34] S. Mirjalili, The ant lion optimizer, Adv. Eng. Softw. 83 (2015) 80–98.
[35] M. Raju, L.C. Saikia, N. Sinha, Automatic generation control of a multi-area system using ant lion optimizer algorithm based PID plus second order derivative controller, Int. J. Electr. Power Energy Syst. 80 (2016) 52–63.
[36] V.K. Kamboj, A. Bhadoria, S.K. Bath, Solution of non-convex economic load dispatch problem for small-scale power systems using ant lion optimizer, Neural Comput. Appl. 28 (8) (2017) 2181–2192.
[37] P. Yao, H. Wang, Dynamic adaptive ant lion optimizer applied to route planning for unmanned aerial vehicle, Soft Comput. 21 (18) (2017) 5475–5488.
[38] M.M. Mafarja, S. Mirjalili, Hybrid binary ant lion optimizer with rough set and approximate entropy reducts for feature selection, Soft Comput. 1 (2018) 1–17.
[39] E. Emary, H.M. Zawbaa, Feature selection via Lévy Antlion optimization, PAA Pattern Anal. Appl. 3 (2018) 1–20.
[40] P. Barthelemy, J. Bertolotti, D.S. Wiersma, A Lévy flight for light., Nature 453 (7194) (2008) 495–498.
[41] H. Hakli, H. Uğuz, A novel particle swarm optimization algorithm with Lévy flight, Appl. Soft Comput. 23 (2014) 333–345.
[42] A.A. Heidari, P. Pahlavani, An efficient modified grey wolf optimizer with Lévy flight for optimization tasks, Appl. Soft Comput. 60 (2017) 115–134.
[43] A. Heinzel, V.M. Barragøn, A review of the state-of-the-art of the methanol crossover in direct methanol fuel cells, J. Power Sources 84 (1) (1999) 70–74.
[44] B.E. Boser, I.M. Guyon, V.N. Vapnik, A training algorithm for optimal margin classifiers, in: The Workshop on Computational Learning Theory, 1992, pp. 144–152.
[45] A.J. Smola, B. Scholkopf, K.R. Muller, The connection between regularization operators and support vector kernels, Neural Netw. 11 (4) (1998) 637–649.
[46] Y. Zhou, T.S. Hou, Detection of line spectrum signal detection based on harmonic wavelet kernel-support vector regression algorithm, Comput. Simul. 30 (1) (2013) 263–267.
[47] L. Liu, B. Liu, H. Huang, A.C. Bovik, No-reference image quality assessment based on spatial and spectral entropies, Signal Process. 29 (8) (2014) 856–863.
[48] D. Sudholt, C. Witt, Runtime analysis of a binary particle swarm optimizer, Theoret. Comput. Sci. 411 (21) (2010) 2084–2100.
[49] J. Feng, L. Jiao, F. Liu, T. Sun, X. Zhang, Mutual-information-based semi-supervised hyperspectral band selection with high discrimination, high information, and low redundancy, IEEE Trans. Geosci. Remote Sens. 53 (5) (2015) 2956–2969.
[50] E. Sarhrouni, A. Hammouch, D. Aboutajdine, Band selection and classification of hyperspectral images using mutual information: An algorithm based on minimizing the error probability using the inequality of Fano, in: Multimedia Computing and Systems, ICMCS, 2012 International Conference on, 2012, pp. 155–159.
[51] M. Bennasar, Y. Hicks, R. Setchi, Feature selection using joint mutual information maximisation, Expert Syst. Appl. 42 (22) (2015) 8520–8532.
[52] J. Jia, N. Yang, C. Zhang, A. Yue, J. Yang, D. Zhu, Object-oriented feature selection of high spatial resolution images using an improved relief algorithm, Math. Comput. Modelling 58 (3–4) (2013) 619–626.
[53] K. Mistry, L. Zhang, S.C. Neoh, C.P. Lim, B. Fielding, A micro-GA embedded PSO feature selection approach to intelligent facial emotion recognition, IEEE Trans. Cybern. 47 (6) (2017) 1496–1509.