A multiple classifier system based on Ant-Colony Optimization for Hyperspectral image classification

Ke Tang, Li Xie and Guangyao Li

No 4800, Cao'an Road, Tongji Universirt, Shanghai, China

E-mail: ke_tang1991@163.com

Abstract. Hyperspectral images which hold a large quantity of land information enables image classification. Traditional classification methods usually works on multispectral images. However, the high dimensionality in feature space influences the accuracy while using these classification algorithms, such as statistical classifiers or decision trees. This paper proposes a multiple classifier system (MCS) based on ant colony optimization (ACO) algorithm to improve the classification ability. ACO method has been implemented on multispectral images in researches, but seldom to hyperspectral images. In order to overcome the limitation of ACO method on dealing with high dimensionality, MCS is introduced to combine the outputs of each single ACO classifier based on the credibility of rules. Mutual information is applied to discretizing features from the data set and provides the criterion of band selection and band grouping algorithms. The performance of the proposed method is validated with ROSIS Pavia data set, and compared to k-nearest neighbour (KNN) algorithm. Experimental results prove that the proposed method is feasible to classify hyperspectral images.

1. Introduction

With the development of remote-sensing image technology, high dimensional data are easily accessed, meaning hyperspectral images are available for characterization, identification, and classification of the land-covers with improved accuracy and efficiency[1]. The information contained in hyperspectral data is performed in more than hundreds of spectral bands, which provide the possibility of distinguishing more classes of the ground type.

A large number of methods have been applied to classify multispectral images over the past decades. Statistical method such as maximum likely hood classifier and Bayesian classifier count on the assumption that the number of each class preforms a normal distribution in the feature space. Decision trees was firstly introduced in remote sensing images in [2,3,4], and provide an acceptable accuracy. Ant colony optimization (ACO) was firstly proposed to mine classification rules by [4]. ACO method is inspired by natural biological systems[5]. By simulating the behavior of ants in their process of searching food, ACO methods can utilize the feedback system of the whole ant colony. The interactional information exchange of ants, which is exactly pheromone system, helps ants learn from past experiences. One advantage of ACO are the easy-understanding rules since the classification rule produced by this induction algorithm is more explicit than mathematical equations. ACO has been demonstrated to be effective and robust when working on multispectral images[6,7].

Unfortunately, these methods above can not be implemented on hyperspectral images directly. Because of the high dimensionality in feature area, the complexity of both traditional classification

Published under licence by IOP Publishing Ltd

method and ACO method will dramatically increase which resulting in the curse of dimensionality. [8,9,10] indicated it's beneficial to remove bands with little or no discriminatory information. However, dimensional reduction brings about loss in information of hyperspectral images. Multiple classifier systems(MCS) can overcome this problem. MCS combines same classifiers' results or different classification algorithms[11]. With the help of this system, high dimensionality feature space can be divided into s set of low dimensionality data sets. After a single classifier is produced, MCS combine these classifiers with a specific fusion method. Multiple classifier system can be utilized to improve classification accuracy in hyperspectral images.

In this paper, a new hyperspectral image classification method is developed in this paper. Firstly, a discretization algorithm based on mutual information is implemented to divide each band into several part, which can be regarded as a feature. Then the feature space are split into a set of band groups to overcome high dimensionality problem. Band selection and band grouping method are introduced to reduce the dimensionality. Next ACO method is applied to each group to find rule sets. Finally use fusion method to combine the output of each classifier to produce a classification decision. This paper is composed of four sections. The related algorithms are introduced in Section 2. Section 3 describes the data set of remote sensing image and the designed experiment. The experimental results and discussion are presented in Section 4. Section 5 concludes the paper. Figure 1 presents the flowchart of the proposed system.

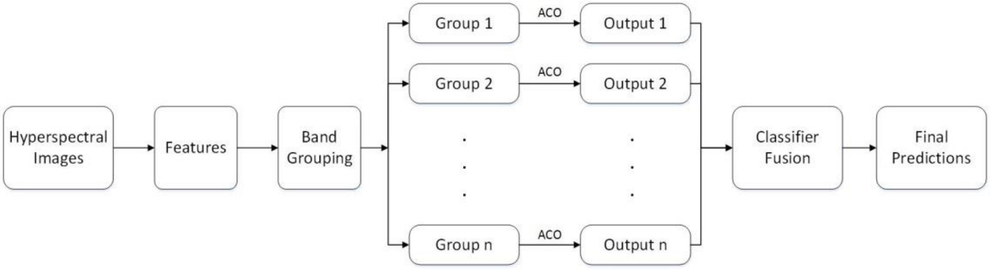

Figure 1. The multiple classifier system based on ACO method

2. Methodology

Ant colony optimization algorithm is innovated by observing food seeking behaviors of ants. When ants are seeking food, they will release pheromone on their route. Ants tend to choose a path that has the largest pheromone. Since pheromone's disappearing with time, the shortest path will collect more pheromone. This process provides a positive feedback, in which the probability of an ant choosing a path is proportional to the number of ants that have passed through that path[7].

First, the hyperspectral remote sensing image pixels are divided into a training data set and a testing data set. Then it comes to data discretization, which is employed to segment the brightness value of each band[7][12]. Since there are more than hundred bands in hyperspectral image and ant colony algorithm can not adapt to high dimensionality, original spectral space can not use this method directly. Besides, not all of the bands are useful to classification, some bands refer to classes those are not interested. This paper proposes a band grouping method based on entropy information to select effective bands and group the remaining bands. After band grouping, ant mining method are used to find the best rules covering the training samples. Then use credibility to describe a rule's confidence degree on classification outputs. Finally, a fusion method based on rules' credibility is applied to establish a multiple classifier system.

2.1. Data discretization

Since the brightness value of each band of a specific pixel in a hyperspectral image ranges from 0 to 65536, conventional classification method will fail when dealing with these continuous values. Continuous attributes must be separated into multiple sections to adapt the decision rules. This will influence the efficiency of the ant miner algorithm and quality of rules. Before a further multi system classifier, a data discretization algorithm is adopted based on information entropy[7].

There are four symbols as (U, R, V, F) to describe the image information. U represents the set of objects, namely the image pixels. R stands for the set of attributes, and $R = C \cup D$, where C and D refer to a condition attribute set and a decision attribute set, respectively. On this condition, C and D refer to the bands of hyperspectral image and the type of lands, respectively. V denotes the domain of the range of value of each band and the value of each class. F is the information function[7].

For a certain condition attribute(a band) $c \in C$, assume its domain is $V_a = [l_a, r_a]$. If there exists a group of points: $l_a < c_1^a < c_2^a < \cdots < c_{m_a}^a < r_a$, then $V_a = [l_a, c_1^a) \cup [c_1^a, c_2^a) \cdots [c_{m_a-1}^a, c_{m_a}^a) \cup [c_{m_a}^a, r_a)$. Attribution a is separated as $m_a + 1$ equivalent classes. c_k^a is called a breakpoint. The goal of discretization is to find a set of breakpoints to represent the data.

In this paper, information entropy is introduced to evaluate the quality of discretization. Supposing $\mathrm{X} \subseteq \mathrm{U}$ as a subset, $|\mathrm{X}|$ is the number of object, k_j is the number of object whose decision attribute is $\mathrm{j}(\mathrm{j} = 1,2,\cdots,\mathrm{r(d)})$, then the information entropy of this subset is defined as:

$$H(x) = -\sum_{j=1}^{r(d)} p_j log_2 p_j \, , p_j = \frac{k_j}{|X|} \tag{1}$$

X is separated into two sections by a breakpoint b_i^a. The number of sample belonging to X and its value is smaller than value of breakpoint b_i^a is $l_j^X(c_i^a)$, $r_j^X(c_i^a)$ means the number of samples whose value are greater than b_i^a:

$$l^X(b_i^a) = \sum_{j=1}^{r(d)} r_j^X(b_i^a) \tag{2}$$

$$r^X(b_i^a) = \sum_{j=1}^{r(d)} r_j^X(b_i^a) \tag{3}$$

Breakpoint c_i^a divide X into two subsets X_l and X_r, and

$$H(X_l) = -\sum_{j=1}^{r(d)} p_j log_2 p_j \, , p_j = \frac{l_j^X(b_i^a)}{l^X(b_i^a)} \tag{4}$$

$$H(X_r) = -\sum_{j=1}^{r(d)} p_j log_2 p_j \, , p_j = \frac{r_j^X(b_i^a)}{r^X(b_i^a)} \tag{5}$$

For a breakpoint b_i^a, information entropy of X is

$$H^X(b_i^a) = \frac{|X_l|}{|U|} H(X_l) + \frac{|X_r|}{|U|} H(X_r) \tag{6}$$

Supposing S is the selected breakpoint set and B is the set of candidate breakpoints. $S = \{X_1, X_2, \cdots, X_m\}$ is considered as the equivalent class of set X. The new information entropy after adding breakpoint $b \notin S$,

$$H(b, L) = H^{X_1}(b) + H^{X_2}(b) + \cdots + H^{X_m}(b) \tag{7}$$

With adding new breakpoints, the decision attribute becomes simplification if $H(b, S)$ decreases.

Let P be the selected breakpoint set, L is the equivalent class of set X, B is the set of candidate breakpoints set. H is the information entropy. The discretization algorithm can be concluded as follows:

1) $P = \phi; S = U; H = H(U)$;

2) For each $b \in B$, calculate $H(b, L)$;

3) If $H \leq minH(b, L)$, terminate the loop;

4) Select breakpoint b_{min}, which contributes the minimum value of $H(b, L)$; $H = H(b, L)$; $B = B - b$;

5) For every $X \in L$, if b_{min} divides equivalent class X into X_1 and X_2, then remove X from S, put equivalent X_1 and X_2 into S;

6) If every sample in equivalent class has the same decision attribute, then end the loop. Turn to step 2 otherwise.

2.2. Band grouping

To split original data space to small groups, information entropy is introduced to evaluate quality of bands. Mutual information can be used to measure the similarity between two random variables[13].

Considering two random variables X and Y, $p(X)$ and $p(Y)$ are marginal probability distributions of X and Y, respectively. $p(X,Y)$ is the joint probability distribution. Then MI is defined as follow:

$$MI(X,Y) = \sum_{X\epsilon\boldsymbol{X},Y\in\boldsymbol{Y}} log\frac{p(X,Y)}{p(X)p(Y)} \tag{8}$$

X represents one of candidate bands. Y represents class type in decision attributes. MI can be used to measure the dependency between spectral image and ground type. Higher MI value means more dependency. Such bands with higher MI value have better ability in classification. Fig.2 shows the MI value for all bands.

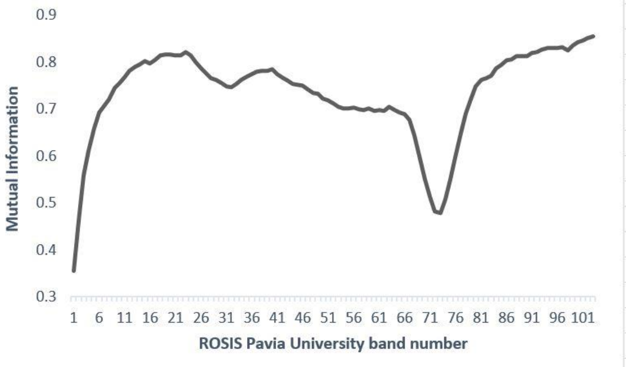

Figure 2. Mutual Information value

Bands with low MI value (lower than a threshold) should be excluded from the condition attributes since these bands hold less information than others. Besides, there are more redundancies between two adjacent bands if the MI is relatively high[9]. Bands with high correlation have weak ability in classification. These two bands need to be separated into different groups. Thus the two steps in this band grouping process can be described as follows:

2.3. Ant Miner

2.3.1. Rule searching. After discretization and band grouping, hyperspectral image are divide into several groups, each group contains several bands. Ant miner algorithm is used to search for classification rules in each group. Rule construction process is similar to the behaviour of ant's seeking food. Ant chooses nodes attributes by attributes until an entire path is constructed. Theoretically, nodes are selected randomly, but this process may take an insufferable time. A heuristic method is designed to help ants find the path in order to shorten the searching time. [7] defines the value of heuristic function of one condition attribute $term_{ij}$ according to statistical data as:

$$\eta_i^j = \frac{\max(\sum freqT_{ij}^1, \sum freqT_{ij}^2, \cdots, \sum freqT_{ij}^k)}{\sum T_{ij}} \tag{9}$$

where T_{ij} is the number of samples which match condition attribute $term_{ij}$, $freqT_{ij}^w$ is the frequency number of class w in T_{ij}. In the process of data mining, data that satisfies rules should be removed from the original data set, which will leading to the change of $\max(\sum freqT_{ij}^1, \sum freqT_{ij}^2, \cdots, \sum freqT_{ij}^k)$ and T_{ij}. Therefore, η needs to be updated after getting a final rule. From the beginning, the information density of each path nodes is initialized as the same value:

$$\tau_{ij}(t=0) = \frac{1}{\sum_{i=1}^{a} b_i} \tag{10}$$

where τ_{ij} is the information density of $term_{ij}$, a is the total condition attributes number, b_i is the potential value that attribute i may hold.

Each node in one attribute are selected out on the probability calculated by following formula:

$$P_{ij} = \frac{\tau_{ij}(t)*\eta_{ij}(t)}{\sum_{i=1}^{a}\sum_{j=1}^{b}\tau_{ij}(t)*\eta_{ij}(t)} \tag{11}$$

Quality of a rule can be defined as follow:

$$Q = \left(\frac{TruePos}{TruePos+FalseNeg}\right) * \left(\frac{TrueNeg}{FalsePos+TrueNeg}\right) \quad (12)$$

where $TruePos$ represents number of samples which entirely match the selecting rule, both condition attributes and decision attributes, $FalsePos$ represents number of samples which satisfy rule's condition attributes, while stand against predict class, $FalseNeg$ represents number of samples which don't match rule's condition attributes while are the same class as prediction, $TrueNeg$ are the number of samples which don't match rule either condition attributes or decision attributes[7].

To provide a basis to the fusion system, credibility of a rule can be defined as follows:

$$C = \left(\frac{TruePos}{TruePos+FalseNeg}\right) \quad (13)$$

2.3.2. Pheromone updating. The next step is updating pheromone. In each ant's searching period, once a complete rule is constructed, the pheromone of nodes which are inside the path should increase, while the others' should decrease. The updating degree can be defined as follows:

$$\tau_{ij}(t+1) = (1-\rho) * \tau_{ij}(t) + \frac{Q}{1+Q} * \tau_{ij}(t) \quad (14)$$

where ρ is pheromone evaporation coefficient, and Q is the quality of a final rule. In every iterations, numbers of ants will find numbers of rules, the rule with highest qualification can be seen as the candidate of final rule, and the other rules are dropped out. If the rule can match specific number of samples, that rule can be added to the final rule sets, otherwise should be dropped out either. Record the credibility of rule. Then it comes to the next iteration, until the number of remaining samples is less than a predetermined value.

2.4. Multi classifier.

The final step is to make a prediction by the rules. [14] utilizes the classifier and confidence measurement to combine prediction results. First, use each classifier to output a unique class so that a vector of classes is produced for each sample. Next, associate a confidence measurement for each class and produces a vector for every classifier and a matrix for ensemble of classifier.

However, this method doesn’t consider the differences of abilities of each rule in classification. Besides, there is a possibility that a sample is in accord with multiple rules resulting in a difficulty in making prediction. A multi classifier system based on the credibility of rules is proposed. For each testing sample, testify whether it satisfies the condition attributes of rules in each group of classifier. If so, add up the credibility of the relative rule to the class according to the rule’s decision attribute. Finally, the vector of prediction for one sample P is calculated out and forms as follows:

$$P_i = (C_1, C_2, \cdots, C_n), 1 < i < ||test\_size|| \quad (15)$$

where C_i is the sum of credibility of one class, n is the total number of class. Choose the class which has the maximum possibility value from P_i as the final prediction.

3. Experiments

3.1. Data sets

The method proposed in this paper is applied to ROSIS, a well-known hyperspectral data sets. The first data set is acquired by the ROSIS sensor during a flight campaign over Pavia, which is made of 103 spectral bands, covering the wavelength range from 0.43 to 0.86 μm. The geometric resolution is 1.3 meters. This data set holds 610*340 pixels, while some of the samples contain no information and have to be discarded. Pavia University is one of the two scenes in the data set, whose ground-truths differentiate 9 classes. In this paper, totally 42776 pixels were selected out to be made up of the data set. Figure 3 shows the origin view of Pavia University. The types and number of samples are represented in Table 1.

Figure 3. Pavia University

Table 1. Ground-truth of Pavia University data set

#	Class	Samples
1	Asphalt	6631
2	Meadows	18649
3	Gravel	2099
4	Trees	3064
5	Painted metal sheets	1345
6	Bare Soil	5029
7	Bitumen	1330
8	Self-Blocking Bricks	3682
9	Shadows	947

3.2. Results and discussion

In this section, two groups of experiments designed to evaluate the effectiveness of the proposed method. The classification accuracy of the proposed method was evaluated by several experiments in the first part. The purpose of this experiment was to study the effects of discretization iteration and max number of bands when grouping bands. In the second experiment, KNN (k-nearest neighbour) algorithm was a to make a comparison to the proposed method, setting training set size as the control variable.

There are a few parameters in this system: (1) max_bands_per_group: This decides a single classifier's dimensionalit. This amount of condition attributes plus the decision attributes to make up of a sub-training set. (2) discretization_iteration: iteration times of discretization, equalling to the number of intervals in one band. (3) max_iteration: ACO method terminates when the number of iterations exceeds this number. (4) min_cover_per_rule: Abandon one rule if the number of samples covered by this rule is less than the threshold. (5) max_uncovered_sample: ACO method terminates when the remaining training samples is less than this threshold. (6) Min_MI: Bands with MI value lower than this number should be excluded. The default parameters setting are as follows: $max\_iteration = 1000, \min\_cover\_per\_rule = 2, \max\_uncovered\_sample = 20, Min\_MI = 0.55$.

The relationship between classification accuracy and discretization iteration was shown in Figure 4. The classification accuracies are always above 81%. It rises when the iteration increases from 5 to around 12. The highest accuracy is 84.3%. However, there are continuous decreases when iterations keep going up. The reason for this drop is shown in Figure 5. Figure 5 represents the cover rate representing the percentage of covering sample size of training size for each generated rule sets. The cover rate suffers a sharp plunge after the 11th discretization iteration. Lower cover rate indicates that the generated rules can not represent the training set effectively, which finally has an influence on overall accuracy.

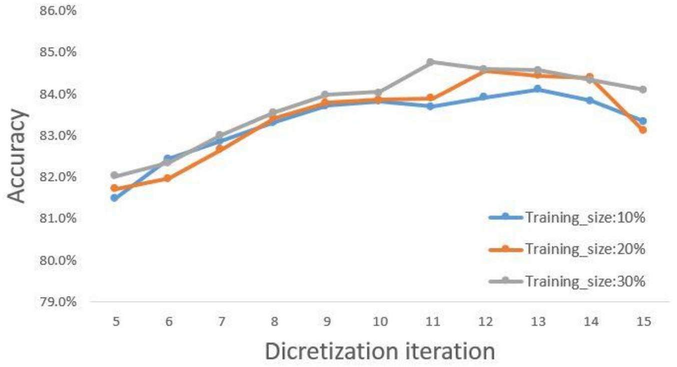

Figure 4. Influence of discretization iteration on accuracy

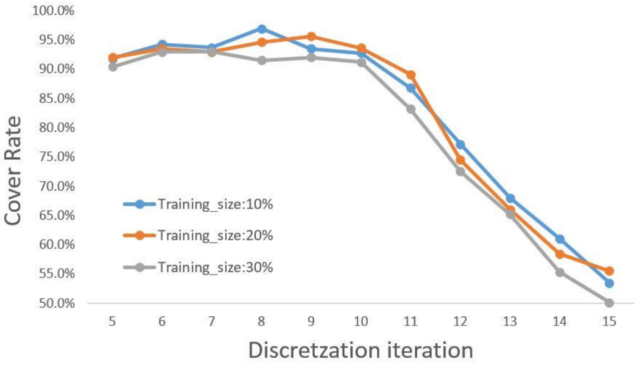

Figure 5. Cover rate per classifier in training set

Figure 6 shows the difference between with or without band selection method before grouping all the bands. This figure illustrates the effectiveness of the proposed band selection method, which will make progress on band grouping and final classification performance. There are improvements on classification accuracy with bands selection.

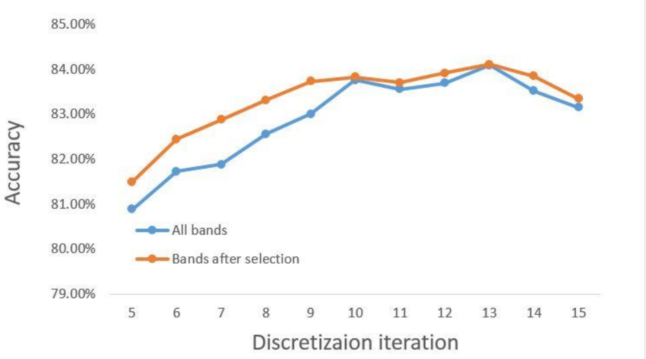

Figure 6. Differences of band grouping on all bands and selected bands

The maximum number of bands in a separate group contributes to classification accuracy and training speed. Figure 7 illustrates the classification ability when setting different size of bands in a group. The classification accuracy will decline with the number of bands in each group increases from 3 to 5. It is reasonable that the number of bands in each group should be more than 2. If the number of bands in each group is too small, the rules are not so believable because it may not be enough to completely represent the sample. On the other hand, too many bands will result in the difficulty on digging rules, and causes either accuracy decrease or time cost rise. The ACO classifiers' performance for this data set are shown in Figure 8 when $max\_bands\_per\_group = 4$ and

$discretization\_iteration = 10$ on different training size. The classification accuracies of every individual ACO classifier are demonstrated in Figure 8.

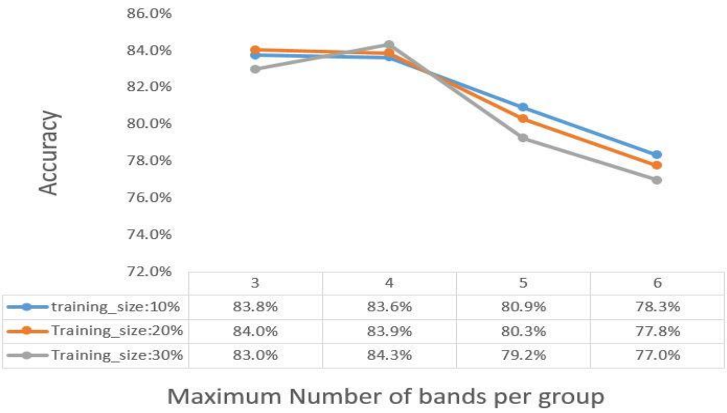

Figure 7. Influence of different size of grouped bands on classification accuracy

The peak of the classification accuracy curve of each classifiers is 78.4% when the lowest is 42.3%. This suggests that the dimensionality reduction method for hyperspectral images in this paper represents an effcacious method of improving the overall classification accuracy.

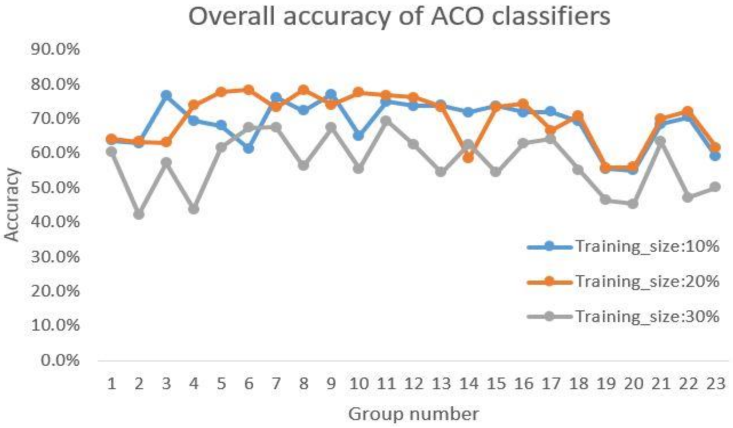

Figure 8. Overall accuracy of ACO classifier on band groups for different training size

To validate the effectiveness of the proposed method, k-Nearest Neighbour (KNN) was introduced to provide a comparison. KNN is a non-parametric approach used for classification and regression[15]. A sample is classified by a majority vote of its neighbours[16,17]. KNN needs a user-defined constant k meaning k nearest neighbours. A drawback of KNN method algorithm is that it is sensitive to the local structure of the training data. The reason to choose KNN as the comparison is that KNN can be directly applied to high dimensionality problem[18] in contrast to the proposed multiple classifier system. In Table 2, the overall accuracy with different training size based on proposed method and KNN method were shown. The results shows that the proposed multiple classifier system outperformed KNN method.

Table 2. Comparison between proposed method and KNN method

Training size(%)	10	20	30
ACO MCS(%)	84.30	83.87	84.32
KNN(%)	83.75	82.90	79.04

4. Conclusion

In this paper, a new multiple classifier system based on ant colony algorithm for hyperspectral images was carried out. A band selection and band grouping system based on mutual information of hyperspectral images was utilized in the proposed method aiming to split the origin images into several band groups. Then ACO method is applied to establish a classifier on each group to generate a multiple classifier system. Finally, a fusion strategy is taken to synthesize all the classifiers. The experiment results demonstrated that the proposed method performed a good classification ability.

Multiple classifier system was introduced to overcome the curse of dimensionality which was the main problem in hyperspectral remote sensing images classification. Feature selection methods were the traditional way of dimensionality reduction for hyperspectral images. It chooses several bands from the original space to represent the data set, which will suffer some information loss. Band grouping method decomposes the whole feature space to subspaces in order to apply ACO method in the later process. ACO method performs better than traditional methods on constructing proper classifier for hyperspectral images.

To combine each single classifier, a fusion method based on the rule confidence is implemented to improve the overall ability. By using the credibility of a specific rule, not just a output class, ACO classifiers' feature are fully employed since the credibility of a rule can be calculated during the rule searching period.

The proposed multiple classifier system has been applied to the classification of Pavia University, ROSIS data set. KNN method is carried out to provide a comparison in terms of classification accuracy. The overall accuracy of the proposed method is 84.11% with Kappa coefficient 80.23% when the KNN method has classification accuracy 83.75% and Kappa coefficient 79.80%.

Further study will focus on a flexible band grouping method that doesn't need a supervision to find the best group size and can adapt to ACO method as well. Moreover, a new fusion strategy considering not only rules' credibility but also single classifiers' accuracy deserves a try. Computation complexity is another possible topic.

References

[1] Camps V G and Lorenzo B 2005 *J. Trans. on Geosci. and Rem. Sens.* **43** 1351-62
[2] Friedl M A and Carla E 1997 *J. Rem. Sens. of Env.* **61** 399-409
[3] Pal M and Paul M 2003 *J. Rem. Sens. of Env.* **86** 554-65
[4] Parpinelli R S, Heitor S and Alex F 2002 *J. Trans. on Evol. Comput.* **6** 321-32
[5] Dorigo M, Mauro B and Thomas S 2006 *J. Comput. Intel. Mag.* **1** 28-39
[6] Dai Q, and Jianbo L 2007 *Int. Symp. on Multispectral Image Processing and Pattern Recognition* vol 7
[7] Xiaoping L 2008 *J. Trans. on Geosci. and Rem. Sens.* **46** 4198-208
[8] Baofeng G 2006 *J. Geosci. and Rem. Sens. Letters* **3** 522-6
[9] Shijin L 2011 *J. Knowledge-Based Sys.* **24** 40-8
[10] Shuang Z, Junping Z and Baoku S 2009 *2nd Int. Cong. On IEEE* vol 10
[11] Bigdeli B, Farhad S and Peter R 2013 *J. Ind. Soc. of Rem. Sens.* **41** 763-76
[12] Li X 2016 *J. Comp.* & Geosci. **89** 252-9
[13] Novovicova J 2007 *Iberoamerican Cong. on Pattern Recognition (Springer Berlin Heidelberg)* vol 13
[14] Kuncheva L I 2004 *Combining pattern classifiers: methods and algorithms* vol 15 (John Wiley & Sons)
[15] Altman N S 1992 *J. the American Sta.* **46** 175-85
[16] Blanzieri E and Farid M 2008 *J. Trans. on Geosci. and Rem. Sens.* **46** 1804-11
[17] Lefei Z 2012 *J. Trans. on Geosci. and Rem. Sens.* **50** 879-93
[18] Li Y and Bo C 2009 *17th Int. Conf. on Geoinformatics* vol 14