Research on Image Classification Algorithm Based on Artificial Immune Learning

Chengming Zhang[1], Yong Liang[1], ShuJing Wan[1] , Jinping Sun[1], Dalei Zhang[2]

[1]School of Information Science and Engineering, Shandong Agricultural University, Taian, Shandong Province, P. R. China, 271018;
[2]Taian tongli Computer Software Co., Ltd, , Taian, Shandong Province, P. R. China, 271018;

E-mail: chming@sdau.edu.cn

Abstract. On the basis of analyzing immune learning mechanism, by modeling for image classification, we can solve the problem of remote sensing image classification by using the basic principles of the use of immune learning. We have realized a classification algorithm with a function of the immune learning. Classification algorithm divides each major category into a number of small categories and the antigen population evolutionary process of each category is considered separately, therefore the convergence time is greatly decreased. When classifying, we use a variety of different ways to discriminate and introduce artificial priori knowledge to improve the classification accuracy. The results show that the algorithm can be well applied in remote sensing image classification.

Keywords: Image Classification, Artificial Immune Learning, Remote Sensing

1 Introduction

Classification is the primary means of using remote sensing images[1]. Traditional non-supervised classification approach requires adequate ground information to iterate which often be hindered classified or without high classification accuracy as lacking of enough calibration information. As the traditional supervised classification methods always consider too much about the local characteristics, it is easy to fall into local optimum but difficult to obtain high precision requirements [2,3].

Immune learning is an important characteristic of the immune system. When the immune system meets a certain antigen at the first time, the system will adjust its composition and structure to identify the antigen better [4,5]. The optimal antibody will be saved as memory after identified[6]. When the immune system meets the same or similar antigen again, the speed of the identification to it will be faster and the accuracy will be higher as the system already has the memory of it [7].

If we regard the surface feature to be classified as antigen and the characteristics of corresponding surface feature as antibody, we can achieve a supervised classification algorithm used for remote sensing images based on the immune learning mechanism

of artificial immune system to improve the computation speed and classification accuracy[8-10].

2 Modeling of Remote Sensing Image Classification

2.1 Description of the Problem

It supposes that the images to be classified have b bands totally and a class of surface feature in a certain band has only one range of gray value. For the class c, we use li and hi to express the lower bounds and upper bounds on its gray value, then its center mi is:

$$m_i = l_i + \frac{(h_i - l_i)}{2} = \frac{(h_i + l_i)}{2} \quad (1)$$

Definition 1: Class center feature vector. The vector formed in the center of all the bands by a certain class is called the class center feature vector.

$m=(m_1,m_2,\ldots,m_b)^T$ (2)

We use ri to represent the corresponding allowable error radius of band I. The value of ri is $|hi-li|/2$. If all the bands' errors are organized in order, it will form an effective radius vector as definition 2. Actually, this vector is the maximum permissible error of this class in every band.

Definition 2: the effective radius vector. For a category, the vector composed by the maximum permissible error vector in each band is called the effective radius vector:

$r=(r_1,r_2,\ldots,r_b)^T$ (3)

Every component indicates the absolute value of the error from corresponding wave band to the centre vector.

Definition 3: mode. The binary group like (4) composed by a central feature vector of a class and the effective radius vector is called the mode of this class.

$p=<m,r>$ (4)

For pixel x that supposed to be divided into class c, comparing the distance between the gray value of each wave band and the central feature vector of the corresponding component one by one. If the distance does not exceed the corresponding allowable radius, then pixel x belongs to class c certainly.

Through the above analysis, a remote sensing image classification problem can be divided into the following sub-problems:

Determine the categories and subcategories needed distinguished;

Determine the mode of each category one by one;

Classify according to the determined mode.

2.2 The Function of Affinity Discriminant

Affinity discriminate function is the function to judge the degree of similarity between antigen and antibody, antibody and antibody. The higher is the affinity; the better is the approximation of both.

Euclidean distance is a commonly used function to judge the similarity in remote sensing image classification, but the traditional Euclidean distance believes that all the bands' influence on distance is equivalent. As different bands play different roles in classification, we should quantify them by a certain rule to get a weight sequence: (w_1, w2, …, wb) and improve the Euclidean distance to make the formula (5) as the affinity calculated function of two feature vectors.

$$affinity\ (x,\ y) = \frac{1}{\sqrt{\sum_{i=1}^{b} w_i (x_i - y_i)^2}} \tag{5}$$

3 The Remote Sensing Image Classification Algorithm Based on Immune Learning

3.1 The Symbol of Algorithm

The classification of remote sensing image based on self-determination Evolution, CRSSE proposed by this article is based on immune learning mechanism. In the process of achieving the immune learning, it is believed that the antigen will cause the immune response without considering the decomposable process of antigen. As both the antigen and antibody are the feature vectors of pixel, the antibody will variety and generate new antibody directly in the algorithm. First of all, we define the symbols used in the algorithm**:**

b: the wave band number that participate the classification actually;

C: sets of classification. Each element represents a category determined in advance and each category contains several sub-categories. Symbol Ci represents category i and Cij represents the sub-categories j of i. c represents a certain category.

Q: linear classifier set. Each linear classifier represents a mode. The symbol Qi represents all the linear classifier sets of category i and Qij represents the corresponding linear classifier of category i's sub-category j. If there is only one element in the set Qj, Qj is equal to Qj1. q represents a linear classifier;

q.mc: represents the classification property of linear classifier and mc.mc $\in$ C;

q.sc: represents sub-category property of linear classifier;

q.cp: feature vector of class center, which is a b-dimensional feature vector;

qr: effective radius vector (the maximum permissible error vector), which is a isomorphic vector with vector cp. Each component represents the maximum deviation of element corresponding weight value of the same type in the class central feature vector;

q.times: the times scale factor, which magnifies the actual identification radius and changes into q.r;

AB: the antibody set (population). ab represents a single antibody and ab $\in$ AB. AB corresponds to the set of feature vectors, while ab corresponds to a specific feature vector. In the description of algorithm, in order to correspond to the noun of biological immune system conveniently we also use MC to represent AB when it mentions to the set of immune cells.

AG: a set of all antigen, in which we use ag to express a single antigen. In the classification algorithm, it represents a pixel to be classified or the pixel's feature vector in the sample.

Rh: empirical parameter, which is used to control the number of cells generated in a variation.

d: After a clone, the number of memory cells which removed from the current MC set.

w: a b-dimensional vector, which the number i component represents the classifying contribution of the number i wave band.

Affinitymin: affinity threshold.

DN_imin 、DN_imax: The minimum and maximum of the number i band in all samples when we train a sub-category,.

3.2 Training Process

It needs to train each type separately in the training. After the train, the training results will be saved into q as a basis for classification in the future. Now we describe the training process of sub- category Cij.

(1)We do the unitary processing of the feature vectors to make the distance values all in the area from 0 to 1.

(2)CRSSE tries to find the memory cells which are most similar to ag in the memory cells MC for the antigen ag transfused. Because the algorithm CRSSE uses the antibody to vary directly, this step is actually to find the most similar antibody ab in AB.

$$ab_m = \begin{cases} ag & \text{MC=null} \\ ag & \text{Max (Affinity (ag,ab}_i\text{)< Affinity}_{min} \\ ab_i & \text{Affinity (ag,ab}_i\text{)>Affinity}_{min} \end{cases} \quad (6)$$

Affinity (ag, abi) represents the affinity of antigen ag and antibody abi. Here, the distance is calculated by (5).

(3)Population updating. Population updates in three ways.

(A)Clone the obtained abm. In the view of immunology, the number of antibody clone cells is not only determined by the affinity of antibody and antigen, but also determined by the concentration of the antibody. When the antibody concentration is too high, the cloning process of antibody will be inhibited [11-12]. Therefore, the clone number of each antibody is calculated by the following way:

Calculate the concentration of antibodies abm. To each antibody to be calculated, it's concentration:

$$Con(ab_i) = \frac{\sum_{j=1}^{total-1} Si}{total - 1} \quad (7)$$

Among them, total is the total number of the antibody within species and Si is a scalar function. Determine their values as following way:

$$s_i = \begin{cases} 1 & \text{If the similarity of } ab_m \text{ and antibody } ab_j \text{ is} \\ 0 & \text{less than a given threshold} \\ & \text{Others} \end{cases} \quad (8)$$

If the concentration of abm is lower than the given threshold Conmax, we determine the number of variation according to the formula (9) and else by the formula (10).

$$Nc = \sum_{i=1}^{n} round(\frac{Rh * N}{i}) \quad (9)$$

$$Nc = round(Rh * N * \frac{1}{Con(ab_m)}) \quad (10)$$

(B)Cell variation. Cell variation is to increase the diversity of the population and the number of variable cells is mainly depended on the affinity of the antibody to the antigen. The antibody with higher affinity can produce more variable cells, while the lower may produce less.

(C)Randomly generated antibody with the same number of the variation added into the population is to improve the diversity of the species groups.

(4)Species competition.

(4-1) Calculate the sum of population resources. First calculate the stimulation level of each antibody and do the unitary processing. Stimulation level Si of antibody abi is calculated as the following way:

$$Si = \frac{affinity\ (ab_i)}{\sum_{j=1}^{t} affinity\ (ab_j)} \quad (11)$$

Calculate each antibody resources as the following way:

$$\mathrm{Re}source(ab_i) = \mathrm{S}(\mathrm{ab}_i) * Nc_i \quad (12)$$

Sum all of the antibody resources as the total resources of the population. If the summation of resources is above the given threshold Resourcemax, first remove the least stimulated antibody and its resources until it meets the requirement.

(4-2) Calculate the average stimulation level of the current population and judge whether it is above the given average threshold Stimulatemax. If it is not more than the given threshold, the evolution is end and then we prepare to determine the candidate memory cell, otherwise, compete again.

(5)For each antigen transfused, repeat steps (2) - Step (4) until the entire antigen are processed completely.

(6)Calculate the average affinity of each memory cell to the entire antigen and select the immune cells (antibodies) of best affinity.

(7)Take the feature vector of immune cell as the centre feature of class into q.

For each sub-category to be learned, repeat step (1) - (7) until complete all the categories process.

3.3 Image Classification

After training, we will get a corresponding linear classifier to each sub-category. We can classify to the element with using the linear classifier. In the classification, consider each element to be classified as antigen ag. To element ag, the classification process is:

(1)The initial classification. Initial classification is mainly used for classifying the more "pure" pixel and then we seek out these pixels which cannot be classified currently.

(1-1)Calculate the absolute distance between ag and each component of category central feature vector in all sub-categories. If for a certain Cij, ag and its absolute value in each band are both less than the corresponding distance of the maximum limits of errors, then record it into queue LC and LC represents that all the sub-categories that may contain ag.

(1-2)For each sub-category in LC, calculate the distance between ag and center feature vector of this class separately and select the biggest in c then classify ag into this class.

(1-3)If LC is empty, it may be caused by the maximum effective radius little. According to the formula (2.21), we can calculate the Affinity between ag and all class feature vector and select the biggest affinity Cij. If Affinity is bigger than Affinitymin, we classfy ag into Cij and revise the corresponding q.r again.

(2)Repeat the first step until complete the process of all the elements.

4 The Analysis of Experimental Results

4.1 Data Source

The data source used in the experiment is ETM + image with a resolution of 30m and eight bands totally. The area of the image of which size is 1500 pixels *1200 pixels and the acquisition time is on the May 31, 2009 is Xueye Reservoir, in laiwu of Shandong province. As is shown in the Figure 1.

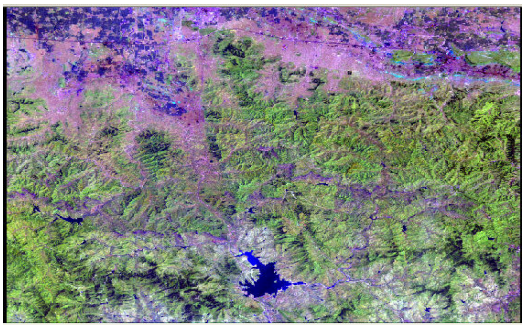


Figure 1 (a) ETM + image

4.2 Sample Data

After a manual interpretation the image, it is due to be divided in four categories: water bodies, vegetation, bare land and field. Figure 2 shows the chart of spectral curve.

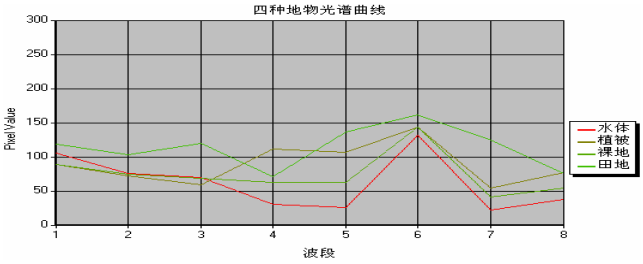


Figure 2, the spectrum curve of the four classes

We can see that each category is too close in the band 6 and band 8 in the spectrum curve chart. As the difference is not clear, we used band 1 to 5 and the band 7 to classify in practical application. We choose the spectrum sample of four categories training sample region as the initial antigen sample population of the algorithm. The sample of each type feature is shown in the Table 1.

Table 1 Object classes and the number of samples in experiment

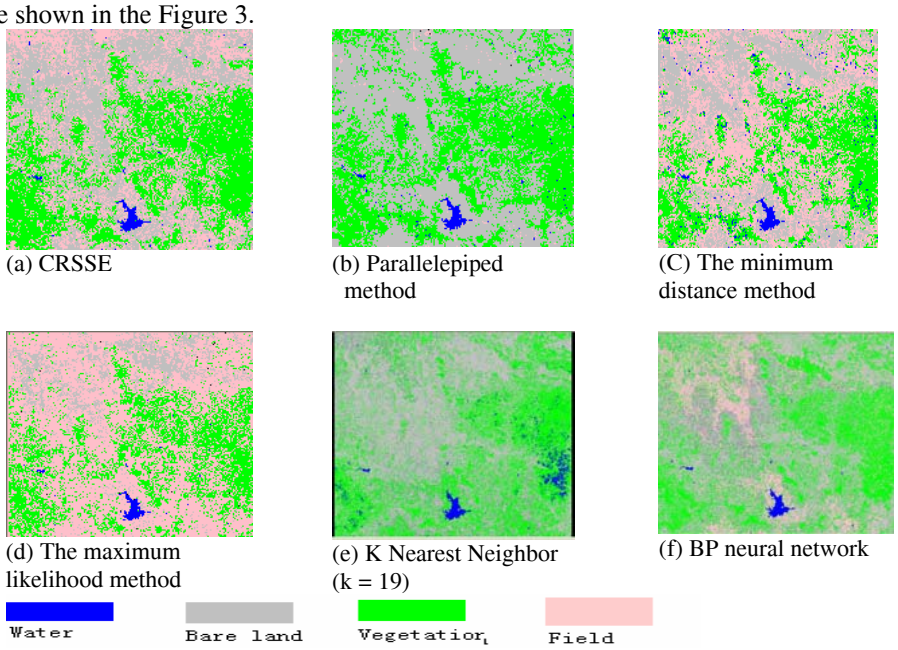
Name of object classes	Samples
Water bodies	1302
Vegetation	3086
Bare land	5793
Field	4766
Total	14937

4.3 Experimental Condition

When experimenting, the value of each parameter in CRSSE algorithm is as follows:

times：1；Rh：0.1；Aberrance：0.2；d：0.2；w: All the components take the average value of 1/b, we think that all the bands has the same right; Resourcemax：50；Stimulatemax：0.9；Affinity max：0.8；

For the convenience of comparison, we give the classification results of parallelepiped classification method, the minimum distance from the average classification method, K nearest neighbor method, maximum likelihood classification method and the classification of BP algorithm. Among them, k takes the value of 19 in the K nearest neighbor algorithm. BP algorithm uses a hidden layer and the number of hidden layer nodes is 20, while the learning rate is 0.3. The classification results are shown in the Figure 3.



(a) CRSSE (b) Parallelepiped method (C) The minimum distance method

(d) The maximum likelihood method (e) K Nearest Neighbor (k = 19) (f) BP neural network

Figure 3 Classification Results

We can see from the Figure 3 that there are always many wrong classification phenomena with using parallelepiped classification method. Through the investigation and interpretation, this method divides a number of land surface features into vegetation and fails to classify water bodies correctly with a result of a part little water body having not been recognized. When we use the minimum distance classification method, the identification of vegetation is better, but the mixed classification phenomenon of bare land and field is quite serious. The maximum likelihood method classifies water body better, but there are more wrong phenomena when classifies the vegetation and bare land. K nearest neighbor method can distinguish water bodies and field better, but cannot distinguish vegetation preferably. BP neural network method has better effects on the classification of each category,

but there is still any vegetation not distinguished. When using CRSSE classification method, it can classify water body, vegetation and bare land with greater accuracy and the result of classification is satisfactory through the investigation and interpretation. Therefore, we can see that CRSSE has a better classification result from manual visual interpretation and compared with the original image and investigation, its classification is improved in accuracy from the traditional classification methods.

To further verify the correctness of this classification method, we compare CRSSE method with traditional remote sensing image classification methods in classification accuracy. We use a common total precision and Kappa coefficient as the comparative evaluation indicators and count the computation time required for different algorithms meanwhile. The results is shown in the Table 2.

Table 2 comparison results of the Classification method

method	Parallel pipeline method	Minimum Distance method	Maximum likelihood	K Nearest Neighbor （k=19）	BP(1 Hidden layer）	CRSSE
Total accuracy	73.00%	81.00%	81.00%	81.77%	84.22%	87.00%
Kappa coefficient	0.6520	0.7133	0.7422	0.7867	0.8023	0.8306
Time (s)	24	25	56	123	1202	346

As can be seen from Table 2.3, the overall accuracy and Kappa coefficient of CRSSE are all better than the conventional methods with overall accuracy of 91.33% and Kappa coefficient of 0.891. The other followed are BP neural network, maximum likelihood, K nearest neighbor method, the minimum distance from the average value and the parallelepiped method. The calculating time of CRSSE is longer than the maximum likelihood method, but its accuracy is significantly higher. Almost each index of RLCRSI is better than the traditional classification methods. It is mainly because parallelepiped method and minimum distance from the average method only consider partial characteristics during the process of training and classification, so it always falls into local optimum in the training process, which will result in the wrong sub-classification of images. While the premise of the maximum likelihood method is that the probability density distribution function of each class is the normal distribution. If the sample data deviates from the normal distribution too much, it will affect the classification accuracy of maximum likelihood method greatly. BP neural network requires more computing time. Because the CRSSE classification method with high classification accuracy has the characteristics of self-organizing of artificial immune system and self-learning ability, the requirements of the conditions on the sample distribution is not too high.

5 Summary

According to immune learning theory, we have built model for the classification problem of remote sensing image and put forward remote sensing image classification

algorithm based on immune learning. The algorithm divides each big category into a number of small categories and the evolution process of each category's antigen population is considered separately, thus it can greatly reduce the convergent time to make it more suitable for processing remote sensing image. We use various methods to determine categories' property to improve the classification accuracy in the classification. Compared with traditional method, CRSSE has higher classification accuracy and can be well applied in remote sensing image.

Acknowledgements

This study has been Funded by Key Laboratory of Geo-informatics of State Bureau of Surveying and Mapping (Contract Number: 200811).

Reference

1.Atkinson PM, Lewis P. Geostatistical Classification for Remote Sensing: an Introduction.Computers & Geosciences, Issue 26, pp.361-371.(2000)
2.ZHANG Liangpei,HUANG Xin. Advanced processing techniques for remotely sensed imagery. Journal of Remote Sensing, Issue 13, pp.560-572 (2009)
3.Dundar,M.and Landgrebe.D. A Model-based Mixture-Supervised Classi-fication Approach in Hyperspectral Date Analysis. IEEE Transactions on Geoscience and Remote Sensing, Issue 40, pp.2692-2699 (2002)
4.V.P. Sakthivel, R. Bhuvaneswari, S. Subramanian. Artificial immune system for parameter estimation of induction motor.Expert Systems with Applications, Volume 37, Issue 8, pp. 6109-6115 (2010)
5.C.A. Laurentys, R.M. Palhares, W.M. Caminhas. Design of an artificial immune system based on Danger Model for fault detection.Expert Systems with Applications, Volume 37, Issue 7, pp. 5145-5152 (2010)
6.L.N.de Castro，J.Timmis. Artificial Immune Systems: A New Computational Intelligence Approach. Berlin, Springer-Verlag, 2002.
7.J. Kim and P. Bentley. Towards an Artificial Immune System for Network Intrusion Detection: An investigation of Clonal Selection with a negative Selection Operator. In the proceedings of the Congress on Evolutionary Computation., Seoul, Korea, May 27-30, 2001.
8.ZHONG Yanfei,ZHANG Liangpei,LI Pingxiang. A Clonal Selection Algorithm Based on Non-uniform Adaptive Mutation. Geomatics and Information Science of Wuhan University, Issue 34, pp.308-313 (2009)
9.ZHONG Yanfei,ZHANG Liangpei,LI Pingxiang. Remote Sensing Image Classification Based on Artif icial Immune System. JOURNAL OF REMOTE SENSING, Issue 9, pp.374-382 (2005)
10.XIAO Renbin, WANG Lei. Artificial Immune System: Principle Models, Analysis and Perspectives. Chinese Joural of Computers. Issue 25, pp.1281～1293 (2002)
11.DU Haifeng, JIAO Licheng, WANG Sunan. Clonal operator and antibody clone algorithms [A]. In: Proceeding of IEEE the First International Conference on Machine Learning and Cybernetics ,Beijing, China, pp.506～509. (2002)
12.Giles M. Foody. Status of land cover classification accuracy assessment. Remote Sensing of Environment. Issue 80, pp.185-201.(2002)