Research Article

A New Fuzzy Clustering Algorithm Based on Clonal Selection for Land Cover Classification

Yanfei Zhong and Liangpei Zhang

State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, 129 Luoyu Road, Hubei 430079, China

Correspondence should be addressed to Yanfei Zhong, zhongyanfei@lmars.whu.edu.cn

A new fuzzy clustering algorithm based on clonal selection theory from artificial immune systems (AIS), namely, FCSA, is proposed to obtain the optimal clustering result of land cover classification without *a priori* assumptions on the number of clusters. FCSA can adaptively find the optimal number of clusters and is designed as a two-layer system: the classification layer and the optimization layer. The classification layer of FCSA, inspired by clonal selection theory, generates the optimal classification result with a fixed cluster number by utilizing the clone, mutation, and selection of immune operators. The optimization layer of FCSA evaluates the optimal solutions according to performance measures for cluster validity and then adjusts the cluster number to output the final optimal cluster number. Two experiments with different types of image evince that FCSA not only finds the optimal number of clusters, but also consistently outperforms the traditional clustering algorithms, such as K-means and Fuzzy C-means. Hence, FCSA provides an effective option for performing the task of land cover classification.

1. Introduction

Land cover classification from remotely sensed images is considered to be a cost-effective and reliable method for generating up-to-date land cover information [1]. Clustering algorithms, or unsupervised classification algorithms, are built to solve the site labeling problem without the need for training samples for land cover classification [2]. For example, the familiar K-means [3] and Iterative Self-Organizing Data (ISODATA) [4] algorithms iteratively assign the pixels of an image to one of the classes. K-means finds an optimal partition of the data distribution into the requested number of subdivisions, while ISODATA is a modified version of the K-means algorithm. Both of them first assign an arbitrary initial cluster vector. The mean vectors and covariance matrix of clusters are then calculated based on the pixels in the initial cluster; pixels in the image are assigned to the closest cluster to form a new cluster and the label of each pixel is updated. The mean vectors and covariance matrix of clusters

are recalculated subsequently based on the new clusters. In every iteration of the classical K-means and ISODATA algorithms, each image pixel is assumed to be in exactly one cluster; an alternative to the crisp membership association uses fuzzy sets to describe the relationship between the data points and the cluster centers.

For instance, Fuzzy C-means (FCM) [3] is an approach to clustering those partitions of an image data set into C fuzzy subsets using fuzzy membership. In addition to the afore-mentioned algorithms, Bayesian classifiers [5] and Markov Random Field [6] have also been employed for the unsupervised classification for remote sensing images. Recently, there has been considerable interest in applying unsupervised neural networks [7], such as Kohonen's Self-organizing Maps (SOM), to multi/hyper-spectral remote sensing image classification. SOM was investigated as a possible tool for automated knowledge acquisition. In addition, with the emergence of genetic algorithm (GA), some GA-based clustering algorithms have been proposed, which can converge to the global optima with high probability [8].

Since geographical information (including remotely sensed data) for land cover classification is imprecise, meaning that the boundaries between different phenomena are fuzzy, fuzzy clustering algorithms, for example, FCM, are better suited for dealing with real-world problems of land cover classification than classical crisp classification models, such as K-means. However, FCM has two major limitations. On the one hand, it requires the *a priori* specification of the number of clusters. When the number of clusters is specified incorrectly, serious problems may arise. On the other hand, FCM is much more sensitive to the initialization and easily falls into a local optimum [9]. To overcome these obstacles, this paper proposes a fuzzy clustering algorithm based on clonal selection theory from artificial immune systems (AIS), namely, FCSA, to automatically evolve the fuzzy partitions of land cover data such that some measure of goodness of the partitions is optimized. Clonal selection theory [10, 11] is a basic theory in immune systems to explain the basic features of an adaptive immune response to an antigenic stimulus. The clonal selection algorithm (CSA) [12], derived from clonal selection theory, is proposed as an important model in artificial immune systems (AISs), which are inspired by the vertebrate immune systems, and use the immunological properties to support a wide range of applications [13–15]. CSA has been successfully applied to pattern recognition, multimodal optimization, feature selection, and classification by utilizing its biological properties such as immune evolution and immune memory [12, 14, 16].

To automatically evolve the optimal number of clusters as well as the fuzzy partitioning of the data, the proposed fuzzy clustering algorithm (FCSA) is designed as a two-layer system: the classification and optimization layers. The classification layer of FCSA can quickly obtain the global optimum and has the better classification results with fixed cluster numbers, since FCSA utilizes different immune operators, such as the clonal operator, mutation operator, selection operator, and those operators can combine the evolutionary search and random search and incorporate the global search with a local search by the clonal operation on candidate solutions. The optimization layer of FCSA controls the process of the classification layer, evaluates the optimal solutions according to performance measures for cluster validity and then adjusts cluster numbers to output the final solution. In this paper, the Xie-Beni (XB) [17] cluster validity index is selected as the underlying optimizing criterion since it has shown to be better able to indicate the correct number of clusters in several experiments [18, 19]. The FCSA is evolution-like and has several interesting features: (1) the cluster number is dynamically adjustable and automatically obtained; (2) has the capability of maintaining local optima solutions; (3) explores the global optimal. The Flightline C1 (FLC1) and TM remote sensing images have been used for demonstrating the effectiveness of the

developed unsupervised fuzzy artificial immune classifier by automatically segmenting the images into unknown regions. Experimental results demonstrate that the proposed algorithm outperforms the traditional methods, that is, FCM, and thus provide an effective option for unsupervised land cover classification.

The remainder of the paper is structured as follows. Section 2 gives an overview of the clonal selection theory and the clonal selection algorithm. Section 3 describes the proposed method and algorithm in detail, while Section 4 illustrates the performance of the proposed algorithm as compared to the traditional algorithms. Finally, Section 5 concludes the paper.

2. Clonal Selection Algorithm (CSA)

2.1. Clonal Selection Theory

The human immune system, a complex system of cells, molecules, and organs, symbolizes an identification mechanism capable of perceiving and combating dysfunction from our own cells and the action of exogenous infectious microorganisms. This immune system protects the body from infectious agents such as viruses, bacteria, fungi, and other parasites. Any molecule that can be recognized by the adaptive immune system is known as an antigen. The basic component of the immune system is the lymphocytes or white blood cells. Lymphocytes exist in two forms, B cells and T cells. These two types of cell are rather similar, but differ in how they recognize antigens and in their functional roles. B cells are capable of recognizing antigens free in solution, while T cells require antigens to be presented by other accessory cells. They have distinct chemical structures and produce many Y-shaped antibodies from their surfaces to kill the antigens. Antibodies are molecules attached primarily to the surface of B cells with the aim of recognizing and coping with antigens [20].

In order to clarify how an immune response is mounted when a nonself antigenic pattern is recognized by a B cell, clonal selection theory has been developed [21, 22]. The main features of clonal selection theory are concerned with (1) proliferation and differentiation on simulation of cells with antigens; (2) generation of new random genetic changes, expressed subsequently as diverse antibody patterns, by a form of accelerated somatic mutation; (3) estimation of newly differentiated lymphocytes carrying low-affinity antigenic receptors. These will be utilized in this paper.

The principle can be detailed as follows. When a B-cell receptor recognizes a nonself antigen with a certain affinity, it is selected to proliferate and produce antibodies in high volumes. The antibodies are soluble forms of the B-cell receptors that are released from the B-cell surface to cope with the invading nonself antigens. Antibodies bind antigens leading to their eventual elimination by other immune cells. Proliferation in the case of immune cells is asexual and it is a mitotic process, in which the cells divide themselves. During reproduction, the B-cell clones undergo a hypermutation process in that the antigen stimulates the B cell to proliferate and mature into terminal antibody secreting cells, named plasma cells. The process of cell division generates a clone. In addition to proliferating and differentiating into plasma cells, the activated B cells with high antigenic affinities are selected to become memory cells with long life spans. These memory cells circulate through the blood, lymph, and tissues. When exposed to a second antigenic stimulus, memory cells commence to differentiate into plasma cells capable of producing high-affinity antibodies, preselected for the specific antigen that had stimulated the primary response [12]. Figure 1 illustrates the clonal selection, expansion, and affinity maturation processes.
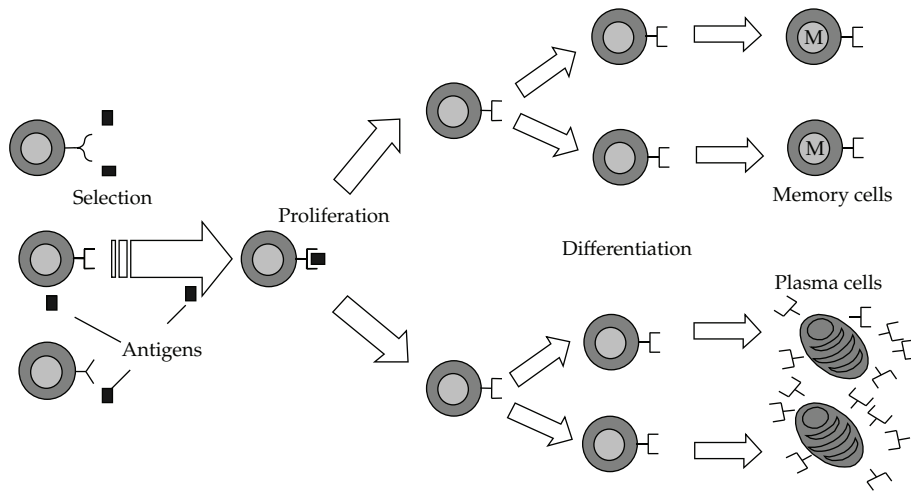
Figure 1: Clonal selection principle.

2.2. Clonal Selection Algorithm (CSA)

Based on the clonal selection theory and the shape space model of the immune system, De Castro and Von Zuben (2002) developed the Clonal Selection Algorithm (CSA) [12]. It has been applied to support pattern recognition and solve multimodal optimization problems. The algorithm can be described as follows.

Step 1. Randomly initialize a population of individuals, M.

Step 2. For each input pattern P, present it to the population M and determine its affinity with each element of M.

Step 3. Select n of the best highest affinity elements of M and clone these individuals proportionally to their affinity with the antigen. The higher the affinity, the higher the number of copies, and *vice versa*.

Step 4. Mutate all these copies with a rate proportional to their affinity with the input pattern–the higher the affinity, the smaller the mutation rate.

Step 5. Add these mutated individuals to the population M and reselect m of these matured individuals to be kept as memory cells of the systems.

Step 6. Repeat Steps 2–5 until a certain criterion is met.

Similar to CSA, the genetic algorithm (GA) is also a heuristic algorithm. However, their underlying mechanisms and methods of evolutionary search significantly differ in terms of inspiration, vocabulary, and fundamentals. While GA uses a vocabulary borrowed from natural genetics and is inspired by the Darwinian evolution theory, CSA makes use of the shape space formalism, along with immunological terminology to describe antigen-antibody interactions and cellular evolution in immune systems. GA performs a search through genetic operators including reproduction, crossover, and mutation, while CSA performs

its search through the mechanisms of somatic mutation and receptor editing, balancing the exploitation of the best solutions with the exploration of the searchspace. The CSA maintains a diverse set of local optimal solutions, while the GA tends to polarize the whole population of individuals towards the best one. This mainly occurs because of the selection and reproduction schemes adopted by the CSA (described in Step 3). Essentially, their coding schemes and evaluation functions are not different, but their evolutionary search differs from the viewpoint of inspiration, vocabulary, and fundamentals [23]. In addition, CSA inherits the memory property of human immune systems to build a memory cell population and can recognize the same or similar antigens quickly at different times [14, 24].

3. Fuzzy Clustering Algorithm Based on Clonal Selection (FCSA)

A fuzzy clustering algorithm based on clonal selection, namely, FCSA, is proposed to perform the task of land cover classification by automatically evolving the optimal fuzzy partition matrix. A main objective of the proposed algorithm is to get closer to a more natural classification of land cover.

A remote sensing image dataset $X = \{x_1, x_2, \ldots, x_N\}$ is observed, where each object x_i will be an earth surface unit or picture element (pixel), $j = 1, \ldots, N$. N represents the total number in an unlabeled image, $N = N_{\text{row}} \times N_{\text{col}}$ where N_{row} and N_{col} represent the image's row number and column number, respectively. In addition, each pixel x_j contains the attributes vector with p bands, $x_j = \{x_j^1, x_j^2, \ldots, x_j^p\} \in \mathfrak{R}^p$. The image dataset is partitioned into a set of nc clusters $CS = \{CS_1, CS_2, \ldots, CS_{nc}\}$, where nc represents the number of clusters. In the fuzzy cluster analysis, each pixel in the dataset can be assigned to more than one cluster, according to a membership value $u_{ij} = \mu_{CS_i}(x_j)$, which defines the membership of the pixel x_j to the cluster CS_i.

To find adaptively the optimal number of clusters, FCSA is designed as a two-layer system: the classification and optimization of FCSA. The optimization layer of FCSA controls the process of the classification layer and evaluates the optimal number of clusters according to performance measures using the Xie-Beni index. Each Xie-Beni index with the different number of clusters may be calculated after the process of classification. The best partition is considered to be the one that corresponds to the minimum value of the Xie-Beni index. The classification layer of FCSA with the fixed number of clusters classifies the image dataset by exploring the optimal membership degrees matrix and centers of clusters with minimum objective function shown in (3.1), under the constraint shown in (3.2), where $\mu_{ij} \in [0, 1]$ indicates the membership of data vector x_j assigned to cluster CS_i. v_i is the ith center of CS and d_{ij} is the Euclidean distance between data vector x_j and center v_i. $m \in [1, \infty)$ is a parameter to control the fuzziness clustering result

$$J(U, v_1, v_2, \ldots, v_{nc}) = \sum_{i=1}^{nc} J_i = \sum_{i=1}^{nc} \sum_{j=1}^{N} u_{ij}^m d_{ij}^2, \tag{3.1}$$

$$\sum_{i=1}^{nc} u_{ij} = 1, \quad \forall j = 1, \ldots, N. \tag{3.2}$$

To obtain the optimal minimum objective function, it is feasible to encode either the center matrix CS or membership function matrix U. The relationship between CS and U can

be denoted as in (3.3). In this paper, we encode CS into antibodies of the proposed algorithm to calculate their values

$$v_i = \frac{\sum_{j=1}^{N} u_{ij}^m x_j}{\sum_{j=1}^{N} u_{ij}^m}, \qquad u_{ij} = \frac{1}{\sum_{k=1}^{nc} \left(d_{ij}/d_{kj}\right)^{2/(m-1)}}. \tag{3.3}$$

To better describe the FCSA, the following notations are used.

(i) AB denotes the set of antibodies and ab represents a single antibody, where $AB = \{ab_1, ab_2, \ldots, ab_M\}$, M is the number of the antibody population. Each antibody $ab_i = (ab_i^1, ab_i^2, \ldots ab_i^{p'})$ $(i = 1, 2, \ldots M)$ represents a possible solution of the cluster result, p' is the number of ab's features, $p' = p \times nc$, $AB \subset \mathfrak{R}^{p'}$.

(ii) AG denotes the set of antigens, which represent unlabeled data or image pixels. $AG = \{ag_1, ag_2, \ldots, ag_N\}$, N is the number of the antigen population, $ag_i = (ag_i^1, ag_i^2, \ldots, ag_i^p)$, and p is the dimension of features. For land cover classification, N represents the total number of unlabeled remote sensing image pixels and p the bands for each pixel and the image. $AG \subset \mathfrak{R}^p$.

(iii) mc denotes the memory cell. mc indicates the best antibody with the highest membership value in each iteration and mc is a candidate solution.

The FCSA algorithm consists of the following steps.

3.1. The Classification Layer of FCSA

The classification layer of FCSA is used to find the best fuzzy partition of the image dataset with the fixed number of clusters.

Step 1 (*initialization and encoding*). In FCSA, the antibodies are made up of real numbers; each antibody ab_i represents a group of clustering centers with nc prototypes as in

$$ab_i = \left\{ \underbrace{ab_i^1, ab_i^2, \ldots, ab_i^p}_{v_1} \ldots, ab_i^j, \ldots, \underbrace{ab_i^{(nc-1)p+1}, ab_i^{(nc-1)p+2}, \ldots, ab_i^{p*nc}}_{v_{nc}} \right\}. \tag{3.4}$$

A first antibody population AB including M antibodies is randomly generated by selecting distinct points from the AG dataset. $AB = \{ab_1, ab_2, \ldots, ab_M\}$ and M is the number of the antibody population.

Step 2 (*cycle the generations*). After initialization, the simulation of the clonal selection process begins. One generation after another is created and each must prove its affinity to the criterion function. In each iteration, a number of possible solutions are generated by applying the immune operators such as clone, mutation, selection in a stochastic process guided by an affinity measure. The algorithm seeks to evolve an optimal solution to the clustering problem.

(1) Calculation of Affinity

According to the initial antibody population, the affinity of all M *ab*s in the antibody population AB is calculated using the criterion function $F = F(ab_i)$. The higher the criterion function, the better the antibody. However, an optimal fuzzy partition should minimize the objective function J in (3.1), which is the generalized least squares error function. To maximize the criterion function F, the function may be defined as follows:

$$F(ab_i) = \frac{1}{J(U, v_1, v_2, \ldots, v_{nc}) + 1}. \tag{3.5}$$

(2) Selection

From AB, the "n" highest affinity antibodies are selected to compose a new set $AB_{\{n\}}$ of high-affinity antibodies and the highest affinity memory cell (mc) is found.

(3) Clone

After receiving antibody individuals closer to the solution, the next generation should mainly be derived from the better-fitting individuals. Thus, the n selected *ab*s are cloned based on their antigenic affinities, generating the clone set C. In the FCSA, the number of clones for each subpopulation is no longer a free parameter but instead a fixed number $n + 1$. This is an interesting feature, since the performance of the CSA algorithm is very sensitive to variations in the number of clones [16, 25].

The total number of clones generated N_c is defined as follows:

$$N_c = \sum_{i=1}^{n} (n + 1) = n^2 + n. \tag{3.6}$$

This step draws the evolutionary process closer to the goal. It raises the average affinity value and gives the following steps a good chance to further move towards the solution.

(4) Mutation

Provide each *ab* in the clone set C with the opportunity to produce mutated offspring C^*. The higher the affinity, the smaller the mutation rate. To adaptively determine the mutation rate according to the affinity of each *ab*, the process is as follows.

Firstly, for each $ab_i \in AB$, normalize its affinity $F(ab_i)$ into the range $[0, 1]$:

$$F'(ab_i) = \frac{F(ab_i) - \min(F(ab_i))}{\max(F(ab_i)) - \min(F(ab_i))} \quad i = 1, 2, \ldots, M. \tag{3.7}$$

Then, let each ab_i have the chance to mutate; the mutation rate is adaptively calculated as

$$p_m = \exp(-2 * F'(ab_i)), \tag{3.8}$$

where p_m is the mutation rate of each *ab*, 2 is the empirical value to control the decay, and $F'(ab_i)$ is the affinity according to (3.7). In (3.8), the range of the mutation rate is $[0, 1]$.

Finally, the cloned antibodies are mutated with probability p_m.

The mutation process to each ab in the clone set C is as shown in Algorithm 1. The function mutation (B) with mutation rate p_m, is defined in Algorithm 1. The function *random* (minimum, maximum) generates a random real value using a uniform distribution in the range from the minimum to the maximum. Function $\Delta(\text{Ite}, u)$ is defined as

$$\Delta(\text{Ite}, u) = u\left(1 - r^{(1-m/\text{Ite})^\lambda}\right), \tag{3.9}$$

where m is the iteration number, Ite is the maximal iteration number, r is a random value within the range $[0, 1]$, and λ is a parameter to decide the nonconforming degree.

This step is crucial in the proposed algorithm. It generates random changes of single features of the individual solutions. The value of these changes can be found at the criterion function calculation within the next generation cycle. This helps avoid local maxima and produces new properties of mutated antibodies that can remain if they are successful, while traditional fuzzy clustering algorithms, such as FCM, often get stuck at suboptimal solutions based on the initial configuration of the system.

To avoid chaotic development and maintain the best abs for each clone during evolution, one original ab for each clone without mutation during the maturation process is kept, else it would destroy the positive development of the previous step and disable any major development towards the solution.

(5) Recalculation of Affinity

Calculate the affinity $F^*(ab_i)$ of the matured clones C^*.

(6) Reselection

From the mature clone set C^*, reselect the nabs with the highest affinity to replace the nabs with the lowest affinity in AB. Select the highest affinity ab in C^* to be a candidate memory cell, $mc_{\text{candidate}}$. If the affinity of $mc_{\text{candidate}}$ is higher than the memory cell, mc, then $mc_{\text{candidate}}$ will replace mc and become a new memory cell.

(7) Displace

In order to replace the d lowest affinity abs from AB, d new antibodies are produced by a random process. This step may increase the diversity of the antibody population.

Step 3 (*stopping criteria*). The stopping criteria for the algorithm are as follows. One option is to set a fixed number of iterations as the stopping condition. The other criterion is that if after a few iterations, there is no improvement of the criterion function (F) value as shown in (3.10), then the optimal clustering result has been found. Otherwise, return to Step 2 until the stop criteria are satisfied.

$$\left\|F^{(t+1)} - F^t\right\| < \varepsilon, \tag{3.10}$$

where the change threshold ε is a user-defined parameter and selected according to different applications.

```
mutate (B)
{
   for each (B · v_i)
   do
       a_i = min v_i
       b_i = max v_i
       rd_mr = random (0, 1)
       rd_to = random (−1, 1)
       if (rd _mr < p_m)
         if (rd _to >= 0)
           B · v_i =  B · v_i + Δ(Ite, b_i − B · v_i)
         else
           B · v_i =  B · v_i − Δ(Ite, b_i − B · v_i)
   done
   return B
}
```

Algorithm 1: Real-value mutation (p_m represents mutation rate).

Finally, the proposed algorithm outputs the value of the memory cell and obtains the optimal fuzzy partition with the current number of classes, nc, $nc = 1, 2, \ldots, C_{\max}$.

3.2. The Optimization Layer of FCSA

Determining the optimal number of clusters is an important issue in FCSA. To evaluate the optimal solutions, FCSA evaluates the validity measure of the c-partition for a range of nc values using the Xie-Beni (XB) index [8, 17] and then selects the optimal number of clusters with the minimum value of the XB index. Here, nc is an estimate of the upper bound of the number of clusters.

The XB index is defined as a function of the ratio of the compactness π to the separation s. Here π and s can be written as follows:

$$
\pi = \frac{\sigma}{N} = \frac{\sum_{i=1}^{nc} \sum_{j=1}^{N} u_{ij}^m d_{ij}^2}{N},
$$
$$
s = \min_{i,k} \|v_i - v_k\|^2,
$$
$$
XB = \frac{\pi}{s} = \frac{\sum_{i=1}^{nc} \sum_{j=1}^{N} u_{ij}^m d_{ij}^2}{N \min_{i,k} \|v_i - v_k\|^2},
$$
$$
d_{ij} = \|v_i - x_j\|,
$$

(3.11)

where d_{ij} represents the distance between the ith center v_i and the jth antigen x_j, N represents the total number of the antigens, nc represents the number of classes.

A smaller XB indicates a partition in which all the clusters are compact and separate from each other. Thus, FCSA has to find adaptively the optimal number of clusters with the smallest XB calculated by the corresponding classification result.
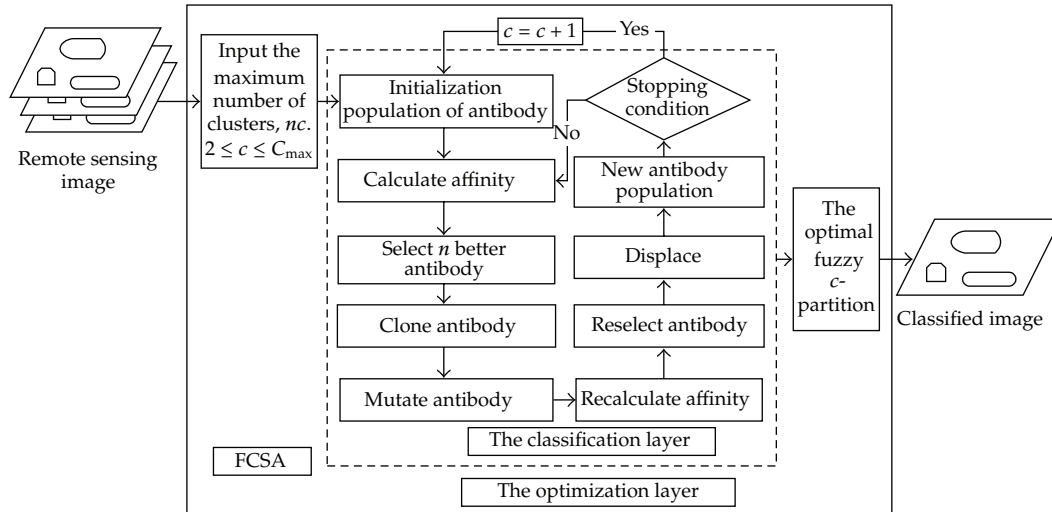
The flowchart for FCSA is shown in Figure 2 .

Figure 2: The process for the fuzzy clustering algorithm based clonal selection (FCSA).

4. Experiments and Analysis

The proposed FCSA and traditional clustering algorithms for land cover classification were all implemented using Visual C++ 6.0 and tested on different types of remote sensing image. Two experiments were conducted to test the performance of classification. Only FCSA can classify the image without *a priori* assumptions on the number of clusters and finally output the optimal number of clusters. To better assess the performance of FCSA, consistent comparisons of classification results with the optimal number of clusters were also performed among FCSA, K-means, ISODATA, and Fuzzy C-means (FCM) using the classification accuracy of the Flightline C1 and Landsat TM images.

4.1. Experiment 1: Flightline C1

This experiment was conducted using a data set designated Flightline C1 (FLC1) [26], which was 12-band multispectral data taken over Tippecanoe County, IN, by the M7 scanner in June, 1966. Figure 3 shows the experimental FLC1 image (92×107 pixels) with spectral ranges from 0.40 to 1.00 μm.

The primary parameters to be provided by users for the classification were the maximum number of classes $C_{\max}$, the maximum iteration MaxIte, antibody population size M, the number of selected antibodies n (see also Step 3 in Section 3), and the number of displaced antibodies, d (see also step (7) in Section 3). Generally, to conveniently apply FCSA, n is often set to N. The affinity function was determined by (3.5). The values of these parameters were set as follows: $M = 20$, MaxIte $= 100$, $n = M = 20$, $d = 5$, and $C_{\max} = 6$. The weighting exponent m, used by the FCSA and FCM, was set to 2, which was the optimal value of m within [1.5, 2.5] in practical applications [18].

FCSA automatically provides four clusters for this image dataset. Figure 4 shows the variation of the XB index with the number of clusters when FCSA is used as the underlying clustering technique. As can be seen from the figure, the minimum value of the XB index

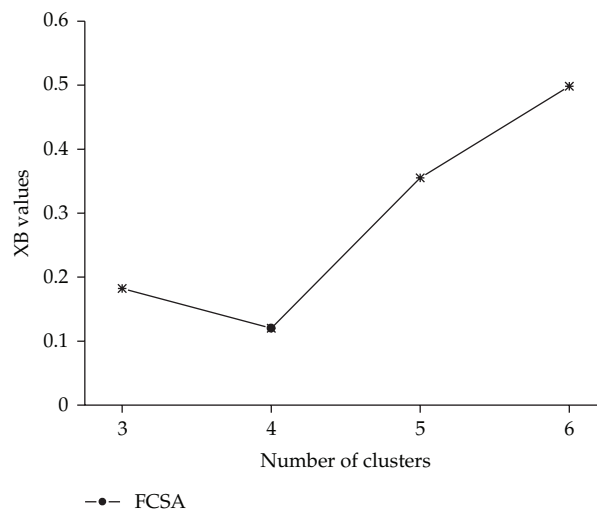**Figure 3:** Flightline C1 Image RGB (9,7,4).

Figure 4: Variation of the XB index with the number of classes for the FLC1 image using FCSA.

is obtained for four clusters with the FCSA algorithm. In fact, from our ground knowledge, the survey area is an agricultural area that is expected to fall into four classes: corn, oats, red clover, and wheat. Hence, it is evident that FCSA correctly finds the optimal number of clusters in this case. The list of classes and the number of labeled samples for each class are given in Table 1. The field map is shown in Figure 5 based on ground truth data and Figure 6 displays the spectral curves of the above four land cover classes.

To better evaluate the classification performance of FCSA, three traditional clustering algorithms for land cover classification are used in this experiment: K-means, ISODATA, FCM, when the optimal number of clusters is set to 4. Figures 7(a), 7(b), 7(c), and 7(d) illustrate the classification results using K-means, ISODATA, FCM, and FCSA, respectively.

The visual comparisons of the four clustering results in Figure 7 show varying degrees of accuracy in pixel assignment. It can be seen from the classification images that the four classifiers have similar classification results in the corn class. For the other classes, K-means

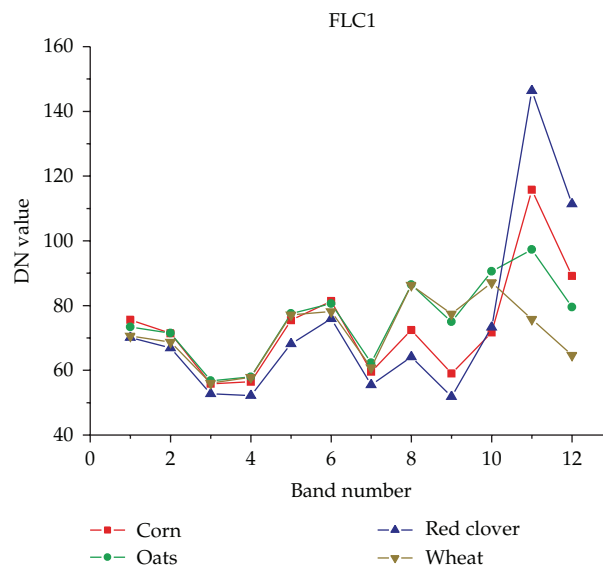**Figure 5:** The ground truth data of FLC1.

Figure 6: The spectral curves of four land cover classes.

and ISODATA create similar classification results and cannot correctly obtain four clustering partitions. In the classification images of K-means and ISODATA, the oats class disappears and is misclassified as wheat. The reason for the incorrect results is that the spectral curves of the oats class (green) and the wheat class (yellow) shown in Figure 6 are too similar to allow differentiation by the K-means and ISODATA algorithms, which have only little differences in the 11th and 12th bands. FCM and FCSA may correctly find the oats class by the corresponding fuzzy partition (Figures 7(c) and 7(d)) . Comparing FCSA with FCM, they

Table 1: List of classes and number of labeled samples in each class for experiment 1.

Class name	Number of labeled samples
Corn	1743
Oats	715
Red clover	1700
Wheat	1472
Total number of samples	5630

have similar results in the corn, oats, and wheat classes. However, FCM fares the worst in the red clover class because many red clover pixels are misclassified to the corn class at the bottom of the classification image. In contrast, FCSA achieves the best visual accuracy in the red clover class and also performs satisfactorily in the oats and wheat classes. As a result, the use of FCSA gives better results for all four classes.

For a more detailed verification of the results, we compared ground truth data (Table 1) with the classified images and assessed the accuracy of each clustering algorithm for land cover classification quantitatively using two statistics, Overall Accuracy (OA), and Kappa Coefficient based on the confusion matrix [2]. Columns in a confusion matrix typically represent the reference data and rows represent the classification data. Overall Accuracy is simply the sum of the pixels classified correctly (e.g., the diagonal elements) divided by the total number of samples in the comparison. The Kappa coefficient can be defined in terms of the confusion matrix as follows:

$$\text{Kappa} = \frac{N \sum_{k=1}^{r} x_{kk} - \sum_{k=1}^{r}(x_{k+} \times x_{+k})}{N^2 - \sum_{k=1}^{r}(x_{k+} \times x_{+k})}, \tag{4.1}$$

where r is the number of rows in the matrix, x_{kk} is the number of observations in row i and column j, x_{k+} and x_{+k} are the marginal totals for row i and column j, respectively, and N is the total number of observations.

Tables 2 and 3 list the results of the comparisons between the ground truth data and the classified images obtained by four clustering algorithms: K-means, ISODATA, FCM, and FCSA. It was noted that FCSA is evolutionary and the results obtained are unlikely to be similar twice, that is, FCSA is nondeterministic; the experiment described above was performed 10 times and the final result obtained were again averaged in tables. From Tables 2 and 3, it is apparent that the FCSA classifier provides better classification results than the other classifiers. The details are as follows: the four classifiers have similar results for the corn class, for which the difference is in the range of 10 pixels. Consistent with the visual classification results, the K-means and ISODATA algorithms have the lowest classification accuracy since they cannot correctly partition the image. FCSA achieves a better classification result for wheat and red clover classes than does FCM, while FCM slightly exceeds FCSA in the oats class by 4 pixels. As a whole, FCSA exhibits the best overall classification accuracy of 92.08% with a gain of 17.3%, 17.3%, and 4.99% over the K-means, ISODATA, and FCM algorithms, respectively. FCSA improves the Kappa Coefficient from 0.6463 to 0.8912, an improvement of 0.2449. One reason for this is that the conventional clustering algorithms often becomes stuck at suboptimal solutions based on the initial configuration of their systems and have a low precision, such as K-means, FCM. Being different from traditional clustering algorithms, FCSA, inspired by immune systems and based on clonal selection
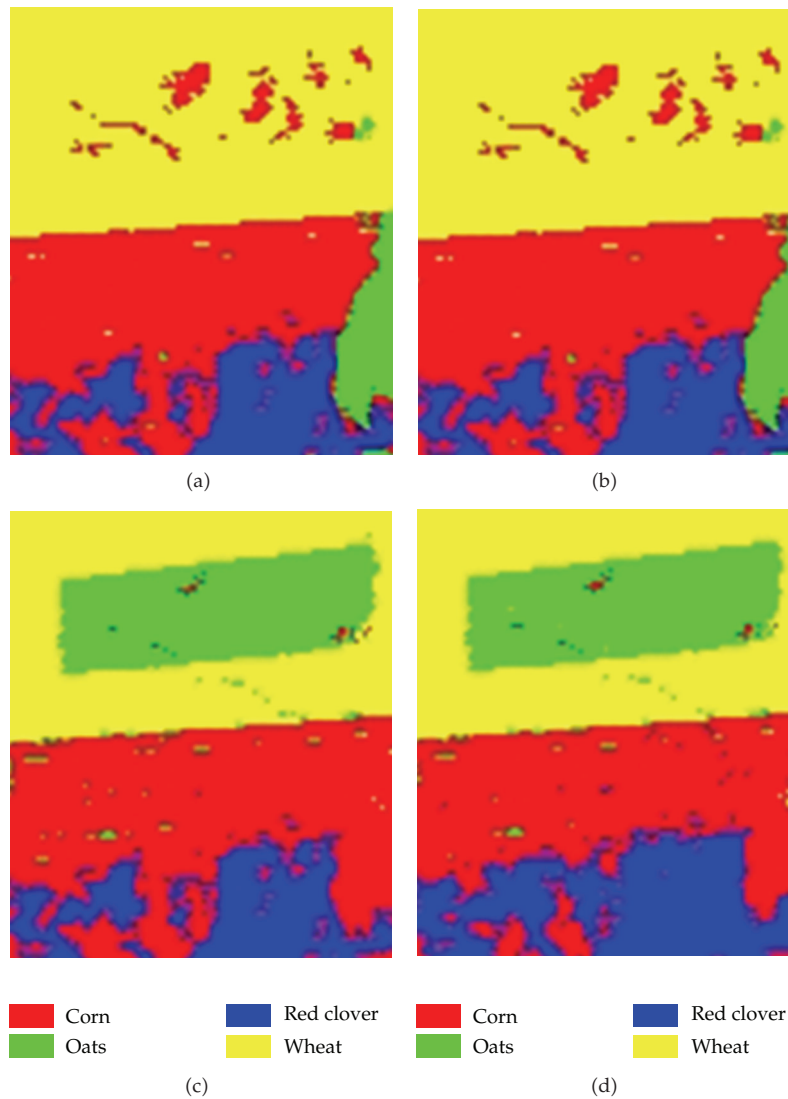
Figure 7: The classification results with four clusters for the Flight C1 image. (a) K-means. (b) ISODATA. (c) FCM. (d) FCSA.

algorithm, is a data-driven, self-adaptive method that can adjust itself to the data without any explicit specification of functional or distributional form for the underlying model. FCSA extends the search space by the process of cloning and quickly finds the optimal solution by the mutation steps. Therefore, FCSA can generate the optimal clustering results to make it flexible in modeling real, complex relationships, which is an important advantage that can adapt to the complex distributions in land cover classification.

In addition, the conventional clustering algorithms require ideal conditions and are sensitive to the initial clustering centers, for example, *a priori* assumptions on the number of clusters. However, because of the complexity of ground substances and the diversity of disturbance, ideal conditions are not often met in real classification calculations. When the number of clusters is incorrectly defined by users, traditional clustering algorithms

Table 2: Comparison of four clustering algorithms in experiment 1.

Methods		Corn	Oats	Red clover	Wheat	Total
K-means	Corn	**1714**	120	583	0	2417
	Oats	16	**0**	93	0	109
	Red Clover	5	0	**1024**	0	1029
	Wheat	8	595	0	**1472**	2075
ISODATA	Corn	**1714**	120	583	0	2417
	Oats	16	**0**	93	0	109
	Red Clover	5	0	**1024**	0	1029
	Wheat	8	595	0	**1472**	2075
FCM	Corn	**1706**	8	674	0	2388
	Oats	31	**707**	0	8	746
	Red Clover	6	0	**1026**	0	1032
	Wheat	0	0	0	**1464**	1464
FCSA	Corn	**1704**	9	388	0	2101
	Oats	20	**703**	0	7	730
	Red Clover	19	0	**1312**	0	1331
	Wheat	0	3	0	**1465**	1468
Total		1743	715	1700	1472	**5630**

Table 3: Comparison of four clustering algorithm performances with four classes in experiment 1.

Accuracy	K-means	ISODATA	FCM	FCSA
Overall accuracy	74.78%	74.78%	87.09%	92.08%
Kappa coefficient	0.6463	0.6463	0.8228	0.8912

find it difficult to obtain satisfactory classification results. By the two-layer system, FCSA can adaptively find the optimal number of clusters to make it appropriate in different real complex conditions, which is another important advantage that adapts to the complex distribution in land cover classification. Therefore, FCSA has the capacity of self-learning and is robust. Based on the above, we can conclude that FCSA is a better clustering algorithm for land cover classification.

4.2. Experiment 2: Wuhan TM

The image data used in this experiment refers to the city area of Wuhan in the central part of China. The image (400×400 pixels) with a spatial resolution of 30 meters was acquired by Landsat-5 on October 26, 1998. The dataset is composed of six spectral bands and their spectral ranges are from 0.45 to 2.35 μm. Figure 8 shows the standard false color composite image of Wuhan TM using bands 4, 3, 2. The values of parameters are set as $M = 20$, MaxIte = 100, $n = M = 20$, $d = 5$. Unlike experiment 1, C_{max} is set to 7 according to the distribution of land cover classes in the image. The parameters in the other traditional algorithms are the same as in experiment 1.

Figure 9 shows the variation of the XB index with the number of clusters when FCSA is used. As can be seen, the minimum value of the XB index is obtained for five clusters with the FCSA algorithm. FCSA automatically yields five clusters and five is the optimal number
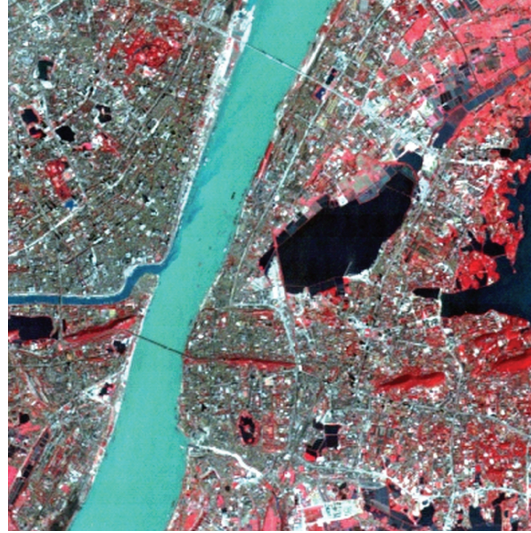
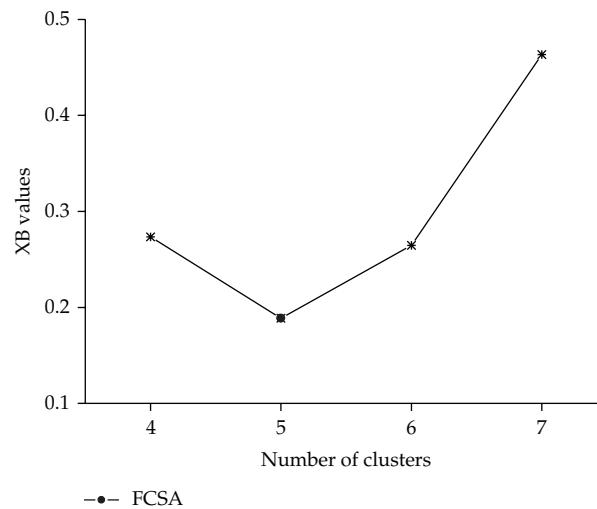**Figure 8:** Wuhan TM Image RGB (4,3,2).

$-\bullet-$ FCSA

Figure 9: Variation of the XB index with the number of classes for the FLC1 image using FCSA.

of clusters. The result is consistent with the real class distribution of land cover based on the ground information available to us. As shown in Figure 8 , some characteristic regions in the image are the well-known Yangtze River cutting across the middle of the image, a city, Wuhan, to both sides of the river. Two parallel lines observed in the middle of the image are the First and Second Bridge over the Yangtze River in Wuhan. The red pixels depict the vegetation classes according to the principles of the standard false color composite. The lakes are found in the right side of the image. The white pixels are known to be roads or open spaces according to visual interpretation experience. Apart from these, there are several water bodies, rare soils, and so forth in the image. Based on the above information, the image is expected to fall into five classes: river, vegetation, lake, building, and road. It can be noted
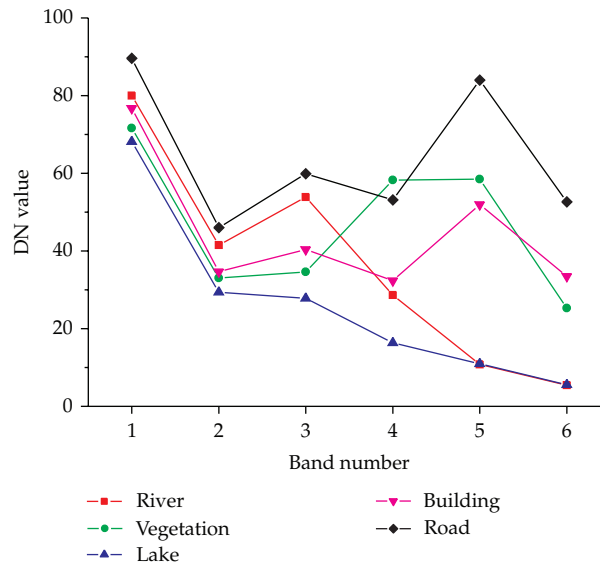
Figure 10: The spectral curves of five land cover classes.
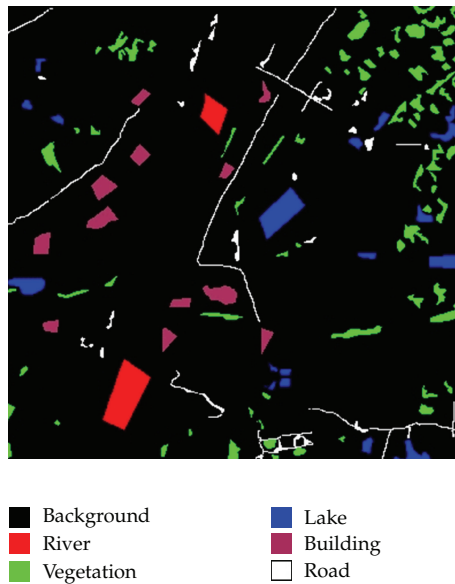

Figure 11: The ground truth data of TM.

that the water class has been differentiated into river (Yangtze River) and lake classes because of a difference in their spectral properties shown in Figure 10 , which displays the spectral curves of the above five land cover classes. Figure 11 displays the field map of the image based on ground truth data. The list of classes and the number of labeled samples for each class are given in Table 4.

Figures 12(a), 12(b), 12(c), and 12(d) show the TM image partitioned using K-means, ISODATA, FCM, and FCSA, respectively, when the number of clusters is set to 5. As can be
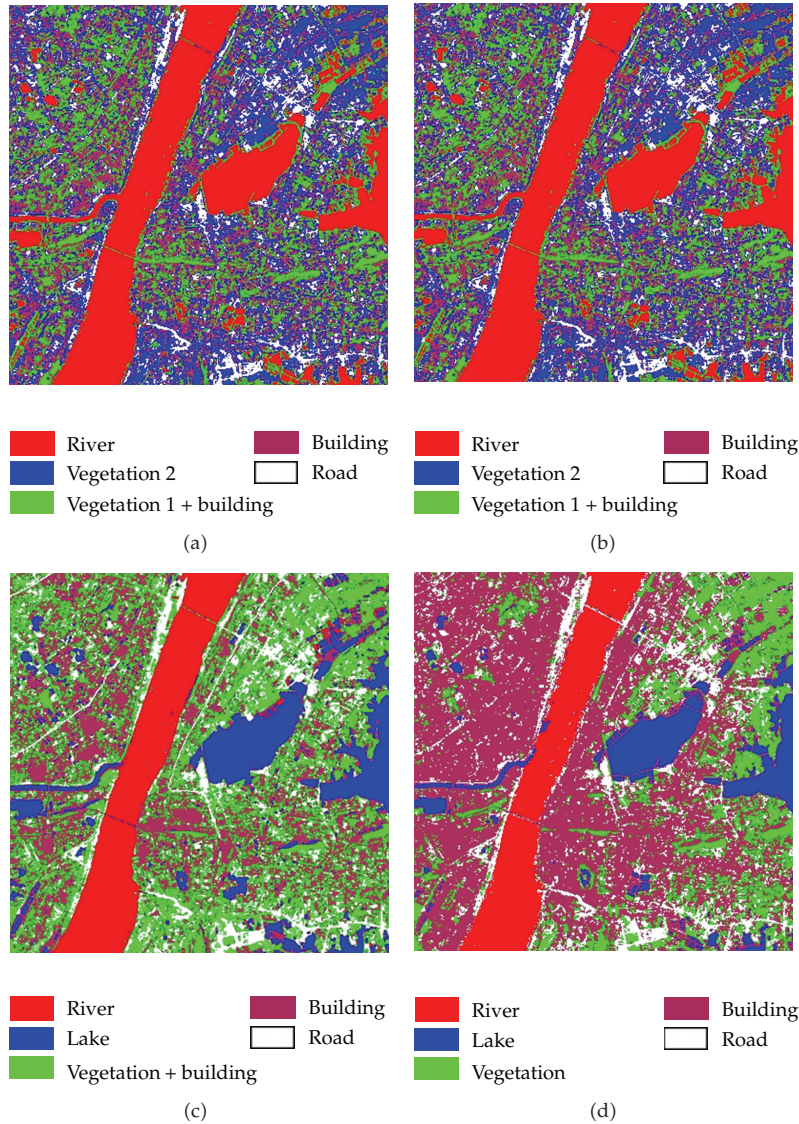
Figure 12: The classification results with for clusters for the TM image. (a) K-means. (b) ISODATA. (c) FCM. (d) FCSA.

seen, the river and lake classes have been incorrectly classified as belonging to the same class in the classification images using K-means and ISODATA. That is, K-means and ISODATA cannot successfully find the lake class. Furthermore, they partition the vegetation class into two subclasses denoted by vegetation 1 and vegetation 2 in Figures 12(a) and 12(b). In addition, K-means, ISODATA, and FCM fare the worst in building classification because many building pixels are misclassified as vegetation. In the classification images, we have put the corresponding label as vegetation+building. Therefore, we can conclude that although some regions, rivers, parts of buildings, and vegetation, and so forth, have been correctly identified, a significant amount of confusion is evident in the clustering results using the

Table 4: List of classes and number of labeled samples in each class for experiment 2.

Class name	Number of labeled samples
River	1774
Vegetation	5007
Lake	2664
Building	2161
Road	2099
Total number of samples	13705

Table 5: Comparison of four clustering algorithm performances with five classes in experiment 2.

Accuracy	K-means	ISODATA	FCM	FCSA
Overall accuracy	41.13%	41.13%	74.27%	80.43%
Kappa coefficient	0.2811	0.2811	0.6703	0.7346

three traditional clustering algorithms. However, FCSA achieves the best visual accuracy in the vegetation and building class and also performs satisfactorily for other classes.

Table 5 lists the results of the comparisons between the ground truth data and classified images obtained by four clustering algorithms: K-means, ISODATA, FCM, and FCSA. From Table 5, it is apparent that FCSA produces better classification results than the other clustering algorithms. The details are as follows: K-mean and ISODATA have the lowest accuracy because the lake class disappears in their classification images. This is evidence again that they are sensitive to the initiation steps. FCSA exhibits the best overall classification accuracy of 80.43%, that is, the best percentage of correctly classified pixels among all the tested pixels, with a gain of 39.3%, 39.3%, and 6.16% over K-means, ISODATA, and FCM, respectively. FCSA improves the Kappa coefficient from 0.2811 to 0.7346, an improvement of 0.4535. These evince that FCSA is a very competent clustering algorithm, which makes it promising for land cover classification.

5. Conclusions

A fuzzy clustering algorithm based on clonal selection for land cover classification, namely, FCSA, was proposed in this paper. Traditional clustering algorithms, such as fuzzy c-means, require the *a priori* specification of the number of clusters and easily fall into a local optimum. The proposed algorithm has attempted to tackle the problems of FCM by use of the clonal selection algorithm to provide near-optimal solutions without *a priori* assumptions of the number of clusters. For this purpose, FCSA is designed as a two-layer system: the classification layer and the optimization layer. In the classification layer, FCSA is used to find the optimal fuzzy partition to a fixed number of classes by the immune operators of the clonal selection algorithm, clone, selection, mutation operators, and so forth, while in the optimization layer FCSA uses the Xie-Beni index as a measure of the validity of the corresponding partition to find the optimal number of classes.

Two experiments were carried out to test the performance of FCSA using Flightline C1 and TM remote sensing images. Compared with three traditional clustering algorithms, K-means, ISODATA, and FCM, only FCSA can adaptively find the optimal number of clusters and FCSA has consistently demonstrated its better performance with the optimal number of clusters. Since K-means and ISODATA cannot correctly partition the image because one class

often disappears and a significant amount of confusion is provided in their classification results, their average Overall Accuracy (OA) and Kappa Coefficient are worst, 57.96% and 0.4637, respectively. FCM improves the average OA and Kappa Coefficient to 80.68% and 0.7466, respectively. The best classification result is provided by FCSA, its average OA and Kappa Coefficient being 86.26% and 0.8129, respectively. These evince that FCSA is applicable for performing the task of land cover classification and has high classification precision. In future work, we will analyze the sensitivity of the proposed algorithm in relation to the parameters, for example, population size, for improving the classification performance and may test FCSA using high-dimensional datasets, such as hyperspectral remote sensing imagery.

Acknowledgments

References

[1] A. Laha, N. R. Pal, and J. Das, "Land cover classification using fuzzy rules and aggregation of contextual information through evidence theory," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 44, no. 6, pp. 1633–1641, 2006.

[2] J. B. Campbell, *Introduction to Remote Sensing*, Taylor & Francis, London, UK, 2000.

[3] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, Wiley-Interscience, New York, NY, USA, 2nd edition, 2001.

[4] D. Hall and G. Ball, "Isodata: a novel method of data analysis and pattern classification," Tech. Rep., Stanford Research Institute, 1965.

[5] G. Storvik, R. Fjørtoft, and A. H. S. Solberg, "A Bayesian approach to classification of multiresolution remote sensing data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 43, no. 3, pp. 539–547, 2005.

[6] T. N. Tran, R. Wehrens, D. H. Hoekman, and L. M. C. Buydens, "Initialization of Markov random field clustering of large remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 43, no. 8, pp. 1912–1919, 2005.

[7] S. Lee and R. G. Lathrop, "Subpixel analysis of landsat ETM using Self-Organizing Map (SOM) neural networks for urban land cover characterization," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 44, no. 6, pp. 1642–1654, 2006.

[8] U. Maulik and S. Bandyopadhyay, "Fuzzy partitioning using a real-coded variable-length genetic algorithm for pixel classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 41, no. 5, pp. 1075–1081, 2003.

[9] J. R. Jensen, *Introductory Digital Image Processing: A Remote Sensing Perspective*, Prentice Hall, Upper Saddle River, NJ, USA, 3rd edition, 2005.

[10] F. M. Burnet, *The Clonal Selection Theory of Acquired Immunity*, Cambridge University Press, Cambridge, UK, 1959.

[11] F. M. Burnet, "Clonal selection and after," in *Theoretical Immunology*, G. I. Bell, A. S. Perelson, and G. H. Pimbley, Eds., pp. 63–85, Marcel Dekker, New York, NY, USA, 1978.

[12] L. N. De Castro and F. J. Von Zuben, "Learning and optimization using the clonal selection principle," *IEEE Transactions on Evolutionary Computation*, vol. 6, no. 3, pp. 239–251, 2002.

[13] D. Dasgupta, *Artificial Immune Systems and Their Applications*, Springer, Berlin, Germany, 1999.

[14] L. N. de Castro and J. Timmis, *Artificial Immune Systems: A New Computational Intelligence Approach*, Springer, London, UK, 2002.

[15] Y. Zhong, L. Zhang, BO. Huang, and P. Li, "An unsupervised artificial immune classifier for multi/hyperspectral remote sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 44, no. 2, pp. 420–431, 2006.

[16] L. Zhang, Y. Zhong, B. Huang, J. Gong, and P. Li, "Dimensionality reduction based on clonal selection for hyperspectral imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 12, pp. 4172–4186, 2007.

[17] X. L. Xie and G. Beni, "A validity measure for fuzzy clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 8, pp. 841–847, 1991.

[18] N. R. Pal and J. C. Bezdek, "On cluster validity for the fuzzy c-means model," *IEEE Transactions on Fuzzy Systems*, vol. 3, no. 3, pp. 370–379, 1995.

[19] Y. Zhong, L. Zhang, and P. Li, "Fuzzy cluster validation for remote sensing image classification," *Geomatics and Information Science of Wuhan University*, vol. 34, no. 4, pp. 391–394, 2009.

[20] N. K. Jerne, "The immune system," *Scientific American*, vol. 229, no. 1, pp. 52–60, 1973.

[21] N. K. Jerne, "Towards a network theory of the immune system," *Annual Immunology*, vol. 125, pp. 51–60, 1974.

[22] A. S. Perelson, "Immune network theory," *Immunological Reviews*, no. 110, pp. 5–36, 1989.

[23] L. N. De Castro and F. J. Von Zuben, "Artificial immune systems: part I-basic theory and application," Tech. Rep. RT DCA 01/99, State Univerity of Campinas, Campinas, Brazil, 1999.

[24] J.-S. Chun, H.-K. Jung, and S.-Y. Hahn, "A study on comparison of optimization performances between immune algorithm and other heuristic algorithms," *IEEE Transactions on Magnetics*, vol. 34, no. 5, pp. 2968–2971, 1998.

[25] F. Campelo, F. G. Guimarães, H. Igarashi, J. A. Ramírez, and S. Noguchi, "A modified immune network algorithm for multimodal electromagnetic problems," *IEEE Transactions on Magnetics*, vol. 42, no. 4, pp. 1111–1114, 2006.

[26] S. Tadjudin and D. A. Landgrebe, "Robust parameter estimation for mixture model," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 38, no. 1, pp. 439–445, 2000.