Land Cover Classification using GA based Fuzzy Clustering Techniques for Remotely Sensed Data

Dr. N. Sujatha

Assistant Professor, Department of Computer Science, Raja Doraisingham Govt. Arts College, Sivagangai, Tamil Nadu, India
sujamurugan@gmail.com

ABSTRACT:
Remote Sensing Imagery is used by the Government and private agencies for the wide range of applications from military to farm development. Fuzzy c-means clustering is an effective algorithm, but the random selection in center points makes iterative process falling into the local optimal solution easily. In this Paper, a novel clustering method is developed using GA based clustering techniques. This technique enables the clustering to be performed by taking the initial centroid using mode function which allows the iterative algorithm to meet to a “better” local minimum. Then the GA based improvement algorithm to get better cluster quality. The study area taken here is the Theni region, Tamil Nadu.

Keywords: Fuzzy C-means, Membership, Mode, Euclidean Distance, Population, Chromosomes, Mutation, Crossover, Selection.

1. INTRODUCTION
Remote sensing [1] can be defined as any process whereby information is gathered about an object, area or phenomenon without being in contact with it. The output of a remote sensing system is usually an image representing the scene being observed. Image classification [2] is an important part of the remote sensing, image analysis and pattern recognition. In some instances, the classification itself may be the object of the analysis. The image classification therefore forms an important tool for examination of the digital images.

Classification strategies are basically divided into three. Supervised Classification [3] techniques require training areas to be defined by the analyst in order to determine the characteristics of each category. Unsupervised Classification searches for natural groups of pixels, called clusters, present within the data by means of assessing the relative locations of the pixels in the feature space. Hybrid Classification takes the advantage of both the supervised classification and unsupervised classification.

2. Clustering Analysis
Clustering analysis [5] is classifying samples according to their similarity by means of unsupervised training. It makes the samples, which have greater similarity, as a class, and occupies the partial area of feature space. The clustering center of each partial area is respectively acting as a representative of the corresponding type. There are various types of clustering. One of its type is Partitional Clustering. Based on Partitional Clustering, many algorithms are used. One of the familiar algorithms is the Fuzzy C-means clustering algorithm.

3. Fuzzy C-means Algorithm
It is an extension of k-means. Fuzzy C-Means [6] allows data points to be assigned into more than one cluster. Each data point has a degree of membership (or probability) of belonging to each cluster. In fuzzy clustering, each point has a
degree of belonging to clusters, as in fuzzy logic, rather than belonging completely too just one cluster. Thus, points on the edge of a cluster may be in the cluster to a lesser degree than points in the center of cluster. For each point \( x \) we have a coefficient giving the degree of being in the \( k \)th cluster \( u_k(x) \). Usually, the sum of those coefficients is defined to be

\[
\forall x \sum_{k=1}^{\text{min. clusters}} u_k(x) = 1.
\]

With fuzzy \( k \)-means, the centroid of a cluster is the mean of all points, weighted by their degree of belonging to the cluster:

\[
\text{center}_k = \frac{\sum_x u_k(x)^m x}{\sum_x u_k(x)^m}.
\]

The degree of belonging is related to the inverse of the distance to the cluster center:

\[
u_k(x) = \frac{1}{d(\text{center}_k, x)},
\]

then the coefficients are normalized and fuzzyfied with a real parameter \( m > 1 \) so that their sum is 1. So

\[
u_k(x) = \frac{1}{\sum_j \left( \frac{d(\text{center}_k, x)}{d(\text{center}_j, x)} \right)^{2/(m-1)}}.
\]

For \( m \) equal to 2, this is equivalent to normalizing the coefficient linearly to make their sum 1. When \( m \) is close to 1, then cluster center closest to the point is given much more weight than the others, and the algorithm is similar to \( k \)-means.

The algorithm [7] steps are as follows

- Select \( m (m > 1) \); initialize the membership function values \( \mu_{ij} \), \( i = 1, 2, \ldots, n; j = 1, 2, \ldots, c \).
- Compute the cluster centers \( z_j \), \( j = 1, 2, \ldots, c \).
- Compute Euclidian distance \( d_{ij} \), \( i = 1, 2, \ldots, n; j = 1, 2, \ldots, c \).
- Update the membership function \( \mu_{ij} \), \( i = 1, 2, \ldots, n; j = 1, 2, \ldots, c \)

\[
\mu_{ij} = \frac{1}{\sum_{l=1}^{c-1} \left( \frac{d_{il}}{d_{ij}} \right)^{2/(m-1)}}
\]

- If not converged, go to step 2.

Several stopping rules can be used. One is to terminate the algorithm when the relative change in the centroid values becomes small or when the objective function cannot be minimized more. The FCM algorithm is sensitive to initial values and it is likely to fall into local optima.

4. Genetic Algorithm

American scholar, J. Holland, first raised the Genetic Algorithm (GA) concept in 1975. It is based on “survival of the fittest” in Darwin’s theory of evolution. The basic genetic operations, which are repetitively utilized for the groups possibly containing solution, make the new groups generated then make them evolved constantly. At the same time, the optimization individuals in optimized groups are searched based on the global parallel search technique so as to obtain the global optimum solution fulfilled demands. GA [4] generates valuable solutions for hard optimization problems using techniques that are inspired by natural evolutionary operators such as inheritance, mutation, selection, and crossover. A common genetic algorithm involves two main parts:

- All solutions should have a genetic representation (in a shape of chromosome).
• There should be a fitness function to assess the solutions.

4.1. Chromosomes

The chromosomes represent set of genes, which code the independent variables. Every chromosome represents a solution of the given problem. Individual and vector of variables will be used as other words for chromosomes. A set of different chromosomes (individuals) forms a generation. By means of evolutionary operators, like selection, recombination and mutation an offspring population is created.

4.2. Selection

The selection of the best individuals is based on an evaluation of fitness function or fitness functions. Fitness value is a quality measurement of each solution. Better fitness values belong to better individuals in each population. When termination criteria are satisfied, algorithm reaches to better fitness value. In the final generation, a solution with better fitness value among others is found as the desired solution.

4.3. Crossover

The first step in the reproduction process is the recombination (crossover). In it the genes of the parents are used to form an entirely new chromosome. The typical recombination for the GA is an operation requiring two parents, but schemes with more parent area also possible. Two of the most widely used algorithms are Conventional (Scattered) Crossover and Blending (Intermediate) Crossover.

4.4. Mutation

The newly created by means of selection and crossover population can be further applied to mutation. Mutation means random change of the value of a gene in the population.

• Generate the initial population.
• Calculate the values of the function that we want to minimize or maximize.
• Check for termination of the algorithm.
• Selection is done between all individuals in the current population are chose those, who will continue and by means of crossover and mutation will produce offspring population
• Crossover – the individuals chosen by selection recombine with each other and new individuals will be created. The aim is to get offspring individuals that inherit the best possible combination of the characteristics (genes) of their parents.
• Mutation – by means of random change of some of the genes, it is guaranteed that even if none of the individuals contain the necessary gene value for the extreme, it is still possible to reach the extreme.
• New generation – the best individuals chosen from the selection are combined with those who passed the crossover and mutation, and form the next generation.

Figure 1. The GA Flowchart
5. Methodology

Land cover is an important component in understanding the interactions of the human activities with the environment and thus it is necessary to be able to simulate changes. This proposal aims at developing a novel land cover clustering method using GA based clustering techniques.

The proposed method has two phases: the first step computes a refined starting condition from a given initial one that is based on an efficient technique for estimating the modes of a distribution. The refined initial starting condition allows the iterative algorithm to converge to a "better" local minimum. And in the second step, a novel method has been proposed to improve to cluster quality by GA based refinement algorithm.

5.1. Preprocessing

Satellite images cannot be given directly as the input for the proposed technique. Thus, it is indispensable to perform pre-processing on the input image, so that the image gets transformed to be relevant for the further processing. In proposed technique, A Median filter which is a non linear filter is used in the R, G and B layers for filtering noise. It is used because, under certain conditions, it preserves edges while removing noise.

5.2. GA Based Clustering

A major problem with Fuzzy C-means algorithm is that it is sensitive to the selection of initial partition and may converge to a local minimum of variation if the initial partition is not properly chosen. So in the proposed method, we estimate the mode value as an initial partition.
Performance optimization using genetic algorithms is given by a sequence of steps, which are:

1. Generate initial population.  
2. Evaluate population  
3. Selection.  
5. Mutation.  
6. Reinsertion of new individuals to the population.

From step 2 to step 6, it performs an iterative process until a stopping criterion is met, in Fig. 3 we can see the Scheme of GA for optimization of the Fuzzy C-Means algorithm (FCM). In this figure we can observe that population evaluation is done by FCM algorithm, but for us to know how good some individuals need something that does not indicate the fitness of these, to measure aptitude of individuals evaluated by FCM, we use the proposed validation index mentioned in section III. Individuals evaluated by the FCM algorithm, are formed only by two parameters which are the number of clusters and the exponent of weight.

Even though the visual comparison gives detailed information for Fuzzy C-means clustering, to further evaluate the performance of proposed work the accuracy assessment has been done. The confusion matrix in terms of pixels and percentage is given in Table 1 and Table 2. The overall classification accuracy is 96.04%.

Table 1. Confusion Matrix (Pixels)

<table>
<thead>
<tr>
<th>Class</th>
<th>Urban</th>
<th>Vegetation</th>
<th>Hilly Region</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Urban</td>
<td>2239</td>
<td>21</td>
<td>4</td>
<td>22642</td>
</tr>
<tr>
<td>Vegetation</td>
<td>20</td>
<td>1680</td>
<td>100</td>
<td>1800</td>
</tr>
<tr>
<td>Hilly Region</td>
<td>7</td>
<td>90</td>
<td>1963</td>
<td>2060</td>
</tr>
<tr>
<td>Total</td>
<td>2266</td>
<td>1791</td>
<td>2067</td>
<td>6124</td>
</tr>
</tbody>
</table>

6. Results

The proposed algorithm is applied to the remotely sensed data (Survey of India toposheets and IRS-1C satellite imageries) of Theni region. Two Theni region images are used to implement the proposed algorithm. The first image is 1152x1152 size tiff image. The second image is 1153x1153 size tiff image. Both images are color image. The Original and the clustered images are shown.
From the Confusion Matrix, it is clear that urban yields a maximum classification accuracy of 98.81% when compared to Vegetation and Hilly Region.

Table 2. Confusion Matrix (Percentage)

<table>
<thead>
<tr>
<th>Class</th>
<th>Urban</th>
<th>Vegetation</th>
<th>Hilly Region</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Urban</td>
<td>98.81</td>
<td>1.17</td>
<td>0.19</td>
<td>36.97</td>
</tr>
<tr>
<td>Vegetation</td>
<td>0.88</td>
<td>93.80</td>
<td>4.84</td>
<td>29.39</td>
</tr>
<tr>
<td>Hilly Region</td>
<td>0.31</td>
<td>5.03</td>
<td>94.97</td>
<td>33.64</td>
</tr>
<tr>
<td>Total</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

Table 3 gives the producer and user accuracy for individual classes. By reducing the misclassification between the Vegetation and Hilly region the overall accuracy can be further improved.

Table 3. Accuracy Assesment

<table>
<thead>
<tr>
<th>Class</th>
<th>Producer accuracy (%)</th>
<th>User Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Urban</td>
<td>98.81</td>
<td>98.90</td>
</tr>
<tr>
<td>Vegetation</td>
<td>93.80</td>
<td>93.33</td>
</tr>
<tr>
<td>Hilly Region</td>
<td>94.97</td>
<td>95.29</td>
</tr>
</tbody>
</table>

7. Conclusion

The proposed method has two phases: the first step computes a refined starting condition from a given initial one that is based on an efficient technique for estimating the modes of a distribution. The developed initial starting condition allows the iterative algorithm to converge to a “better” local minimum. And in the second step, a novel method has been proposed to improve to cluster quality by GA based improvement algorithm.

8. References


