International Journal of Multidisciplinary Research and Development

Volume :2, Issue :4, 406-409
April 2015
www.allsubjectjournal.com
e-ISSN: 2349-4182
p-ISSN: 2349-5979
Impact Factor: 3.762

N.Sujatha
Assistant Professor,
Department of Computer Science, Raja Dhoraisingham Govt. Arts College, Sivagangai, TamilNadu, India

Correspondence:
N.Sujatha
Assistant Professor,
Department of Computer Science, Raja Dhoraisingham Govt. Arts College, Sivagangai, TamilNadu, India

Refinement of land cover classification of satellite images using GA based k-means clustering algorithm

N.Sujatha

Abstract
Today government and private agencies use remote sensing imagery for a wide range of applications from military applications to farm development. Remote sensing image classification is one amongst the most significant application for remote sensing. In this Paper, a novel classification method is developed using GA based clustering techniques. This method enables the clustering to be performed by taking the initial centroid using mode function which allows the iterative algorithm to converge to a "better" local minimum. Then the GA based refinement algorithm to improve the cluster quality. The study area taken here is the Theni region, Tamil Nadu

Keywords: kmeans, Mode, Euclidean Distance, Population, Chromosomes, Mutation, Crossover

1. Introduction

Remote Sensing (RS) [1] refers to the science of identification of earth surface features by measuring portion of reflected or emitted electromagnetic radiation from earth's surface by sensors onboard manmade satellites orbiting around the earth. 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.

There are three basic classification strategies. 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

Cluster Analysis [5] groups data objects based on information found in the data that describes objects and its relationship. The main goal of clustering is that an object within a group be related to one another and different from the objects in other groups. The relationship within a group and the difference between the groups defines the clustering. 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 kmeans clustering algorithm.

3. Kmeans Algorithm

The k-means method [6] aims to minimize the sum of squared distances between all points and the cluster center. This procedure consists of the following steps

Input: k : the number of desired clusters

Output: A set of k clusters Processing

- Select k objects from D as initial cluster centers
- Form k clusters by assigning each object to its closest center
- Recomputed the center of each cluster
- Until centers do not change

Calculate the distance between each object x_i and each cluster center, and then assign each object to the nearest cluster, formula for calculating distance as:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2 ,$$

Where $\| x_i^{(j)} - c_j \|^2$ is a chosen distance measure between a data point $x_i^{(j)}$and the cluster centre C_j, is an indicator of the distance of the n data points from their respective cluster centres.

$$d(x_i, m_i) = \sqrt{\sum_{j=1}^{d} (x_{i1} - m_{j1})}$$

i = 1,2.....N
j = 1,2.....k
$d(x_i, m_j)$ is the distance between data i and cluster j;
Calculate the mean of objects in each cluster as the new cluster centers,

$$m_i = \frac{1}{N_i} \sum_{j=1}^{N_i} x_{ij}$$

i=l, 2 . . . k; N_i is the number of samples of current cluster i.

4. Genetic Algorithm

Genetic Algorithms (GA) are stochastic search procedures introduced by J.Holland in the 70's [7]. These algorithms are based on ideas and techniques from genetic and evolutionary theory which is a field of artificial intelligence and is a kind of searching for good solutions that mimics the process of natural evolution, 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.

4.1 Population

In GAs, there is a population containing a number of solutions which are represented by strings (called chromosomes or the genotype) that evolve in the direction of better solutions. Each string is an encoded candidate solution. Conventionally, solutions are encoded in binary strings of 0s and 1s, but other kinds of encoding models are also probable. The evolution starts by generating several individuals to create an initial population. Then, the fitness function is computed for each individual to produce a selection priority for individuals throughout the generations. Therefore, individuals are preferred from the present population according to their fitness values and modified to a number of offspring. The new population replaces the current population and is used as an input to the next iteration of the algorithm. Usually, the algorithm will be terminated when either maximum number of generations is reached, or a reasonable fitness value is attained.

4.2 Criteria

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.

4.3 Fitness Function

Regarding optimization problems, to produce better solutions from the current one, there should be a fitness function to evaluate the quality of each encoded solution through the generations.

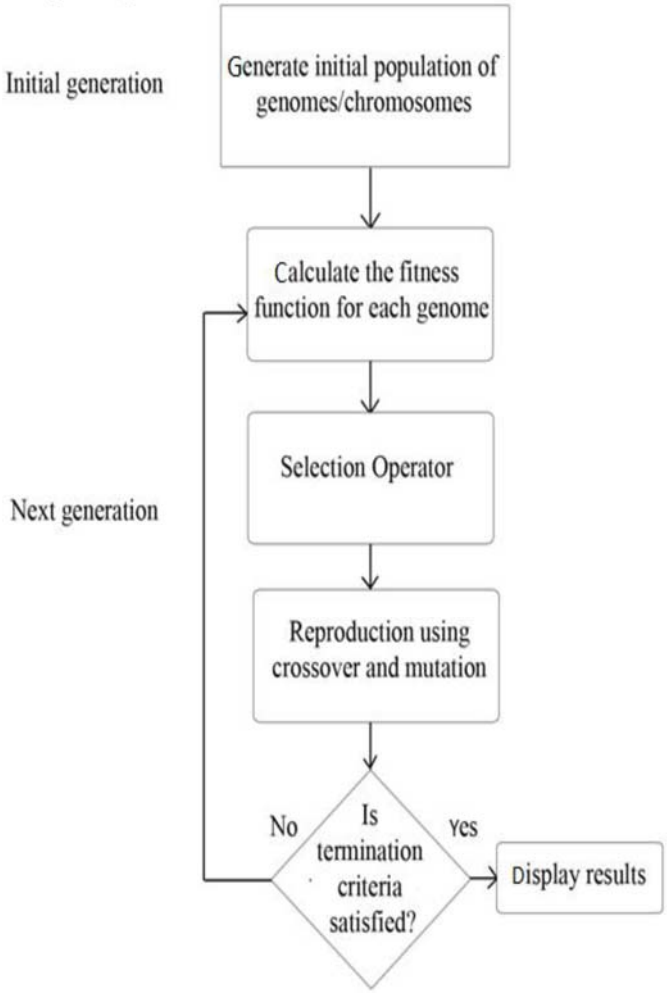

Fig 1: The GA Flowchart

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. 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 classification 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.

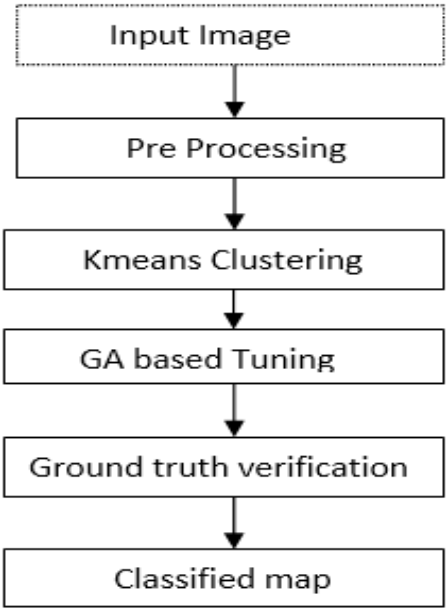

Fig 2: Methodology

Genetic algorithm (GA) is randomized search and optimization techniques guided by the principles of evolution and natural genetics, having a large amount of implicit parallelism. The basic reason for the refinement is, in any clustering algorithm the obtained clusters will never give us 100% quality. There will be some errors known as misclustered. That is, a data item can be wrongly clustered. These kinds of errors can be avoided by using the refinement algorithm.

The genetic algorithms based clustering may not be able to handle large amounts of data. The K-means algorithm does not lend itself well to adaptive clustering. And an important point is, so far, the researchers are not contributed to improve the cluster quality after grouping. In this proposed method, a new framework has been introduced to improve the cluster quality from k-means algorithm. The proposed algorithm is applied to the remotely sensed data (Survey of India toposheets and IRS-1C satellite imageries) of Theni region.

4.4 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.

4.5 GA Based Clustering

A major problem with Kmeans 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.

Initialization:

Way of selecting initial population is the mode value.

Selection:

Selection operator randomly selects a chromosome from the previous population.

Mutation:

The Mutation changes an allele value depending on the distances of the cluster centroids from the corresponding data point. It may be recalled that each allele corresponds to a data point and its value represents the cluster to which the data point belongs operator is defined such that the probability of changing an allele value to a cluster number is more if the corresponding cluster center is closer to the data point. The algorithm steps are

- Construct the initial population (p1)
- Calculate the global minimum (Gmin)
- For i = 1 to N do
- Perform reproduction
- Apply the crossover operator
- Between each parent.

Perform mutation and get the new population. (p2)
Calculate the local minimum (Lmin).
If Gmin < Lmin then
Gmin = Lmin; p1 = p2;
Repeat

After the genetic operators are applied, the local minimum fitness value is calculated and compared with global minimum. If the local minimum is less than the global minimum then the global minimum is assigned with the local minimum, and the next iteration is continued with the new population. The cluster points will be repositioned corresponding to the chromosome having global minimum. Otherwise, the next iteration is continued with the same old population. This process is repeated for N number of iterations.

5 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.

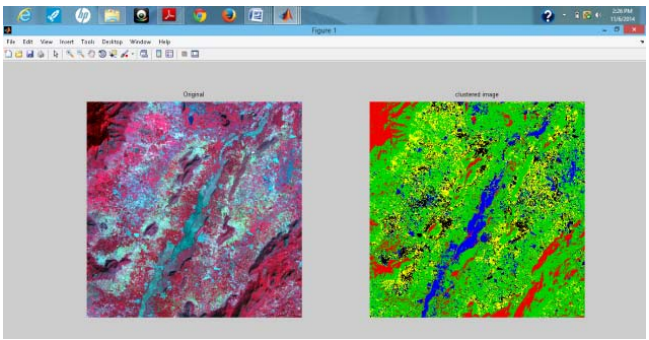

Fig 3: Original & Clustered Theni Region Image1

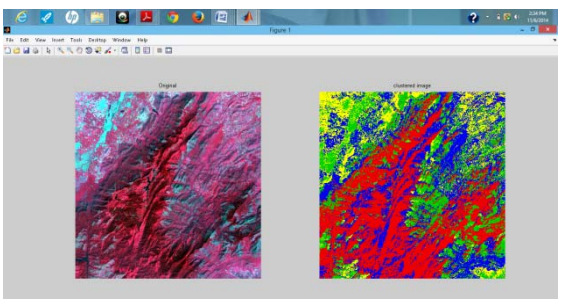

Fig 4: Original & Clustered Theni Region Image2

Even though the visual comparison gives detailed information for Kmeans 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 92.14%.

Table 1: Confusion Matrix (Pixels)

Class	Urban	Vegetation	Hilly Region	Total
Urban	2229	11	2	22542
Vegetation	10	1650	80	1700
Hilly Region	5	70	1863	2120
Total	2166	1591	1967	5124

From the Confusion Matrix, it is clear that urban yields a maximum classification accuracy of 95.75% when compared to Vegetation and Hilly Region.

Table 2: Confusion Matrix (Percentage)

Class	Urban	Vegetation	Hilly Region	Total
Urban	95.75	4.05	4.03	33.97
Vegetation	2.11	90.85	3.10	31.39
Hilly Region	2.14	5.10	92.97	34.64
Total	100	100	100	100

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 Assessment

Class	Producer accuracy (%)	User Accuracy (%)
Urban	95.21	95.90
Vegetation	91.80	91.33
Hilly Region	92.87	94.59

6 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 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.

7. References

1 Minakshi Kumar & R. K.Singh, "Digital Image Processing of Remotely Sensed Satellite Images for Information Extraction".

2 Aykut Akgun et al., "Comparing Different Satellite Image Classification Methods: An Application In Ayvalik District,Western Turkey".

3 Ratika Pradhan et al., "Land Cover Classification of Remotely Sensed Satellite Data using Bayesian and Hybrid classifier", International Journal of Computer Applications (0975 – 8887) Volume 7– No.11, October 2010.

4 K. Moje Ravindra et al., "Classification of Satellite images based on SVM classifier using Genetic Algorithm", IJIREEICE, Vol 2, issue 5, May 2014.

5 R. Balakrishnan et al., "An Application of Genetic Algorithm with Iterative Chromosomes for Image Clustering Problems", IJCSI, Vol. 9, Issue 1, No 1, January 2012.

6 Kitti Koonsanit et al., "Determination of the Initialization Number of Clusters in Kmeans Clustering Application using Co-Occurrence Statistics Techniques for Multispectral Satellite Imagery", IJIEE, Col. 2, No. 5, September 2012.

7 Maryam Gholami Doborjeh, "Genetic Optimization for Image Segmentation".