Image Clustering using GA based Fuzzy c-means Algorithm

A.J.Patil
E&TC Department,
SGDCOE (Jalgaon),
Maharashtra, India

R.R.Karhe
E&TC Department,
SGDCOE (Jalgaon),
Maharashtra, India

M.A.Aher
E&TC Department,
SGDCOE (Jalgaon),
Maharashtra, India

Abstract

Clustering or grouping, is a process that searches for “natural” structure within a data set. It has great theoretical and practical importance in the fields of image processing and pattern recognition applications. The process involves sorting the objects into groups of clusters. The traditional soft (fuzzy c-means) FCM algorithm is sensitive to noise and intensity in heterogeneity. So, by using a hybrid GA-FCM algorithm the limitations of FCM are overcome by the optimum solution of the evolutionary genetic algorithm. Genetic algorithms determine the optimal value of a criterion by simulating the evolution of a population until survival of best fitted individuals. The results of the GA-FCM algorithm are compared with the traditional FCM algorithm. The performance is evaluated by the cluster accuracy rate (CAR).

Keywords: Clustering, FCM, GA, GA-FCM, CAR.

1. Introduction

Clustering of objects is as ancient as the human need for describing the salient characteristics of men and objects and identifying them with a type. Therefore, it embraces various scientific disciplines: from mathematics and statistics to biology and genetics, each of which uses different terms to describe the topologies formed using this analysis. From biological “taxonomies”, to medical “syndromes” and genetic “phenotypes” to “manufacturing “group technology”—the problem is identical: forming categories of entities and assigning individuals to the proper groups within it. The fuzzy C-means algorithm (FCM) has been utilized in a wide variety of image processing applications such as medical imaging, and remote sensing. The algorithm and various modifications of it with focus on practical applications in both industry and science are discussed. Its advantages include a straightforward implementation, its robust behavior, the multi-channel applicability, and the ability to properly model the uncertainty present in the data. FCM is more effective to the fuzzy boundary region segment, but the biggest disadvantage is that no better way to determine the C value of clustering and the initial cluster centers, essentially, FCM is a local search optimization algorithm, it will converge to the local minimum point and this clustering effect would have a greater impact if the initial selection value are not selected properly.

The most important components of the proposed methods concern both the modeling of the problem with GA and the definition of the fitness function. GA can be used in finding the optimal label of each pixel, to determine the optimal parameters of a segmentation method (number of regions for example), or to merge regions of a fine segmentation result. By considering the fitness function, it can be an unsupervised quantitative measure of a segmentation result or a supervised one using some a priori knowledge. Genetic algorithms determine the optimal value of a criterion by simulating the evolution of a population until survival of best fitted individuals. The survivors are individuals obtained by selection, cross-over, and mutation of individuals from the previous generation. We think that GA is a good candidate to find out the optimal combination of segmentation results for two main reasons. The first one is due to the fact an evaluation criterion is not very easy to differentiate. GA is an optimization method that does not necessitate to differentiate the fitness function but only to evaluate it. Second, if the population is enough important considering the size of the search space, there are guarantees that one can reach the optimal value of the fitness. The most widely used method for segmentation of textured image clustering is the Fuzzy C-Means (FCM) algorithm, which is fuzzy logic based when compared to the simple c-means technique. The fuzzy c-means clustering algorithm has been utilized in a wide variety of image processing applications such as medical imaging and remote sensing. It performs a fuzzy partition of a given data set. The advantages of this method are its straightforward implementation, robust in nature, is applicable to multichannel data and the uncertainty data modeling ability. The organization of rest of this paper is as follows: Section 2 gives a detailed discussion of fuzzy c-means clustering algorithm. Section 3 discusses the evolutionary genetic
algorithm. Section 4 proposes the GA based FCM clustering algorithm. Section 5 gives the results evaluated by cluster accuracy rate. Section 6 concludes the clustering result.

2. Fuzzy c-means (FCM)

Fuzzy C-means is a clustering method which allows a piece of data to belong to more than one cluster, which is frequently used in the fields of object recognition, computer vision, and medical image processing. The segmentation results are obtained by using fuzzy classification such as FCM. Unlike the k-means which is a hard classification method which allows a data object to be strictly belong to only single cluster. FCM allows a data object (or pixel) to belong to multiple classes with varying degree of memberships. FCM approach is quite effective for image segmentation. The fuzzy c-means (FCM) clustering algorithm was first introduced by Dunn, (1974) and later was extended by Bezdek, (1981). The algorithm is an iterative clustering method that produces an optimal c partition by minimizing the weighted within group sum of squared error objective function. In FCM, given a data set of size n, X={x1,…,xn} and FCM groups x into c clusters minimizing the weighted within group sum of squared error objective function. In FCM, given a data set of size n, X={x1,…,xn} and FCM groups x into c clusters by minimizing the weighted distance between the data and the cluster centers defined by the objective function as given in equation.

\[ Q = \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}^m \| x_j - o_i \|^2 \]

uij is the membership of data xj belonging to cluster i. Where:

n: the number of patterns in X
c: the number of clusters
uij: the degree of membership Xi of in the jth cluster m: the fuzzification parameter

The FCM iteratively updates the prototypes [o1, o2, …, oc] and membership’s uij through the equations below:

\[ O_i = \frac{\sum_{i=1}^{N} u_{ij}^m \cdot x_i}{\sum_{i=1}^{N} u_{ij}^m} \]

\[ U_{ij} = \frac{1}{\sum_{k=1}^{c} \left( \frac{\| x_i - o_i \|^2}{\| x_i - o_k \|^2} \right)^{\frac{1}{m-1}}} \]

This iteration is stopped until the difference between old membership values and the updated new ones is small enough. Finally based on resulting uij, we assign data xj into the cluster k, where Uk is the largest membership value in all Uij (i=1 to c).

Fuzzy c-means 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. Fuzzy c-means has been a very important tool for image processing in clustering objects in an image. Traditional clustering approaches generate partitions; in a partition, each instance belongs to only cluster. Hence, there exist disjointed clusters in a hard clustering. Fuzzy clustering (see for instance (Hoppner, 2005)) extends this notion and suggests a soft clustering schema. In this case, each pattern is associated with every cluster using some sort of membership function, therefore, each cluster belongs to all patterns related fuzzy set. Larger membership values indicate higher confidence in the assignment of the pattern to the cluster. A hard clustering can be obtained from a fuzzy partition by using a threshold of the membership value.

3. Genetic algorithm

Evolutionary techniques are stochastic general purpose methods for solving optimization problems. Since clustering problem can be defined as an optimization problem, evolutionary approaches may be appropriate here. The idea is to use evolutionary operators and a population of clustering structures to converge into a globally optimal clustering. Candidate clustering are encoded as chromosomes. The most commonly used evolutionary operators are: selection, recombination, and mutation. A fitness function evaluated on a chromosome determines a chromosome’s likelihood of surviving into the next generation. The most frequently used evolutionary technique in clustering problems is genetic algorithms (GAs). A fitness value is associated with each clusters structure. A higher fitness value indicates a better cluster structure. A suitable fitness function is the inverse of the squared error value. Cluster structures with a small squared error will have a larger fitness value. A genetic algorithm for clustering is as follows:

Input:S (instance set), K (number of clusters), n (population size)
Output:clusters

1: Randomly create a population of n structures; each corresponds to valid K- clusters of the data.
2: Repeat.
3: Associate a fitness value structure 2 population.
4: Regenerate a new generation of structures.
5: Until some termination condition is satisfied.
GA is a good candidate to find out the optimal combination of segmentation results for two main reasons. The first one is due to the fact an evaluation criterion is not very easy to differentiate. GA is an optimization method that does not necessitate to differentiate the fitness function but only to evaluate it. Second, if the population is enough important considering the size of the search space, there exists good guarantees that the optimal value of the fitness will be reached.

There are three major design decisions to consider when implementing a GA to solve a particular problem. A representation for candidate solutions must be chosen and encoded on the GA chromosome, fitness function must be specified to evaluate the quality of each candidate solution, and finally the GA run parameters must be specified, that includes which genetic operators to use, such as crossover, mutation, selection, and their possibilities of occurrence. A genetic algorithm is defined by considering four essential data:

1. **genotype**: the segmentation result of an image I is considered as an individual described by the class of each pixel,
2. **initial population**: a set of individuals characterized by their genotypes. It is composed of the segmentation results to combine,
3. **fitness function**: this function enables us to quantify the fitness of an individual to the environment by considering its genotype. The evaluation criteria can be used as a fitness function in the unsupervised case or in and in the semi-supervised cases,
4. **operators on genotypes**: they define alterations on genotypes in order to make the population evolve during generations. Three types of operators are used; selection, mutation and crossover.

### 4. GA-FCM algorithm

GA is a stochastic, powerful and non-linear optimization tool based on the principles of natural selection and evolution. To find the optimum fuzzy partitions of a synthetic test image, a new GA based fuzzy c mean clustering method has been proposed. Clustering using GAFCM can be achieved using the following steps. Here each chromosome in the population of GA encodes a possible partition of image and the goodness of the chromosome is computed by using a fitness function. The system design of a hybrid GA-FCM Algorithm is as shown in figure 2:
The step wise procedure of the implementation of the GA based FCM technique is as follows:

1) The complete feature set extracted from the raw images are given as input to the Genetic Algorithm.
2) An initial population of 30 chromosomes is used in this work. Each chromosome is the candidate solution of this work. Each chromosome is made up of genes of 8 bits corresponding to the input features.
3) The fitness value is estimated for each chromosome using the fitness function which is a maximization function.
4) The least fit chromosomes are removed from the population and new off-springs are generated using the cross-over operation and mutation operation.
5) The cross-over operations are performed between two or more parent chromosomes and the mutation operation is performed within a single chromosome. In both the techniques, swapping of bits is performed to generate the new off-springs.
6) The generated new off-spring may be one of the already existing chromosome (or) a new chromosome.
7) This process is repeated for 100 iterations and the chromosome which is available for the maximum number of iterations is selected as the optimal chromosome.
8) The features corresponding to bit position with the value of ‘1’ is selected as the optimal feature set.
9) These optimal feature set are further given as input to the the fuzzy C-means algorithm.
10) These are further trained and tested in case of neural using these feature set to estimate the performance measures of the classifiers if required.
11) In the case of fuzzy clustering, the FCM algorithm is used with these optimal fitness features and the output centroid values of FCM algorithm are observed.
12) Further, the distance between the unknown testing input’s centroid value and the centroid values of the stored categories are determined. The input is categorized to the class for which the distance value is minimum.

The various parameters involved in this algorithm must be properly fixed for enhancing the success rate of the subsequent steps. The size of the complete feature set is 8. An initial population size of 30 is used in this work. Though, 216 combinations are available, the size of the initial population is reduced for avoiding the computational complexity. The unused combinations are possibly may be generated as one of the off-spring in the subsequent generations. The number of genes used in each chromosome is 8 corresponding to the number of features.

Each bit position is associated with a feature which is extremely important to estimate the optimal feature set at the end of the iterations. Binary representations are adopted in this work for each chromosome. The fitness function carried out is to minimize the objective function by providing optimal threshold values and the higher fitness values indicate the optimal chromosomes. The FCM accuracy in the fitness function for each chromosome is estimated by training the classifier with the features of ‘1’ value in the corresponding chromosome. Since more importance has to be given to the FCM accuracy, a higher value of α is used.

The number of rejected chromosomes per iteration is 9 which is roughly one-third of the total population size. So, the probability of the success for each chromosome is 0.7 which is fair for the entire population. The cross-over rate used for generating the off-springs is 0.7 and the mutation rate used in this implementation is 0.01. The number of iterations employed for stopping criterion is the generation and in this work it is 200.

The error tolerance value used for FCM clustering is 0.001. The number of clusters used in this work is 8. The average number of iterations required for convergence is 200. Thus, these parameters are used in the implementation of the hybrid technique with an objective to enhance the performance of the clustering.

5. Results and discussions

To evaluate performance of the clustering algorithms, we use Clustering Accuracy Rate (CAR) which is defined as,
\[ \text{CAR} = \sum_{i=1}^{C} \frac{A_i \cap C_j}{\sum_{i=1}^{C} C_i} \times 100\% \]

Where \( C \) is the number of clusters, \( A_i \) represents the set of pixels belonging to the \( j \)th class found by the algorithm, while \( C_i \) represents the set of pixels belonging to the \( j \)th class in the reference segmented image. \( \text{CAR} \in [0,1] \), thus the clustering performance is getting better when the value of \( \text{CAR} \) is higher. [7]

Clustering Accuracy Rate can also be calculated as formula below:

\[ \text{CAR} = \frac{\text{Number of correctly classified pixels}}{\text{Total number of pixels}}. \]

5.1. Experiment

Firstly we take a synthetic test image then we corrupt it with 15% salt and pepper noise. In the first image the result of FCM image segmentation is shown. And in the second image the result of GA-FCM image segmentation is shown.

“Fig 3: FCM results for a synthetic test image”.

The figure 3 shows that for the synthetic image corrupted by 15% salt and pepper noise the FCM gives the image clustering with an accuracy of 84.61%.

“Fig 4: GA-FCM results for a synthetic test image”.

The figure 4 shows that for the synthetic image corrupted by 15% salt and pepper noise the GA-FCM gives the image clustering with an accuracy of 100% which results into good clustering.

The performance evaluation by using cluster accuracy rate (CAR) is shown in table below.

“Table 1 Comparison of clustering accuracy rate”.

<table>
<thead>
<tr>
<th></th>
<th>Digital image</th>
<th>FCM</th>
<th>GA-FCM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Salt &amp; Pepper (15%)</td>
<td>84.61</td>
<td>100</td>
<td></td>
</tr>
</tbody>
</table>

Table 1 above shows the cluster accuracy rate of the two algorithms. From the table we can say that GA based FCM provides good cluster accuracy rate when compared with the FCM algorithm.

6. Conclusion

FCM image clustering has been observed to consume a lot of computational resources, has more simulation time and involves a lot of iteration which makes it more expensive to use for image segmentation. To eliminate these drawbacks, a GA approach is used to select features that are most favorable to FCM clustering, which is named as GA-FCM. The result above indicates that Hybrid GA-FCM image clustering has a remarkable improvement on the cluster accuracy rate. FCM in the face of image segmentation showing that GA-FCM has eliminated the drawbacks of FCM.
FCM provides good segmentation results when compared with the other hard and soft clustering algorithms. For comparison we have used the cluster accuracy rate (CAR) which shows the best performance of our GA-FCM algorithm.

7. References


