Generalized Fuzzy Clustering Model with Fuzzy C-Means

Hong Jiang[1]

[1] Computer Science and Engineering, University of South Carolina,
Columbia, SC 29208, US
jiangh@cse.sc.edu
http://www.cse.sc.edu/~jiangh/

Abstract. This paper extends the traditional Fuzzy C-Means clustering method to a generalized fuzzy clustering model. According to most applications, this fuzzy clustering model briefly includes 3 parts: feature extractor transfers original objects information to desired feature data; fuzzy cluster analyzer gets cluster information from the feature data; and post treatment obtains the final results based on the cluster information. Among them, fuzzy cluster analyzer is encapsulated to 5 parts instead of traditional E-step and M-step: Initialization, U^{α}, E-step, Distance Calculation, and M-step. This model makes each part keep relatively independent and easy to improve by just replacing one or several parts in needs. An implementation of this model is supplied, and 3 examples are given to test the properties. Moreover, potential optimizations are analyzed and listed.

1 Introduction

The capacity of classifying patterns is one of the most fundamental characteristics of human intelligence. Cluster Analysis or clustering thus becomes to one of the most fundamental issues in it. Since the early 1950s, pattern recognition has become to a field of study. In the mid-1960s, fuzzy set theory started to be used in pattern recognition and cluster analysis. Now, the literatures involving fuzzy clustering are already quite extensive, partly because the vague boundaries are desirable for most categories we commonly encounter.

On the other hand, the applicability of cluster analysis is not necessarily restricted to the pattern recognition. It could be used to classify documents in information retrieval, or used in social groupings based on various criteria. In one word, it is applicable in most areas only if you want to classify some objects into several categories.

Further more, there are lots of researches on the fuzzy clustering. However, most of them focus on the optimization on some fuzzy clustering algorithms or application in some special cases. Personally, I did not find any literatures or research involving generalizing the fuzzy clustering model in convenience of applications and researches.

Thus, based on most applications and researches, this project focuses on building up a generalized model for fuzzy clustering. This model reorganizes the traditional idea of fuzzy clustering, and extent it to make it suitable for most applications. It is well capsulated so that each part keeps enough independence and flexibility. This design also takes most potential optimizations into account. That is, the way that this model is capsulated is suitable for most possible optimizations. Based on this model, researchers can focus on the study more by simply replacing one small part in this model.

To test the working status of this model, an implementation is supplied, and 3 experiment results are given in this paper as well. According to the structure of the model and the experiment results, potential optimizations are also analyzed and listed in this paper.

2 Generalized Fuzzy Clustering Model

This section will introduce the basic idea of the generalized fuzzy clustering model, and the details of each part in this model.

Before that, I would like to give some useful definitions first:

- Pattern Recognition: A process by which we search for structures in data and classify these structures into categories such that the degree of association is high among structures of the same category and low between structures of different categories [3].
- Clustering: Given a finite set of data X, the problem of clustering in X is to find several cluster centers that can properly characterize relevant classes of X [3].

First of all, this generalized fuzzy clustering model is based on one of the fuzzy clustering algorithm ---- Fuzzy C-means. The goal of building this model is to extend the traditional fuzzy c-means to a generalized model in convenience of application and research.

2.1 Fuzzy C-Means

The basic idea of fuzzy c-means is to find a fuzzy pseudo-partition to minimize the cost function.

A brief description is as follows:

$$\begin{aligned}&\text{Cost function}: \sum_{i=1}^{n}\sum_{k=1}^{K} u_{ik}^{\alpha}\left\|\mathbf{x}_i - \mathbf{m}_k\right\|^2, \alpha > 1 \\ &s.t. \quad \sum_{k=1}^{K} u_{ik} = 1, \quad 0 \le u_{ik} \le 1\end{aligned} \qquad (1)$$

In above formula, x_i is the feature data to be clustered; m_k is the center of each cluster; u_{ik} is the fuzzy partition corresponding to the feature data; n describes the number of the feature data; K is the number of the clusters; and a is the exponent used to adjust the fuzzy degree. Generally, a should be greater than 1, and when a is tend to infinity, the fuzzy degree is increasing. This cost function is used as a control on the updating. That is, we get final result u_{ik} and stop the updating by minimizing the cost function. Moreover, the u_{ik} has the range from 0 to 1 is the main difference with hard c-means which can only have value 0 or 1.

The updating steps are defined as:

$$E-Step: \quad \mathbf{m}_k = \frac{\sum_{i=1}^{n} u_{ik}^{\alpha} \mathbf{x}_i}{\sum_{i=1}^{n} u_{ik}^{\alpha}} \tag{2}$$

$$M-Step: \quad u_{ik} = \frac{1}{\sum_{l=1}^{k} \left(\frac{\|\mathbf{x}_i - \mathbf{m}_k\|}{\|\mathbf{x}_i - \mathbf{m}_l\|} \right)^{\frac{1}{\alpha-1}}} \tag{3}$$

E-Step is used to obtain the new center of each cluster and M-Step is used to update the fuzzy partition. By repeating E-step and M-step, cluster center m and fuzzy partition u are updated, until the cost function reaches the minimal value, or can't be reduced anymore, we can get the final cluster information.

2.2 Generalized Fuzzy Clustering Model with FCM

To make the model suitable for most applications and convenient to optimization, the generalized fuzzy clustering model with fuzzy c-means is separated to 3 function parts and 4 data flow parts as in figure 1.

For the functions parts, the first one is feature extractor, which is used to transfer original objects information to desired feature data; the second one is the fuzzy cluster analyzer, which gets cluster information from the feature data; the last part is post treatment, which is used to obtain the desired final results based on the cluster information.

The data flow parts are connected by above 3 function parts. The first one is original objects information, or the representation of input data, which is obtained by measurements on objects that are to be recognized. It may be any kind of data information in any kind of data structure. The next is feature information; it is the characteristic features extracted from the input data in terms of which the dimensionality of pattern vectors can be reduced. The features should be characterizing attributes by which the given pattern classes are well discriminated. The third part is cluster information, or category information, which is obtained through cluster analysis. The last one is goal object information, it is the final desired result, for some special cases, it may not be necessary always.

In this model, the fuzzy cluster analyzer is the key part and most work is also focused on this part. In convenience of optimization, it is encapsulated to 5 parts instead of traditional E-step and M-step only. It is as in figure 1 as well.

For the fuzzy cluster analyzer, the input data mainly includes 3 data information. The first part is the feature data to be clustered, with dimension (n x d). n describes how many data want to be clustered, and d is the number of dimension of each feature data. The second input is cluster number K, which describes how many clusters are desired for the clustering. The next is exponent, which is used to adjust the degree of fuzzy, usually grater than one. If it's one, the result is equal to the hard c-means or k-means. In most applications, the exponent value is 2.

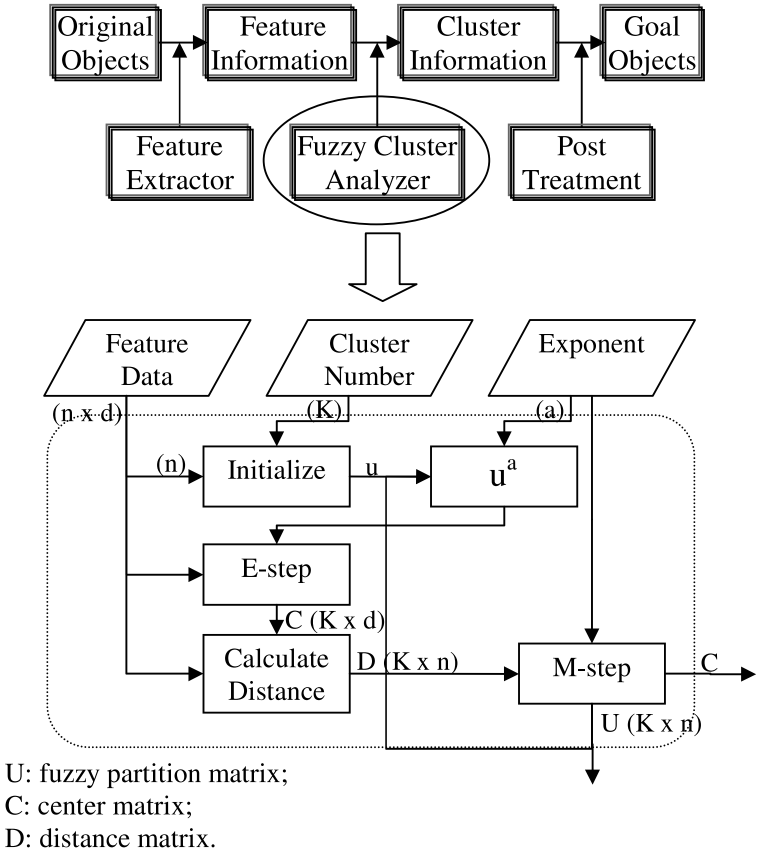


Fig. 1. Structure of Generalized Fuzzy Clustering Model with FCM. The lower part is the flowchart to describe the fuzzy cluster analyzer. Among it, n describes the number of the feature data; K is the number of the clusters; and a is the exponent used to adjust the fuzzy degree.

The function parts for the fuzzy clustering analyzer are separated to 5 parts as follows. The first one is initialization, which generate initial fuzzy partition matrix for clustering. The second one is u^a, which get the matrix after exponential modification. E-step is used to get the new center matrix, as described in formula (2). The forth part is distance calculation, which calculates the distance of the cluster center with input feature data. In general case, the Euclidean distance is used. The fifth part is M-step, which get new fuzzy partition matrix and cost function value. Among them, the cost function value is used to control the iterations in the implementation. The final fuzzy partition matrix U is the result that we want, which includes the information of cluster and with dimension (K x n). K is the cluster number, and n is the number of the feature data.

2.3 How Does the Model Work?

Assume what we have is the original object information. It could be some images with digital information or some continues signals or some other kind of information. Based on some criteria or prior knowledge, we need to classify it to several categories or several regions.

The detail steps for how this model works are described as follows:

Step 1. Input the original object information to the feature extractor. The feature extractor extracts feature data based on the criteria or prior knowledge. It could be briefly considered as a kind of mathematical transformation. The desired feature data should with exact form ---- a matrix with dimension: n (feature data number) x d (feature dimension). The features should be characterizing attributes by which the given pattern classes are well discriminated.

Step 2. Input the obtained feature data and desired cluster number into the fuzzy cluster analyzer. If needs be, you could also adjust the fuzzy degree by inputting the exponent value, and set termination condition to control the iterations.

Step 2.1. Initialize the fuzzy partition matrix U, small letter u is used to describe the element of the matrix here;

Step 2.2. Calculate the matrix after exponential modification ---- U^a;

Step 2.3. Calculate the cluster center matrix C with dimension (K x d), among which K is the desired cluster number and d is the dimension of feature;

Step 2.4. Calculate the distance between center and input feature data. Obtained distance matrix is with dimensions same with the fuzzy partition matrix.

Step 2.5. Calculate new fuzzy partition matrix and cost function value. If the difference of the cost function value is less than or equal to termination condition, stop; otherwise, repeat step2.2. to 2.5. until cost function value satisfies the termination condition.

Step 3. Based on above obtained fuzzy partition matrix and requirements for each special case, do post treatment to get the goal object. In most cases, this post treatment could be considered as the inverse transform of the first step.

3 Potential Optimizations

Following lists the main potential optimizations:

1. Optimizations on feature data:
   - Feature data obtaining: well discriminated feature data is also the key to obtain good clustering result. Thus, feature data should best describe the difference between the clusters. How to obtain this feature data is always an important research issue.

- Normalization problem: un-normalized data often makes the clustering time-consuming, and some time even make the cost function very hard to reach the minimal point.

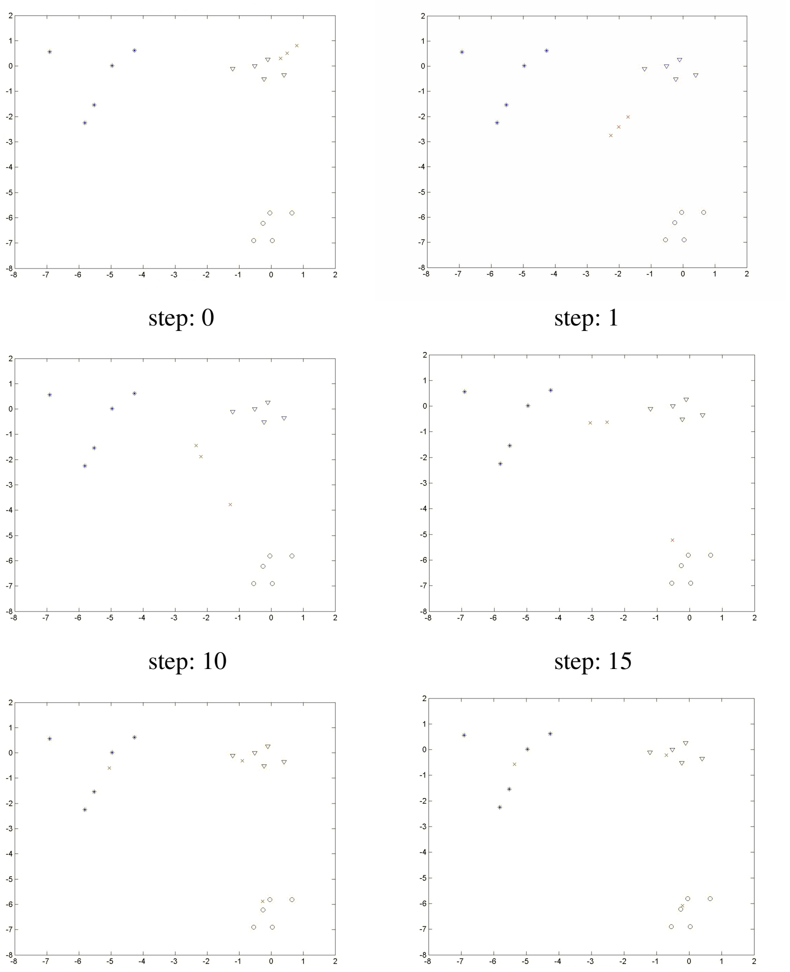
2. Optimization on cluster number:
   - How to determine the cluster number is always a problem. In many cases, the desired cluster number is not so clear. We could do some research to determine the cluster number.
3. Optimizations on U^a
   - The exponent is used to adjust the degree of fuzzy. Generally, when it is close to 1, the fuzzy c-means converges to hard c-means. When it tends to infinity, all cluster centers tend towards the center of the data set. However, there is no theoretical basis for an optimal choice. Thus, some research could focus on it.
   - On the other hand, to calculate an item with exponent is time-consuming in computer, there is probably another way to define it.
4. Optimization on distance computation:
   - Generally, we use the Euclidean distance in traditional fuzzy c-mean, but you might also define other kind of distance.

4 Application Example and Result

This section shows 3 examples to test the working status of this generalized fuzzy clustering model.

4.1 Example 1

This example focuses on testing the working of the fuzzy cluster analyzer. 15 feature data with 2 dimension feature are supplied directly. This data set are displayed as in following figures with sign “*”, “o” and “v” to show the desired clusters. The cluster number is 3, exponent is 2 here. The cluster center is described by “x” sign. The step 0, 1… show the iterations number of the updating. The data sets are with 3 initialized centers in step 0, and the others subfigure with steps show the centers’ moving with the repeat steps, or the iteration of the updating. The details as in following figure 2.



step: 0

step: 1

step: 10

step: 15

step: 20

step: 25

Fig. 2. This figure shows how the cluster center moves to get desired clusters. The data set are displayed as in the figures with sign "*", "o" and "v" to show the desired clusters. "x" signs describe the 3 cluster center. The cluster number is 3, exponent is 2 here. Step 0 shows the data sets with 3 initialized centers, the others show the centers' moving with the repeat steps.

The final fuzzy partition result is as follows:

0.0031	0.9952	0.0017
0.0161	0.9735	0.0105
0.0230	0.9650	0.0120
0.0006	0.9991	0.0004

0.0175	0.9701	0.0124
0.0856	0.0562	0.8583
0.0829	0.0365	0.8806
0.1562	0.0343	0.8096
0.0272	0.0083	0.9645
0.0362	0.0185	0.9453
0.9942	0.0023	0.0035
0.9660	0.0141	0.0200
0.9308	0.0347	0.0345
0.9777	0.0072	0.0151
0.9788	0.0097	0.0114

From above figure and results, we can find that: first, a data point may belong to multiple clusters with different degrees of membership. That is, it reflects some kind of fuzzy. Next the sum of the degrees of memberships is 1. This reflects the requirement of clustering. Moreover, the fuzzy partition value reflects the how close it belongs to the cluster. Thus, from the greatest value, we can get the cluster result.

Above all, the result shows that this model works great on this case.

4.2 Example 2

This example focuses on the working of the whole parts together. It works on an image, and this image displays some information of bone structure. In this case, the feature data is the histogram extracted based on the gray value of each pixel in the image. That is, the feature extractor here works just to get the histogram information of the image. The fuzzy clustering analyzer is using the reorganized fuzzy c-means as described in this paper. The post treatment can be looks as the inverse of the first step, just redefine each pixel with a new gray value based on the cluster information obtained through the fuzzy clustering analyzer.

The testing results are as in following figure 3. The original image is as in subfigure (a). Subfigure (b) shows the results with cluster number 2, and the subfigure (c) shows the results with cluster number 4. The value of exponent is 2 here also.

From the results, we can see that the whole parts do work great. While, there are also some noise existed in the results. This is due to the noise in the original image. If we could do some pretreatment to smooth it before we extract the feature information, the result might be improved a lot. So, in the real applications, the pretreatment sometime is very important too. However, in this example, I just want to simply test the whole model's working. Thus, the result is ok.

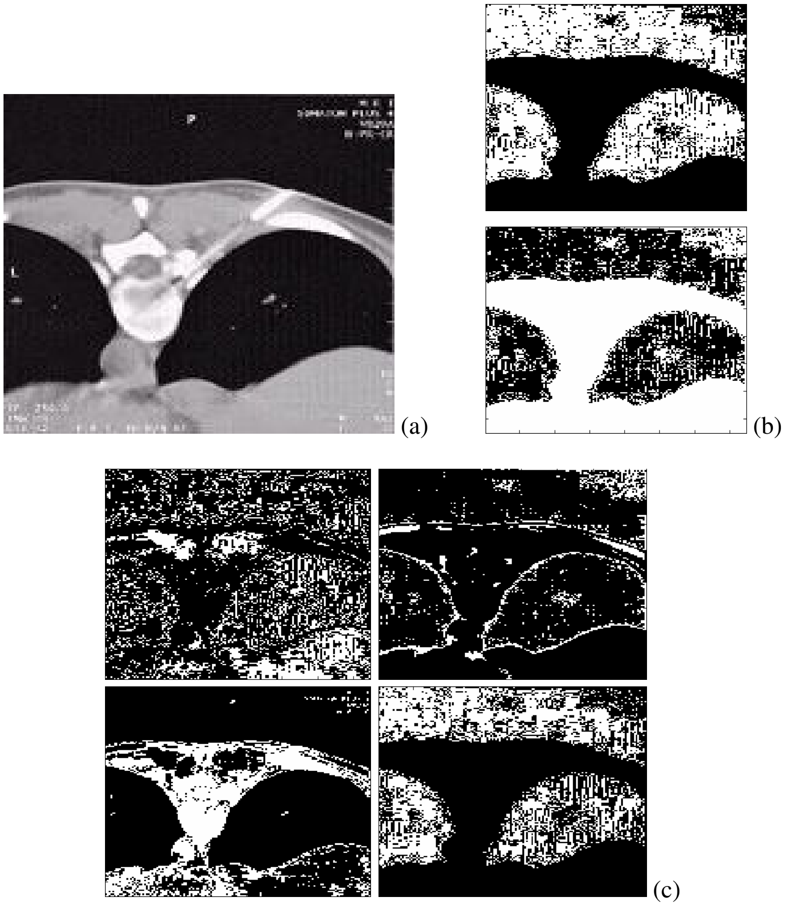


Fig. 3. (a) Original image; (b) Clustering result with 2 clusters; (c) Clustering result with 4 clusters.

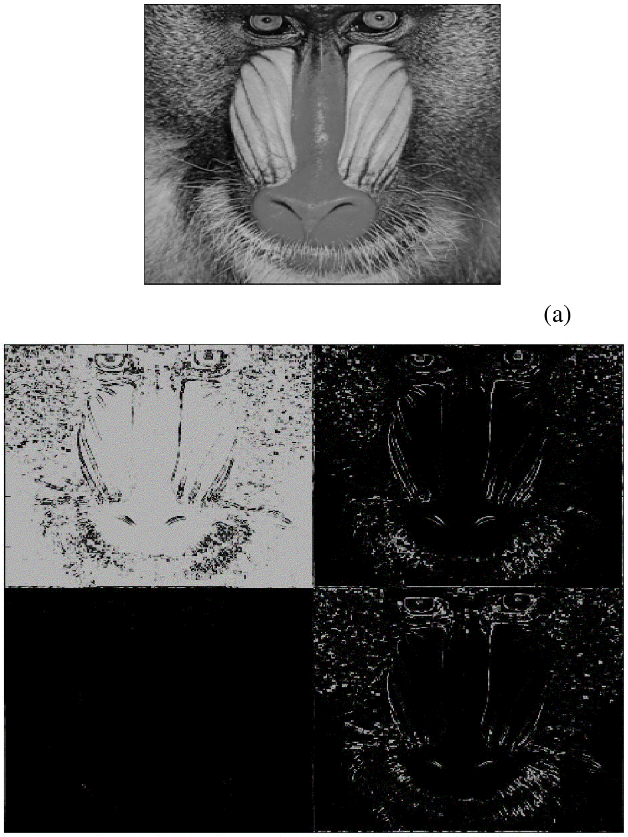
4.3 Example 3

This example is used to test whether each part in this model keeps flexible and independent enough, so that each part can be replaced by other similar part. Results are as follows figure 4.

The image, the feature data and the post treatment method used in this test are directly obtained from internet[1]. The feature is based on some texture instead of the pure gray value. That is, comparing with example 2, the feature extractor is replaced, and

[1] http://vulcan.ee.iastate.edu/~dickerson/classes/ee571x/homework/hw4soln/hw4.html

so is the post treatment part. Though we don't know exactly how the feature is obtained based on the texture information, we can still replace the corresponding part, and test the working status.


(a)

(b)

Fig. 4. (a) Original image; (b) Clustering result with cluster number 4. The image, the feature data and the post treatment method used in this test are directly obtained from internet, but the result shows that it does work fine with the replaced parts.

From above results, we can see that the model also works great with replaced parts. On this way, the result does reflect some kind of texture information. For example, comparing the image on the upper-right corner with the one on the lower-right corner, the former focuses on the texture information of the face, while the later focuses more on the texture information of the eyes.

Above all it does keep some kind of independence, and works fine with the replaced parts.

5 Conclusion

In this paper, to simplify applications and research, I extend the traditional Fuzzy C-Means clustering method to a generalized fuzzy clustering model. This generalized fuzzy clustering model is simplified to 3 function parts: Feature extractor, Fuzzy cluster analyzer and the Post treatment, and 4 data flow parts: original object information, feature data, cluster information and goal object information. Among the 3 function parts, the fuzzy cluster analyzer is based on Fuzzy C-Means, and encapsulated to 5 parts instead of traditional E-step and M-step.

An implementation of this model is supplied, and 3 examples are given to test the working status of it. The properties of this model and the test results show us that:

1. This model could be used to most applications;

2. Each part of this model is well capsulated, keeps independent and flexible. Each part could be replaced by some other corresponding function parts with similar function.

3. It is convenient to potential optimizations. We just need to change a simple unit in this model to realize some optimization.

More over, some major potential optimizations are analyzed and listed in this paper, future optimizations or research may be based on them.

References

1. A. Baraldi, and P. Blonda, "A survey of fuzzy clustering algorithms for pattern recognition (1998)", IEEE Transactions on Systems, Man and Cybernetics, Part B (Cybernetics). http://citeseer.nj.nec.com/baraldi98survey.html
2. F. Masulli, A. Schenone, and A.M. Massone, "Fuzzy clustering methods for the segmentation of multimodal medical images", www.ge.infm.it/~masulli/papers/masulli-fsm2000.pdf
3. J.K.George, and Y. bo, Fuzzy sets and fuzzy logic theory and applications, New Jersey: Prentice Hall, 1995
4. J. Jantzen: "Neurofuzzy Modelling", Technical University of Denmark: Oersted-DTU, Tech report no 98-H-874 (nfmod), 1998. http://fuzzy.iau.dtu.dk/download/nfmod.pdf
5. Yingkang Hu, and Richard J. Hathaway, “On Efficiency of Optimization in Fuzzy c-Means", http://www.cs.gasou.edu/faculty/hu/publications/OptFCM-NPSC.pdf