CLUSTER VALIDITY FOR FCM CLUSTERING ALGORITHM USING UNIFORM DATA

Sergio López
Telefónica I+D
Emilio Vargas 6, 28043, Madrid, (SPAIN)
slg@tid.es

Luis Magdalena **Juan R. Velasco**
ETSI Telecomunicación (UPM)
Ciudad Universitaria s/n. 28080, Madrid , (SPAIN)
llayos@mat.upm.es juanra@gsi.dit.upm.es

Abstract

One of the main drawbacks of the FCM clustering algorithm is that it does not calculate the suitable number of clusters. This paper presents a method to solve this problem, by means of an equalization function (using uniform data) for the FCM functional *J*. The results for 2 and 3 dimensional data tests are also presented.

Keywords: Clustering, fuzzy logic, uniform data

1 INTRODUCTION

A clustering algorithm is a mathematical tool that detects similarities in a collection of data. One of the most widely used in different fields, such as pattern recognition, data analysis and image processing is the fuzzy c-means [1] clustering algorithm.

However, FCM has several drawbacks, such as the dependence of the results on the initialization and the need to predefine the number of clusters to be generated. In [2] the authors have presented a genetic fuzzy c-means (GFCM) clustering algorithm, that solves the initialization problem using genetic algorithms. In addition to this, [2] introduced a way of finding the suitable number of clusters for an unidimensional problem. In this paper the authors will deal with the problem of finding the number of clusters for higher dimensions.

2 STANDARD FCM CLUSTERING ALGORITHM

Consider a set of n objects X = {x_1, x_2,...,x_n} where $x_i \in R^S$. Each x_i is an object that is described by *s* real-valued measurements of their features. A fuzzy c-partition of *X* is a class of c fuzzy sets V_1, V_2, ..., V_C, where *c* is an integer in the range [2,n]. Then a fuzzy c-partition space for *X* is the set:

$$M_{fcn} = \left\{ U \in R^{cxn} \middle| u_{ik} \in [0,1], \forall i,k; \sum_{i=1}^{c} u_{ik} = 1, \forall k; 0 < \sum_{k=1}^{n} u_{ik} < n, \forall i \right\} \quad (1)$$

The aim of the algorithm is to find the best partition matrix U in M_{fcn}. This objective is reached when the following function is minimized:

$$J_M(U,V) = \sum_{k=1}^{n} \sum_{i=1}^{c} u_{ik}^m \cdot d_{ik}^2(v_i, x_k), \quad U \in M_{fcn}, 1 < m < \infty \quad (2)$$

In this function, v_i are the prototypes (or cluster centroids) of each class, *m* is a weighting exponent and *d* is the Euclidean distance. The cluster centroids are obtained with the following expression:

$$v_i = \frac{\sum_{k=1}^{n} u_{ik}^m \cdot x_k}{\sum_{k=1}^{n} u_{ik}^m} \quad (3)$$

The steps of FCM algorithm are the following:

1. Select an initial partition $U^{(0)}$.
2. Compute the cluster centroids v_i.
3. Update the fuzzy membership

$$u_{ik} = \frac{\left(\frac{1}{d_{ik}^2(v_i, x_k)} \right)^{1/(m-1)}}{\sum_{j=1}^{c} \left(\frac{1}{d_{jk}^2(v_j, x_k)} \right)^{1/(m-1)}} \quad (4)$$

4. Repeat steps (2) and (3) until the value of J_M is no longer decreasing.

3 CONCEPT OF CLUSTER SPECIES

One of the main drawbacks of the FCM clustering algorithm is that the number of clusters c is fixed in the beginning of the algorithm. It can be assumed in several problems, where the number of clusters is known. However, there are many problems where is necessary to find the number of clusters.

In [5] the authors have presented the idea of cluster species, based on natural genetics. Given a certain set of solutions to a clustering problem, a cluster species is composed of the solutions having the same number of clusters.

The last step of the process is to apply a quality function to the better individual of each species. This function determines how strong each species is. It will be presented in the next section.

4 QUALITY FUNCTION FOR CLUSTER SPECIES

In FCM, the value of the functional J_M determines how good a clustering partition is. However, due to the definition of J_M, this value decreases as the number of clusters increases.

The solution to this problem is to find a reference case, where there are not any substructures in data. In this way, as the results for any number of clusters must be equivalent, it can be found an equalization function g such as:

$$g(c) \cdot J_m^c(U,V) = K \tag{5}$$

In [2] it was found that the uniform distribution of data is the solution for the unidimensional case. The equalization function g, considering the case of $m=1$, is

$$g(c) = c^2 \tag{6}$$

In the next subsections the 2 and 3-dimensional cases will be dealt with.

4.1. 2-DIMENSIONAL CASE

First of all, it is needed to establish the geometric figure that will define the uniform distribution of points. According to [6], we have selected the circle[1].

In order to find g, let us start from the general expression of J_M

$$J_M(U,V) = \sum_{k=1}^{n} \sum_{i=1}^{c} u_{ik}^m \cdot d_{ik}^2(v_i, x_k) \tag{7}$$

considering the case of $m=1$ and X being a family of objects uniformly distributed inside the unit circle with a infinite number of elements. The first assumption implies that the solutions will be crisp clusters (see [2]) while the second assumption transforms the finite sum on k to a double integral inside the circle.

As the clusters are crisp, u_{ik} is 1 for the cluster k, 0 for the other clusters and ½ in the intersection between clusters.

[1] Another shapes (for instance, a square) will need a further study.

We have studied two situations:

1. For c between 2 and 5, the area or each cluster has the shape shown in Figure 1.
2. For c=6 and c=7, a cluster in the origin turns up. Figure 2 shows an example.

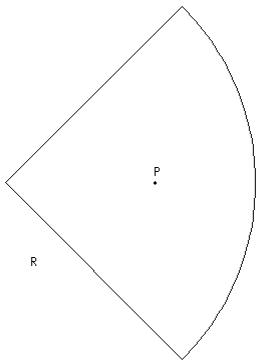

Figure 1: Shape of clusters in situation 1

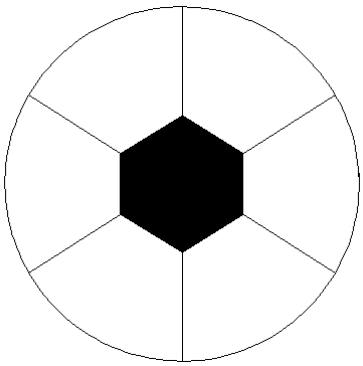

Figure 2: Regions in situation 2

Let us work in the situation 1. The formulation is (where R is the region where the cluster has points):

$$J_2(U,V) = \sum_{i=1}^{c} \iint_R d_{ik}^2(v_i, x_k) \tag{8}$$

As all the clusters are equal, we can eliminate the sum, having:

$$J_2(U,V) = c \iint_R d_{ik}^2(v_i, x_k) \tag{9}$$

Then the centroid P can be placed in the x axis, so that:

$$J_2(U,V) = c \iint_R \left((x-P)^2 + y^2\right) dxdy \tag{10}$$

Changing to polar co-ordinates:

$$\begin{aligned} J_2(U,V) &= 2c \int_{r=0}^{1} \int_{\vartheta=0}^{\pi/c} (r^2 + P^2 - 2rP\cos(\vartheta)) r dr d\vartheta = \\ &= \pi/2 + P^2\pi - 4cP\,\mathrm{sen}(\pi/c)/3 \end{aligned} \tag{11}$$

The minimum of the function is reached when

$$P = \frac{2c\,\mathrm{sen}(\pi/c)}{3\pi} \tag{12}$$

In this way, we have:

$$J_2(U,V) = \pi/2 - 4c^2\,\mathrm{sen}(\pi/c)^2/9\pi \tag{13}$$

For the situation 2 the result is:

$$J_2(U,V) = \frac{\pi}{2} + P^2\pi - \frac{4(c-1)P\,\mathrm{sen}(\frac{\pi}{c-1})}{3} - \frac{(c-1)P^4\,\mathrm{tg}(\frac{\pi}{c-1})}{12} \tag{14}$$

where P can be solved in the same way as before.

To obtain the equalization function, we should solve the equation

$$g(c) = K / J_2(U,V) \tag{15}$$

Nevertheless, as the expression of J_2 is not simple, we directly use the coefficients. The next table shows the *1/g* coefficients. The need of studying situation 2 is proved due to the fact that for 6 and 7 clusters the values of (13) are greater than the values of (14).

Table 1: Coefficients for 1/g in the 2-D case

C	*2*	*3*	*4*	*5*	*6*	*7*
1/g	1.00	0.62	0.44	0.35	0.29	0.24

In the next section it will be presented some results using these coefficients.

4.2. 3-DIMENSIONAL CASE

For the 3-dimensional case, the situation is similar to the previous one studied. In this case, the reference case studied is the sphere of radius equal to 1.

The way of solving J_3 is also similar, but the situations found were different. These are the following:

1. For *c* from 2 to 4, the area of each cluster has the shape shown in Figure 3.
2. For *c* from 5 to 7, each cluster has the shape shown in Figure 4.

In the first situation the integral is the following:

$$J_3(U,V) = 2c \int_{r=0}^{1} \int_{\vartheta=\pi/2}^{\pi/2+\pi/c} \int_{\varphi=0}^{\pi} (r^2 + P^2 - 2rP\cos\vartheta\,\mathrm{sen}\,\varphi)\cdot \cdot r^2\cos\vartheta\cdot dr\cdot d\vartheta\cdot d\varphi \tag{16}$$

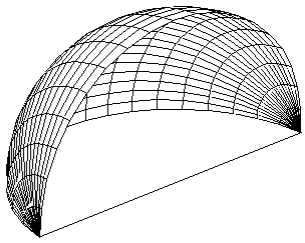

Figure 3: Shape of clusters in situation 1

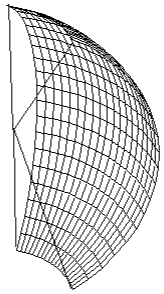

Figure 4: Shape of clusters in situation 2

It is minimized when:

$$P = \frac{3c\,\mathrm{sen}(\pi/c)}{16} \tag{17}$$

And the final expression of J_3 is:

$$J_3(U,V) = 4\pi/5 - 3\pi\cdot c^2\,\mathrm{sen}(\pi/c)^2/64 \tag{18}$$

About the situation 2, we have found that the centroids are the vertex of regular polyhedrons. These polyhedrons have symmetry with respect to a plane that crosses the center of the sphere. Then, for 6 sets the centroids are vertex of an octahedron, and the symmetry plane have four centroids that make up a square. Respectively, for 5 sets there are 3 centroids in the symmetry plane that compose an equilateral triangle and for 7 sets there are 5 centroids in the symmetry plane that make up a pentagon.

About the equalization function, g will be defined by coefficients. They are shown in the following table.

Table 2: Coefficients for 1/g in the 3-D case

c	*2*	*3*	*4*	*5*	*6*	*7*
g^{-1}	1.92	1.52	1.33	1.04	0.88	0.80

5 RESULTS

In this section it will be shown the application of the equalization function g to a practical case.

5.1. 2-DIMENSIONAL CASE

Figure 5 shows the data set used for testing. It can be seen that there are four well separate clusters. Figure 6 presents the value of J_2 normalized by g. Minimum is found in 4 clusters, as it was expected.

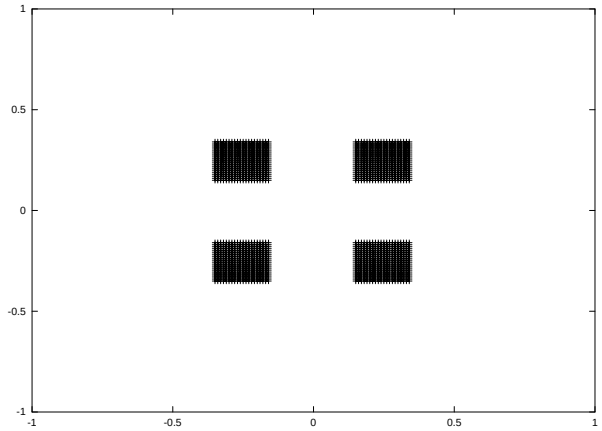

Figure 5: Data set for the 2-D case

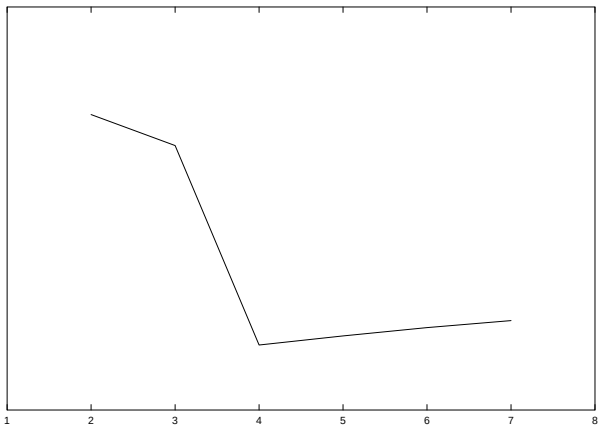

Figure 6: Results for the 2-D case

5.2. 3-DIMENSIONAL CASE

Data set used for testing is shown in Figure 7. It can be seen that there are four well separate clusters too. The value of J_3 normalized by g is presented in Figure 8. The same as before, the result of 4 clusters is correct.

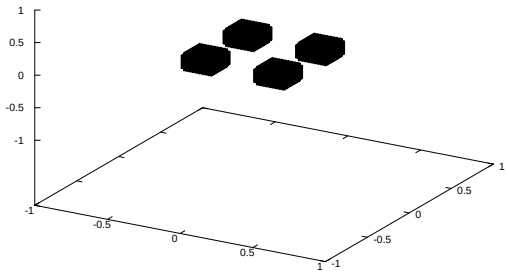

Figure 7: Data set for the 3-D case

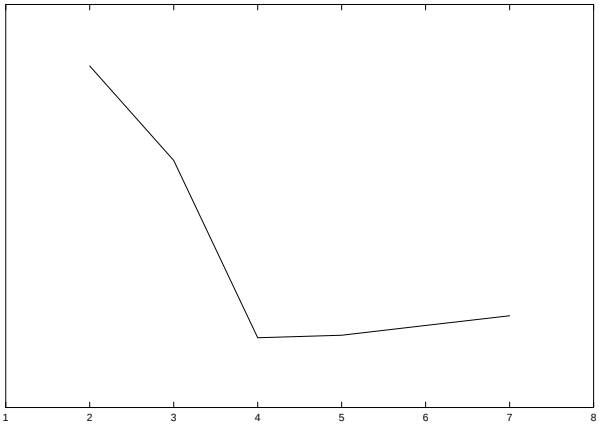

Figure 8: Results for the 3-D case

6 REFERENCES

[1] J.Dunn. A fuzzy relative of the isodata process and its use in detecting compact well separated clusters. J. Cybernetics, 3(3):32-57, 1973.

[2] S. López, L. Magdalena, J.R. Velasco. Genetic fuzzy c-means algorithm for the automatic generation of fuzzy partitions. IPMU'98.

[3] G.J. Klir, B. Yuan. Fuzzy sets and fuzzy logic theory and applications. Prentice-Hall, 1995.

[4] N.R. Pal, J. Bezdek. On cluster validity for the fuzzy c-means model. IEEE Trans. On Fuzzy Systems, vol 3, Num 3, pp. 370-379, 1995.

[5] J.R. Velasco, S. López, L. Magdalena. Genetic fuzzy clustering for the definition of the fuzzy sets. Proc. FUZZ-IEEE'97, pp. 1665-1670, 1997.

[6] M. Windham. Cluster Validity for the fuzzy c-means clustering algorithm. IEEE Trans. On Patt. Anal. And Mach. Int. vol PAMI-4, no. 4, pp. 357-363, 1982.