A NEW FUZZY CLUSTERING METHOD WITH ADJUSTABLE MEMBERSHIP CHARACTERISTICS

Dian-Rong Yang, Leu-Shing Lan, and *Wei-Cheng Pao*

Department of Electronics Engineering
National Yunlin University of Science and Technology, Taiwan

ABSTRACT

Clustering is an unsupervised procedure to group objects in accordance with their similarities. For non-separable clusters, the concept of fuzziness is incorporated. Among other approaches, the fuzzy c-means algorithm is the most well-known fuzzy clustering method. In this work, we present a modified form of the fuzzy c-means based on a new definition of distance measure which can be considered as an extension of the conventional one. The key advantage of this new fuzzy clustering schemem is its ability to flexibly control the membership function curves. Analytical formulae have been derived for both cluster centers and the fuzzy partition matrix. Parameter effects related to the membership function curves have also been analyzed. Examples are given to demonstrate the clustering results of the newly presented scheme.

Keywords: clustering, fuzzy clustering, fuzzy c-means

1. INTRODUCTION

The process of subdividing a data set into distinct subsets with homogeneous elements is called *clustering*. For hard clustering, a membership value of zero or one is assigned to each pattern data. With fuzzy clustering, each datum belongs to all clusters simultaneously, but to different degrees. Cluster analysis has been extensively studied in various fields of engineering [1]. A variety of fuzzy clustering methods have been proposed, including fuzzy c-means [2], possibilistic c-means [3], noise clustering [4], fuzzy entropy clustering [5], credibilistic fuzzy c-means [6], convex-set-based fuzzy clustering [7], generalized weighted conditional fuzzy clustering [8], etc. This paper focuses on the extension of the well-known fuzzy c-means algorithm. Specifically, we propose a new distance measure which adds a higher-term to the original one. The net effect is that the membership curves can be controlled by adjusting two parameters a_1 and a_2. Thus a more flexible clustering method is obtained. To facilitate the algorithm development, we have analytically derived the formulae for cluster centers and the fuzzy partition matrix. Explicit formulae for the membership curves in a 1-D two-cluster environment have also been derived to characterize the effects of the a_1 and a_2 parameters. Some examples are given to demonstrate the clustering results of this new scheme.

The rest of this paper is organized as follows. In Section 2, we give a brief summary of the conventional fuzzy c-means algorithm. Then in Section 3, we present the new fuzzy clustering method with adjustable membership characteristics. The succeeding section presents analysis of membership functions. Some examples are given in Section 5. Finally Section 6 concludes this paper.

2. THE FUZZY C-MEANS ALGORITHM

The fuzzy c-means algorithm [2] originated from an optimization problem

$$\min_{\boldsymbol{U},\boldsymbol{V}} \quad J(\boldsymbol{U},\boldsymbol{V}) \triangleq \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ki}^m \|\boldsymbol{x}_k - \boldsymbol{v}_i\|^2 \quad (1)$$

$$\text{subject to} \quad \sum_{i=1}^{c} u_{ki} = 1, \quad \forall k \quad (2)$$

where $\boldsymbol{U}$ is the fuzzy partition matrix, $\boldsymbol{V}$ is the collection of cluster centers, n is the number of data samples, c is the number of clusters, m is the weighting exponent, and u_{ki} is the membership value of $\boldsymbol{x}_k$ with respect to cluster i. Optimality conditions at the stationary points require that

$$\boldsymbol{v}_i = \frac{\sum_{k=1}^{n} u_{ki}^m \boldsymbol{x}_k}{\sum_{k=1}^{n} u_{ki}^m}, \quad \forall i \quad (3)$$

$$u_{ki} = \left[\sum_{j=1}^{c}\left(\frac{\|\boldsymbol{x}_k - \boldsymbol{v}_i\|^2}{\|\boldsymbol{x}_k - \boldsymbol{v}_j\|^2}\right)^{\frac{1}{m-1}}\right]^{-1}, \forall k, i \quad (4)$$

An alternating optimization procedure is commonly ulilized to solve the fuzzy c-means problem. The alternating procedure consists of two steps: (1) fix cluster centers and find the fuzzy partition matrix, and (2) fix the fuzzy partition matrix and update cluster centers. Steps (1) and (2) are alternately executed until convergence is achieved. Note that the algorithm may converge to a local minimum or even a saddle point.

3. THE NEW FUZZY CLUSTERING METHOD WITH ADJUSTABLE MEMBERSHIP CHARACTERISTICS

In convectional fuzzy c-means algorithm, the distance measure is defined by $\|\boldsymbol{x}_k - \boldsymbol{v}_i\|^2$. We now extend this definition to include an additional higher-order term, namely, let us define the new distance measure by

$$d_{ki} \triangleq a_1\|\boldsymbol{x}_k - \boldsymbol{v}_i\|^2 + a_2\|\boldsymbol{x}_k - \boldsymbol{v}_i\|^4 \tag{5}$$

where $\boldsymbol{x}_k$ is the kth input datum, $\boldsymbol{v}_i$ is the ith cluster center, and a_1 and a_2 are two parameters specified by the user. With this new definition of distance measure, the fuzzy clustering problem can be reformulated as a new optimization problem

$$\min_{\boldsymbol{U},\boldsymbol{V}} \quad J(\boldsymbol{U},\boldsymbol{V}) \triangleq \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ki}^m (a_1\|\boldsymbol{x}_k - \boldsymbol{v}_i\|^2 + a_2\|\boldsymbol{x}_k - \boldsymbol{v}_i\|^4) \tag{6}$$

$$\text{subject to} \quad \sum_{i=1}^{c} u_{ki} = 1, \quad \forall k \tag{7}$$

where $\boldsymbol{U}$ is the fuzzy partition matrix, $\boldsymbol{V}$ is the collection of cluster centers, n is the number of data samples, c is the number of clusters, m is the weighting exponent, and u_{ki} is the membership value of $\boldsymbol{x}_k$ with respect to cluster i, for $k = 1, \cdots, n$ and $i = 1, \cdots, c$. The necessary conditions for this optimization problem can be found using the method of Lagrange multipliers. First we define the corresponding Lagrangian function as

$$\mathcal{L} \triangleq \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ki}^m (a_1\|\boldsymbol{x}_k - \boldsymbol{v}_i\|^2 + a_2\|\boldsymbol{x}_k - \boldsymbol{v}_i\|^4) + \sum_{k=1}^{n}\alpha_k\left(\sum_{i=1}^{c} u_{ki} - 1\right) \tag{8}$$

where α_k is the corresponding Lagrange multiplier. At the optimal points of solutions, the partial derivatives of $\mathcal{L}$ with respect to all related variables should be equal to zero. From $\frac{\partial \mathcal{L}}{\partial u_{ki}} = 0$, we obtain

$$m u_{ki}^{m-1}(a_1\|\boldsymbol{x}_k - \boldsymbol{v}_i\|^2 + a_2\|\boldsymbol{x}_k - \boldsymbol{v}_i\|^4) + \alpha_k = 0, \quad \forall k, i \tag{9}$$

From $\frac{\partial \mathcal{L}}{\partial \boldsymbol{v}_i} = \boldsymbol{0}$, we have

$$\boldsymbol{v}_i = \frac{\sum_{k=1}^{n} u_{ki}^m (a_1 + 2a_2\|\boldsymbol{x}_k - \boldsymbol{v}_i\|^2)\boldsymbol{x}_k}{\sum_{k=1}^{n} u_{ki}^m (a_1 + 2a_2\|\boldsymbol{x}_k - \boldsymbol{v}_i\|^2)}, \quad \forall i \tag{10}$$

Substituting Eq. (9) into Eq. (7), one obtains

$$\alpha_k = -\left(\sum_{i=1}^{c} \frac{1}{[m(a_1\|\boldsymbol{x}_k - \boldsymbol{v}_i\|^2 + a_2\|\boldsymbol{x}_k - \boldsymbol{v}_i\|^4)]^{\frac{1}{m-1}}}\right)^{-m+1}, \forall k \tag{11}$$

Eqs. (11) and (9) together yield

$$u_{ki} = \left[\sum_{j=1}^{c}\left(\frac{a_1\|\boldsymbol{x}_k - \boldsymbol{v}_i\|^2 + a_2\|\boldsymbol{x}_k - \boldsymbol{v}_i\|^4}{a_1\|\boldsymbol{x}_k - \boldsymbol{v}_j\|^2 + a_2\|\boldsymbol{x}_k - \boldsymbol{v}_j\|^4}\right)^{\frac{1}{m-1}}\right]^{-1}, \forall k, i \tag{12}$$

The *alternating algorithm* of the conventional fuzzy c-means method can also be applied to the new fuzzy clustering method with different updating functions for the cluster centers and fuzzy partition matrix. We summarize the solution algorithm as follows:

1. Initialize cluster centers. Usually this is performed by random assignment.
2. Update the fuzzy partition matrix using Eq. (12).
3. Update the cluster centers using Eq. (10).
4. Check for convergence. Usually this is done by checking $\|\boldsymbol{U}^{(k+1)} - \boldsymbol{U}^{(k)}\| \leq \varepsilon$. If not yet converged, go to Step 2 and proceed.

4. ANALYSIS OF MEMBERSHIP FUNCTIONS

In accordance with Eq. (12) one can define the membership function by

$$f_i(\boldsymbol{x}) = \left[\sum_{j=1}^{c}\left(\frac{a_1\|\boldsymbol{x} - \boldsymbol{v}_i\|^2 + a_2\|\boldsymbol{x} - \boldsymbol{v}_i\|^4}{a_1\|\boldsymbol{x} - \boldsymbol{v}_j\|^2 + a_2\|\boldsymbol{x} - \boldsymbol{v}_j\|^4}\right)^{\frac{1}{m-1}}\right]^{-1}, \forall i \tag{13}$$

where $\boldsymbol{v}_i$ and $\boldsymbol{v}_j$ are the centers for cluster i and j respectively. To gain more insight regarding how the parameters a_1 and a_2 affect the membership fuctions, let us focus on a specific situation with only two clusters where the feature vector possesses a single dimension and $m = 2$. Let the cluster centers of these two clusters be denoted by $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ respectively, and let x denote the Euclidean distance between the input datum and $\boldsymbol{v}_1$. Here we consider two different scenarios: (1) the input datum is located between $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$; (2) $\boldsymbol{v}_1$ is

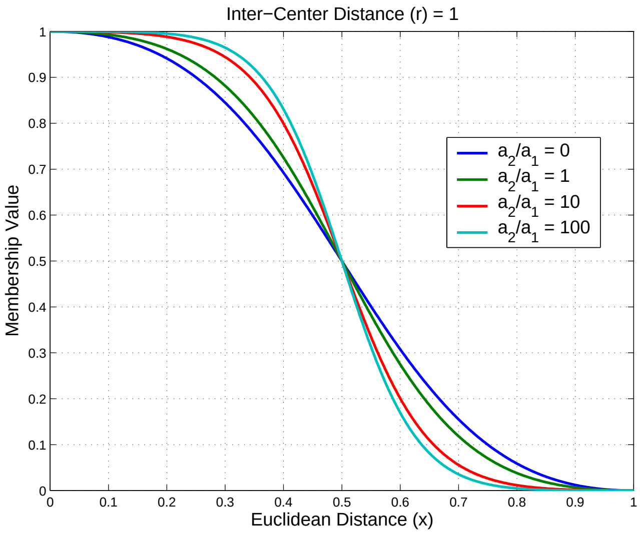

Fig. 1. The membership vs. Euclidean distance curve. Shown in this figure is the first scenario where the input datum is located between $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$.

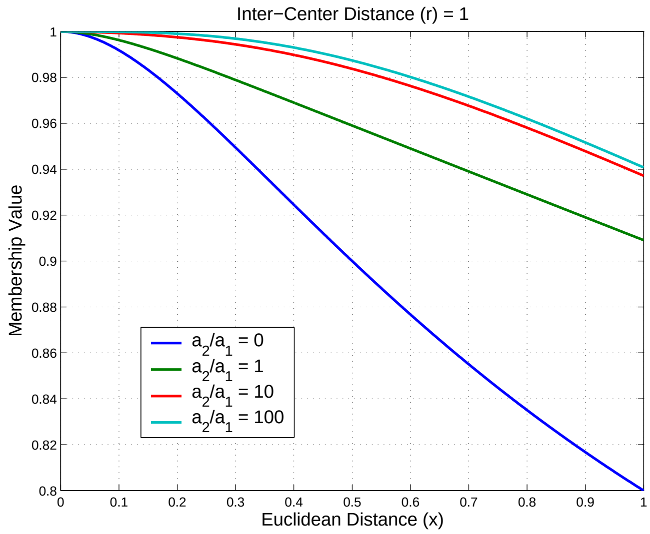

Fig. 2. The membership vs. Euclidean distance curve. Shown in this figure is the second scenario where $\boldsymbol{v}_1$ is located between the input datum and $\boldsymbol{v}_2$.

located between the input datum and $\boldsymbol{v}_2$. For the first scenario, we have

$$\begin{aligned} f_1(x) &= \left\{1+[\frac{a_1x^2+a_2x^4}{a_1(r-x)^2+a_2(r-x)^4}]^{\frac{1}{m-1}}\right\}^{-1} \\ &= \frac{(r-x)^2+\frac{a_2}{a_1}(r-x)^4}{[x^2+(r-x)^2]+\frac{a_2}{a_1}[x^4+(r-x)^4]} \end{aligned} \quad (14)$$

where r is the Euclidean distance between $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$. Similarly, for the second scenario we easily obtain

$$\begin{aligned} f_1(x) &= \left\{1+[\frac{a_1x^2+a_2x^4}{a_1(r+x)^2+a_2(r+x)^4}]^{\frac{1}{m-1}}\right\}^{-1} \\ &= \frac{(r+x)^2+\frac{a_2}{a_1}(r+x)^4}{[x^2+(r+x)^2]+\frac{a_2}{a_1}[x^4+(r+x)^4]} \end{aligned} \quad (15)$$

From Eqs. (14) and (15) it is clearly seen that in both scenarios the memberships are affected by r and the ratio a_2/a_1. Figs. 1 and 2 show the membership vs. x curves. For the first scenario, with an increasing a_2/a_1 ratio, the clustering gradually changes from soft partition toward harder partitions, as illustrated by Fig. 1. For the second scenario, we also observe that with an increasing a_2/a_1 ratio, the clustering gradually changes from soft partition toward harder partitions, which is illustrated by Fig. 2. However, in the latter scenario, the changes occur at a slower pace.

5. EXAMPLES

In this study we used an artificially synthesized data set for experimentation. This data set with 2000 data points comprises three non-overlapped clusters. Since the clusters are separable, the clustering task can be easily accomplished. However, with different fuzzy clustering strategies, varying fuzzy partition matrices can be generated. Figs. 3 through 6 show the experimental results. Prior to clustering, the data set went through a linear normalization operation so that each component of the feature vector is scaled to be in the range $[0, 1]$. In each figure, the scatter plot of input data are drawn with cluster centers and iso-membership contours overlaid. Each figure possesses a different combination of a_1 and a_2 values. From these figures, it is seen that cluster centers change slightly with varying combinations of a_1 and a_2; however, the iso-membership contours exhibit distinct changes. In principle, the larger the a_2/a_1 ratio is, the higher the membership value is observed around the same neighborhood of a cluster center. Refer to Figs. 3 through 6 for the detailed contour plots.

6. CONCLUSION

Clustering plays a very important role in almost all branches of science and engineering. The conventional k-means and fuzzy c-means algorithms have been most popular methods for solving separable and non-separable clustering tasks. Based on a new definition of distance measure, we propose a new form of fuzzy clustering method. The key distinct property of this new fuzzy clustering scheme is that it is capable of controlling the membership curve through adjusting the values of a_1 and a_2. Further research will be necessary in order to characterize the choice of a_1 and a_2 values in a practical setting.

7. REFERENCES

[1] A. Jain and R. Dubes, *Algorithms for Clustering Data*, Englewood Cliffs, NJ: Prentice-Hall, 1988.

[2] J. bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, New York: Plenum, 1981.

[3] R. Krishnapuram and J. Keller, “A Possibilistic Approach to Clustering,” *IEEE Tr. Fuzzy Systems*, Vol. 1, pp. 98-110, May 1993.

[4] R. N. Dave, “Characterization and Detection of Noise in Clustering,” *Pattern Recognition Letters*, Vol. 12, pp. 657-664, 1991.

[5] D. Tran and M. Wagner, “Fuzzy Entropy Clustering,” *Proc. IEEE 2000 Int’l Conf. Fuzzy Systems*, pp. 152-157.

[6] K. K. Chintalapudi and M. Kam, “A Noise-Resistant Fuzzy c-Means Algorithm for Clustering,” *Proc. IEEE 1998 Int’l Conf. on Fuzzy Systems*, pp. 1458-1463.

[7] I. H. Suh, J.-H. Kim, and F. C.-H. Rhee, “Convex-Set-Based Fuzzy Clustering,” *IEEE Tr. Fuzzy Systems*, Vol. 7, No. 3, pp. 271-285, June 1999.

[8] J. M. Leski, “Generalized Weighted Conditional Fuzzy Clustering,” *IEEE Tr. Fuzzy Systems*, Vol. 11, No. 6, pp. 709-715, Dec. 2003.

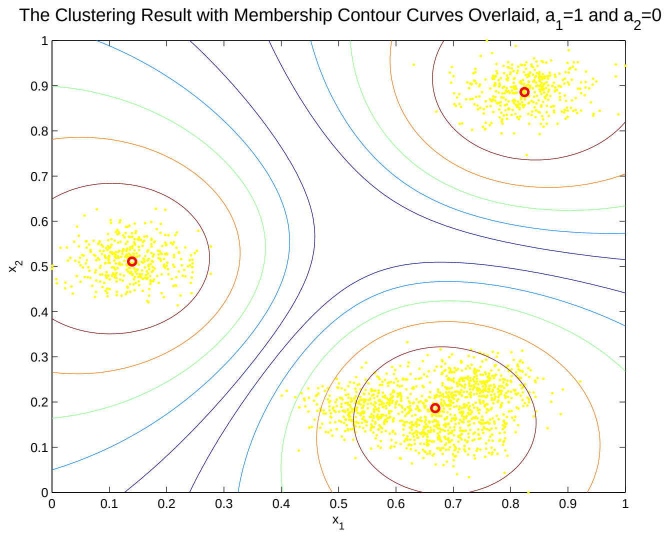

Fig. 3. An example to demonstrate the fuzzy clustering result. In this figure, $a_1 = 1$ and $a_2 = 0$, and cluster centers are shown as small red circles. For each cluster, five iso-membership contours are illustrated, which correspond to membership values of 0.9, 0.8, 0.7, 0.6 , and 0.5.

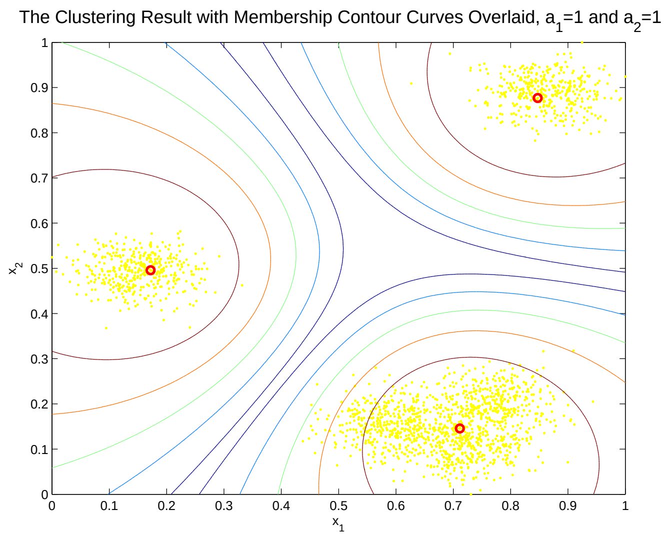

Fig. 4. An example to demonstrate the fuzzy clustering result. In this figure, $a_1 = 1$ and $a_2 = 1$, and cluster centers are shown as small red circles. For each cluster, five iso-membership contours are illustrated, which correspond to membership values of 0.9, 0.8, 0.7, 0.6 , and 0.5.

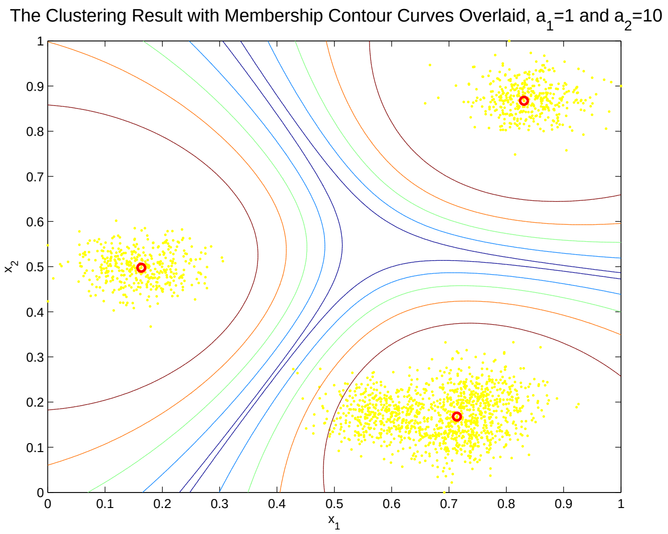

Fig. 5. An example to demonstrate the fuzzy clustering result. In this figure, $a_1 = 1$ and $a_2 = 10$, and cluster centers are shown as small red circles. For each cluster, five iso-membership contours are illustrated, which correspond to membership values of 0.9, 0.8, 0.7, 0.6 , and 0.5.

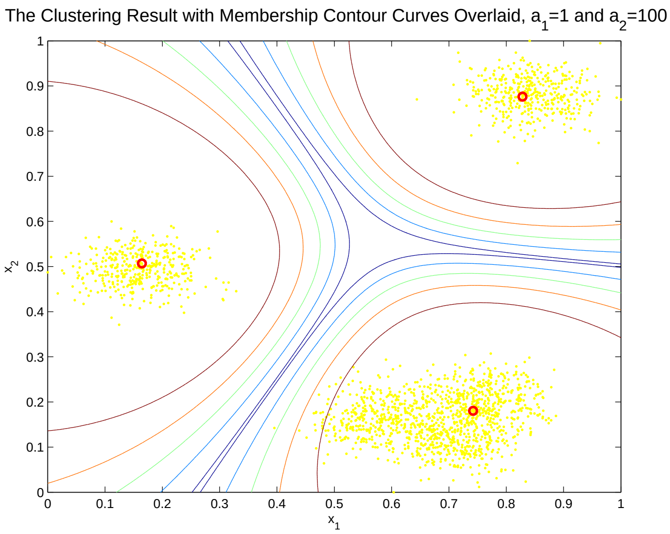

Fig. 6. An example to demonstrate the fuzzy clustering result. In this figure, $a_1 = 1$ and $a_2 = 100$, and cluster centers are shown as small red circles. For each cluster, five iso-membership contours are illustrated, which correspond to membership values of 0.9, 0.8, 0.7, 0.6 , and 0.5.