AN INTEGRITY-BASED FUZZY C-MEANS METHOD RESOLVING CLUSTER SIZE SENSITIVITY PROBLEM

Y. H. LAI1, P. W. HUANG2, P. L. LIN3

1,2 Computer Science and Engineering, National Chung Hsing University, Taichung 40227, Taiwan, R.O.C
3 Computer Science and Information Engineering, Providence University, Shalu, Taichung 43301, Taiwan, R.O.C
E-MAIL: yanhao.lai@gmail.com, powhei.huang@msa.hinet.net, lan@pu.edu.tw

Abstract:
Cluster size insensitive FCM (csiFCM) dynamically adjusts the membership value of each object based on the size of the cluster to which it is assigned after defuzzification to resolve the size sensitivity problem. Our investigation indicates that csiFCM cannot correctly partition datasets containing clusters with dispersive data distribution or insignificant distinction from others, if initial cluster centers are not properly selected. In this paper, we present a concept of cluster integrity and propose an enhanced conditional FCM, itgFCM, based on both cluster integrity and cluster size. For objects classified to a cluster of high integrity after defuzzification, itgFCM assigns their condition values predominantly depending on the size of that cluster. If an object is assigned to a cluster of low integrity, itgFCM adjusts the size-dependent condition value with a multiplicative weight that grows with both the complement of cluster integrity and the object's purity. Experimental results demonstrate that itgFCM can partition numerical datasets as well as synthetic and real images of various number of classes to clusters that are more conforming to human perception than csiFCM can, regardless of both initial cluster centers and data distribution of the datasets.

Keywords:
Clustering; Fuzzy c-means; Conditional fuzzy c-means; Integrity-based fuzzy c-means; Unequal cluster size

1. Introduction

Clustering is an unsupervised learning method used in feature space for grouping unlabeled data or objects into a number of clusters [1]. Each cluster is a collection of objects that are “similar” in the same cluster and “dissimilar” to those in different clusters.

Fuzzy C-Means (FCM) clustering [2] is a technique reformed from the traditional hard C-means model by an introduction of fuzziness of the object membership to clusters. Although FCM yields good results in many image segmentation and object classification applications, it tends to draw the centers of smaller clusters to the adjacent larger one in unbalanced sized cases, because the minimum sum of a squared-errors objective function can usually be obtained from equal cluster populations [2]. Tran et al. [3] pointed out that the unequal cluster size is one of the main problems encountered when clustering. A small cluster can be very important but not found, because the larger clusters often determine the clustering result.

To overcome the size sensitivity problem, conditional memberships are used to weaken the influences from objects within larger clusters and to strengthen the influences from objects within smaller clusters [4]. Semi-supervised FCM (ssFCM) [5] was proposed to make objects in small clusters receive higher membership than objects in larger clusters by manually assigning cluster-size dependent weights to objects, where the cluster size is obtained from a training sample dataset. Noordam et al. [6] presented a cluster size insensitive version of FCM (csiFCM) that automatically assigns a conditional value to each object inversely proportional to the size-ratio of the cluster to which it is assigned after the defuzzification step in each iterative cycle. Despite that csiFCM can automatically adjust the influences of objects to prevent from grouping the dataset to approximately equal cluster populations, our investigation indicates that it still unsatisfactorily partitions datasets containing clusters with dispersive data distribution or insignificant distinction from others, if initial cluster centers are not properly selected.

Several FCM-based algorithms [7-9] have been presented in recent years. However, they mainly focus on increasing the robustness against noises by incorporating local spatial information of neighbors into the squared-errors objective function. Thus, the unequal cluster size problem has not yet been totally taken care.

In this paper, we propose an integrity-based FCM method, itgFCM, in which the contribution of each data point is based on both the integrity of the cluster to which it is assigned and its purity with respect to that cluster. For objects classified to a cluster of high integrity after defuzzification, itgFCM assigns their condition values predominantly depending on the size of that cluster. If an object is assigned
to a cluster of low integrity, itgFCM adjusts the size-dependent condition value with a multiplicative weight that grows with both the complement of cluster integrity and the object’s purity.

The remainder of this paper is organized as follows. In Section 2, the unequal cluster size problem of FCM is described and the effectiveness of csiFCM for solving this problem is investigated. In Section 3, the definitions of cluster integrity and purity along with the proposed integrity-based FCM are detailed. Experimental results demonstrating the effectiveness of itgFCM and the comparisons to csiFCM are provided in Section 4. Finally, conclusions and future works are given in Section 5.

2. Preliminary

2.1. Fuzzy c-means

Bezdek presented the Fuzzy C-Means algorithm (FCM) [2] to partition a set of \( N \) measured objects into \( C \) clusters through an iterative minimization of an objective function \( J(U,V) \) with respect to a fuzzy partition matrix \( U \) and a set of prototypes \( V \). The objective function is defined as:

\[
J(U,V) = \sum_{i=1}^{C} \sum_{j=1}^{N} u_{ij}^m \|x_j - v_i\| \tag{1}
\]

where \( x_j \) is the \( j \)th measured data point, \( v_i \) is the center of cluster \( i \), \( u_{ij} \) is the membership value of \( x_j \) with respect to cluster \( i \), \( m \) is the controller of fuzziness, and \( \| \| \) stands for the Euclidean norm.

For the minimization of the objective function, \( u_{ij} \) can be obtained by using Lagrange Multiplier:

\[
u_{ij} = \frac{1}{\sum_{j=1}^{C} \left( \|x_j - v_i\| \right)^{(m-1)}/},\tag{2}
\]

with a constraint \( \sum_{i=1}^{C} u_{ij} = 1, \ \forall j = 1,2,...,N \), and a new set of prototypes \( V \):

\[
v_i = \frac{\sum_{j=1}^{N} u_{ij}^m x_j}{\sum_{j=1}^{N} u_{ij}^m}, \ \text{l} \leq i \leq C.\tag{3}
\]

2.2. Unequal cluster size problem of FCM algorithm

A major drawback of FCM is the sensitivity to unequal cluster size, because the minimum sum of a squared-errors function can usually be obtained from approximately equal cluster populations. According to the updating formula for the cluster center in (3), objects with low membership value for a particular cluster have less contribution to the final center position of that cluster. In other words, the final cluster center of a particular cluster should drift towards objects with higher membership value for that cluster. However, in equally sized cases, the center of the small cluster tends to be pulled towards the adjacent larger cluster even if most of the objects in the larger cluster have low membership value for the small cluster, because the total contribution from the objects in the large cluster is much higher than the total contribution from the small cluster.

Figure 1(a) is an uneven illuminated synthetic image with background (intensity range from 0 to 128) and a square foreground object (intensity range from 200 to 228). Based on the normalized histogram shown in Fig. 1(b), the image should obviously be partitioned into two clusters from human’s perception. However, FCM with the initial cluster centers set at the positions near the true cluster centers (0.3, 0.95) gradually drifts the centers towards the data of lower intensity. Figure 2 shows the defuzzification results at various stages by applying the FCM algorithm.

2.3. Conditional FCM and cluster size insensitive FCM

Conditional FCM (c-FCM) [4] was proposed to overcome the cluster size sensitivity problem. Its principle is to weaken the contribution of the objects from the larger cluster while retaining the contribution of the objects from the smaller cluster, so that the larger cluster will not have enough influences to pull the center of the smaller cluster towards itself. Cluster size insensitive FCM (csiFCM) [6] introduced a condition value \( f_j \) for each object \( x_j \) after defuzzification step in the iteration as follows:

- Proceedings of the Ninth International Conference on Machine Learning and Cybernetics, Qingdao, 11-14 July 2010

2713
\[ f_j = \frac{1}{1 - S_{\text{min}}} \cdot (1 - S_i), \quad 0 \leq f_j \leq 1. \]  

(4)

where cluster \( i \) is the cluster to which \( x_j \) is classified, \( S_i = N_i / N \), \( N_i \) is the number of objects classified to cluster \( i \), \( N \) is the total number of objects in the dataset, and \( S_{\text{min}} \) is the smallest \( S_i \) among all clusters. With this definition, the condition value of each \( x_j \) is based only on the size ratio of the cluster to which it is assigned. Applying the condition value in (4) to FCM, the degree of the membership of object \( x_j \) to cluster \( i \) is modified as

\[
u_j = \frac{f_j}{\sum_{i=1}^{C} \left( \frac{\|x_j - v_i\|}{\|x_j - y_i\|} \right)^{2/(m-1)}}.
\]

(5)

2.4. Deficiency of cluster size insensitive FCM

Our investigation indicates that csiFCM is not completely cluster size insensitive. At first, we applied csiFCM to the same image dataset that was used in testing FCM, as shown in Fig. 1(a) with (0.3, 0.6) as the initial center pair of the two clusters. Thus, an effective size-insensitive FCM should gradually pull the center of Cluster 1 from 0.6 towards 0.95. Unfortunately, csiFCM gradually pulls both center positions leftwards and finally partitions the dataset to two clusters of almost equal population as FCM does. Let’s go through a couple of iterations of csiFCM. As shown in Fig. 1(c), the dataset was divided to Cluster 1 comprising region I and Cluster 2 comprising regions II and III at the beginning, because region II is closer to the initial center of Cluster 2 (0.6). According to csiFCM, the size-weights for region II and region III are the same after stage-0, as both regions are in Cluster 2. However, the total contribution from region II for computing the new center of Cluster 2 is much larger than that from region III, because (a) the distance between each data point in region II and the center of Cluster 2 is much smaller than the distance between each data point in region III and the center of Cluster 2, and (b) the size of region II is much larger than that of region III. Thus, the new center of Cluster 2 in stage-1 will definitely drift towards region II, and those data points that are closer to the new center of Cluster 2 within region I will be assigned to Cluster 2. The same situation repeats for the subsequent iterations. In order to overcome the problem, we conjecture that the data distribution of the cluster must be considered besides cluster size, when determining condition values of objects.

In our second experiment, we used dataset-1 and dataset-2 that are almost identical to those used in the experiment 1 in [6] and ran FCM with both initial centers set to the true groups’ centers, followed by csiFCM with its initial centers set to the final centers resulted from FCM, as shown in symbol (“◆”) in Fig. 3. The initial cluster size ratios of both datasets for csiFCM are (20:23) and (17:26:6), respectively. The “★’s and the dotted line(s) in both Figs. 3(a) and (c), which are the final cluster centers and the cluster separation boundaries, show that csiFCM indeed clusters both datasets correctly. We then moved all data in the large group 10 units up in both datasets, respectively. Unfortunately, the final cluster results of both datasets, as shown in Figs. 3(b) and (d), were not satisfactory to human perception, even though the initial cluster size ratios of both datasets, (21:22) and (18:25:6), were only slightly different from (20:23) and (17:26:6), respectively. Analyzing these datasets, we notice that all data points of Cluster 2 in Fig. 3(b) are closer to the center of Cluster 1 than those data points in Fig. 3(a). The same situation also occurs in Figs. 3(d) and (c). Thus, we conjecture that the data distance between two adjacent clusters must be considered besides cluster size, when determining condition values of objects.

3. Fuzzy c-means based on integrity and size

Our previous investigation indicates that csiFCM cannot always correctly partition datasets of unequal population. To overcome its problem, we introduce a concept of cluster integrity and use it as the other factor besides cluster size for determining the condition value of a given object.

3.1. Definition

3.1.1. Cluster compactness

The compactness of cluster \( i \) is defined as the
complement of the standard deviation of the distance between each data point assigned to cluster $i$ and the center of that cluster after defuzzification. The formulas are:

$$I_{ci} = 1 - \frac{1}{\sqrt{|A_i|}} \sum_{j \in A_i} (\|x_j - v_i\| - \mu_i)^2. \quad (6)$$

$$\mu_i = \frac{1}{\sqrt{|A_i|}} \sum_{j \in A_i} \|x_j - v_i\|. \quad (7)$$

3.1.2. Cluster purity

The purity of cluster $i$ is defined as the average purity of all objects that are assigned to cluster $i$ after one defuzzification iteration. The formula is:

$$I_{pi} = \frac{1}{|A_i|} \sum_{j \in A_i} P_{ij} = \frac{1}{|A_i|} \sum_{j \in A_i} abs(\frac{\|x_j - v_i\| - \|x_j - v_{i'}\|}{\|v_i - v_{i'}\|}) \quad (8)$$

where all data points $x_j$’s and cluster centers $v_k$’s are normalized to $[0,1]$. $P_{ij}$ is the purity of $x_j$ with respect to cluster $i$, can be obtained as the normalized difference between the distance from $x_j$ to the center of its assigned cluster ($v_k$) and the distance from $x_j$ to $v_{i'}$’s nearest cluster ($v_{i'}$). $A_i$ is the subset containing data points that are classified to cluster $i$ after defuzzification, and $|A_i|$ is the size of $A_i$. Thus, the larger the normalized difference is, the bigger the purity of $x_j$ with respect to its assigned cluster $i$ will be. Notice that applying normalization (i.e., divided by $\|v_i - v_{i'}\|$ in (8)) is to handle the distance variation between each cluster $i$ and its nearest cluster.

3.1.3. Cluster integrity

The integrity of cluster $i$ is defined as the average of both cluster compactness and cluster purity. The formula is:

$$I_i = \frac{1}{2} (I_{ci} + I_{pi}). \quad (9)$$

3.2. Integrity-based fuzzy c-means (itgFCM)

The integrity-based FCM (itgFCM) is an enhanced version of csiFCM in which the condition value is based on both cluster integrity and cluster size. For objects within clusters of high integrity, cluster size plays the key role in computing their condition values; whereas for objects within clusters of low integrity, the effect of cluster size will be strengthened by integrity.

Our design rationale is as follows. After defuzzification in iterations when using a condition value based only on cluster size, a cluster of high integrity means that most objects assigned to the cluster are packed and pure. Thus, the total contribution from the impure data is not significant to affect the computation of the new cluster center for the next iteration. In other words, integrity should not be too much a concern in determining the condition value; instead, cluster size should play the key role.

On the contrary, after defuzzification in iterations, a cluster of low integrity means that many objects assigned to this cluster are either impure or are dispersive against the cluster center. In order to push out these impure objects from the cluster, the contribution from these impure objects needs to be weakened; in other words, the contribution from the pure objects needs to be further strengthened. Thus, integrity should play a strong role besides cluster size.

Combining both requirements, the weight for cluster-size based condition value is designed as:

$$W_f = \frac{D_{ij}}{\max(D_{ij})} e^{\frac{1}{e}} \leq W_f \leq 1. \quad (10)$$

$$D_{ij} = \exp(1 - x_{ij}), 1 \leq D_{ij} \leq e. \quad (11)$$

$$I_i = \frac{I_{ci} - \min_{j=1,2,...,c}(I_{ij})}{\max_{j=1,2,...,c}(I_{ij}) - \min_{j=1,2,...,c}(I_{ij})}, 0 \leq I_i \leq 1. \quad (12)$$

where $P_{ij}$ is the purity of $x_j$ with respect to cluster $i$ as defined in (8). Thus, the proposed enhanced condition value for each $x_j$ with respect to cluster $i$ is

$$H_{ij} = W_f \cdot f_j, \quad (13)$$

where $f_j$ is the condition value of $x_j$ based only on cluster size as in (4). As a result, the enhanced condition value exponentially strengthens the contribution from the pure objects within small clusters of low integrity while retaining the contribution from the objects within large clusters of high integrity so that the total contribution from these pure objects is strong enough to gradually pull the center of their assigned cluster to the right position. With the new enhanced condition value, the membership value of object $x_j$ with respect to cluster $i$ is modified as

$$u_{ij} = \frac{H_{ij}}{\sum_{j=1}^{n} \left(\frac{\|x_i - v_j\|}{\|x_i - v_j\|^2}\right)^{\frac{1}{n-1}}}. \quad (14)$$

4. Experimental results and comparisons

We use unequal cluster-sized datasets, synthetic and real images to demonstrate that the clustering results using itgFCM are more conforming to human perception, when compared with both traditional FCM and csiFCM.
4.1. Synthetic dataset

4.1.1. Clustering 1-D dataset

The first 1-D dataset for test, as shown in Fig. 4(a), is completely identical to the one used in demonstrating the size sensitivity problem of FCM. We apply itgFCM to this dataset three times with (0.14, 0.49), (0.3, 0.6), and (0.5, 0.8) as the three respective initial cluster center pairs. The results, as shown in Fig. 4(b), are two clusters with the final center pairs as (0.24, 0.94), (0.22, 0.92), and (0.22, 0.92), respectively. Figure 4(c) is the clustering result of csiFCM with (0.14, 0.49), (0.15, 0.49), and (0.22, 0.92) as its three respective final center pairs. The results show that itgFCM can indeed partition the dataset correctly regardless of its initial center positions; whereas csiFCM cannot.

The second 1-D dataset for test, as shown in Fig. 4(d), consists of four clusters: cluster I (the small square with intensity 0), cluster II (the horizontal bar with intensity 85) and cluster III (the vertical bar with intensity 170), and cluster IV (the large square with intensity 255). Each cluster is added with a Gaussian noise of \((\mu=0, \sigma=0.01)\) to simulate the adjacent clusters with different overlap. Again, we apply itgFCM to this dataset three times with (77, 166, 230, 254), (10, 50, 90, 130), and (160, 190, 220, 250) as the three respective initial cluster center quadruples. The results, as shown in Fig. 4(e), are four clusters that are identical to the original clusters I, II, III, and IV with their respective final center quadruples as (13, 87, 168, 245), (7, 89, 165, 245) and (7, 85, 171, 244). Fig. 4(f) is the result of csiFCM with (77, 166, 230, 254), (77, 163, 227, 254) and (6, 87, 127, 248) as its final center quadruples, which clearly are not correct except the last one.

Table 1 summarizes the number of misclassified pixels in each class for both tested datasets after using csiFCM and itgFCM. Comparing the misclassified numbers of both datasets in Table 1 and the images in Figs. 4(a)-(f), we can see that itgFCM indeed improves the clustering performance over csiFCM on unequal cluster-sized synthetic images with uneven illumination or noises.

4.1.2. Clustering 2-D dataset

We repeat the same tests for csiFCM in section 2.4. Like csiFCM, itgFCM also clusters both datasets correctly in the first test, as shown in Fig. 5(a) and (b). However, for both the updated datasets (i.e., the objects in the large cluster of both datasets are moved 10 units up), itgFCM still partitions them correctly regardless of the initial center positions, as shown in Figs. 5(c) and (d).

4.2. Minced Meat Image

In this test, we use a gray-scale sub-image from a disk of minced meat [6], as shown in Fig. 6(a), to demonstrate that itgFCM produces a better clustering result for a gray scale image containing pixel clusters with their intensities not significantly distinct from one another. As shown in Fig. 6(b),

![Figure 4](image_url)

![Figure 5](image_url)

![Table 1](image_url)
the minced meat can be classified into dark meat (dark gray regions), fat (white regions), and light meat (light gray regions surrounding fat). Figures 6(c), (d), and (e) show the clustering results using itgFCM, csiFCM, and FCM, respectively. Observing the results, we can see that the classified dark meat regions (dark gray /blue part) are approximately the same for all three methods, as the data distribution of the dark meat region is compact and distinct enough from the data of both class fat and class light meat. However, the intensity span of the light meat regions almost overlaps with that of the fat regions; that is, the data of class light meat are not very distinct from the data of fats. Thus, the regions of fat class clustered by itgFCM, as shown in white or green in Fig. 6(c), are more conforming to human perception on high resolution monitors than those obtained from both FCM and csiFCM, as shown in Figs. 6(d) and (e), respectively.

5. Conclusions

csiFCM, in which the membership of a given object is conditioned on the size ratio of the cluster, was proposed to overcome unequal cluster size problem of FCM. According to our investigation, it cannot always cluster the dataset in ways that conform to human perception. In this paper, we presented an enhanced conditional FCM, itgFCM, based on both cluster integrity and cluster size. By weakening the contribution of data within large clusters of high integrity while sustaining that within small clusters, and elevating the influence of pure data while reducing that of impure ones within low integrity cluster, itgFCM prevents the center of small cluster from drifting to its adjacent larger cluster. Our experimental results show that (i) itgFCM can partition both synthetic and real datasets to clusters that are more conforming to human perception than csiFCM can; and (ii) the clustering results of itgFCM is independent of the initial cluster centers. Our future works includes incorporating integrity with image spatial information so that itgFCM can be more effective for segmenting noisy images with unequal sized clusters.

Acknowledgements

This work was partially supported by the National Science Council, Taiwan, under the Grant No. NSC 98-2221-E-126-009 and NSC 96-2628-E-005-084-MY3.

References