Front. Electr. Electron. Eng. China (2006) 3: 269–272
DOI 10.1007/s11460-006-0036-x

RESEARCH ARTICLE

**KUANG Ping, ZHU Qing-xin, WANG Ming-wen,
CHEN Xu-dong, QING Li**

A research on fast FCM algorithm based on weighted sample

Abstract To improve the computational performance of the fuzzy C-means (FCM) algorithm used in dataset clustering with large numbers, the concepts of the equivalent samples and the weighting samples based on eigenvalue distribution of the samples in the feature space were introduced and a novel fast cluster algorithm named weighted fuzzy C-means (WFCM) algorithm was put forward, which came from the traditional FCM algorithm. It was proved that the cluster results were equivalent in dataset with two different cluster algorithms: WFCM and FCM. Furthermore, the WFCM algorithm had better computational performance than the ordinary FCM algorithm. The experiment of the gray image segmentation showed that the WFCM algorithm is a fast and effective cluster algorithm.

Keywords fuzzy C-means, weighted fuzzy C-means (WFCM), weighted sample, image segmentation

1 Introduction

FCM algorithm [1–8] is a widely used fuzzy cluster method based on object function optimization. Comparing with the ordinary hard C-means (HCM) algorithm, FCM provides the fuzzy segmentation matrix that contains much cluster information. But, as an unsupervised dynamical optimization technique, FCM needs to calculate the cluster centers and the fuzzy segmentation matrix by iteration until the objective function converges. When the number of the samples grows big, this iteration process becomes very slow, which restricts the practical uses [1, 2] of FCM. To improve its computational performance, Refs. [3–5] propose several kinds of sampling based on fast FCM methods to reduce the number of the clustered samples, and apply them in large-scale image segmentation area such as DDSM image segmentation (image size: 2 000×2 000 pixels); IRS-1A image segmentation (image size: 2 500×2 500 pixels). But, no matter what sampling ratio and sampling criteria to choose, enough samples (also many samples in some situation) must be acquired to describe exactly the image features. In such conditions, it also means a computational bottleneck when FCM is used to cluster. At the same time, the sampling-based method is an incomplete description of the image features, which results in the inaccuracy segmentation results. Reference [6] presents a simplified FCM that uses the one dimensional histogram and applies it in gray image segmentation, but the relationship between the simplified FCM and original FCM has not been strictly proved, and the applicable condition has not been pointed out clearly, so it can not be commonly used.

Using the eigenvalue distribution of the sample set in feature space, this paper introduces the concept of equivalent samples in one specific feature space. With the help of this conception the equivalent sample can be described by this sample and its weight value, and then the weight set can be constructed based on the weighted sample. After that the WFCM algorithm is presented. In weighted set, the number of clustered samples is much smaller than the ordinary set, which means the obvious reduction of the computation time and resources. For comparison and application, the gray image segmentation uses FCM and WFCM algorithms respectively.

Translated from *Journal of Sichuan University* (*Engineering Science Edition*), 2005, 37(6): 130–134 (in Chinese)

KUANG Ping (✉), ZHU Qing-xin, WANG Ming-wen,
CHEN Xu-dong, QING Li
School of Computer Science and Engineering,
University of Electronic Science and Technology of China,
Chengdu 610054, China
E-mail: kuangping@uestc.edu.cn

2 WFCM algorithm based on weighted set

2.1 FCM algorithm description

The FCM algorithm divides a given set with n_s samples into C clusters, value u_{ji} is defined as the membership of the ith sample s_i in the jth cluster and u_{ji} satisfies:

$$u_{ji} \in [0,1], \quad \forall i \quad \sum_{j=1}^{C} u_{ji} = 1$$

$$\text{and} \quad \forall j \quad 0 < \sum_{i=1}^{n_s} u_{ji} < n \tag{1}$$

$d_{ji}(m_j, s_i)$ represents the distance between the sample s_k and the centroid m_j and it is usually defined as Euclidean distance, namely, $d_{ji}(m_j, s_i) = \|x_i - m_j\|^2$, where x_i is the eigenvalue of sample s_i. Mathematically, the standard FCM objective function is given by:

$$J_f = \sum_{j=1}^{C}\sum_{i=1}^{n_s} u_{ji}^b \|x_i - m_j\|^2 \tag{2}$$

For a set of Eqs . (1) and (2), the iteration equation of FCM can be derived by Lagrange method, that is:

$$m_j = \frac{\sum_{i=1}^{n_s} u_{ji}^b x_i}{\sum_{i=1}^{n_s} u_{ji}^b}, \qquad u_{ji} = \frac{\left(1\Big/\left(\|x_i - m_j\|^2\right)\right)^{1/(b-1)}}{\left(1\Big/\sum_{j=1}^{C}\|x_i - m_j\|^2\right)^{1/(b-1)}} \tag{3}$$

where $i = 1, 2, \cdots, n_s$; $j = 1, 2, \cdots, C$.

2.2 WFCM algorithm description

Set $S = \{s_1, s_2, \cdots, s_{n_s}\}$ has n_s samples, $x_i(H_k)$ $(k = 1, 2, \cdots, n_H)$ denotes the eigenvalue of sample s_i in feature space H_k, eigenvector $\boldsymbol{x}_i = (x_i(H_1), x_i(H_2), \cdots, x_i(H_{n_H}))^{\mathrm{T}}$ is constructed by all eigenvalues of sample s_i in different feature space, where n_H is the dimension value of feature space.

Definition 1 The eigenvalue of sample s_i in feature space H_k is defined by characteristic mapping M_k, namely, $M_k : M_k(s_i) = x_i(H_k)$ $(i = 1, 2, \cdots, n_s;\ k = 1, 2, \cdots, n_H)$.

Definition 2 Feature space SF is spanned by the eigenvector $\boldsymbol{x}_i$ $(i = 1, 2, \cdots, n_s)$, where SF_k $(k = 1, 2, \cdots, n_H)$, which is the subspace of feature space SF, denotes the set of the eigenvalues of sample set S in feature space H_k.

In subspace SF_k, $\exists s_m, s_n \in S$ $(m \neq n)$ satisfies $M_k(s_m) = M_k(s_n)$, which means that the eigenvalue of two different samples in one feature space is equal. Then, the sample s_m and s_n are named as the equivalent samples in feature space H_k.

Definition 3 Samples s_m, s_n satisfying $M_k(s_m) = M_k(s_n)$ are called equivalent samples in feature space H_k.

Definition 4 $w_i(H_k)$ denotes the weight of sample s_i in feature space H_k and it is defined as the total number of samples that is equivalent to sample s_i under Definition 3.

Definition 5 $s_i \cdot w_i(H_k)$ denotes the sample s_i with the weight w_i in feature space H_k, in such situation, sample s_i is also called weighted sample s_i.

Definition 6 The weighted samples in feature space H_k can construct the weighted set WS_k, which is: $WS_k = \{s_1 \cdot w_1(H_k), s_2 \cdot w_2(H_k), \cdots, s_{n_{ws}} \cdot w_{n_{ws}}(H_k)\}$, where $n_{ws} \le n_s$, and $\sum_{i=1}^{n_{ws}} w_i = n_s$. the simplified representation of weighted sample is: $WS = \{s_1 \cdot w_1, s_2 \cdot w_2, \cdots, s_{n_{ws}} \cdot w_{n_{ws}}\}$.

Theorem 1 In feature space H_k, the cluster results, which are derived by using FCM alogorithm to cluster in weighted set WS_k and set S, are equivalent.

Proof Set S can be represented in the following format: $S = \{\underbrace{s_1, \cdots}_{w_1}, \underbrace{s_2, \cdots}_{w_2}, \cdots, \underbrace{s_i, \cdots}_{w_i}, \cdots, \underbrace{s_{n_{ws}}, \cdots}_{w_{ns}}\}$, where: w_i denotes the equivalent number of samples to s_i in H_k. With the Definition 6, the weighted set of S in feature space is WS_k. According to Definition 4, WS_k can be rewritten as $WS_k = WS_k = \{\underbrace{s_1, \cdots, s_1}_{w_1}, \underbrace{s_2, \cdots s_2}_{w_2}, \cdots, \underbrace{s_{n_{ws}}, \cdots s_{n_{ws}}}_{w_{ns}}\}$, and it is obvious that $S \neq WS'_k$, but, reviewing the eigenvalue set SF_k and SWS'_k which belong to set S and SWS'_k respectively in feature space H_k, we find that

$$SF_k = SWS'_k = \{\underbrace{x_1(H_k), \cdots}_{w_1}, \underbrace{x_2(H_k), \cdots}_{w_2}, \cdots, \underbrace{x_i(H_k), \cdots}_{w_i}, \cdots, \underbrace{x_{n_{ws}}(H_k), \cdots}_{w_{n_{ws}}}\} \tag{4}$$

Namely, the two eigenvalue sets of S and WS'_k in the same feature space H_k are equivalent. There is no doubt that same eigenvalue sets result in the same cluster result, so set WS_k and S can achieve the equivalent cluster results in feature space H_k.

Theorem 2 The objective function of WFCM is:

$$J_f = \sum_{j=1}^{C}\sum_{i=1}^{n_{ws}} w_i u_{ji}^b \|x_i - m_j\|^2 \tag{5}$$

Proof The weighted set of S in feature space H_k is defined as: $WS_k = \{s_1 \cdot w_1, s_2 \cdot w_2, \cdots, s_{n_{ws}} \cdot w_{n_{ws}}\}$, and according to Definition 3, and Definition 4, WS_k can be rewritten as $WS'_k = \{\underbrace{s_1, \cdots, s_1}_{w_1}, \underbrace{s_2, \cdots, s_2}_{w_2}, \cdots, \underbrace{s_{n_{ws}}, \cdots, s_{n_{ws}}}_{w_{n_{ws}}}\}$, so according to Eq. (2), the objective function is:

$$J_f = \sum_{j=1}^{C}\sum_{i=1}^{n_{ws}} u_{ji}^b \|x_i - m_j\|^2$$

$$= \sum_{j=1}^{C} (\underbrace{u_{j1}^b \|x_1 - m_j\|^2 + \cdots}_{w_1} + \underbrace{u_{j2}^b \|x_2 - m_j\|^2 + \cdots}_{w_2}$$

$$+ \cdots + \underbrace{u_{jn_{ws}}^b \|x_{n_{ws}} - m_j\|^2 + \cdots}_{w_{n_{ws}}})$$

$$= \sum_{j=1}^{C} (w_1 u_{j1}^b \|x_1 - m_j\|^2 + w_2 u_{j2}^b \|x_2 - m_j\|^2 + \cdots)$$

$$= \sum_{j=1}^{C}\sum_{i=1}^{n_{ws}} w_i u_{ji}^b \|x_i - m_j\|^2$$

the result is also Eq. (5)

Theorem 3 The membership u_{ji} between the weighted sample and the cluster center satisfies:

$$u_{ji} \in [0,1], \quad \forall i, \quad \sum_{j=1}^{C} w_i u_{ji} = 1$$
$$\forall j \quad 0 < \sum_{i=1}^{n_{ws}} w_i u_{ji} < n \tag{6}$$

Proof With the help of Eq. (2) and the definition of the weight of the sample in Definition 4, we can achieve this result.

Theorem 4 The iteration equations for calculating the cluster center m_j and the membership matrix $\boldsymbol{U}$ are:

$$m_j = \frac{\sum_{i=1}^{n_{ws}} w_i u_{ji}^b x_i}{\sum_{i=1}^{n_{ws}} w_i u_{ji}^b}$$
$$u_{ji} = \frac{\left(1\Big/\left(w_i \|x_i - m_j\|^2\right)\right)^{1/(b-1)}}{\left(1\Big/\sum_{j=1}^{C} w_i \|x_i - m_j\|^2\right)^{1/(b-1)}} \tag{7}$$

where $i = 1, 2, \cdots, n_{ws}$, $j = 1, 2, \cdots, C$

Proof According to Theorems 2 and 3, using Lagrange can achieve this result.

Discussion:

1) From Theorem 1, the cluster on sample set S in feature space H_k by using FCM has an equal result as that of WFCM on weight set WS_k.

2) The essence of WFCM is to attain less computation by reducing the number of samples. According to Definition 6, the number of samples n_s in set S and that of n_{ws} in weight set WS_k satisfy $n_{ws} \le n_s$. Thus cluster in WS_k always takes less time.

The performance of WFCM is related to the value of $R = n_{sw}/n_s$. A smaller R can achieve a higher performance. In this paper, the experiment of gray image segmentation proves that when R is smaller, WFCM can obviously accelerate the processing of the cluster. However, it also happens that no sample has an equivalent in the feature space. In this circumstance, the performance of WFCM equals to that of FCM. We can revise the determinant condition of Definition 2, for instance, define the samples that satisfy $|M_k(s_m) - M_k(s_n)| \le \varepsilon$ as the equivalents. Then, an equivalent set is built to perform WFCM. We can prove that this result is close to that of FCM. We can also use the result of WFCM as the initial value of FCM to get the final fuzzy membership matrix.

2.3 The iteration steps of WFCM

The iteration steps of WFCM are similar to that of FCM. But first we should compute the weight set WS_k for sample set S in H_k according to the eigenvalue of sample set S in H_k to get the weight matrix. Use Eq. (7) instead of Eq. (3) to compute the cluster center and fuzzy membership matrix iteratively.

3 The application of WFCM: gray image segmentation

Fuzzy set theory is a good tool to describe the uncertainty of an image. Many researchers have applied FCM in gray image processing. But when processing clusters for navigation and satellite pictures, the computation cost of FCM is too large.

Notice that the gray value of a pixel is between 0 and 255. We can use the weight set to describe the sample images to reduce the number of samples. Figure 1 is an image to be segmented. Figure 2(a) is the corresponding eigenvalue distribution. Figure 2(b) is the weight value statistic of the eigenvalue in Fig. 2(a). It's also the one dimension histogram. The weight set WS_k can be computed through Fig. 2(b) for WFCM.

Fig. 1 The gray image to be segmented

In order to judge the performance of WFCM, we do some segmentation experiments under different resolutions and numbers of cluster centers. The results are compared with FCM. The experiments are performed on CII 1 G, 256 M PC and MATLAB 6.5. Each test is done for 3 times, the average value is then recorded. The experiments results are listed in Table 1. The method we use to choose the initial value is as follows:

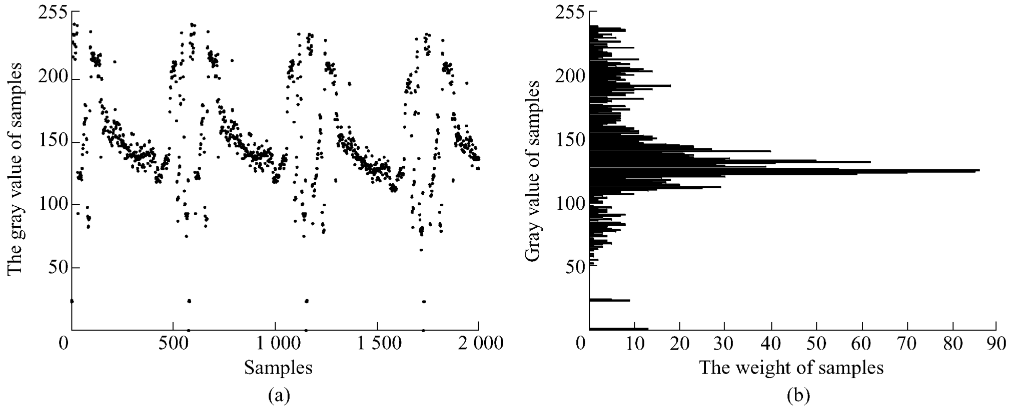

Fig. 2 The feature space. (a) and weight value distribution; (b) of samples

Table 1 The two algorithms' performance comparison used in the same gray image segmentation

Image size /pixel	C	The number of samples		Average iteration times		Objective function value		Average time spent /s		proportion of time spent /%
		WFCM	FCM	WFCM	FCM	WFCM	FCM	WFCM	FCM	
	3			23	25	99.783	99.783	0.141	4.677	3.015
176×144	4	222	2.53×10^4	71	78	60.883	60.883	0.461	20.39	2.271
	5			50	51	35.475	35.475	0.283	16.10	1.739
	3			22	25	340.78	340.78	0.171	16.98	1.007
352×288	4	231	1.01×10^5	38	56	229.23	229.23	0.221	56.27	0.393
	5			52	59	122.94	122.94	0.361	67.69	0.533
	3			25	27	1 372.7	1 372.7	0.191	70.16	0.272
704×576	4	234	4.01×10^5	155	70	910.26	910.26	1.022	258.3	0.396
	5			62.3	66	503.80	503.80	0.511	295.5	0.173

Firstly, construct a two dimensional Gauss template $H(u,v)=\exp(-D^2(u,v)/2\sigma^2)$. $D(u,v)$ is the distance from origin. Use this template and the image to compute the convolution.

Secondly, compute the gray statistic histogram after the smoothing.

Thirdly, use the distribution of the peak value, the number of cluster centers and threshold ε to appoint initial cluster centers $p_j(j=1,\cdots,C)$. Produce the initial cluster center $m_j^{(0)}$ in interval in interval $[p_j-\varepsilon, p_j+\varepsilon]$.

From Table 1, we can see that the average time spent in WFCM is far less than that of FCM. The proportion can reach 3.015 % at the best condition. When the image size increases, this proportion decreases. In the worst condition, it is 0.173 %. These comparisons show the advantage of WFCM. To explain this, we notice that when the image size increases, the number of samples in weight set stays almost the same, while the samples for traditional FCM increase significantly. This leads to the time of iteration increases accordingly.

4 Conclusions

In this paper, we propose WFCM for data clusters. Compared with FCM, our method is faster. Weighted sample-based WFCM can be applied in gray image segmentation as well as other large sample cluster fields. It is more effective and faster than FCM.

References

1. Yong Wonlin, Lee S. U., On the color image segmentation algorithm based on the thresh holding and the fuzzy C-means techqiques, Pattern Recognition, 1990, 23: 935–952
2. Bezdec C. J., Convergence theory for fuzzy C-means: counter examples and repairs, IEEE Transtraction on System, Man and Cybernetics, 1987, 17: 873–877
3. Pal N. R., Bezdek J. C., Complexity reduction for "large image" processing, IEEE Transaction on Systems, Man and Cybernetics, 2002, B32(5): 598–611
4. Thitimajshima P., A new modified fuzzy C-means algorithm for mulitispectral satellite image segmentation, IEEE 2000 International Proceedings, IGARSS, 2000, 4: 1684–1686
5. Eschrich S., Ke Jing-wei, Hall L. O. et al., Fast accurate fuzzy clustering through data reduction, IEEE Transaction on Fuzzy Systems, 2003, 11(2): 262–270
6. Ding Zhen, Hu Zhong-shan, Yang Jing-yu et al., FCM algorithm for the research of intensity image segmentation, Journal of Electronics, 1997, (5): 39–43
7. Bezdek J. C., Pattern Recognition With Fuzzy Objective Function Algorithm. New York: Plenum Press, 1981
8. Pal N. R., Bezdek J. C., On cluster validity for the fuzzy C-means model, IEEE Transaction on Fuzzy Systems, 1995, 3(3): 370–379