Neurocomputing 73 (2010) 2935–2943

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Locality sensitive C-means clustering algorithms

Pengfei Huang, Daoqiang Zhang *

Department of Computer Science and Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China

ARTICLE INFO

Article history:
Received 25 December 2009
Received in revised form
22 July 2010
Accepted 29 July 2010
Communicated by X. Gao
Available online 18 September 2010

Keywords:
Locality sensitive weight
Fuzzy C-means (FCM)
Semi-supervised clustering

ABSTRACT

The concept of preserving locality information in dimensionality reduction and semi-supervised classification have been very popular recently. In this paper, we attempt to use locality sensitive weight for clustering, where the neighborhood structure information between objects are transformed into weights of objects. We develop two novel locality sensitive C-means algorithms, i.e. Locality-weighted Hard C-Means (LHCM) and Locality-weighted Fuzzy C-Means (LFCM), following the standard C-Means and fuzzy C-means, respectively. In addition, two semi-supervised extensions of LFCM are proposed to better use some given partial supervision information in data objects. Experimental results on both artificial and real datasets validate the effectiveness of the proposed algorithms.

1. Introduction

Clustering deals with finding a structure in a collection of unlabeled data. A cluster is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters. At present, clustering algorithms can be categorized into several types, such as partitional method, hierarchical method, density-based method, grid-based method and model-based method [1].

In this paper, we mainly focus on the partitional method. Presently, most clustering algorithms treat all data samples equally in the clustering process, such as hard C-Means (HCM) and its fuzzy extension, i.e. fuzzy C-Means (FCM) [2]. However, different samples may play different roles in the clustering process, because the samples distribute nonuniformly and asymmetrically. Moreover, a sample may contribute to the clustering results differently in different processes. Hence, it is very useful to give an appropriate sample weight in cluster analysis. For that purpose, sample weighting clustering algorithm have been proposed in literature [3–7].

In sample weighting clustering, the weight of each sample is very important, since it determines the impact of the sample on the clustering analysis. Conditional fuzzy C-means [3] and deterministic annealing clustering [4] consider various contributions of different samples and take account of sample weighting. However, the application of the above algorithms are limited because they need users or heuristic principle to weight samples.

To overcome that problem, Nock and Nisseslen proposed a formalized clustering framework, borrowing the idea of the boosting algorithm, which offers penalizing solutions via weights on the samples [5]. In their paper [6], they pointed out the importance of calculating the sample weight automatically during the process of clustering analysis. Li et al. have proposed a typical-weighting clustering algorithm for large datasets. It can obtain original clustering samples using the atom-clustering algorithm, then weight them according to the atom number of samples [7]. Zhang et al. have introduced the document clustering algorithm based on sample weighting, which utilizes PageRank value as the weight of the samples and then assigns different weights to various samples, such that more reasonable centers could be obtained [8]. However, it is only applicable to document clustering and related areas. Gao et al. have presented weighted fuzzy C-means clustering, which considers the appearance probability of the gray levels from the gray histogram in an image as the weight parameter and hence improves the algorithm efficiency [9]. However, it is only suitable for image data. Recently, the weighting idea has also been used for clustering of fuzzy and relational data, respectively [10,11].

On the other hand, in machine learning and pattern recognition community, there have been a recent trend to utilize the local structural information for learning. For example, The concept of preserving locality information in dimensionality reduction and semi-supervised classification have been very popular recently [12,13]. Literatures [14–16] effectively utilizes the structure information by building a graph incorporating neighborhood information of the dataset. Using the notion of the graph Laplacian, a weight matrix which indicates the intrinsic structure is set up. However, to the best of knowledge, it remains unknown whether the local structure information among the clustering objects is also helpful to sample weighting clustering.

In this paper, motivated by the idea of optimally preserving the neighborhood structure in dimensionality reduction and semi-supervised learning, we propose a novel locality preserving

* Corresponding author.
E-mail address: dqzhang@nuaa.edu.cn (D. Zhang).

0925-2312/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2010.07.015

weighting scheme for clustering, from which two new algorithms, i.e. Locality-weighted Hard C-Means (LHCM) and Locality-weighted Fuzzy C-Means (LFCM) are developed. LHCM and LFCM calculate the distance between the samples and the centers to gain a proper weight parameter so that they can primely describe the neighborhood structure of the data. In addition, the proposed methods are extended for semi-supervised cases to use the available supervision information in data, e.g. partial labeled data or pairwise constraints which specify whether a pair of data belong to the same class or not [16].

The rest of this paper is organized as follows: In Section 2 the background on HCM and FCM are briefly described. Section 3 derives the proposed LHCM and LFCM Clustering algorithms in detail. Section 4 gives the semi-supervised extensions on LFCM. The experimental results are given in Section 5. Finally, we conclude this paper in Section 6.

2. Background

2.1. HCM

HCM is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. The algorithm classifies n vectors x_j ($j=1,2\ldots,n$) through a certain number of clusters (assume c clusters G_i ($i=1,2\ldots,c$) fixed a priori, and calculates each centroid v_i aiming at minimizing the objective function. The objective function is defined as follows:

$$J=\sum_{i=1}^{c}\sum_{x_j\in G_i}\|x_j-v_i\|^2 \tag{1}$$

2.2. FCM

FCM is a method of clustering which allows one piece of data to belong to two or more clusters. It is based on minimization of the following objective function:

$$J(U,v_1,\ldots,v_c)=\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}^m\|x_j-v_i\|^2 \tag{2}$$

where x_j is the jth data example, v_i is the ith cluster center, and u_{ij} is the degree of membership of x_j in the cluster i. The weighting exponent m is a real number greater than 1 and the appropriate values depend on datasets. The theoretical analysis on the parameter m can be seen in the Refs. [17,18]. Finally in (2), $\|\cdot\|$ is a norm measuring the distance metric between data examples and the cluster centers. Fuzzy partitioning is carried out through an alternate iterative optimization [2,19] of the objective function shown above, with some properties:

$$\sum_{i=1}^{c}u_{ij}=1,\quad \forall j=1,\ldots,n,\ 0<u_{ij}<1,\ 0<\sum_{j=1}^{n}u_{ij}<n \tag{3}$$

3. Locality sensitive C-means clustering

3.1. Locality-weighted hard C-means (LHCM)

Suppose that $X=x_1,x_2,\ldots,x_n$ is a d-dimensional database with n points, and is divided into c clusters $v=v_1,v_2,\ldots,v_c$, each cluster can be represented by its cluster center v_i. The objective function of LHCM is defined as follows:

$$J=\sum_{i=1}^{c}\sum_{x_j\in G_i}s_{ij}\|x_j-v_i\|^2 \tag{4}$$

where G_i denotes the ith cluster and s_{ij} is the weight between points $\{x_j\}$ and centers $\{v_i\}$. To preserve the neighborhood structure information in the weight, we define the weighting function as follows:

$$s_{ij}=e^{-(\|x_j-v_i\|^2)/t_i} \tag{5}$$

where t_i is a scaling parameter. When $t_i\to 0$, the weight matrix becomes the most important ingredient of the clustering result while the weights are very similar with each other. In this case, the weighted clustering will generate much poorer clustering result. On the other hand, when $t_i\to\infty$, the weight matrix has entries all equal to 1, and thus the weighted clustering is degraded into non-weighted clustering.

In order to choose appropriate values for the weights, we use a local scale for t_i as follows:

$$t_i=\begin{cases}\sigma_i^2 & x_j\in N_{ik}\\ \left(\frac{1}{c}\sum_{i=1}^{c}\sigma_i\right)^2 & \text{otherwise}\end{cases} \tag{6}$$

where $\sigma_i=(1/k)\sum_{j=1}^{k}\|x_j-v_i\|^2$, k is the number of the neighbors of the ith center. N_{ik} is the k Nearest Neighbor (k-NN) neighborhoods of the ith cluster. From (6), we can see that the scale can automatically adapt to the local structure. In practice, usually it is much easier to choose values for k than for t_i.

Let $u_{ij}\in\{0,1\}$ denote whether $x_j\in G_i$ or not, i.e. $u_{ij}=1$ means $x_j\in G_i$, and vice versa. Following the standard HCM, given the locality weight s_{ij} we can easily derive the solutions of LHCM by the following alternate iterations between the indicator u_{ij} and the cluster centers v_i. The detailed pseudo-code of LHCM is listed in Table 1.

$$u_{ij}=\begin{cases}1 & \text{if } \forall k,\|x_j-v_i\|^2\le\|x_j-v_k\|^2\\ 0 & \text{otherwise}\end{cases} \tag{7}$$

$$v_i=\frac{\sum_{j=1}^{n}u_{ij}s_{ij}x_j}{\sum_{j=1}^{n}u_{ij}s_{ij}} \tag{8}$$

3.2. Locality-weighted fuzzy C-Means (LFCM)

As in LHCM, we modify the standard FCM by introducing the locality weight s_{ij}. The objective function of LFCM is defined as follows:

$$J(U,v_1,\ldots,v_c)=\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}^m s_{ij}\|x_j-v_i\|^2 \tag{9}$$

where x_j is the jth of d-dimensional measured data, v_i is the ith cluster center, u_{ij} represents the fuzzy membership of the j-th point with respect to cluster i, s_{ij} is the locality weight between points x_j and centers v_i. The parameter m is a weighting exponent

Table 1
The LHCM algorithm.

Initialize: the cluster centers $v^{(0)}=\{v_1^{(0)},v_2^{(0)},\ldots,v_c^{(0)}\}, l=0, \varepsilon>0$

Step 1: update $s_{ij}^{(l+1)}$ by the equation: $s_{ij}^{(l+1)}=e^{-(\|x_j-v_i^{(l)}\|^2)/t_i}$

Step 2: update $u_{ij}^{(l+1)}$ with the equation:

$$u_{ij}^{(l+1)}=\begin{cases}1 & \text{if } \forall k,\|x_j-v_i\|^2\le\|x_j-v_k\|^2\\ 0 & \text{otherwise}\end{cases}$$

Step 3: update $v_i^{(l+1)}$ with the equation: $v_i^{(l+1)}=\frac{\sum_{j=1}^{n}u_{ij}^{(l+1)}s_{ij}^{(l+1)}x_j}{\sum_{j=1}^{n}u_{ij}^{(l+1)}s_{ij}^{(l+1)}}$

If $\max_i\|v_i^{(l+1)}-v_i^{(l)}\|<\varepsilon$, then stop; else $l=l+1$ and go to step 1.

Table 2
LFCM algorithm.

Input:	the cluster centers $v^{(0)}=\{v_1^{(0)},v_2^{(0)},\ldots,v_c^{(0)}\}, l=0, \varepsilon>0$ the number of clusters c, parameters $m=2$, small change $\varepsilon>0$, and the maximal iterations $T_{\max}$
Output:	the final prototype v and membership U
Step 1:	Compute the weight matrix $S=\{s_{ij}^{(l+1)}\}$ by the equation: $s_{ij}^{(l+1)}=e^{-(\|x_j-v_i^{(l)}\|^2)/t_i}$
Step 2:	update $u_{ij}^{(l+1)}$ with the equation: $u_{ij}^{(l+1)}=\frac{1}{\sum_{k=1}^{c}\left(\frac{s_{ij}^{(l)}\|x_j-v_i^{(l)}\|^2}{s_{ik}^{(l)}\|x_k-v_i^{(l)}\|^2}\right)^{1/(m-1)}}$
Step 3:	update $v_i^{(l+1)}$ with the equation: $v_i^{(l+1)}=\frac{\sum_{j=1}^{n}(u_{ij}^{(l+1)})^m s_{ij}^{(l+1)}x_j}{\sum_{j=1}^{n}(u_{ij}^{(l+1)})^m s_{ij}^{(l+1)}}$
If $\max_i\|v_i^{(l+1)}-v_i^{(l)}\|<\varepsilon$, then stop; else $l=l+1$ and go to step 1.	

on each fuzzy membership that determines the amount of fuzziness of the resulting classification.

By definition, each sample point x_j satisfies the constraint that $\sum_{i=1}^{c}u_{ij}=1$. Hence like in FCM, in order to obtain the solutions of LFCM we can rebuild its objective function by minimizing:

$$\bar{J}(U,S,v_1,\ldots,v_c,\lambda_1,\ldots,\lambda_n)=\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}^m s_{ij}\|x_j-v_i\|^2+\sum_{j=1}^{n}\lambda_j\left(\sum_{i=1}^{c}u_{ij}-1\right) \tag{10}$$

Suppose that $\partial\bar{J}/\partial v_i=0$, we obtain

$$v_i=\frac{\sum_{j=1}^{n}u_{ij}^m s_{ij}x_j}{\sum_{j=1}^{n}u_{ij}^m s_{ij}} \tag{11}$$

In order to get the optimization membership, we compute $\partial\bar{J}/\partial u_{ij}$

$$\frac{\partial\bar{J}}{u_{ij}}=mu_{ij}^{m-1}s_{ij}\|x_j-v_i\|^2+\lambda_j \tag{12}$$

Suppose that $\partial\bar{J}/\partial u_{ij}=0$, then

$$u_{ij}=\left(\frac{-\lambda_j}{ms_{ij}\|x_j-v_i\|^2}\right)^{1/(m-1)} \tag{13}$$

According to $\sum_{i=1}^{c}u_{ij}=1$ and (13), we obtain

$$(-\lambda_j)^{1/(m-1)}=\left(\sum_{k=1}^{c}\frac{1}{ms_{ij}\|x_j-v_i\|^2}\right)^{-1/(m-1)} \tag{14}$$

We use (14) to re-formulate (13), then we obtain

$$u_{ij}=\frac{1}{\sum_{k=1}^{c}\left(\frac{s_{ij}\|x_j-v_i\|^2}{s_{ik}\|x_k-v_i\|^2}\right)^{1/(m-1)}} \tag{15}$$

Similarly as LHCM, given the locality weight s_{ij} in (5) the solutions of LFCM can be found by the following alternate iterations between the indicator the membership function u_{ij} in (15) and the cluster centers v_i in (11). The detailed pseudo-code of LFCM is listed in Table 2.

4. Generalization

4.1. Semi-supervised LFCM (SLFCM)

Traditional clustering algorithms usually rely on a pre-defined similarity measure between unlabeled data to attempt to identify natural classes of items [20]. When compared to what a human expert would present on the same data, the results obtained may be somehow disappointing if the pre-defined similarity measure employed by the system is too different from the one a human expert would use. In order to obtain clusters fitting user expectations better, we can use, in addition to some unlabeled data, some limited form of supervision, such as labels of some data samples, constraints specifying whether two data items belong to a same cluster or not. The main goal of the approach above named semi-supervised clustering [21–23] is to allow a human to bias clustering with a minimum of effort by providing a small amount of knowledge concerning either class labels for some items or pairwise constraints between data items.

In this section, we put forward a modification of LFCM with the idea of semi-supervision called semi-supervised LFCM (SLFCM). The objective function to be minimized by SLFCM combines the feature-based similarity between data points and cost terms for the pairwise constraints.

Suppose that $\mathcal{M}$ is the set of *must-link* pairs such that $(x_i,x_j)\in\mathcal{M}$ implies x_i and x_j should be assigned to the same cluster, and $\mathcal{C}$ is the set of *cannot-link* pairs such that $(x_i,x_j)\in\mathcal{C}$ implies x_i and x_j should be assigned to the different clusters. Using the same notations as for LFCM, we can write the objective function of SLFCM as follows:

$$J_{SLFCM}(X,S,V)=\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}^m s_{ij}\|x_j-v_i\|^2+\alpha\left(\sum_{(x_j,x_k)\in\mathcal{M}}\sum_{i=1}^{c}\sum_{l=1,l\neq i}u_{ij}s_{ij}u_{lk}s_{lk}+\sum_{(x_j,x_k)\in\mathcal{C}}\sum_{i=1}^{c}u_{ij}s_{ij}u_{ik}s_{ik}\right) \tag{16}$$

The first term in (16) is the sum of squared distances to the prototypes weighted by constrained memberships. This term reinforces the compactness of the clusters. The second term is composed of the following two parts:(1) The cost of violating the pairwise *must-link* constraints; (2) The cost of violating the pairwise *cannot-link* constraints. This term is weighted by α, which is a way to specify the relative importance of the supervision. After combining the two terms and choosing the appropriate value of α, the final result will confirm the compactness of the clusters as possible and consider the semi-supervision information as sufficient. To minimize the objective function in (16), we rebuild it with the membership property $\sum_{i=1}^{c}u_{ij}=1$ as follows:

$$\bar{J}_{SLFCM}(X,S,V,\lambda)=\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}^m s_{ij}\|x_j-v_i\|^2+\alpha\left(\sum_{(x_j,x_k)\in\mathcal{M}}\sum_{i=1}^{c}\sum_{l=1,l\neq i}u_{ij}s_{ij}u_{lk}s_{lk}+\sum_{(x_j,x_k)\in\mathcal{C}}\sum_{i=1}^{c}u_{ij}s_{ij}u_{ik}s_{ik}\right)-\sum_{j=1}^{n}\lambda_j\left(\sum_{i=1}^{c}u_{ij}-1\right) \tag{17}$$

When the extreme values are achieved, the following equations should be satisfied:

$$\frac{\partial J}{\partial v_i}=0 \quad \text{and} \quad \frac{\partial J}{\partial u_{ij}}=0 \tag{18}$$

Ignoring the relativeness between α and other parts, we can derive the updating equation of the centers satisfying the first condition as follows:

$$v_i=\frac{\sum_{j=1}^{n}u_{ij}^2 s_{ij}x_j}{\sum_{j=1}^{n}u_{ij}^2 s_{ij}} \tag{19}$$

Assumed that the memberships in the next iteration must not change significantly, then the memberships in the last time can be

used. Hence, we can obtain

$$u_{ij}=\frac{\lambda_j}{2s_{ij}\|x_j-v_i\|^2}-\alpha\frac{\sum_{(x_j,x_k)\in\mathcal{M}}\sum_{l=1,l\neq i}s_{ij}u_{lk}s_{lk}+\sum_{(x_j,x_k)\in\mathcal{C}}s_{ij}u_{ik}s_{ik}}{2s_{ij}\|x_j-v_i\|^2} \tag{20}$$

At last, we obtain the updating equation of the membership as follows:

$$u_{ij}=\frac{1}{\sum_{r=1}^{c}\frac{s_{ij}\|x_j-v_i\|^2}{s_{rj}\|x_j-v_r\|^2}}+\frac{\alpha}{\sum_{r=1}^{c}\frac{s_{ij}\|x_j-v_i\|^2}{s_{rj}\|x_j-v_r\|^2}}\times\sum_{r=1}^{c}\frac{\sum_{(x_j,x_k)\in\mathcal{M}}\sum_{l=1,l\neq i}s_{ij}u_{lk}s_{lk}+\sum_{(x_j,x_k)\in\mathcal{C}}s_{ij}u_{ik}s_{ik}}{2s_{ij}\|x_j-v_i\|^2}-\alpha\frac{\sum_{(x_j,x_k)\in\mathcal{M}}\sum_{l=1,l\neq i}s_{ij}u_{lk}s_{lk}+\sum_{(x_j,x_k)\in\mathcal{C}}s_{ij}u_{ik}s_{ik}}{2s_{ij}\|x_j-v_i\|^2} \tag{21}$$

Parameter α determines the weight of the semi-supervision. So it is important to choose an appropriate value for α. Because the number of the pairwise constraints is usually small compared to the number of all the samples. To let the constraints take react in the clustering process effectively, we set α as follows:

$$\alpha=\frac{N}{M}\frac{\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}^2s_{ij}\|x_j-v_i\|^2}{\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}^2s_{ij}} \tag{22}$$

4.2. Semi-supervised LFCM with spatial constraints

In this section, we apply semi-supervised LFCM (SLFCM) for image segmentation tasks. It has been shown that introducing spatial constraints can effectively improve the robustness of clustering methods for image segmentation [24–28]. In this subsection we introduce the spatial constraints in SLFCM following [28] and develop an approach called Semi-supervised LFCM with Spatial constraints (SLFCM-S) which considers the spatial context as well as some pair-wise constraints for image segmentation. We restrict the membership functions integrating some pair-wise constraints in LFCM to be spatially smooth, and the objective function is as follows:

$$\begin{aligned}J=&\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}^2s_{ij}\|x_j-v_i\|^2\\&+\frac{\alpha}{N_R+M_R+C_R}\sum_{i=1}^{c}\sum_{j=1}^{n}s_{ij}u_{ij}^2\left(\sum_{r\in N_j}(1-u_{ir})^2\right.\\&\left.+\sum_{r\in M_j}\sum_{k\in N_r}(1-u_{ik})^2+\sum_{r\in C_j}\sum_{k\in N_r}u_{ik}^2\right)\\&\text{s.t}\quad\sum_{i=1}^{c}u_{ij}=1,\quad\forall j=1,\ldots,n\end{aligned} \tag{23}$$

where x_j is the observation at pixel j, N_j is the set of neighbors of pixel j(without the pixel j itself) and N_R represents the domain; M_j is the set of pixels which have the *must-link* relation with pixel j and M_R represents the domain; C_j is the set of pixels which have the *cannot-link* relation with pixel j and C_R represents the domain. α is the regularization coefficient which controls the trade-off between minimizing the LFCM objective function and obtaining smooth membership functions. When $\alpha=0$, the algorithm degrades into LFCM.

5. Experimental evaluation

In this section, we compare the effectiveness and efficiency of the HCM, FCM, LHCM and LFCM algorithms on two artificial datasets and the standard UCI datasets, validate the effectiveness of the SLFCM and SLFCM-S algorithms and make experiments on some medical images to test the performance of the proposed algorithm for image segmentation. In all the following experiment, we set the parameter $m=2$, and others are detailed in specific experiments.

5.1. Experiments on synthetic datasets

We use two synthetic datasets D1 and D2. The dataset D1 contains 100 sample points in two dimensions which are divided into two clusters. One of the clusters contains 50 points in a Gaussian distribution with the center (0,0), and the other cluster contains 49 points in a Gaussian distribution with the center (3,0). Besides, there's an outlier at (200,0).

Fig. 1 presents the clustering performance of HCM, LHCM, FCM and LFCM on the dataset D1 which only contains 99 points without the outlier. It shows that HCM and FCM are interfered by the noise so seriously that take the two clusters as one cluster and the outlier as the other one, while the two weighted clustering approximately avoid the effect of the noise and roughly divide the points into two cluster correctly. The results ensure that the algorithm proposed can effectively overcome the problem of the dataset with noise. By introducing the weight matrix, the algorithm assigns a small weight to the outlier for it is far away from the centroid, hence the robustness of the algorithm is improved and better clustering performance is achieved.

Dataset D2 is class-unbalanced. One cluster contains 25 points and the other contains 125 points. The clustering results of FCM and LFCM on D2 are shown in Fig. 2. Fig. 2(a) shows that FCM takes partial points of the second cluster as the first cluster while LFCM correctly divide the points completely. Obviously, on this dataset, the performance of LFCM is significantly superior to FCM. By introducing the weight matrix, the within-class becomes more tight and the between-class's distance is enlarged. Therefore, we can see that the proposed algorithm can effectively handle the class-unbalanced dataset.

5.2. Experiments on UCI datasets

In this section we assess the relative performance of our algorithms over seven UCI datasets. We use *F-measure* as the criteria to assess the clustering performance. Here *F-measure* is defined as

$$F=\frac{(1+\beta^2)\times p\times r}{\beta^2\times p\times r} \tag{24}$$

where

$$p=\frac{t_p}{t_p+f_p} \tag{25}$$

$$r=\frac{t_p}{t_p+f_n} \tag{26}$$

Here p is precision, r is recall and β allows one to weight either precision or recall more heavily. They are balanced when $\beta=1$. In most experiments, there is no particular reason to favor precision

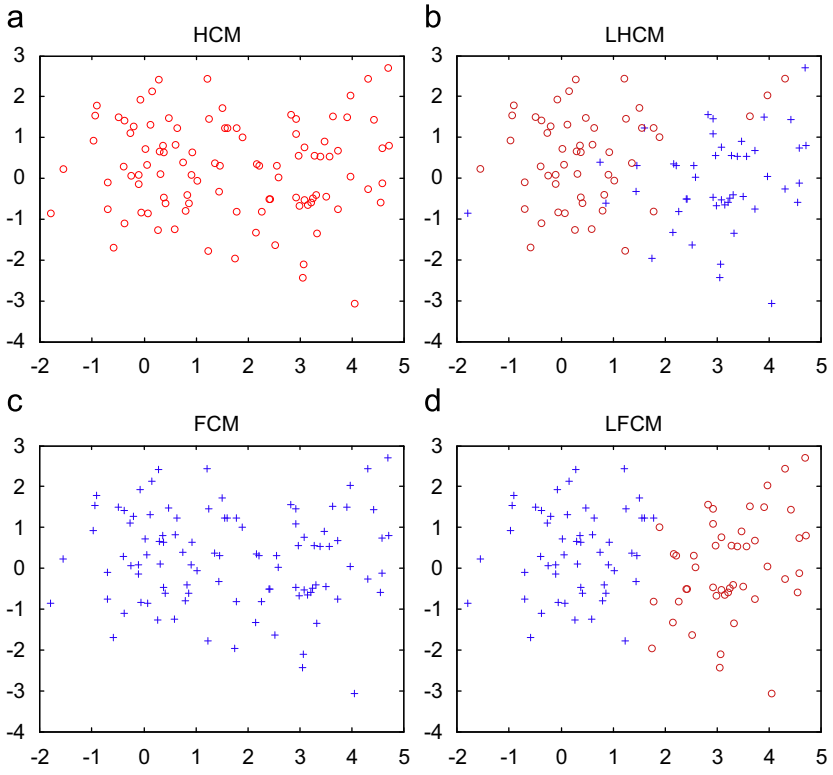

Fig. 1. Clustering results on artificial dataset D1: (a) HCM; (b) LHCM; (c) FCM; (d) LFCM (without the single noise).

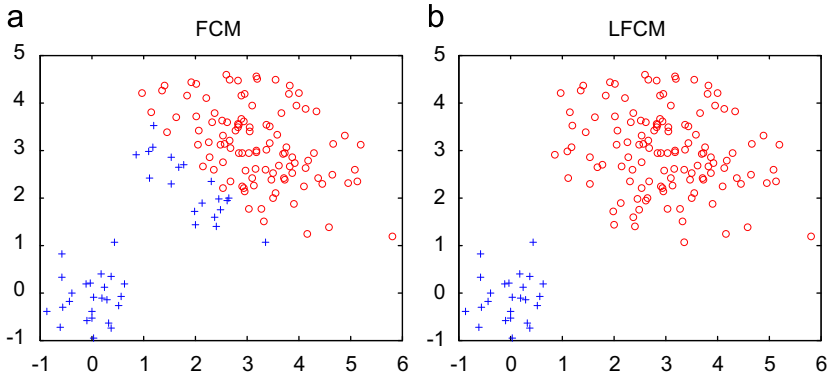

Fig. 2. Clustering results on artificial dataset D2: (a) FCM; (b) LFCM.

or recall, so most researchers use $\beta = 1$ and we choose $\beta = 1$ in our experiments.

Table 3 shows the datasets and the experimental results, where n, c and a, respectively, denotes number of samples, number of classes and number of attributes or dimensions of each dataset. Taking account of the conditions of initializations and others' effect, we test every dataset 50 times and compute the average results. The bold values in the table is the largest one among all algorithms. According to Table 3, the performance of LFCM always seems superior to other three algorithms. Moreover, LHCM is superior to HCM. Hence, it is clear that the locality sensitive weight works effectively.

Most clustering algorithms need users to give supervision to determine the number c of clusters and always appear to be sensitive to the value of c. We test the sensitivity of our approach over Iris and Soybean. Results in Fig. 3 show that LHCM and LFCM always achieve good performance in spite of the value c, while HCM and FCM turn worse when the value c go away from the real number of clusters. Introducing the weight matrix which can effectively describe the neighborhood information, points in a

Table 3
The clustering performance of HCM, LHCM, FCM and LFCM.

Dataset				HCM	LHCM	FCM	LFCM
Name	n	c	a				
Iris	150	3	4	0.7901	0.8344	0.8506	**0.8622**
Glass	214	7	9	0.4375	0.4426	0.4882	**0.4928**
Soybean	47	4	35	0.3855	0.4398	0.3994	**0.4565**
Hayesroth	132	3	2	0.4530	0.4802	0.4958	**0.4998**
Ionosphere	351	2	34	0.5811	0.5978	0.5996	**0.6278**
Balance-scale	625	3	4	0.4657	0.5813	0.5311	**0.5969**
Image-segment	210	7	19	0.4637	0.4718	0.4687	**0.4726**

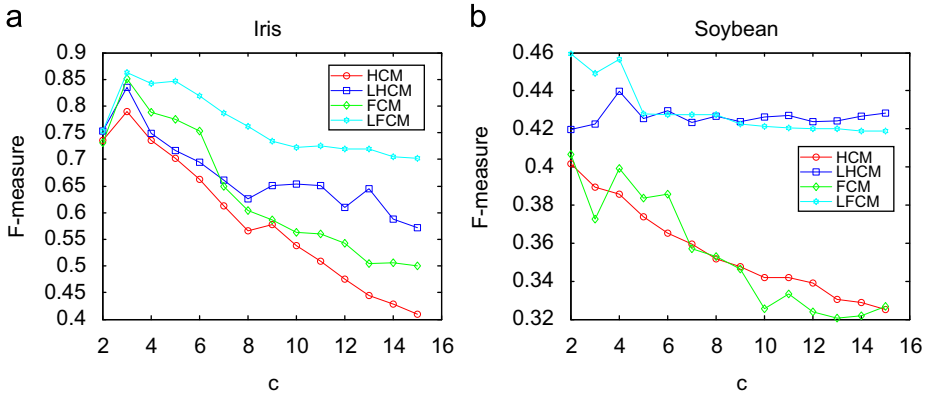

Fig. 3. Performance of clustering of the algorithms on datasets: (a) Soybean; (b) Iris.

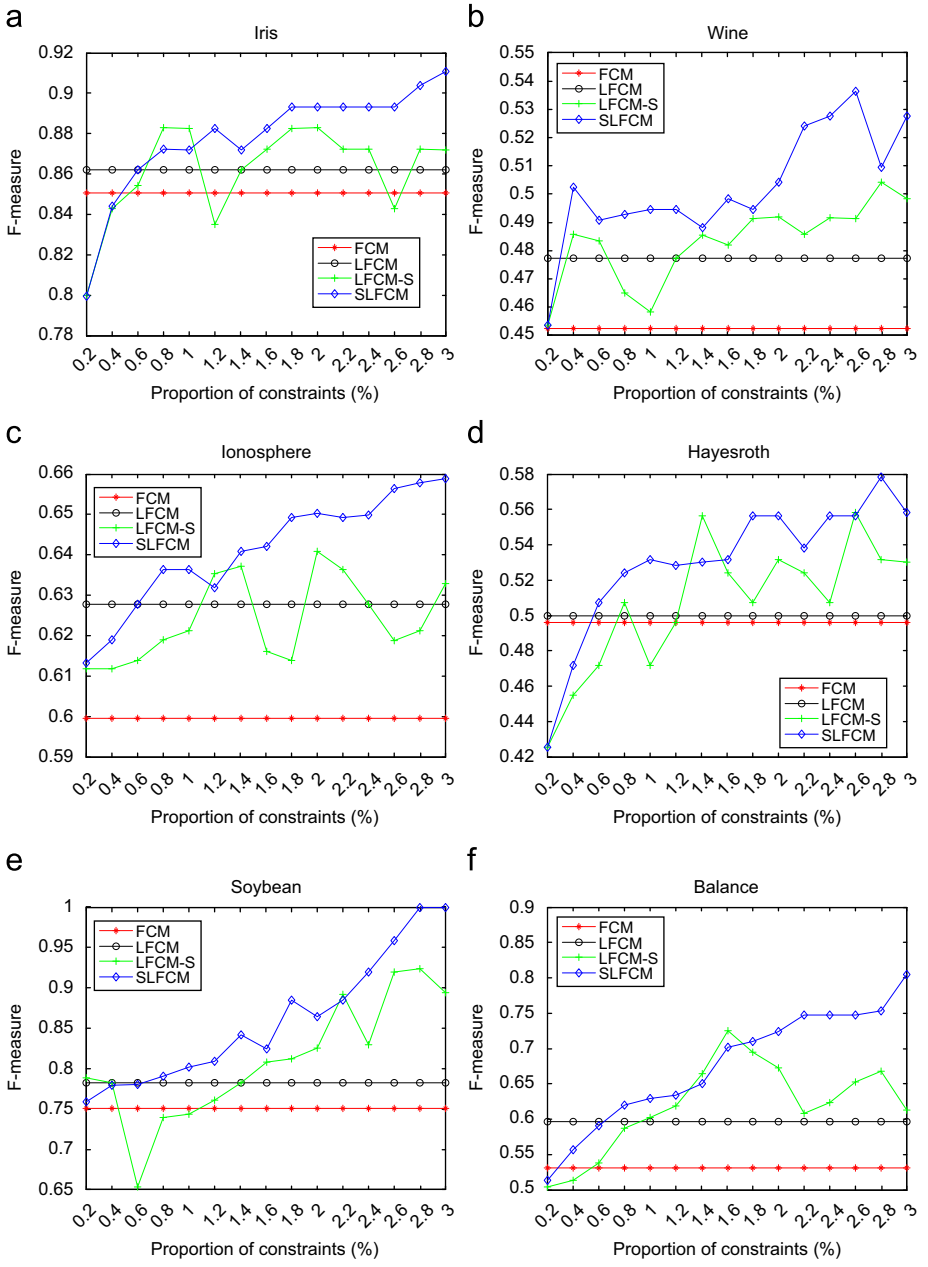

Fig. 4. Clustering results of FCM, LFCM, LFCM-S, SLFCM on 6 UCI datasets.

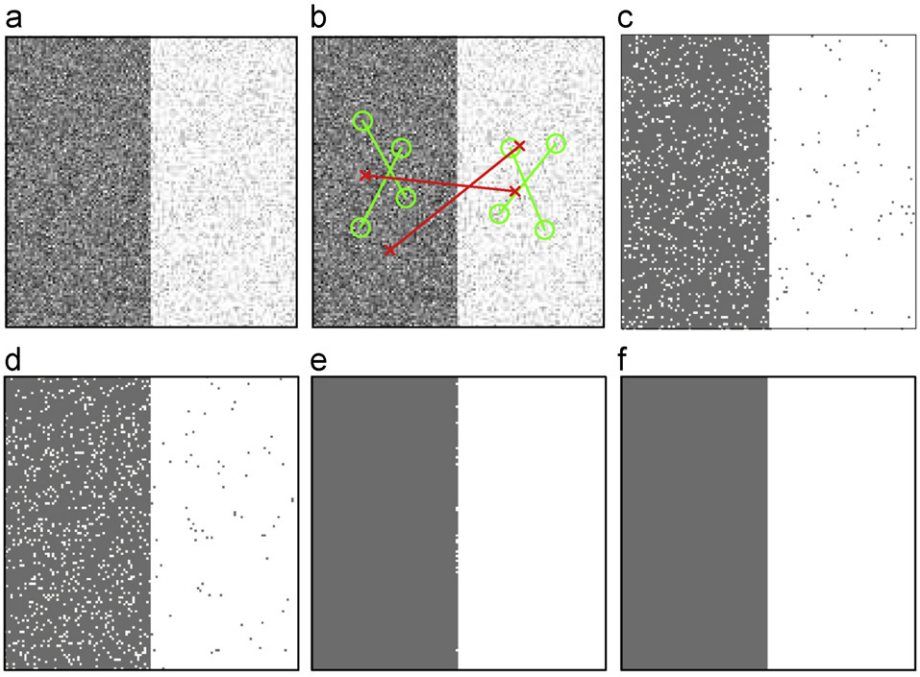

Fig. 5. Segmentation results with gaussian noise on artificial images: (a) image with noise; (b) noisy image adding some constraints; (c) FCM; (d) LFCM; (e) SFCM; (f) SLFCM-S.

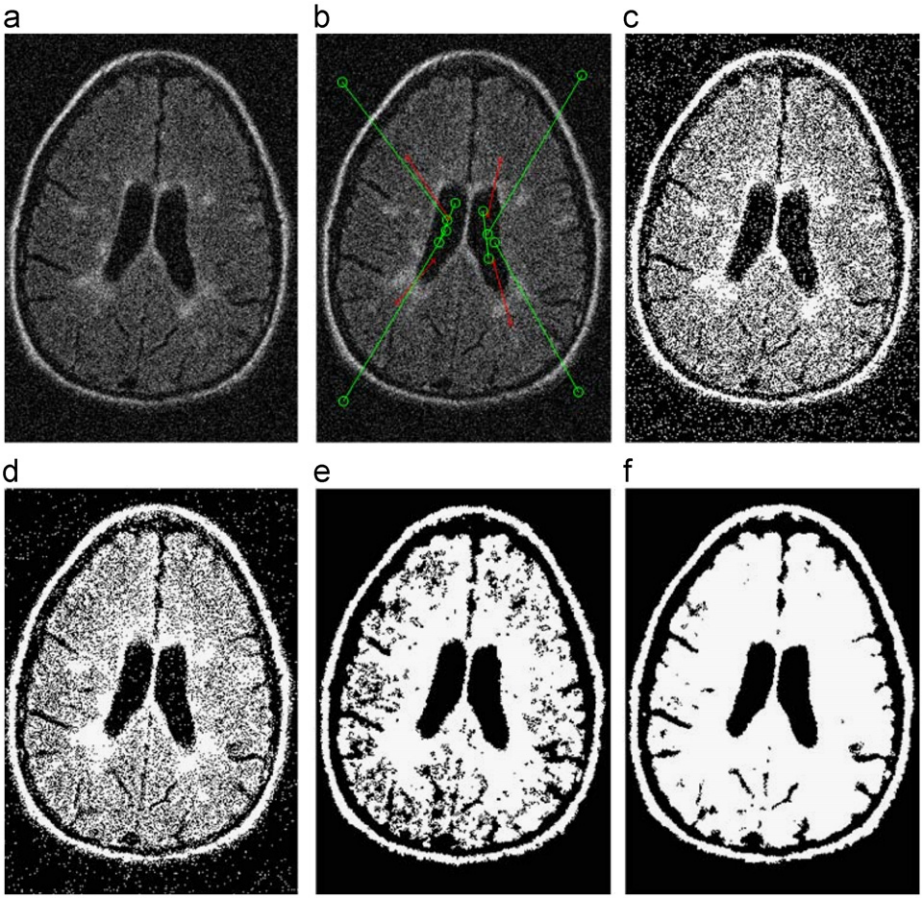

Fig. 6. Segmentation results with gaussian noise on a brain MRI image: (a) image with noise; (b) noisy image adding some constraints; (c) FCM; (d) LFCM; (e) SFCM; (f) SLFCM-S.

same class keep in a same class as far as possible while points in different classes still go into different classes, then the algorithm can achieve higher *F-measure*.

Finally in this section, we evaluate the performance of SLFCM when partial supervision information is available. We build the centers of the clusters in the following method. For the relations between pair-wise constraints with *must-link* are transmissible, we can obtain a connected set $\{N_p\}^{\lambda} p=1$ in the sets of *must-link*, where the λ is the number of the connected sets. Each samples in the connected set N_p belongs to one cluster. Assumed that the number of the cluster is c, if $\lambda \geq c$, then the cluster centers can be initialized by the c lager connected sets; Otherwise, using the connected sets to initialize the λ centers, and then finding a sample point that has a *cannot-link* with all the connected sets to initialize the $\lambda+1$ center. If there is no samples fitting the situation, choose samples without any supervision to initialize the rest centers in any way.

During the process of updating the membership, some adjustments are made to promise the membership satisfy $\sum_{ij}^{c} u_{ij}=1, \forall j=1,\ldots,n$:

$$u_{ij}^{new}=\frac{u_{ij}+|\min(u_{kj})|}{c|\min(u_{kj})|+\sum_{k=1}^{c} u_{kj}} \tag{27}$$

Since the weight matrix of the proposed algorithms is closely related to the cluster centers, proper and superior centers can obtain better results. We compare the algorithm LFCM only using the pair-wise constraints to utilize its centers called LFCM-S(LFCM-Semi-intialized) and SLFCM on some UCI datasets to validate the effectiveness of SLFCM.

From Fig. 4, we can find that SLFCM can effectively uses the available pair-wise constraint to improve performance, and roughly the more pair-wise constraints the higher clustering performance. In contrast, for LFCM-S which only uses the semi-supervision to initialize the centers, the performance somehow is affected by the pair-wise constraints. In nearly all cases, LFCM-S is inferior to SLFCM. On the other hand, both LFCM-S and SLFCM are superior to FCM and LFCM in most cases, which validate the usefulness of supervision information for clustering.

5.3. Experiments on images

In this section, we give the segmentation results of FCM, LFCM, SFCM and SLFCM-S on an artificial image and brain MRI images. Fig. 5(a) shows an artificial image with gaussian noise. Fig. 5(b) shows the way to add pair-wise constraints while o represents the *Must-link* connection and x represents the *Cannot-link* connection. Image segmentation results of FCM, LFCM, SFCM and SLFCM-S are shown in Fig. 5(c)–(f), respectively. It can be seen that FCM and LFCM cannot obtain good segmentation results without the spatial constraints. On the other hand, considering the spatial context, SFCM can roughly segment the image into two parts while SLFCM-S utilizes both spatial constraints and semi-supervision and preserves the neighborhood information, and thus two parts are correctly separated.

Figs. 6 and 7 give the image segmentation results of FCM, LFCM, SFCM, SLFCM-S on brain MRI images. Figs. 6(a) and 7(a) display the image with gaussian noise. Figs. 6(b) and 7(b) show the way to add pair-wise constraints while o represents the *Must-link* connection and x represents the *Cannot-link* connection. Image segmentation results of FCM, LFCM, SFCM and SLFCM-S show in Figs. 6(c)–(f) and 7(c)–(f), respectively. It can be clearly seen from the figures that SLFCM-S achieves the best image segmentation results on both images.

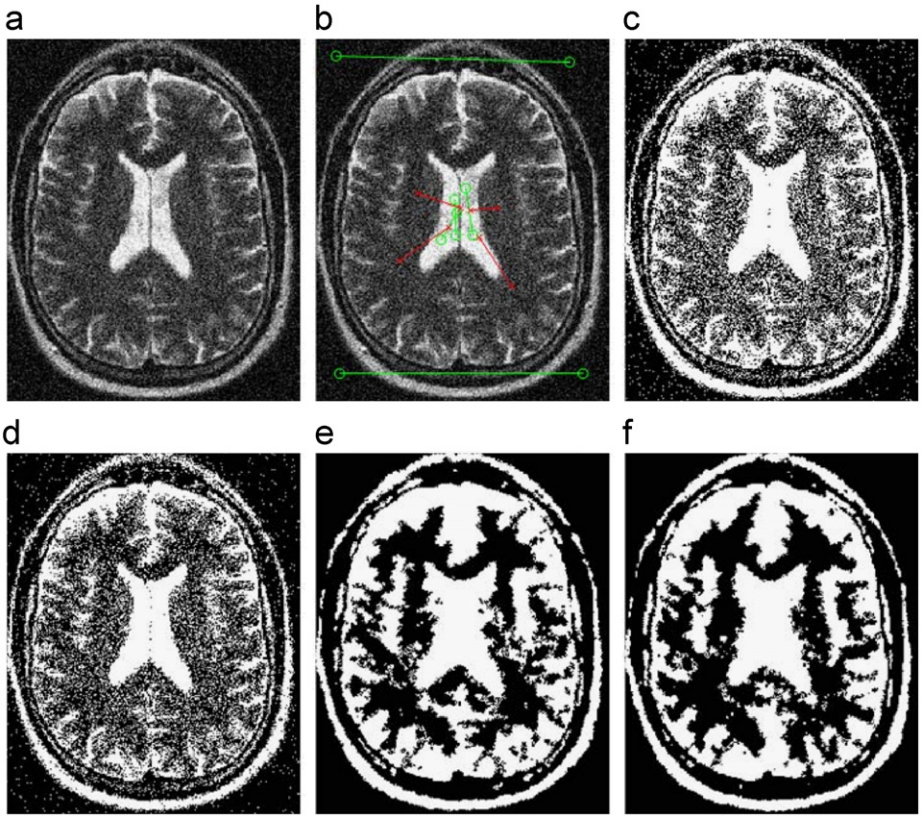

Fig. 7. Segmentation results with gaussian noise on another brain MRI image: (a) image with noise; (b) noisy image adding some constraints; (c) FCM; (d) LFCM; (e) SFCM; (f) SLFCM-S.

6. Conclusion

In this paper, we proposed two new locality sensitive C-means clustering algorithms called LHCM and LFCM, which integrates sample weighting into HCM and FCM, respectively. By borrowing the idea of optimally preserving the neighborhood structure from locality preserving dimensionality reduction, our approach builds a graph indicating the neighborhood information with the weight matrix. As a result, the proposed locality-weighted scheme can effectively improve the clustering performance. In addition, the new algorithms achieve some extra advantages such as robustness to outliers, suitability for class-imbalance data clustering and insensitivity to number of clusters, etc. In addition, we generalize LFCM for semi-supervised cases when partial supervision information in the form of pair-wise constraints and prior spatial constraints in images are available. Experimental results on a lot of datasets validate the effectiveness of the proposed methods.

Acknowledgements

This work is supported by National Science Foundation of China under Grant no. 60875030 and the Open Projects Program of National Laboratory of Pattern Recognition (20090044).

References

[1] J. Han, M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann, 2000.
[2] J.C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Kluwer, Norwell, MA, 1981.
[3] W. Pedrycz, Conditional fuzzy C-means, Pattern Recognition Letters 17 (1996) 625–632.
[4] K. Rose, Deterministic annealing for clustering, compression, classification, regression, and related optimization problems, Proceedings of the IEEE 86 (11) (1998) 2210–2239.
[5] R. Nock, F. Nielsen, An abstract weighting framework for clustering algorithms, in: Proceedings of the Fourth International SIAM Conference on Data Mining, 2004, pp. 200–209.
[6] R. Nock, F. Nielsen, On weighting exponent, IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (8) (2006) 1223–1235.
[7] J. Li, X.B. Gao, L.C. Jiao, A novel typical-sample-weighting clustering algorithm for large datasets, LANI, vol. 3801, 2005, pp. 696–703.
[8] C.Z. Zhang, Q.H. Shi, D.J. Xue, Document clustering algorithm based on sample weighting, Journal of the China Society for Scientific and Technical Information 27 (1) (2008) 42–48.
[9] X.B. Gao, J. Li, H.B. J, A multi-threshold image segmentation algorithm based on weighting fuzzy C-means clustering and statistical test, Acta Electronica Sinica 32 (4) (2004) 661–665.
[10] P. D'Urso, P. Giordani, A weighted fuzzy C-means clustering model for fuzzy data, Computational Statistics & Data Analysis 50 (6) (2006) 1496–1523.
[11] J. Mei, L. Chen, Fuzzy clustering with weighted medoids for relational data, Pattern Recognition 43 (5) (2010) 1964–1974.
[12] L.K. Saul, K.Q. Weinberger, et al., Spectral Methods for Dimensionality Reduction, Book Chapter, MIT Press, Cambridge, MA, 2005.
[13] X. Zhu, Semi-supervised learning literature survey, Technical Report 1530, Department of Computer Sciences, University of Wisconsin, Madison, 2009.
[14] M. Belkin, P. Niyogi, Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering, Advances in Neural Information Processing Systems, vol. 14, MIT Press, Cambridge, MA, 2002, pp. 585–591.
[15] X.F. He, P. Niyogi, Locality Preserving Projections, Advances in Neural Information Processing Systems, vol. 16, MIT Press, Cambridge, MA, 2004.
[16] D.Q. Zhang, Z.H. Zhou, S.C. Chen, Semi-supervised dimensionality reduction, in: Proceedings of the 2007 SIAM Conference on Data Mining, Minneapolis, MN, 2007, pp. 629–634.
[17] J. Yu, Q. Cheng, H. Huang, Analysis of the weighting exponent in the FCM, IEEE Transactions on Systems, Man and Cybernetics-part B: Cybernetics 34 (1) (2004) 634–639.
[18] J. Yu, General C-means clustering model, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (8) (2005) 1197–1211.
[19] W. Zangwill, Nonlinear Programming: A Unified Approach, Prentice-Hall, Englewood Cliffs, 1969.
[20] N. Grira, M. Crucianu, N. Boujemaa, Active semi-supervised fuzzy clustering, Pattern Recognition 41 (5) (2008) 1834–1844.
[21] K. Wagstaff, C. Cardie, S. Rogers, S. Schroedl, Constrained k-means clustering with background knowledge, in: ICML01, Williamstown, MA, 2001, pp. 577–584.
[22] E.P. Xing, A.Y. Ng, M.I. Jordan, S. Russell, Distance metric learning, with application to clustering with side-information, in: NIPS 15, MIT Press, Cambridge, MA, 2003, pp. 505–512.
[23] A. Bar-Hillel, T. Hertz, N. Shental, D. Weinshall, Learning a mahalanobis metric from equivalence constraints, Journal of Machine Learning Research 6 (2005) 937–965.
[24] D.L. Pham, Spatial models for fuzzy clustering, Computer Vision and Image Understanding 84 (2001) 285–297.
[25] D.L. Pham, Fuzzy clustering with spatial constraints, Proceedings of the IEEE International Conference on Image Processing, USA 2 (2002) 65–68.
[26] W. Cai, S. Chen, D. Zhang, Fast and robust fuzzy C-means clustering algorithms incorporating local information for image segmentation, Pattern Recognition 40 (3) (2007) 825–838.
[27] S. Chen, D. Zhang, Robust image segmentation using FCM with spatial constraints based on new kernel-induced distance metric, IEEE Transactions on System, Man and Cybernetics-Part B 34 (4) (2004) 1907–1916.
[28] D. Zhang, S. Chen, A novel kernelised fuzzy C-means algorithm with application in medical image segmentation, Artificial Intelligence in Medicine 32 (1) (2004) 37–50.

Pengfei Huang received the B.Sc. in Information and Communication Engineering from Jilin University, China, in 2006. From 2006, he was a graduate student in the Department of Computer Science and Engineering of Nanjing University of Aeronautics and Astronautics. His research interests include pattern recognition and image processing.

Daoqiang Zhang received the B.Sc. and Ph.D. degrees in Computer Science from Nanjing University of Aeronautics and Astronautics, China, in 1999 and 2004, respectively. From 2004 to 2006, he was a postdoctoral fellow in the Department of Computer Science & Technology at Nanjing University. He joined the Department of Computer Science and Engineering of Nanjing University of Aeronautics and Astronautics as a Lecturer in 2004, and is a professor at present. His research interests include machine learning, pattern recognition, data mining, and image processing. In these areas he has published over 40 technical papers in refereed international journals or conference proceedings. He was nominated for the National Excellent Doctoral Dissertation Award of China in 2006, and won the best paper award at the 9th Pacific Rim International Conference on Artificial Intelligence (PRICAI'06). He has served as a program committee member for several international and native conferences. He is a member of Chinese Association of Artificial Intelligence (CAAI) Machine Learning Society.