

Partially Horizontal Collaborative Fuzzy C-Means

Fusheng Yu, Juan Tang, and Ruiqiong Cai

Abstract

Horizontal collaborative clustering is such a clustering method that carries clustering on one data set describing a pattern set in one feature space with collaborative introducing of outer partition information obtained by clustering on another data set but describing the same pattern set in another feature space. In order to implement the collaborative clustering, horizontal collaborative fuzzy c-means (HC-FCM) was proposed by W. Pedrycz. In HC-FCM, the outer partition matrix is incorporated with the objective function in FCM. This manner of making use of the outer partition matrix emphasizes on the use of total collaborative clustering information provided by the outer partition matrix, thus this method can be called completely horizontal collaborative fuzzy c-means (CHC-FCM). In reality, on many occasions of collaborative clustering, we may be interested only in the cluster information of some special patterns, say the patterns with distinct membership for example. In this paper, we implement the horizontal collaborative clustering with the partial supervision clustering approach where the clustering is carried by the guidance of some labeled patterns. In this approach, we can select the patterns we are interested in to provide FCM with collaborative information and control the degree of the influence of the selected patterns on the clustering. This new method is called partially horizontal collaborative fuzzy c-means (PHC-FCM). After presenting two approaches to realizing the selection of the labeled patterns, named cut-set based approach and entropy based approach, we give the detailed algorithm of PHC-FCM. Experiments are carried and show the performance of the new method.

Keywords: fuzzy c-means, horizontal collaborative clustering, partial supervision clustering, partially horizontal collaborative clustering

1. Introduction

Fuzzy C-Means (FCM) [1], [4] is a typical one of fuzzy clustering methods on a single data set. While, independent clustering for a data set is incomplete. A group of patterns have diverse information in different fields. So we can get another data set describing the same set of patterns in different feature space. In order to seek a comprehensive study of these patterns, knowledge-based clustering is highly recommended recently and collaborative clustering between data sets is appreciated [5], [7], [8].

Due to problem of privacy-preserving, outer field usually unwillingly offers the whole data set to share. The final clustering result obtained by FCM may be available and can be provided with the partition matrix. How to make use of the outer clustering information is the key to the collaborative clustering problem. W. Pedrycz proposed a new fuzzy clustering method, named horizontal collaborative fuzzy c-means (HC-FCM) [7], [8], where the outer clustering information was made use of in such a manner where the outer partition matrix was incorporated in the objective function of FCM. With the characteristic of totally using the clustering information provided by the outer partition matrix, HC-FCM can be called completely horizontal collaborative fuzzy c-means (CHC-FCM). It has been shown that such an approach is of great use in actual collaborative fuzzy clustering problems.

Whereas, there may occur such occasions where only part of the patterns are concerned and only the impact of them on the clustering is expected to focus on. How to deal with such a kind of collaborative clustering problems? In this paper, we will give a partial supervision clustering [5], [7] based method to solve this problem. In the new fuzzy clustering method, the selection of the concerned patterns, who is the crucial step in the partial supervision clustering, is considered and implemented by giving some labeled patterns. Two approaches, the cut-set based approach and the entropy based approach, are proposed for the selection of labeled patterns. With the characteristic of partially use of the clustering information provided by the outer partition matrix, the new proposed horizontal collaborative fuzzy clustering method may be called partially horizontal collaborative fuzzy c-means (PHC-FCM).

The study of this paper is organized in the following

Corresponding Author: Fusheng Yu is with the School of Mathematical Sciences, Beijing Normal University, Laboratory of Mathematics and Complex Systems, Ministry of Education, Beijing 100875, The People's Republic of China.

E-mail: yufusheng@263.net

Manuscript received 21 Sep. 2007; revised 1 Nov. 2007; accepted 21 Nov. 2007.

way: Section 2 gives a brief review of the concerned fuzzy clustering methods. In Section 3, we present the partial supervision clustering based collaborative fuzzy c-means algorithm and two approaches for the selection of labeled patterns. Experimental studies and corresponding results are given in Section 4. Section 5 concludes the study of this paper.

2. Preliminaries

In this section, we briefly introduce the concerned fuzzy clustering methods which are used in the discussion of the later sections.

A. Fuzzy C-Means

Fuzzy c-means (FCM) [1], [8] is an effective means of clustering based on one fuzzy partition. It optimizes the clustering through cyclic iteration of partition matrix and ends the cycle when the objective function reaches an appropriate threshold. The corresponding objective function is as follows:

$$\text{Min } J(U, V) = \sum_{i=1}^c \sum_{k=1}^n u_{ik}^m d_{ik}^2 \quad (1)$$

$$\text{s.t. } \sum_{i=1}^c u_{ik} = 1$$

Where $U = (u_{ij})_{c \times n}$ is the partition matrix, $V = (v_{ij})_{c \times l}$ is the matrix of centroids of clusters, d_{ik} is the distance between the k th patterns x_k and the centroid v_i of the i th cluster ($i=1,2,\dots,c$; $k=1,2,\dots,n$), the fuzzification factor m assumes values bigger than 1 and is usually set to be 2. By Lagrange multipliers method, we can obtain the optimal solution of the optimization problem resulting in a group of iteration formulas. Thus the clustering is realized by an iteratively calculation in term of the iteration formulas [1], [2].

B. Horizontal Collaborative Fuzzy Clustering Based on Partition Matrix

The background of horizontal collaborative clustering may be outlined as follows: horizontal collaborative clustering deals with collaboration among data sets describing the same group of patterns in different feature spaces. Choose one data set $X[ii]$ to be the central party, other parties $X[jj]$ only offer referential information in the form of partition matrices ($U[jj]$, named outer partition matrix) to guide the central party's clustering because of potential security and privacy restrictions. In order to implement the collaborative clustering, the outer partition matrix is incorporated with the objective function. The refined objective function is as fol-

lows:

$$\text{Min } Q = \sum_{i=1}^c \sum_{k=1}^n u_{ik} [ii]^2 d_{ik} [ii]^2 + \sum_{\substack{jj=1 \\ jj \neq ii}}^p \alpha [ii, jj] \sum_{i=1}^c \sum_{k=1}^n (u_{ik} [ii] - u_{ik} [jj])^2 d_{ik} [ii]^2 \quad (2)$$

With the constraint $\sum_{i=1}^c u_{ik} = 1$. Where $U = (u_{ij})_{c \times n}$

is the partition matrix of reference data set X ; p denotes the number of external data sets $X[1], X[2], \dots, X[p]$; $U[jj]$ denotes the corresponding partition matrix of external data set $X[jj]$ ($jj=1,2,\dots,p, jj \neq ii$); ii is fixed. $U[ii]$ denotes the corresponding partition matrix of reference data set $X[ii]$; $d_{ik} [ii]$ is the distance between the k th pattern and i th cluster in reference data set $X[ii]$. $\alpha [ii, jj]$, the value of the interactive coefficient arranged in the collaborative clustering, sets up the collaboration level. It assumes nonnegative values, the higher the value of interaction coefficient is, the stronger the effect of data set $X[jj]$ on data set $X[ii]$ will be.

The optimal solution to the minimization problem can be obtained by Lagrange multiplier method. We omit the formulas of optimal partition matrix and centroids of clusters here, readers may refer to papers [7], [8]. The algorithm consists of two main phases that are realized in interleaved manner. The first phase is data driven and is primarily the standard FCM applied to the patterns. The second phase concerns an accommodation of external partition matrices.

C. Partial Supervision Fuzzy Clustering

This form of collaboration is concerned with clustering carried out in presence of labeled patterns, so optimizing objective function needs to augment the standard fuzzy c-means algorithm [6], [8] by extending supervision component offered by labeled patterns. The objective function is:

$$Q = \sum_{i=1}^c \sum_{k=1}^n u_{ik}^2 d_{ik}^2 + \alpha \sum_{i=1}^c \sum_{k=1}^n (u_{ik} - f_{ik})^2 b_k d_{ik}^2 \quad (3)$$

$$\text{s.t. } \sum_{i=1}^c u_{ik} = 1$$

The additive form of objective function has a plausible interpretation. The first term is used to discover the structure of data and is the same as that in the standard FCM. The second term addresses the effect of partial supervision offered by labeled patterns. Optimizing sum of these two terms will lead to reconciliation between the structure residing in central party (being primarily dis-

covered by the underlying clustering mechanism in one feature space) and the labels of the patterns of outer party in other feature space.

The terms in (3) and their meanings are given below:

α : the nonnegative weight factor, helps set up a suitable balance between the supervised and unsupervised modes of learning and expresses collaboration strengthen.

b : the vector of labels, denoted by $b = [b_1, b_2, \dots, b_n]^T$.

Each pattern x_k comes with a Boolean indicator: we make b_k equal to 1 if the pattern has been already labeled and equal to 0 otherwise.

$F = [f_{ik}]$, the partition matrix, $i = 1, 2, \dots, c$, $k = 1, 2, \dots, n$, which contains membership degrees assigned to the selected patterns (already identified by the nonzero value of b). If $b_k = 1$, then the corresponding column shows the provided membership degrees of the labeled pattern. If $b_k = 0$, then the entries of the corresponding k th column of F don't matter; technically, we could set them equal to 0.

Using knowledge of partial derivatives, the resulting entries of the partition matrix and prototypes are given as the following forms:

$$u_{ik} = \frac{1}{1 + \alpha} \left[\frac{1 + \alpha \left(1 - b_k \sum_{i=1}^c f_{ik} \right)}{\sum_{j=1}^c \left(\frac{d_{ik}}{d_{jk}} \right)^2} + \alpha f_{ik} b_k \right] \quad (4)$$

$$v_i = \frac{\sum_{k=1}^n \psi_{ik} x_k}{\sum_{k=1}^n \psi_{ik}}, \psi_{ik} = u_{ik}^2 + \alpha (u_{ik} - f_{ik} b_k)^2 \quad (5)$$

$k = 1, 2, \dots, n$, $i = 1, 2, \dots, c$ (n, c are the numbers of individuals and clusters respectively.)

3. Partially Horizontal Collaborative Fuzzy C-Means

Generally speaking, the more the information of the outer party is shared, the more the impact on the central party will be. Theoretically speaking, collaborative clustering based on whole partition matrix offers adequate granular information and can certainly leads to thoroughly collaborative clustering. But in the applications of the HC-FCM, sometimes we may meet with some cases that do not satisfy out expectations; examples for such cases are as follows:

- Disguise the original data structure of central

party

Sharing the whole external partition matrix means entering of much more outer information. It is followed that the outer party will play an important role and may affect the collaborative clustering to large extend unexpectedly. To be worse, it may destroy the structure of central party's clusters and make it be separated far away from the original one produced by independent clustering of the central party.

- Low efficiency
Sharing too much information must result in huge quantity of computation and prolong the clustering time during the collaboration.

The examples given above indicate that sharing whole partition matrix is not ideal in some cases. They drive us to consider a new approach to partially sharing the partition matrix. Partial supervision fuzzy clustering offers us a good idea. We can label the important part of the external matrix and sharing the membership information of only these labeled patterns. Of course, the labeled patterns must be typical ones of the outer party and can reveal the main structure of the data set in the outer party. Under this condition, collaboration will satisfy both comparatively adequate sharing between parties because of specialty and representative of typical labeled patterns and comparatively higher efficiency because of the decreased computations. We call this clustering method partially horizontal collaborative fuzzy c-means. In this method, how to choose the labeled patterns from the outer party becomes a crucial step. So, we will first give two approaches to find the labeled patterns, and then give the algorithm of partial supervision based horizontal collaborative fuzzy clustering.

A. Approaches to Labeling Patterns

We will introduce two approaches to choosing labeled patterns. One is based on cut-set, and the other is based on fuzzy entropy.

a) Cut-set based approach

In collaborative clustering, we are usually interested in patterns with distinct membership and expect to label them to collaborate with. These patterns usually correspond to distributions with one membership degree close to 1 while others close to 0. So we can set a rational threshold (λ) as the criteria for labeling patterns. For a given pattern, if the largest membership degree among the pattern's memberships to all clusters is larger than λ , then we'll label the pattern and consider its guidance in collaboration. This is similar to the manner of obtaining the cut-set of a fuzzy set. So, this approach is named cut-set based approach, and the algorithm of cut-set based approach is illustrated as follows:

```

Initialize the threshold  $\lambda$ 
for each pattern  $x_k$ 
    calculate  $t = \max_{i=1, \dots, c} (u_{ik})$ 
    if  $t > \lambda$ 
        label the pattern
    end {if}
end {for}

```

From the given algorithm, it is not difficult to see that: the larger the threshold is; the small the number of the labeled patterns will be. Furthermore, the large the threshold is, the better the labeled pattern as a representative pattern of this cluster is. Thus, λ is generally set to be larger than 0.5.

b) Entropy based approach

For a given distribution $[x_1, x_2, \dots, x_n]$, where x_1, x_2, \dots, x_n fall in the interval $[0.0, 1.0]$ and satisfy $\sum_{i=1}^n x_i = 1$, the entropy of $[x_1, x_2, \dots, x_n]$ is defined in the following form [3]:

$$E = -\sum_{i=1}^n [x_i \ln x_i + (1 - x_i) \ln (1 - x_i)] \quad (6)$$

The entropy trends to a large value when the corresponding distribution approaches to the uniform and a small value when all the entries in the corresponding distribution are close to binary value (for sake of simplification, we call such distribution binary distribution). Thus, we can introduce the idea of entropy to the process of labeling patterns. Each pattern x_k corresponds to a distribution consisting of its membership degrees to c clusters $[u_{1k}, u_{2k}, \dots, u_{ck}]$. When x_k has a binary distribution $[u_{1k}, u_{2k}, \dots, u_{ck}]$, the corresponding entropy is small according to formula (6). Similarly, when x_k has a uniform distribution, the entropy is high. In collaborative clustering, the important patterns are those with binary distribution. Sometimes, we want to consider first percentage of patterns with good classification. We can introduce a threshold (γ), which is a percentage of the labeled patterns to all patterns, to control the number of labeled patterns. The algorithm of entropy based approach to determine the labeled patterns is illustrated as follows:

111

```

Initialize the threshold  $\gamma$ ; (N is the number of all patterns)
denote  $nm = N * \gamma$ 
for each pattern
    compute correspondent entropy for the pattern ac-

```

```

    cording to formula (6)
    end
    sort patterns from small to large according to their entropies
    label the first  $nm$  patterns (the patterns have been sorted)

```

Based on above analysis, it is not difficult to get the following conclusion: the smaller the threshold is, the smaller the number of the labeled patterns will be. Furthermore, the smaller the threshold is, the smaller the entropy of each labeled pattern is, and the better the labeled pattern as representative patterns of this cluster is.

B. Algorithm of Partially Horizontal Collaborative Fuzzy C-Means

Based upon the above discussion, we present here the algorithm of the partially collaborative fuzzy c-means. It includes four steps as follows:

Given: data set X in central party, referential partition matrix in outer party

Select: distance function, number of clusters (c), termination criterion, threshold (λ, γ), collaboration strength (α)

Goal: clustering result of X after collaboration based on the labeled patterns

Phase I : Independent clustering of X

For data set X

Repeat

 Compute prototypes v_i ($i=1,2,\dots,c$) and partition matrices $(u_{ik})_{c \times N}$ by the formulas of FCM

 Until the termination criterion has been satisfied.

Phase II : Choose the labeled patterns

 Use the cut-set based approach or entropy based approach to determine the labeled patterns.

 Compute vectors b, F through the membership information of labeled patterns.

Phase III: Partially collaborative clustering based on the labeled patterns

Repeat

 for the given collaboration strength α , compute prototypes and partition matrix using formula (4) and (5)

 until a termination criterion has been satisfied.

Phase IV: Evaluation of partially collaborative clustering

 Let $\delta = \|u_{ref} - u\|, \phi = \|u - u[jj]\|$, where u_{ref} is the partition matrix produced after independent clustering, u is the partition matrix produced after collaborating with the labeled patterns, and $u[jj]$ is the external partition matrix in the outer party.

 The two indexes are used to evaluate the result of

partially collaborative clustering. Where δ quantifies the difference between collaborative clustering and independent clustering without collaboration, and ϕ expresses the difference between the partition matrix of the central party and that of the outer party.

In the algorithm of partially horizontal collaborative FCM, there are several important parameters: α, λ, γ . The influence of the last two parameters will be discussed and shown in the next section, while the function of the first parameter has been discussed [6], [9].

4. Experimental Studies

In this section, we'll show the performance of the partially collaborative FCM by presenting two experiments which are carried with the guidance of the algorithms of the partially collaborative FCM equipped with the cut-set based approach and the entropy based approach respectively. In collaboration, we'll take different thresholds (λ, γ) to show the influence on clustering.

A. Partially Horizontal Collaborative Clustering with Cut-Set Based Approach

In this experiment, data set X consisting of 600 patterns is distributed in two-dimension subspaces and required to do the clustering of three clusters: $c=3$. Let $u[jj]$ be the external partition matrix in the outer party and the labeled patterns are determined from $u[jj]$ according to cut-set based approach. The result of clustering under different thresholds (λ) is shown in Figure 1.

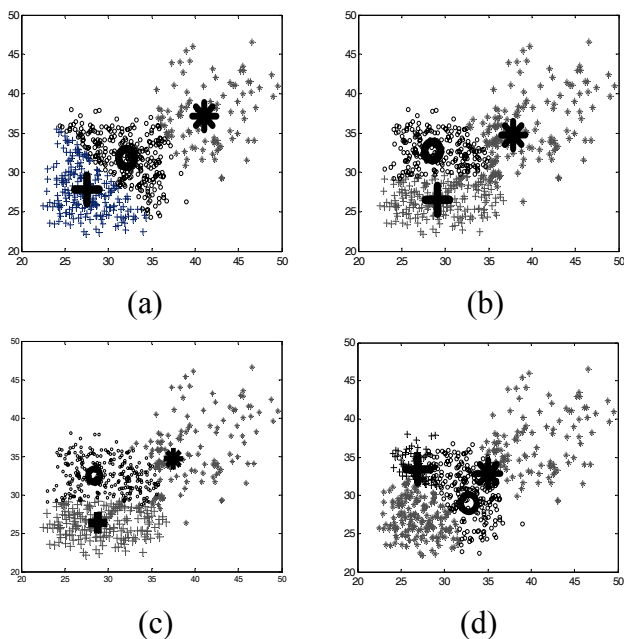


Figure 1. Result of collaborative clustering of X under different thresholds: (a) $\lambda=1.0$; (b) $\lambda=0.8$; (c) $\lambda=0.7$; (d) $\lambda=0.6$

Since no pattern has membership degree larger than 1, so there is no pattern labeled when $\lambda=1$. In this case, the result of collaborative clustering is same to that of independent clustering of X without collaboration. When λ is equal to 0.8 or 0.7, the results of clustering have subtle difference from that when $\lambda=1.0$. (Visualized in Figure 1 (a),(b),(c)) When $\lambda=0.6$, the structure is destroyed to large extent. In Figure 1(d), we can see that, two separated clusters are combined to one cluster, and the middle cluster are separated into two clusters (referring to (a)). It illustrates that the collaboration is affected to large extent by the labeled patterns. This phenomenon results from comparatively low value of λ that allows much more labeled patterns to be collaborated with. Much more labeled patterns to collaborate with means too many constrains on the clustering of X in the central party. In the process of optimization, the revealed structure of X in central party will finally conforms to the membership degrees of these labeled patterns in the outer party. Thus large difference between collaborative clustering and independent clustering without collaboration is achieved. This is also illustrated in Table 1 (where N is the corresponding number of labeled patterns in three clusters).

Table 1. Evaluation of partially horizontal collaborative FCM

Threshold λ	$\delta = \ u_{ref} - u\ $	$\phi = \ u - u[jj]\ $	N
$\lambda=1.0$	0.01	11.0960	[0, 0, 0]
$\lambda=0.8$	8.6260	10.9292	[0, 42, 77]
$\lambda=0.7$	7.8763	10.7482	[0, 63, 129]
$\lambda=0.6$	11.1789	6.7849	[0, 244, 190]

From Table 1, it is not difficult to find that the smaller λ is, the more labeled patterns will be collaborated with, and it will induce better reconciliation effect between structure of clustering in the central party and the outer party (measured by ϕ). We can also see that ϕ decreases and N increases with λ decrease. But no regular conclusion about the trend of δ can be asserted.

B. Partially Horizontal Collaborative Clustering With Entropy Based Approach

In this experiment, the 600 patterns and the collaboration environment are completely same to those in Example I. But the labeled patterns are determined by the entropy based approach. The result under different thresholds is shown in Figure 2.

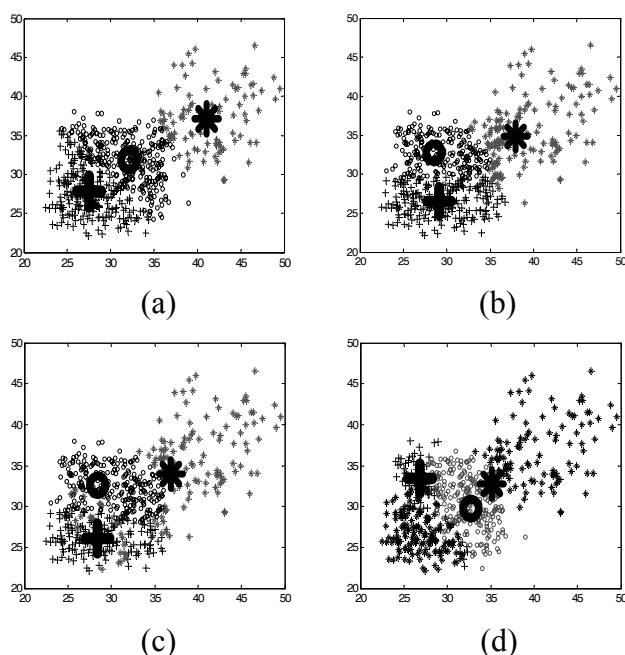


Figure 2. Result of collaborative clustering of X under different thresholds: (a) $\gamma=0$; (b) $\gamma=20\%$; (c) $\gamma=50\%$; (d) $\gamma=80\%$.

In Figure 2(a), the threshold is equal to 0, it means no pattern is labeled and the result of collaborative clustering on X is same to that of clustering on X without collaboration. For thresholds 20%, 50% and 80%, 120, 300 and 480 patterns are labeled out respectively. With the collaboration of the corresponding labeled patterns, different clustering results are obtained, see Figure 2(b), Figure 2(c) and Figure 2(d). In Figure 2(b) and Figure 2(c), we can find subtle shift in prototypes and visible change in clusters' distribution by comparing to Figure 2(a). While, in Figure 2(d), great change in prototypes and clusters' distribution appears there. The original data structure is seriously destroyed. Specifically, two clusters in Figure(a) are united to one new cluster, while the third cluster in Figure(a) is separated into two new clusters.

We can also quantify the effect of collaborative clustering with the two evaluation indexes δ, ϕ and N which stands for the number of the labeled patterns, see Table 2.

From Table 2, we can conclude that: The larger the threshold is, the more labeled patterns will be collaborated with, and better reconciliation effect between the structure of clustering in central party and that in the outer party (measured by ϕ) will be induced. Furthermore, ϕ decreases and N increases following the decrease of γ . But no certain variation trend of δ exists.

Table 2. Evaluation of partially horizontal collaborative FCM

Threshold γ	$\delta = \ u_{ref} - u\ $	$\phi = \ u - u[jj]\ $	N
0	0.01	11.0960	[0, 0, 0]
20%	8.2763	10.8659	[0, 41, 79]
50%	6.9770	9.5175	[6, 116, 178]
80%	10.0998	4.9959	[36, 232, 212]

5. Conclusions

In this paper, we present a new approach to dealing with the knowledge-based horizontal collaborative clustering (HC-FCM), which is named partially horizontal collaborative FCM (PHC-FCM). Instead of completely use of the knowledge provided by the outer partition matrix in CHC-FCM, the new version method focuses on partially use of the knowledge provided by the outer partition matrix. Exactly speaking, in the new method, the knowledge is provided by some labeled patterns selected from all the patterns in terms of the partition matrix. In this paper, two approaches to realize the selection of the labeled patterns, named cut-set based approach and entropy based approach are proposed there. By these two approaches, we can select the desired patterns and label them effectively. Based on these two approaches, a detailed algorithm of PHC-FCM is designed. From the results of the carried experiments, we can say that the new method is effective and flexible, and may meet the need of reality on some occasions. One thing needed to say is that in this paper we confine our study to such horizontal collaborative clustering problems where there is one central party and only one outer party. How to deal with the more general horizontal collaborative clustering problem is our future work.

Acknowledgment

Support from the Project 60775032 supported by National Natural Science Foundation of China (NSFC) is gratefully acknowledged.

References

- [1] J.C. Bezdek, *Pattern recognition with fuzzy objective function algorithms*, Plenum Press, New York, 1981.
- [2] J.C. Bezdek, R.Ehrlich, W.Full, "FCM: the fuzzy C-means clustering algorithm," *Computers and Geosciences*, vol. 10, no. 2-3, pp. 191-203, 1984.
- [3] A. Deluca, S. Termini, "A definition of a non-probabilistic entropy in the setting of fuzzy theory," *Information and Control*, vol. 20, pp. 301-312, 1972.
- [4] Jiulun Fan, Wenzhi Zhen and Weixin Xie, "Sup-

- pressed Fuzzy C-Means Clustering Algorithm [J],” *Pattern Recognition Letters*, vol. 24, pp. 1607-1612, 2003.
- [5] Stanley R.M. Oliveira, Osmar R.Zaiane, “A privacy-preserving clustering approach toward secure and effective data analysis for business collaboration,” *computers & XXX*, pp. 1-13, 2006.
- [6] W. Pedrycz, J. Waletzky, “Fuzzy clustering with partial supervision,” *IEEE Transactions on Systems, Man, and Cybernetics-B*, vol. 27, no. 5, October, pp. 787-795, 1997.
- [7] W. Pedrycz, “Collaborative fuzzy clustering,” *Pattern Recognition Letter*, vol. 23, no. 14, December, pp. 1675-1686, 2002.
- [8] W.Pedrycz, *Knowledge-Based Clustering: From Data to Information Granules*, A John Wiley&Sons, Inc.,Publication, 2005.
- [9] Fusheng Yu, Juan Tang, Ruiqiong Cai, “A Necessary Preprocessing in Horizontal Collaborative Fuzzy Clustering,” 2007 *IEEE International Conference on Granular Computing*, Silicon Valley, USA. Nov., no. 2-4, 2007.



Dr. Fusheng Yu is an associate professor in the School of Mathematics Sciences, Beijing Normal University, Beijing, China. He received the M.S. degree and PhD degree in Applied Mathematics from Beijing Normal University in 1989 and in 1998

respectively. From 2002 to 2004, he was with the Department of Electrical and Computer Engineering, University of Alberta, Edmonton, Alberta, Canada as a visiting scholar. He is pursuing research in computational intelligence, fuzzy systems and fuzzy modeling, knowledge discovery and data mining, fuzzy neural networks, expert systems and fault diagnosis, and knowledge representation.