# Similarity-Driven Cluster Merging Method for Unsupervised Fuzzy Clustering

Xuejian Xiong, Kian Lee Tan
Singapore-MIT Alliance
E4-04-10, 4 Engineering Drive 3
Singapore 117576

***Abstract*— In this paper, a similarity-driven cluster merging method is proposed for unsupervised fuzzy clustering. The cluster merging method is used to resolve the problem of cluster validation. Starting with an overspecified number of clusters in the data, pairs of similar clusters are merged based on the proposed similarity-driven cluster merging criterion. The similarity between clusters is calculated by a fuzzy cluster similarity matrix, while an adaptive threshold is used for merging. In addition, a modified generalized objective function is used for prototype-based fuzzy clustering. The function includes the $p$-norm distance measure as well as principal components of the clusters. The number of the principal components is determined automatically from the data being clustered. The performance of this unsupervised fuzzy clustering algorithm is evaluated by several experiments of an artificial data set and a gene expression data set.**

***Index Terms*— Cluster Merging, Unsupervised Fuzzy Clustering, Cluster Validity, Gene Expression Data**

## I. Introduction

In prototype-based fuzzy clustering methods, for example, the well-known Fuzzy $C$-Means (FCM) algorithm[2], several problems are still open for obtaining a good performance. These concern the number of clusters in the data, the uneven distribution of the data points, the initialization of the clustering algorithm, the large difference of cluster's sizes, the shape of the clusters, etc. Determining the optimal number of clusters is an importance issue in cluster validation for clustering. Traditionally, the optimal number of clusters is determined by evaluating a certain global validity measure of the $c$-partition for a range of $c$ values, and then picking the value of $c$ that optimizes the validity measure in some sense[6], [15], [1]. However, it is difficult to devise a unique measure that takes into account the variability in cluster shape, density, and size. Moreover, these procedures are computationally expensive because they require solving the optimization problem repeatedly for different values of the number of clusters $c$ over a pre-specified range $[c_{min}, c_{max}]$. In addition, the validity measures may not always give the correct number of clusters $c$ [10]. In order to overcome these problems, researchers proposed merge-split or progressive clustering schemes based on the values of validity function. However, the validity function in turn depends on the objective function of the fuzzy clustering algorithm[2], [5], and it is non-convex.

Cluster merging[11] is proposed as a way to select the number of clusters. The data is clustered with an overspecified value of $c$. After the data is partitioned into $c$ clusters, similar clusters are merged together based on a given assessment criterion until no more clusters can be merged. The procedure of cluster validation is independent of the clustering algorithm, and the number of clusters is reduced dynamically. Krishnapuram *et al.* presented the compatible cluster merging method for unsupervised clustering[10]. Kaymak *et al.*[9] also used the cluster merging method to determine the number of clusters in an extended FCM algorithm.

In this paper, a similarity-driven cluster merging method is proposed for unsupervised fuzzy clustering. The clustering starts with a large number of clusters. Pairs of similar clusters are repeatedly merged, based on the proposed similarity-driven cluster merging criterion, until the correct number of clusters are determined. The similarity between clusters is calculated by a proposed fuzzy cluster similarity matrix. The merging threshold can be determined automatically and adaptively. In addition, a modified generalized objective function is used for fuzzy clustering. The function includes the $p$-norm distance measure and the principal components of clusters. The number of the principal components is

Manuscript received November 03, 2003. This work was supported by the Singapore-MIT Alliance.

Xuejian Xiong is with Singapore-MIT Alliance, 3 Science Drive 2, National University of Singapore, Singapore, 117543 (phone: 065-6874-4248; fax: 065-6779-4580; email: smaxx@nus.edu.sg).

Kian Lee Tan is with Singapore-MIT Alliance, 3 Science Drive 2, National University of Singapore, Singapore, 117543 (email: tankl@comp.nus.edu.sg).

determined automatically from the data being clustered.

The organization of this chapter is as follows. Section II presents the similarity-driven cluster merging method for solving the fuzzy cluster number problem in unsupervised fuzzy clustering. In section III, the modified generalized objective function based on the fuzzy $c$-prototype form is described. Experimental results on an artificial data set and a gene expression data set are presented in section V. Finally, conclusion is given in section VI.

## II. Similarity-Driven Cluster Merging Method

The cluster merging approach offers an automatic and computationally less expensive way for cluster validation, but so far, most of the cluster merging methods heavily depend on the clustering procedure. In other words, these methods belong to dynamic cluster validation. They cannot be applied to other clustering algorithms easily. In the process, the intermediate clustering results are also affected by cluster merging. However, the static cluster validation method leads to heavy computation due to repeated clustering. In this section, a cluster merging method is proposed. It combines the advantages of dynamic and static cluster validation approaches. The proposed cluster merging method is based on a new similarity-driven cluster merging criterion. As a result, similarity between two clusters can be measured by two conflicting factors: separation between a pair of clusters and compactness within each cluster of the pair. Based on this criterion, similar clusters can be merged together at a time. As a result, the over-partitioning of the data can be merged to the optimal fuzzy partitioning in a few steps.

### A. Similarity-Driven Cluster Merging Criterion

Let us consider a collection of data $\mathbf{X} = \{\mathbf{x} \in \Re^n\}$, in which there are $c$ clusters $\{\mathbf{P}_1, \mathbf{P}_2, \cdots, \mathbf{P}_c\}$. $\{\mathbf{V}_i \in \Re^n, i = 1, 2, \cdots, c\}$ are the prototypes of the corresponding clusters. If $dp_i$ is the fuzzy dispersion of the cluster $\mathbf{P}_i$, and $dv_{ij}$ denotes the dissimilarity between two clusters $\mathbf{P}_i$ and $\mathbf{P}_j$, then a fuzzy cluster similarity matrix $\mathbf{FR} = \{FR_{ij}, (i, j) = 1, 2, \cdots, c\}$ is defined as:

$$FR_{ij} = \frac{dp_i + dp_j}{dv_{ij}}. \quad (1)$$

The fuzzy dispersion $dp_i$ can be seen as a measure of the radius of $\mathbf{P}_i$,

$$dp_i = \left( \frac{1}{n_i} \sum_{\mathbf{x} \in \mathbf{P}_i} \mu_i^m \parallel \mathbf{x} - \mathbf{V}_i \parallel^2 \right)^{1/2} \quad (2)$$

where $n_i$ is the number of data points in the cluster $\mathbf{P}_i$, $\mu_i = \{\mu_{i1}, \cdots, \mu_{iN}\}$ denotes the $i$-th row in the membership matrix $\mathbf{U} = \{\mu_{ij}\}$, and $m \in [0, \infty)$ is a fuzziness parameter. $dv_{ij}$ describes the dissimilarity between $\mathbf{P}_i$ and $\mathbf{P}_j$,

$$dv_{ij} = \parallel \mathbf{V}_i - \mathbf{V}_j \parallel = (\sum_{s=1}^{n} \mid V_{is} - V_{js} \mid^2)^{1/2}. \quad (3)$$

It can be seen that $FR_{ij}$ reflects the ratio of the sum of the fuzzy dispersion of two clusters, $\mathbf{P}_i$ and $\mathbf{P}_j$, to the distance between these two clusters. It can be concluded that $FR_{ij}$ is nonnegative and symmetric. $FR_{ij}$ reflects the similarity between $\mathbf{P}_i$ and $\mathbf{P}_j$. Hence, it can be used to determine whether two clusters are similar or not, according to the following defined similarity-driven cluster merging criterion.

Considering a data set $\mathbf{X}$, there are $c$ clusters $\{\mathbf{P}_i, i = 1, 2, \cdots, c\}$. In each cluster, for example $\mathbf{P}_i$, $\mu_i$ is the membership vector of all data in $\mathbf{X}$ with respect to $\mathbf{P}_i$, and $\mathbf{V}_i$ denotes the prototype of $\mathbf{P}_i$. For a fuzzy similarity matrix $\mathbf{FR} = \{FR_{ij}\}$ and a given threshold $\tau$, the similarity-driven cluster merging criterion is defined as:

If $FR_{ij} \leq \tau$,
the cluster $\mathbf{P}_i$ and the cluster $\mathbf{P}_j$ are completely separated.
If $FR_{ij} > \tau$,
the cluster $\mathbf{P}_i$ and the cluster $\mathbf{P}_j$ can be merged
to form a new cluster $\mathbf{P}_{i'}$, with $\mu_{i'} = \mu_i + \mu_j$ and $\mathbf{V}_{i'} = \frac{\mathbf{V}_i + \mathbf{V}_j}{2}$,
then $c' = c - 1$, (4)

where $\mathbf{P}_{i'}$ refers to the new cluster after merging, $\mu_{i'}$ and $\mathbf{V}_{i'}$ denote the membership vector and the prototype of $\mathbf{P}_{i'}$, respectively, and $c'$ is the number of clusters after merging. Note that the merging order of pairs of clusters in an iteration is according to the value of $FR_{ij}$ (see Table I).

Furthermore, a corresponding index is defined as:

$$DB_{FR} = \frac{1}{c} \sum_{i=1}^{c} FR_i \quad (5)$$

TABLE I
THE MERGING ORDER OF CLUSTERS BASED ON THE SIMILARITY-DRIVEN CLUSTER MERGING CRITERION.

| |
|---|
| If $[i_1, j_1] = \arg\max_{(i,j)}\{FR_{ij} > \tau\}$, |
| then the cluster $\mathbf{P}_{i_1}$ and the cluster $\mathbf{P}_{j_1}$ are merged first, |
| If $[i_2, j_2] = \arg\max_{(i,j)\neq(i_1,j_1)}\{FR_{ij} > \tau\}$, |
| then the cluster $\mathbf{P}_{i_2}$ and the cluster $\mathbf{P}_{j_2}$ are merged next, |
| ..... |
| Until there is no $FR_{ij} > \tau$. |

where $FR_i = \max_{i \neq j}\{FR_{ij}, (i, j = 1, 2, \cdots, c)\}$. Actually, the minimum $DB_{FR}$ corresponds to the optimal $c_{opt}$. Because $DB_{FR}$ is similar to the well-known DB index[14], it is named as the fuzzy DB index.

### *B. The Determination of the Threshold for Similarity-Driven Cluster Merging Criterion*

In order to define $\tau$, the following definition is given. For the data set $\mathbf{X} = \{\mathbf{x}_k, k = 1, \cdots, N\}$, $\mathbf{P} = \{\mathbf{P}_i, i = 1, \cdots, c\}$ is a set of $c$ clusters of $\mathbf{X}$, and the corresponding prototypes are $\{\mathbf{V}_i, i = 1, \cdots, c\}$. $\forall \mathbf{x}_k \in \mathbf{P}_i$, there is

$$\mathbf{P}_i^{'} = \{\mathbf{x}_k | D(\mathbf{x}_k, \mathbf{V}_i) \leq dp_i, \quad \mathbf{x}_k \in \mathbf{P}_i, \quad k = 1, 2, \cdots, N\} \quad (6)$$

where $D(\mathbf{x}_k, \mathbf{V}_i)$ denotes the distance between $\mathbf{x}_k$ and $\mathbf{V}_i$, and $dp_i$ represents the fuzzy dispersion of $\mathbf{P}_i$.

It can be seen that $\mathbf{P}_i^{'} \subset \mathbf{P}_i$. Nonetheless, $\mathbf{P}_i^{'}$ can be used to represent the cluster $\mathbf{P}_i$, i.e. $\mathbf{P}_i^{'} \approx \mathbf{P}_i$. Therefore, the following criteria can be obtained.

$$\begin{aligned} &\text{if} \quad \mathbf{P}_i^{'} \cap \mathbf{P}_j^{'} = \emptyset, \quad \text{i.e.} \quad \sharp(\mathbf{P}_i^{'} \cap \mathbf{P}_j^{'}) = 0, \\ &\text{then} \quad dp_i + dp_j < dv_{ij} \quad \text{i.e.} \quad FR_{ij} < 1; \end{aligned} \quad (7)$$

$$\begin{aligned} &\text{if} \quad \mathbf{P}_i^{'} \cap \mathbf{P}_j^{'} \neq \emptyset, \quad \text{i.e.} \quad \sharp(\mathbf{P}_i^{'} \cap \mathbf{P}_j^{'}) \geq 1, \\ &\text{then} \quad dp_i + dp_j \geq dv_{ij} \quad \text{i.e.} \quad FR_{ij} \geq 1, \end{aligned} \quad (8)$$

where $\sharp(\mathbf{P}_i)$ stands for the number of data points in the cluster $\mathbf{P}_i$.

The contour of the dispersion of a cluster can be drawn to represent the cluster as shown in Figure 1. If two clusters, $\mathbf{P}_i$ and $\mathbf{P}_j$, are far away from each other, i.e. there is no intersection between two dispersion contours (refer to equation(7)), it is believed that the two clusters are well separated from each other. As shown in Figure 1, $\mathbf{P}_1$ and $\mathbf{P}_4$ are two completely separated clusters. If there is an intersection between the dispersion contours of two clusters, it can be said that these two clusters are overlapped clusters and should be merged together (refer to equation(8)). From Figure 1, it can be considered that $\mathbf{P}_2$ and $\mathbf{P}_3$, $\mathbf{P}_4$ and $\mathbf{P}_5$, $\mathbf{P}_5$ and $\mathbf{P}_6$ are overlapped with each other. However, if two dispersion contours are at tangent, i.e. $dp_i + dp_j = dv_{ij}$ and then $FR_{ij} = 1$, it can be considered that $\mathbf{P}_i$ and $\mathbf{P}_j$ are separated. Therefore, the similarity threshold $\tau$ can be fixed as 1. $\tau$ can also be given other values. If $\tau > 1$, for example $\tau = 2$, it means that two clusters can be seen as separating from each other well even though they overlap much more. Otherwise, if $\tau < 1$, for example $\tau = 0.5$, it means that two clusters should be merged together even though they are well separated.

The value of $\tau$ can affect the final solution and speed of the cluster merging. Thus, the definition of the similarity-driven cluster merging criterion, in equation (4), can be refined. Consider a data set $\mathbf{X}$ with $c$ clusters $\{\mathbf{P}_i, i = 1, 2, \cdots, c\}$. In each cluster, for example $\mathbf{P}_i$, $\mu_i$ is the membership vector of all data in $\mathbf{X}$ with respect to $\mathbf{P}_i$, and $\mathbf{V}_i$ denotes the prototype of $\mathbf{P}_i$. For a fuzzy similarity matrix $\mathbf{FR} = \{FR_{ij}\}$ and two given thresholds $\tau_1$ and $\tau_2$, the refined similarity-driven cluster merging criterion is defined as:

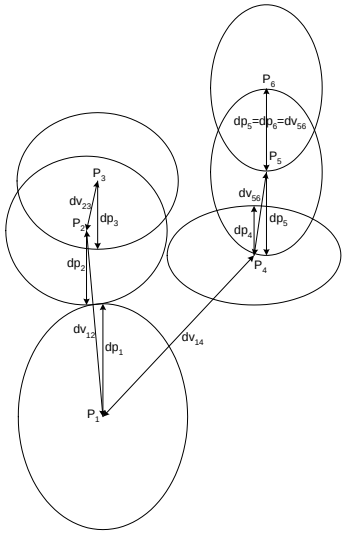



Fig. 1. Intersection between pairs of clusters represented by their dispersion contours.

$$\begin{aligned} &\text{If } FR_{ij} < \tau_1, \\ &\text{two clusters, } \mathbf{P}_i \text{ and } \mathbf{P}_j, \text{ separate from} \\ &\text{each other completely.} \\ &\text{If } \tau_1 \leq FR_{ij} \leq \tau_2, \\ &\text{an annealing technique is needed to} \\ &\text{find the appropriate } \tau \text{ for the equation(4).} \\ &\text{If } FR_{ij} > \tau_2, \\ &\text{two clusters, } \mathbf{P}_i \text{ and } \mathbf{P}_j, \text{ should be merged} \\ &\text{immediately to form a new cluster } \mathbf{P}_{i'}, \\ &\text{with } \mu_{i'} = \mu_i + \mu_j \text{ and } \mathbf{V}_{i'} = \frac{\mathbf{V}_i + \mathbf{V}_j}{2}, \\ &\text{then } c^{'} = c - 1, \end{aligned} \quad (9)$$

where $\mathbf{P}_{i'}$ refers to the new cluster after merging, $\mu_{i'}$ and $\mathbf{V}_{i'}$ denote the membership vector and the prototype of $\mathbf{P}_{i'}$, respectively, and $c^{'}$ is the number of clusters after merging.

Based on the discussion of equations (7) and (8) and Figure 1, it is seen that $\tau_1$ can be reasonably set as 1. Normally, if $FR_{ij} \geq 2$, $\mathbf{P}_i$ and $\mathbf{P}_j$ will be considered as the overlapped clusters to be merged with no doubt.

As a result, $\tau_2$ is set to 2. If $1 \leq FR_{ij} \leq 2$, the appropriate value of the threshold is obtained adaptively and automatically.

## III. A Modified Generalized Objective Function

A modified generalized objective function for the unsupervised fuzzy clustering algorithm is described in this section. The function consists of the $p$-norm distance measure and principal components of clusters.

Consider a collection of $N$ data $\{\mathbf{x}_k \in \Re^n, k = 1, 2, \cdots, N\}$ forming the data set $\mathbf{X}$. There are $c$ clusters whose prototypes are $\mathbf{V} = \{\mathbf{V}_i \in \Re^n, i = 1, \cdots, c\}$. The modified generalized objective function based on [2], [16] is proposed as follows:

$$\begin{aligned} J_{\{m,p\}}(\mathbf{U}, \mathbf{V}; \mathbf{X}) &\triangleq \sum_{i=1}^{c}\sum_{k=1}^{N}(\mu_{ik})^m D_{ik} \\ &= \sum_{i=1}^{c}\sum_{k=1}^{N}(\mu_{ik})^m \left\{ \{\parallel \mathbf{x}_k - \mathbf{V}_i \parallel_p\}^p + g\sum_{s=1}^{r}\mathbf{S}_{is}^T(\mathbf{x}_k - \mathbf{V}_i) \right\} \\ &\triangleq \sum_{i=1}^{c}\sum_{k=1}^{N}(\mu_{ik})^m \{D_p(ik) + gD_r(ik)\} \end{aligned} \quad (10)$$

where $p \geq 1$, $m \in [0, \infty)$ is a fuzziness parameter, and $g \in [0, 1]$ is a weight. $\{\mathbf{S}_{is} \in \Re^n, s = 1, \cdots, r\}$ are $r$ eigenvectors of the generalized within-cluster scatter matrix of the cluster $\mathbf{P}_i$. $\mathbf{U} = \{\mu_{ik}\}$ is the fuzzy membership matrix, and $\mu_{ik}$ should satisfy the following constraints:

$$\begin{aligned} 0 \leq \mu_{ik} \leq 1 \qquad & \forall \quad i, k, \\ \textstyle\sum_{i=1}^{c} \mu_{ik} = 1 \qquad & \forall \quad k, \\ 0 < \textstyle\sum_{k=1}^{N} \mu_{ik} < N \qquad & \forall \quad i. \end{aligned} \quad (11)$$

The first term $D_p(ik)$, in the objective function $J_{\{m,p\}}(\mathbf{U}, \mathbf{V}; \mathbf{X})$, characterizes the distance from a data point $\mathbf{x}_k$ to the cluster $\mathbf{P}_i$, based on the $p$-norm distance measure. The second term $D_r(ik)$ introduces the principal axes of the cluster $\mathbf{P}_i$, which are determined by the collection of $r > 0$ linearly independent vectors $\{\mathbf{S}_{is}, s = 1, 2, \cdots, r\}$.

$\{\mathbf{S}_{i1}, \mathbf{S}_{i2}, \cdots, \mathbf{S}_{ir}\}$ are eigenvectors corresponding to the first $r$ largest eigenvalues of the generalized within-cluster scatter matrix $\mathbf{E}_i$. $\mathbf{E}_i$ is given as follows,

$$\mathbf{E}_i = \sum_{k=1}^{N}(\mu_{ik})^m(\mathbf{x}_k - \mathbf{V}_i)(\mathbf{x}_k - \mathbf{V}_i)^T. \quad (12)$$

$\{\mathbf{S}_{is}, s = 1, 2, \cdots, r\}$ gives the cohesiveness of the cluster $\mathbf{P}_i$. In fact, $\{\mathbf{S}_{is}, s = 1, 2, \cdots, r\}$ are the $r$ principal eigenvectors of the cluster $\mathbf{P}_i$. They give the most important directions, along which most of the data points in the cluster scatter. Through the weighted term $D_r(ik)$, the principal directions of the cluster $\mathbf{P}_i$ can be emphasized. In other words, the search for the prototype $\mathbf{V}_i$ is only along the principal directions. As a result, the speed of the search is improved. Especially for a large number of data points, the appropriate value of $r$ can be selected to significantly improve the convergence speed of the fuzzy clustering algorithm.

Choosing a suitable value of $r$ in different applications is still a problem. For the fuzzy $c$-elliptotypes and fuzzy $c$-variants algorithms, two variations of the FCM[2], [16], $r$ must be specified *a priori* based on the assumed shape of clusters. However, it is difficult to imagine the shape of clusters if the dimension of the data is larger than three, i.e. $n > 3$.

Since the minimum description length (MDL)[7] is one of the well-known criteria for model order selection, the MDL is used here to find the optimal value of $r$. For $N$ input data $\{\mathbf{x}_k \in \Re^n, k = 1, 2, \cdots, N\}$, there is

$$MDL(j) = -(n-j)N \ln \frac{\mathcal{G}(\lambda_{j+1}, \cdots, \lambda_n)}{\mathcal{A}(\lambda_{j+1}, \cdots, \lambda_n)} + \frac{1}{2}j(2n-j)\ln N \quad (13)$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ denote the eigenvalues of $\mathbf{E}_i$, and $j \in [1, 2, \cdots, n]$. $\mathcal{G}(\cdot)$ and $\mathcal{A}(\cdot)$ denote the geometric mean and the arithmetic mean of their arguments, respectively. Hence, the optimal value of $r$ can be determined as:

$$r = \{j | \min_{j=r_1, r_1+1, \cdots, n-1} MDL(j)\}. \quad (14)$$

That is, equation (14) searches for the optimal $r$ from $[r_1, \cdots, n-1]$. Normally, $r_1 = 1$.

Depending on the data set, the involved second term can also affect the shapes of clusters. Because the $r$ principal components of clusters are considered during the clustering, the shape of the clusters can be changed from, for example, if $p = 2$, hyperspherical to hyperellipsoidal.

## IV. The Complete Unsupervised Fuzzy Clustering Algorithm

The unsupervised fuzzy clustering algorithm consists of a modified generalized objective function for fuzzy clustering, and a similarity-driven cluster merging criterion for cluster merging, i.e. the GFC-SD algorithm in short. The complete GFC-SD algorithm is described as follows:

step 1. *Initialization*:

Pre-selecting the maximum value for the number of clusters $c = c_{max}$, obviously $c_{max} < N$; predefining $g$, $p$, $r_1$, $m$, the tolerance $\epsilon$, the merging thresholds $\tau_1$ and $\tau_2$; setting the initial

membership matrix $\mathbf{U}$ subject to constraints in equation (11).

step 2. *Updating*:
updating the cluster prototypes $\mathbf{V}$ and the membership matrix $\mathbf{U}$. The updating formulas can be obtained by differentiating the generalized objective function $J_{\{m,p\}}(\mathbf{U}, \mathbf{V}; \mathbf{X})$ with respect to $\mathbf{V}$ and $\mathbf{U}$, respectively.

step 3. *The penalty rule*:
If the given stopping criterion is satisfied, i.e. $\| \mathbf{U}_{new} - \mathbf{U} \| < \epsilon$, go to next step. Else go back to step 2, replace the old $\mathbf{U}$ with the new partition matrix $\mathbf{U}_{new}$.

step 4. *Cluster merging*:
Merging clusters based on proposed similarity-driven cluster merging criterion. If $c$ is not changed, then stop the procedure. Else go back to the step 2, repeat the whole procedure according to the new number of clusters $c$, and use current corresponding $\mathbf{V}$ and $\mathbf{U}$ as the initialization.

## V. Experiments

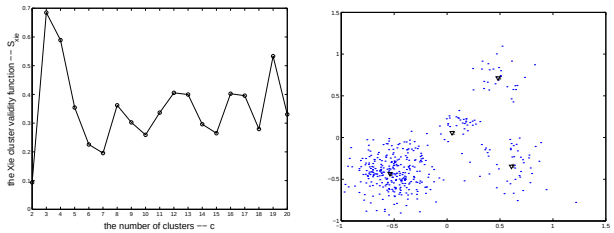
In this section, the performance of the GFC-SD algorithm is studied. For comparison, the GFC-SD algorithm is applied to an artificially generated two-dimensional data set, which was used in [9]. Moreover, the proposed GFC-SD algorithm is applied to a gene expression data set, the serum data set, which is preprocessed in the same way as in [3], [13] by using the variance normalization. All experiments are done with a fuzziness parameter $m = 2$ and a 2-norm distance measure, i.e. $p = 2$. The tolerance for fuzzy clustering $\epsilon$ is selected as $0.001$. The merging threshold $\tau$ is determined adaptively according to equation (9) with $\tau_1 = 1$ and $\tau_2 = 2.$ , All experimental results are obtained on a 1.72GHz Pentium IV machine with 256MB memory, running Matlab 5.3 on Windows XP.

### *A. The Artificial Data Set with Uneven-Distributed Groups*

As mentioned in [9], four groups of data are generated randomly from normal distributions around four centers with the standard deviations given in Table II. The number of sample points in each group is also indicated. It can be seen that the number of sample points in group 1, i.e. 300, is much larger than that of other three groups. That is, the differences in cluster density are quite large.

TABLE II
The group centers and number of samples in each group of the artificial data set with uneven-distributed groups.

| group | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| original center | (-0.5, -0.4) | (0.1, 0.2) | (0.5, 0.7) | (0.6, -0.3) |
| number of samples | 300 | 30 | 30 | 50 |




(a) FCM-Xie (b) GFC-SD

Fig. 2. (a) The FCM-Xie algorithm fails in determining the four clusters in the data set. (b) The GFC-SD algorithm automatically detects the correct number of clusters in the data set. The searched GFC-SD prototypes are denoted by the black triangle and numbers.

In this experiment, the goal is to automatically detect clusters reflecting the underlying structure of the data set. Here, the well-known FCM method with the popular Xie's cluster validity function[15], i.e. FCM-Xie in short, is used for comparison. The range of values of $c$ is $[2, 20]$. From Figure 2(a), it can be observed that the number of clusters is seven referring to the minimum $S_{xie}$. Therefore, the conventional approach, FCM-Xie, fails to determine the correct number of clusters in the data due to the largely uneven distribution of the data. The proposed GFC-SD algorithm, however, finds the four groups present in the data correctly, as shown in Figure 2(b). As a result, the GFC-SD algorithm is more robust for largely uneven-distributed data than the FCM-Xie algorithm, as well as Kaymak's extended FCM algorithm[9].

Like the experimental procedure in [9], the influence of initialization on the GFC-SD algorithm is also studied. The data set is clustered 1000 times with the FCM and the GFC-SD algorithms, respectively. At each time, the randomly initialized fuzzy partitions, $\mathbf{U}$, are input into the algorithms. The FCM algorithm is set to partition the data into four clusters, i.e. $c = 4$, while the GFC-SD algorithm is started with twenty clusters, i.e. $c_{max} = 20$. After 1000 experiments, the mean and standard deviation of obtained cluster prototypes are shown in Table III. Obviously, the cluster prototypes found by the GFC-SD algorithm are closer to the true centers than those found by the FCM algorithm. Moreover, the standard deviation of the GFC-SD found prototypes is much more lower. In fact, it almost equals to zero. The FCM algorithm has difficulty with small data groups, whose prototypes will be attracted by those of large ones. If there are

TABLE III

MEAN AND STANDARD DEVIATION OF CLUSTER PROTOTYPES FOUND BY THE FCM AND GFC-SD ALGORITHMS AFTER 1000 EXPERIMENTS WITH RANDOM INITIALIZATION.

| | | FCM prototype | |
|---|---|---|---|
| group | original center | mean | std. dev. |
| 1 | (-0.5, -0.4) | (-0.5942, -0.4191) | (0.0192, 0.0814) |
| 2 | (0.1, 0.2) | (-0.4073, -0.3904) | (0.0066, 0.1166) |
| 3 | (0.5, 0.7) | (0.4176, 0.6120) | (0.0067, 0.0095) |
| 4 | (0.6, -0.3) | (0.5799, -0.2848) | (0.0032, 0.0038) |

| | | GFC-SD prototype | |
|---|---|---|---|
| group | original center | mean | std. dev. |
| 1 | (-0.5, -0.4) | (-0.5397, -0.4326) | $< 10^{-13}$ |
| 2 | (0.1, 0.2) | (0.0469, 0.0578) | $< 10^{-13}$ |
| 3 | (0.5, 0.7) | (0.4815, 0.7133) | $< 10^{-13}$ |
| 4 | (0.6, -0.3) | (0.6083, -0.3459) | $< 10^{-13}$ |

TABLE IV

AVERAGE COMPUTATIONAL LOAD OVER 1000 TIMES FOR VARIOUS CLUSTERING ALGORITHM.

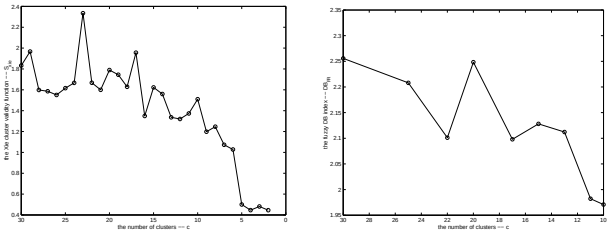
| FCM | FCM-Xie | GFC | GFC-SD |
|---|---|---|---|
| $(c = 4)$ | $(c = [20, 2])$ | $(c = 4)$ | $(c_{max} = 20)$ |
| 7.9630s | 467.4100s | 11.1670s | 243.0890s |

much more data points in the large group than those in the small group, the later one will be missed when bad initialization is given. Therefore, its obtained mean cluster prototype is far away from the true center and the corresponding standard deviation is very large. It can be concluded that the GFC-SD algorithm is much more robust to the initialization.

To compare the computational load of various algorithms, different algorithms have been run 1000 times. Similarly, the algorithms are initialized randomly at each time. Table IV gives the results, where the GFC means the fuzzy clustering algorithm only with generalized objective function. For $c = 4$, the computational load of the GFC algorithm is larger than that of the FCM algorithm because of the additional calculation of the second term in the generalized objective function. However, by using the merging method to find the optimal partitions, i.e. GFC-SD, the computational load is only half of that using the conventional FCM-Xie approach (see Table IV).

### *B. The Gene Expression Data - The Serum Data*

The serum data[8] contains expression levels of 8613 human genes by studying the response of human fibroblasts to serum. A subset of 517 genes whose expression levels changed substantially across samples was analyzed in [3], [12], [4], [8]. By using the proposed GFC-SD algorithm, the 517 genes can be partitioned into clusters



(a) FCM-Xie (b) GFC-SD

Fig. 3. The number of clusters of the serum data is determined as ten and two, by using the FCM-Xie and the proposed GFC-SD algorithms, respectively.

whose entities share some measure of similarity in their expression pattern. Unlike other previously used clustering methods, the GFC-SD provides a "completely" unsupervised clustering of the gene expression data, because even the number of serum clusters is determined automatically.

Figure 3 presents the clustering results from using the proposed GFC-SD and the FCM-Xie algorithms. The vertical axis in Figure 3(b) represents the values of the fuzzy DB index $DB_{FR}$ (seeing equation (10)), while the horizontal one represents the number of clusters in each clustering iteration step. Using the GFC-SD, the number of clusters is determined when there is no more similar fuzzy clusters to be merged. It can be observed from Figure 3(b) that, starting with 30 clusters, the number of clusters is reduced to 25, 22, 20, 17, 15, 13, 11, and finally 10 in only nine steps, based on the proposed similarity-driven cluster merging method. As a result, the number of clusters $c$ is determined as 10, which also corresponds to the minimal value of $DB_{FR}$. For using the FCM-Xie, the number of clusters can only be found after the exhausting search from the all possible values of $c$. In this case, the range of $c$ is from 30 to 2. After 29 clustering iterations, in Figure 3(a), the number of clusters is fixed as two referring to the minimum $S_{xie}$. In [3], [12], [8], it is consistently agreed that there are 10 clusters in the serum data set with 517 genes. As a result, the proposed GFC-SD algorithm is efficient for finding the number of gene clusters automatically and correctly.

Obviously, repeated clustering leads to a heavy computation, especially for gene expression data which have high dimensionality and a large number of genes. The consumed time for running the GFC-SD and FCM-Xie is $1.1911 \times 10^2$ seconds and $3.1029 \times 10^3$ seconds, respectively. It can be seen that running the FCM-Xie spends almost 30 times longer than running the GFC-SD. Furthermore, if the given $c_{max}$ is increased, e.g. $c_{max} = 40$, the time gap between these two algorithms will

TABLE V

THE NUMBER OF CLUSTERS $c$ AND THE NUMBER OF PRINCIPAL COMPONENTS $r$ OF THE SERUM CLUSTERS IN EACH CLUSTERING ITERATION.

| iteration | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| $c$ | 30 | 25 | 22 | 20 | 17 | 15 | 13 | 11 | 10 |
| $r$ | 8 | 9 | 10 | 10 | 11 | 11 | 12 | 11 | 11 |

be enlarged quickly. Therefore, the computation time is decreased largely by using the similarity-driven cluster merging method for unsupervised fuzzy clustering of gene expression data.

An additional advantage of the proposed GFC-SD algorithm is that the optimal value of $r$ is set automatically (refer to equation (14)). Therefore, the number of principal components of each cluster can be adaptively determined. For the serum data, the values of $r$ and $c$ in each clustering iteration are listed in Table V. It is observed that there are around ten principal components constructing the serum clusters. Therefore, the GFC-SD algorithm can do the feature selection of gene expression data to some extent.

## VI. CONCLUSION

In this paper, a similarity-driven cluster merging method is proposed for unsupervised fuzzy clustering. The cluster merging method is used to resolve the problem of cluster validation. The data is clustered initially with an overspecified number of clusters. Pairs of similar clusters are merged based on the proposed similarity-driven cluster merging criterion. The similarity between clusters is calculated by a fuzzy cluster similarity matrix, while an adaptive threshold is used for merging. Therefore, a few iterations are needed to find the optimal number of clusters $c$, and more precise partitions can be obtained. Moveover, the dependency of the clustering results on the random initialization is decreased. For prototype-based fuzzy clustering, a modified generalized objective function is used. The function introduces the principal components of clusters by including an additional term. Because the data are grouped into different clusters along the principal directions of the clusters, the computational precision can be improved while the computation time can be decreased.

Two data sets are used to evaluate the performance of the GFC-SD algorithm. It can be concluded from the experiments that clustering using the GFC-SD algorithm is far less sensitive to initialization and more reliable than the compared methods. This is because the GFC-SD algorithm not only can start with a large number of small clusters, but also converge towards the optimal partitions of the data set. Moreover, because the partitions after one merging step are always the initialization of the next iteration of clustering, the total time of the fuzzy clustering is reduced. Thus, by using the GFC-SD algorithm, the optimal number of clusters and the optimal partitions of the data set can be obtained.

## REFERENCES


[1] J. C. Bezdek. Numerical taxonomy with fuzzy sets. *Journal of Math. Biol.*, 1:57–71, 1974.

[2] J. C. Bezdek, J. Keller, R. Krisnapuram, and N. R. Pal. *Fuzzy Models and Algorithms for Pattern Recognition and Image Processing*. Kluwer Academic Publishers, 1999.

[3] D. Dembele and P. Kastner. Fuzzy c-means method for clustering microarray data. *bioinformatics*, 19(8):973–980, 2003.

[4] M. Eisen, P. Spellman, P. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. In *Proceedings of the National Academy of Sciences USA*, volume 95, pages 14863–14868, December 1998.

[5] H. Frigui and R. Krishnapuram. Clustering by competitive agglomeration. *Pattern Recognition*, 30(7):1109–1119, July 1997.

[6] R. E. Hammah and J. H. Curran. Validity measures for the fuzzy cluster analysis of orientations. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(12):1467–1472, 2000.

[7] A. Hyvarinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. John Wiley & Sons, Inc., 2001.

[8] V. Iyer, M. Eisen, D. Ross, G. Schuler, T. Moore, J. Lee, J. Trent, L. Staudt, J. J. Hudson, M. Boguski, D. Lashkari, D. Shalon, D. Botstein, and P. Brown. The transcriptional program in the response of human fibroblasts to serum. *Science*, 283:83–7, 1999.

[9] U. Kaymak and M. Setnes. Fuzzy clustering with volume prototypes and adaptive cluster merging. *IEEE Transaction on Fuzzy Systems*, 10(6):705–712, 2002.

[10] R. Krishnapuram. Generation of membership functions via possibilistic clustering. In *Proceedings of the Third IEEE Conference on Fuzzy Systems and IEEE World Congress on Computational Intelligence*, 1994.

[11] R. Krishnapuram, O. Nasraoui, and H. Frigui. The fuzzy c spherical shells algorithm: A new approach. *IEEE Transactions on Neural Networks*, 3(5), September 1992.

[12] R. Sharan and R. Shamir. CLICK: A clustering algorithm with applications to gene expression analysis. In *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology (ISMB 2000)*, pages 307–316, La Jolla, August 2000.

[13] S. Tavazoie, J. D. Hughes, M. J. Campbell, R. J. Cho, and G. M. Church. Systematic determination of genetic network architecture. *Nature Genetics*, 22(3):281 – 285, 1999.

[14] S. Theodoridis and K. Koutroumbas. *Pattern Recognition*. Academic Press, 1999.

[15] X. L. Xie and G. Beni. A validity measure for fuzzy clustering. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 13(8):841–847, 1991.

[16] Y. Yoshinari, W. Pedrycz, and K. Hirota. Construction of fuzzy models through clustering techniques. *Fuzzy Sets and Systems*, 54:157–165, 1993.