Sequential Combination Methods for Data Clustering Analysis

QIAN Yuntao (钱沄涛)[1], Ching Y. Suen[2] and TANG Yuanyan (唐远炎)[3]

[1]*Department of Computer Science, Zhejiang University, Hangzhou 310028, P.R. China*

[2]*Centre for Pattern Recognition and Machine Intelligence, Concordia University, Montreal, Canada*

[3]*Department of Computer Science, Hong Kong Baptist University, Kowloon Tong, Hong Kong, P.R. China*

E-mail: ytqian@mail.hz.zj.cn; suen@cenparmi.concordia.ca; yytang@comp.hkbu.edu.hk

Received April 3, 2001; revised July 9, 2001.

Abstract This paper proposes the use of more than one clustering method to improve clustering performance. Clustering is an optimization procedure based on a specific clustering criterion. Clustering combination can be regarded as a technique that constructs and processes multiple clustering criteria. Since the global and local clustering criteria are complementary rather than competitive, combining these two types of clustering criteria may enhance the clustering performance. In our past work, a multi-objective programming based simultaneous clustering combination algorithm has been proposed, which incorporates multiple criteria into an objective function by a weighting method, and solves this problem with constrained nonlinear optimization programming. But this algorithm has high computational complexity. Here a sequential combination approach is investigated, which first uses the global criterion based clustering to produce an initial result, then uses the local criterion based information to improve the initial result with a probabilistic relaxation algorithm or linear additive model. Compared with the simultaneous combination method, sequential combination has low computational complexity. Results on some simulated data and standard test data are reported. It appears that clustering performance improvement can be achieved at low cost through sequential combination.

Keywords clustering combination, probabilistic relaxation, linear additive model, clustering criterion

1 Introduction

Clustering is an important tool for data analysis. It has been widely used in data compression, data visualization, pattern recognition, economics and other scientific fields. Many algorithms have been developed[1,2], including such popular methods as hierarchical clustering[3,4], graph theoretic clustering[5,6], statistic mixture model estimation[7,8], objective function clustering[9], and neural network clustering[2]. All clustering methods have their own advantages and drawbacks, and they are suitable to different data structures. From the mathematical viewpoint, clustering is an optimization procedure according to the clustering criterion. As there is no unity of clustering criteria that can be respected by all applications, various clustering approaches present different clustering criteria from their own perspectives. In a sense, the more desirable properties the clustering criterion has, the better the algorithm is. Since most clustering criteria are complementary rather than competitive, clustering combination becomes an important method to improve the clustering performance. However, the combination of clustering algorithms is different from the combination of classifiers[10,11], and several difficulties have to be overcome:

• The quality of clustering combination cannot be evaluated as precisely as combining classifiers. Prior knowledge and user's judgement always play a critical role in clustering performance

This work is supported in part by the China State Education Commission Laboratory for Image Processing and Intelligent Control under grant TKLJ9901, and Zhejiang Education Commission under grant No.19990119.

estimation[1]. This problem creates an obstacle to proposing a mathematical theory to design clustering combination schemes.

• As various clustering algorithms always produce results with large differences due to different clustering criteria, directly combining these clustering results with integration rules such as product, sum, median and majority vote may not generate a good result, no matter how effective these rules are in combining classifiers.

• For classifier combination, a group of classifiers must be both diverse and accurate in order to improve the recognition rate of the system[12]. But for clustering combination, a group of clustering algorithms should be chosen by more complicated strategies.

In terms of clustering criteria, all the existing clustering algorithms can be classified into two categories:

• Local criterion based clustering.

• Global criterion based clustering.

In the first method, clustering is realized at the local level, and clusters are formed according to the local structure of the data such as spatial nearest neighbor relationship and the local area statistic features. This kind of clustering method is flexible and non-parametric, and makes very few assumptions on the characteristics of the data. However it will easily run into trouble when the clusters are close to each other and the boundaries are indistinct, moreover it is extremely sensitive to random noise. Most of the graph theoretic clustering and hierarchical clustering belong to the local criterion based method. Different from the local criterion based clustering is the global criterion based clustering that assumes a prototype of data distribution, and assigns the patterns to clusters according to the distance between patterns and prototypes. If the data distribution does not conform to the presumed prototype, it will become less effective. Both the objective function algorithm and the statistic mixture model estimation belong to the global criterion based clustering.

It is obvious that if both the global and local criteria are considered in clustering, the performance of clustering may be improved. In [13,14], we proposed a multi-objective programming based clustering combination algorithm, which incorporates the global and local criteria into an objective function by a weighting method, and solves this problem with constrained nonlinear optimization programming. In most cases, it produces better results than the global or the local criterion based method alone, but it suffers high computational complexity, especially as the size of data set becomes large. To speed up the combination procedure, a sequential combination method is presented in this paper. It first uses the global criterion based fuzzy clustering algorithm to generate an initial result, then uses the local criterion based information to improve the initial result with a probabilistic relaxation algorithm or a linear additive model. The organization of this paper is as follows: in the next section we briefly review a fuzzy objective function clustering that is based on the global criterion and graph theoretic clustering that is based on local criterion. Section 3 describes the sequential combination approach in detail. Experimental results are discussed in section 4. Finally, the summary and conclusion of the study are presented.

2 Objective Function Clustering and Graph Theoretic Clustering

The objective function clustering and the graph theoretic clustering are typical of the global criterion based clustering and the local criterion based clustering respectively. They play an important role in clustering analysis. As their criteria are very complementary, we will use them for combination.

2.1 Objective Function Clustering

In the objective function clustering, the data set is divided into subsets according to their similarities and dissimilarities, the data points belonging to the same subset are close to each other in a specific measurement, whereas those belonging to different subsets are far from each other. The definitions of similarity and dissimilarity are dependent on distance measurement, and the distance measurement

also determines the prototype of clusters. By now many objective function algorithms have been proposed such as the spherical, line, shell and ellipsoid prototype based clustering[9].

In general, a basic objective function can be defined as:

$$J_m(U,H,X)=\sum_{i=1}^{C}\sum_{j=1}^{N}\mu_{ij}^m D^2(x_j,H_i) \tag{1}$$

$$\text{Probabilistic constraint: } \sum_{i=1}^{C}\mu_{ij}=1,\ j=1,\ldots,N \tag{2}$$

where C is the number of clusters, $D(x_j,H_i)$ is the distance from point x_j to the cluster kernel H_i, μ_{ij} is the membership which indicates the degree of point x_j belonging to the ith cluster, m is a constant that controls the fuzzy degree, and if $m=1$, this is a hard clustering. The most common distance measure chosen is the Euclidean distance, i.e. C-mean algorithm.

$$J_m(U,V,X)=\sum_{i=1}^{C}\sum_{j=1}^{N}\mu_{ij}^m\|x_j-v_i\|^2 \tag{3}$$

where v_i is the central position of the ith cluster.

There are various methods to get a solution of the cluster kernel parameters and fuzzy memberships by minimizing the objective function. Among them, mathematical programming and heuristic searching are two main approaches. Readers can refer to [9, 15].

2.2 Graph Theoretic Clustering

Different from the objective function clustering, the graph theoretic clustering uses various kinds of geometric structures or graphs for analyzing data. It can identify irregularly shaped or non-globular clusters according to local area distribution. Several useful algorithms have been proposed such as the nearest neighbor, minimum spanning tree, relative neighborhood and Gabriel graph clustering. These graphs reflect different local structures or inherent visual characteristics in the data set. Clustering divides the graph into connected components by identifying and deleting inconsistent edges, and each subgraph consisting of connected components refers to a cluster. Here we only introduce a modified Gabriel graph method[5], and derive a fuzzy connectivity matrix that will be used for combination.

Suppose a non-directional graph $G=(V,E), V=(v_1,\ldots,v_N)$ is a set of vertices that corresponds to the data set, and $E=(e_1,\ldots,e_M)$ is a set of distinct edges, each edge $e_m=(v_i,v_j)$ connects a pair of vertices. Assume p and q are two data points. The definitions of their Gabriel influence region[16] is given by:

$$\Gamma_{p,q}=B\Big(\frac{p+q}{2},\frac{d(p,q)}{2}\Big) \tag{4}$$

$$d(x,y)=|x-y| \tag{5}$$

$$B(x,r)=\{y:d(x,y)\le r\} \tag{6}$$

Then the edge set E can be computed by:

$$(p,q)\in E,\ \text{if and only if } \Gamma_{p,q}\cap V=\emptyset \tag{7}$$

It can be observed that the Gabriel graph only considers the information of the region between two vertices, and does not consider the regions around these two vertices. Therefore, Urquhart proposed a modified Gabriel graph whose influence region is defined as[5]:

$$\Gamma_{p,q}=B\Big(\frac{p+q}{2},\frac{d(p,q)}{2}\Big)\cup B(p,\alpha d(p,q))\cup B(q,\alpha d(p,q)) \tag{8}$$

where α is a constant, $\alpha = 0.25$ is chosen in our experiments. Then the edge set can be obtained by (7). The clusters can be formed by grouping the connected vertices.

A connectivity matrix $R_{N\times N}$ can also be derived from the edge set, and the value of its element is determined by:

$$r_{ij} = \begin{cases} 1, & \text{if } (v_i, v_j) \in E; \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

As the value of element is 1 or 0, we call this non-directional graph a crisp graph. Crisp graph may cause the loss of useful information. Therefore, it is not suitable for ambiguous and noisy environments, although it is very simple. In this paper, a fuzzy edge set is presented, in which each pair of vertices has an edge with fuzzy membership that is an indicator of connectivity strength. Thus the value of element in the connectivity matrix is redefined as:

$$r_{ij} = \frac{2}{1 + e^{\lambda A_{ij}}} \tag{10}$$

A_{ij} is the number of points in influence region $\Gamma \cap V$. $\lambda > 0$ is a constant that controls the fuzzy degree, if $\lambda \to \infty$, the fuzzy graph degenerates to a crisp graph.

From the above discussion, it can be observed that there is no contradiction between the clustering criteria of the objective function clustering and the graph theoretic clustering, and these criteria actually complement each other, although these two methods are very different. It suggests that the combination of these two methods is possible, and the combination performance may be better.

3 Sequential Combination Algorithms

Sequential combination method means that the global and local criteria are not processed simultaneously, but that they are considered sequentially. In this method, the global criterion is first to be processed, i.e., fuzzy objective function clustering algorithm is used to generate an initial clustering result. Second, the local criterion is processed, i.e., the fuzzy connectivity matrix derived from graph theoretic clustering is used to improve the initial clustering result. The key problem is to find a technique that will efficiently realize this improvement in the second step. In this section we will propose two algorithms — probabilistic relaxation and linear additive model. And their advantages and limits are also discussed.

3.1 Probabilistic Relaxation Based Combination

The second step can be considered as a procedure in which the probabilistic labeling (membership vectors) associated with every node (data point) of a graph is updated according to the statistical relations among the probabilistic labelings at neighboring nodes. So the probabilistic relaxation or the so-called relaxation labeling (RL) scheme[17−19], and other similar methods such as the stochastic relaxation[20] and the Polya Urn model[21] can be used in the second step. The mechanism of RL is based on heuristic arguments, and the local constraints or the so-called compatibility function can be designed flexibly according to our requirements. This gives RL a broader potential application, although RL was originally developed to deal with ambiguity and noise in vision systems. RL is an iterative and parallel technique which uses the local constraints to update the probability of labeling, and its iterative probability updating equations can be written as:

$$p_i^t(l) = \frac{p_i(l)^{t-1}(1 + q_i^t(l))}{\sum_{k=1}^{C} p_i^{t-1}(k)(1 + q_i^t(k))} \tag{11}$$

and

$$q_i^t(l) = \sum_{k=1}^{C} \sum_{j \in \aleph(i)} p_j(k)^{t-1} w_{ij}(l,k) \tag{12}$$

where $p_i^t(l)$ is the probability/fuzzy membership of the ith node belonging to the lth class at the iteration t, and $q_i^t(l)$ is a support function that represents an influence on the labeling by the predefined constraints between class labels. $w_{ij}(l,k)$ is the compatibility value between the pair of ith and jth nodes, and the jth node is a neighbor $\aleph(i)$ of the ith node.

We now turn to the clustering combination problem. In the graph theoretic clustering, we have constructed a non-directional graph based on the modified Gabriel graph method, and this graph contains the useful information about the local constraints of RL. Based on this graph, the neighbor of a node is defined as a node set whose elements have an edge connected to this node, and the compatibility value $w_{ij}(k,l)$ between the ith and jth nodes can be derived from the connectivity strength r_{ij}:

$$w_{ij}(k,l) = \begin{cases} r_{ij}, & \text{if } k = l, \\ -r_{ij}, & \text{if } k \neq l. \end{cases} \tag{13}$$

The initial probability p^0 is the fuzzy membership set U obtained from the objective function clustering in the first step. Based on iterative updating (11)–(12), the clustering result will be modified after each iteration.

Obviously, RL can do something useful to the clustering combination, because the updating equation includes the global and local clustering criteria, the former is reflected in the initial probabilistic distribution, and the latter is reflected in the compatibility value.

In [22], it has been pointed out that the updating sequence converges in the following cases, and if the updating sequence converges, one of the following cases will occur.

Case 1: $p_i^t(l) \neq 0$ for all i and l, and all $q_i^t(l) = \alpha.\alpha$ is a scalar value, i.e. local constraints approach a constant.

Case 2: $p_i^t(l) = 0$ for some (but not all) ls, the corresponding q_i^t can be arbitrary, but for the remaining $p_i^t(l') \neq 0$, the corresponding q_i^t should be a scalar value α.

Case 3: $p_i^t(l) = 1$ for exactly one l, and $p_i^t(l') = 0$ for all $l' \neq l$.

As the convergence conditions of Cases 1 and 2 are very difficult to be satisfied, Cases 1 and 2 seldom occur. Hence case 3 usually happens when RL converges, which means that the initial fuzzy clustering result is changed to the crisp clustering result after RL-based combination.

In practice, we sometimes find that the probabilities of most points in the data set converge so fast to $p(l) = 1$ for exactly one cluster l, and $p(l') = 0$ for all $l' \neq l$, that local connection information cannot impose any further impact on these points in this case. For clustering combination, it makes the local clustering criterion play a lesser role in combination than we expected. To overcome this problem to some extent, we increase the fuzzy degree of the initial fuzzy clustering result. For the fuzzy objective function clustering algorithm, it is only needed that a large value is given to parameter m in (1), because the larger the value of m is, the more fuzzy the clustering result is. Essentially, this method increases the weight of the local criterion by decreasing the weight of the global criterion.

3.2 Linear Additive Model Based Combination

The clustering result with RL is influenced by the predefined constraints between class labels in a neighborhood. Besides RL, other methods can also be used, among them the linear additive model[23] is a good selection. This system is similar to the continuous Hopfield network differing only in (1) the diagonal of compatibility matrix W which is not a null vector, (2) the compatibility matrix W which is not symmetric, and (3) the dynamical property which is not nonlinear. Its additive activation dynamics are:

$$U^{k+1} = U^k W \tag{14}$$

where U is a $C \times N$ membership matrix, and compatibility coefficient W is derived from the connectivity strength r_{ij}.

$$w_{ij} = \begin{cases} \dfrac{r_{ij}}{\left(\sum_{p=1,p\neq i}^{N} r_{ip}\right)+1} & \text{if } i \neq j, \\ \dfrac{1}{\left(\sum_{p=1,p\neq i}^{N} r_{ip}\right)+1} & \text{if } i = j. \end{cases} \tag{15}$$

This definition of W can guarantee that the sum of memberships of each pattern belonging to all clusters remains unchanged in all iterations

$$\sum_{i=1}^{C} \mu_{ij}^{k} = 1 \tag{16}$$

which is necessary to obtain a meaningful result.

Before we discuss the convergence of a linear additive model, a theorem about the convergence of the Markov chain is given[24].

Theorem 1. *Consider a regular Markov chain having a transition matrix P and an initial state x_0. We have* (1) *The sequence of distributions of states $x_0, x_1 = x_0P, x_2 = x_0P^2, \ldots$ approaches a vector x that satisfies $xP = x$. This limit vector is a left eigenvector of P corresponding to the eigenvalue* 1. (2) *The sequence of matrix $P, P^2, P^3, \ldots$ approaches a stochastic matrix T. The rows of T are all identical, a row being a left eigenvector of P corresponding to the eigenvalue* 1.

If we regard the transposition W^t as a specific transition matrix, the following theorem can be obtained.

Theorem 2. *In a linear additive model,* (1) *The sequence of matrix $W, W^2, W^3, \ldots$ approaches a stochastic matrix T'. The columns of T' are all identical, a column being a left eigenvector of W^t corresponding to the eigenvalue* 1. (2) *The sequence of membership sets $U_0, U_1 = U_0W, U_2 = U_0W^2, \ldots$ approaches the transposition of a left eigenvector of W^t corresponding to the eigenvalue* 1.

Theorem 2 shows that the local information in a linear additive model can spread unlimitedly until the memberships of all points are identical, which is different from RL. With this property, the weights of the global and local clustering criteria can be adjusted by the iteration number. The results of clustering combination appear to be getting better for the first few iterations, after which they may become worse, for example, the convergent result of a linear additive system is not our expectation. In order to terminate the iteration before convergence, we propose a method to determine the termination condition, which is based on the measurement of fuzzfication degree[9]. The degree of fuzzification can be measured by:

$$M_F(U^k) = 1 - \frac{1}{N}\sum_{i=1}^{C}\sum_{j=1}^{N}(\mu_{ij})^2 \tag{17}$$

When the updating sequence converges, M_F will reach to its maximum $\dfrac{C-1}{C}$. The determination condition can be defined:

$$\text{if } M_F(U^k) > T, \text{ terminate the iteration process.}$$

T is a threshold, $T = \dfrac{C-1.3}{C}$ is chosen in our experiments. In contrast to the RL-based combination, the final clustering result will be fuzzier than the initial result after the linear additive model based combination.

The probabilistic relaxation and the linear additive model have no essential difference in dealing with clustering combination, although their algorithms have some differences. Compared with the simultaneous combination methods, both of them not only need less computing time, but also provide a more flexible combination strategy. The first step is not limited to the fuzzy objective function clustering, and other global criterion based clustering algorithms can also be used. Just the same, connectivity matrix can also be derived from other local criterion based methods besides the graph theoretic algorithm.

4 Experimental Verification

To test our methods, a lot of experiments on synthetic and real data have been done. For the examples given in this section, the fuzzy C-mean (FCM) algorithm is used as the objective function clustering, and the fuzzy modified Gabriel graph clustering is used as the graph theoretic clustering. Parameter m in (3) is defined as 4 for the probabilistic relaxation based combination, and 2 for the linear additive model based combination. Other parameters used in our experiments have been defined in previous sections. In addition, when experimental results are used for visualization and comparison, they should be defuzzified.

First we apply our algorithms to Fisher's Iris data. Iris data set has four attributes, but only the third and the fourth attributes have good discrimination power, so we only consider these two attributes in the experiments (Fig.1(a)). Figs.1(b)–(d) are the results with $C = 3$ by FCM, RL-based combination, and linear additive model based combination respectively. Because the two subsets of Iris data overlap, the graph theoretic clustering fails in this data set, and the local criterion can not provide useful information for the partition of these two overlapping subsets, but FCM is suitable for this data set. From the experiment we find that the results of the three clustering methods are very similar, and among them only one or two data points have different labels, which illustrates there is not obvious contradiction between the clustering criteria of FCM and the graph theoretic algorithm, even though the latter is not suitable for this data set.

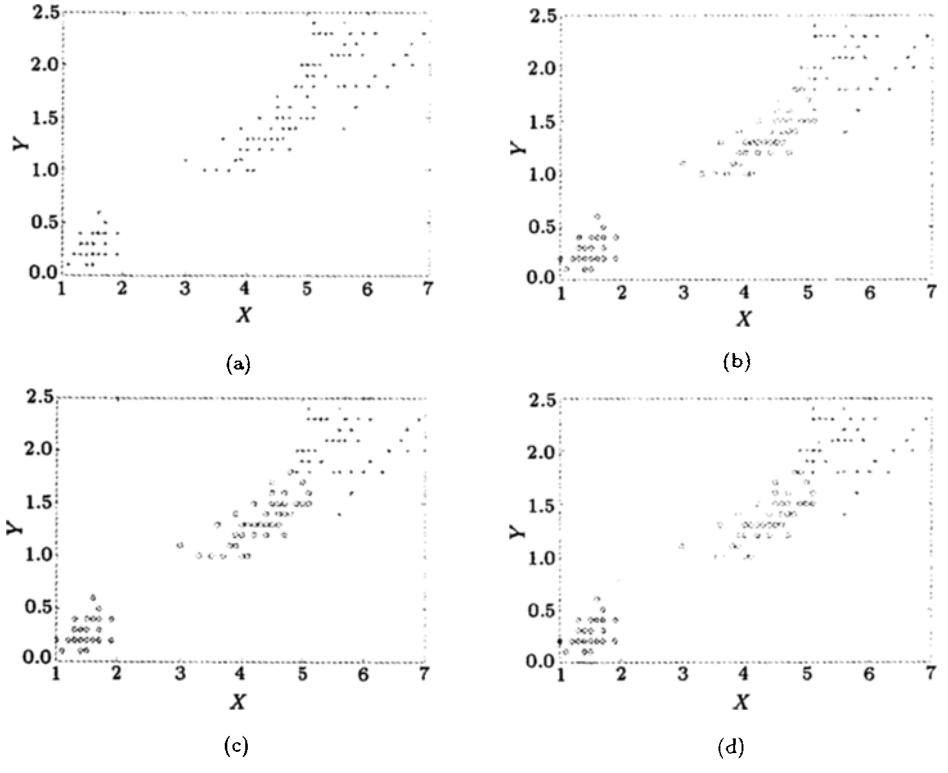

Fig.1. Experiment on Iris data. (a) Original data only including the third and the fourth attributes. (b) FCM. (c) RL-based combination. (d) Linear additive model based combination.

The data set in Fig.2(a) consists of two Gaussian distributions with different parameters, their means are (320, 200) and (320, 300), and variances are 35 and 10 respectively. As the difference between these variances of two Gaussian distributions becomes large, some patterns belonging to the

distribution with a large variance are misclassified by FCM to the distribution with a small variance (Fig.2(b)), but the modified Gabriel graph clustering can deal with this case effectively. Figs.2(c) and (d) show that the results of RL-based combination, and linear additive model based combination respectively. We find that the labels of three data points in the left of Fig.2(c) have not been corrected, which is due to the too fast convergence of RL. This experiment illustrates that the result of FCM can be improved by the local criterion based information

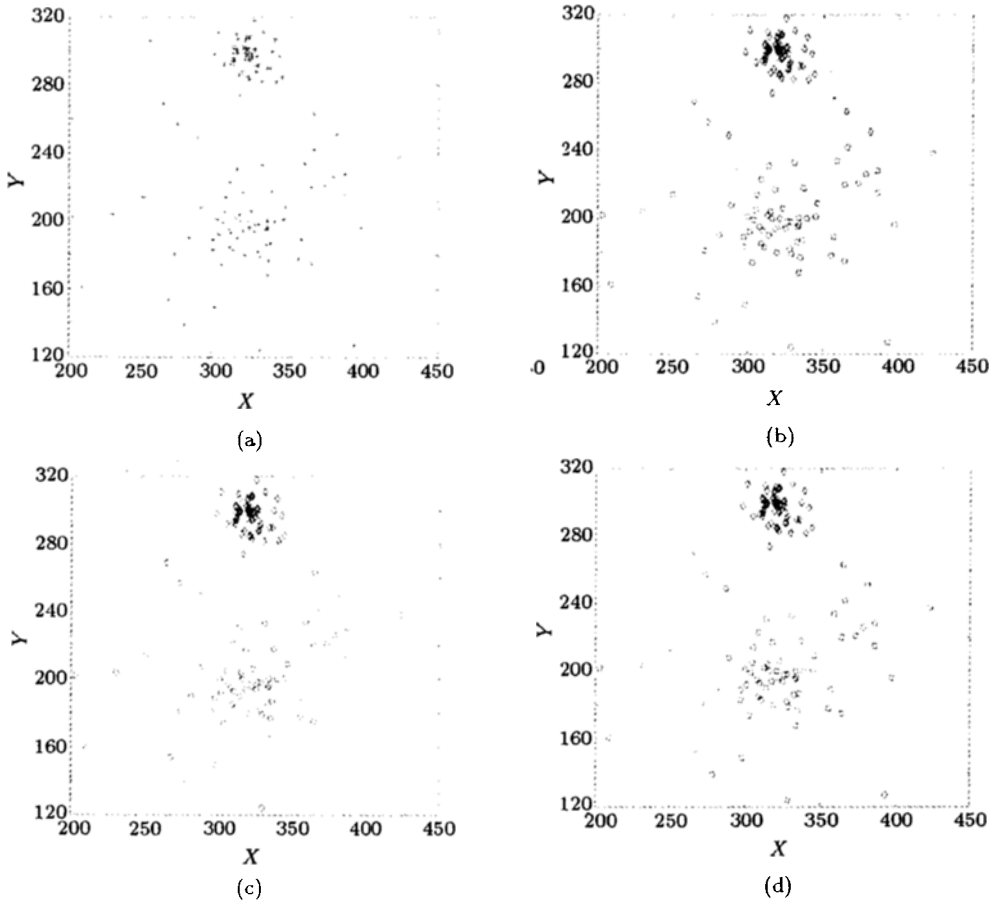

Fig.2. Experiment on the data set including two different Gaussian distributions. (a) Original data. (b) FCM. (c) RL based combination. (d) Linear additive model based combination.

We further evaluate our algorithms on six datasets (Table 1) from the UCI repository of machine learning database[25]. All examples in the dataset have their own classification labels that are not used in clustering process, but can be used to evaluate clustering performance. We give the following definition of error rate as the measure of clustering performance, which is based on the difference between the clustering labels μ_{ij} and classification labels y_{ij}.

$$\text{error rate} = \frac{1}{2}\sum_{i=1}^{C}\sum_{j=1}^{N}\|\mu_{ij} - y_{ij}\|/N \tag{18}$$

where μ is the crisp membership obtained by the defuzzification of the clustering result, and y is the standard classification label. If the error rate of clustering algorithm A is less than that of algorithm B, in most cases we can say that the performance of algorithm A is superior to that of algorithm

B. The error rates of FCM, RL-based combination, and linear additive model based combination are given in Table 2.

Table 1. Description of Datasets

Domain	Data size	Classes	Attributes
Balance scale	625	3	4
Wisconsin breast cancer	699	2	9
Glass	214	6	9
Image segmentation	210	7	19
Pima diabetes	768	2	8
Wine	178	3	13

Table 2. Error Rates of FCM, RL-Based Combination, Linear Additive Model Based Combination, and Nearest Neighbor Classifier

Domain	FCM	RL	Linear additive model	NN classifier
Balance scale	0.3648	0.2288	0.2294	0.2080
Wisconsin breast cancer	0.0443	0.0343	0.0343	0.0572
Glass	0.5514	0.5327	0.5249	0.3410
Image segmentation	0.3524	0.2738	0.2574	0.1381
Pima diabetes	0.2956	0.3021	0.3206	0.3010
Wine	0.0787	0.0562	0.0562	0.0510

We also give the error rates of the nearest neighbor (NN) classifier in Table 2. As we know, the NN classifier is not a clustering algorithm, but its classification performance can reflect the local characteristics of the data set. If the error rate of NN classifier is low for a specific data set, the patterns in this data set belonging to the same class are always closer to each other than the patterns belonging to different classes. In other words, the local criterion based clustering is suitable to this data set, and the performance of the clustering combination algorithm will be better than that of FCM. If the error rate of the NN classifier is high, clustering combination will not make notable improvement on FCM. From Table 2, we find that in all cases the performances of clustering combination algorithms are superior to FCM except for the Pima diabetes and Glass data sets, and the improvements brought by combination algorithms range from 23% to 31%. Furthermore, our results also outperform the results obtained by many other clustering approaches, for example the best clustering performance of the breast cancer data set by other algorithms is about 0.0443[4], but our algorithms can achieve 0.0343. For the Pima diabetes and Glass datasets, the error rates of the NN classifier are 30.1% and 34.1% respectively, which illustrates the local information is not suitable for classifying these two data sets correctly, so it is reasonable that the clustering combination can not produce better results than FCM in these two particular cases. We also find that the combination performance of RL and the linear additive model have only slight difference, so whichever of them can be selected in practice.

For the sequential clustering combination, the initial clustering result produced by the objective function clustering has decisive influence on the performance of clustering combination. In the above experiments, FCM has been chosen as the objective function clustering algorithm, but for some applications, FCM is not suitable, because the distribution of clusters is not always 'spherical'. Therefore, other objective function algorithms such as the linear, ellipsoidal and shell prototype based algorithms should be selected according to specific situation. Although other graphs can replace the modified Gabriel graph, it is noted that the change of the cluster prototype in objective function clustering has not any relation with the choice of the local criterion based clustering algorithm.

Finally, we discuss the computational cost of these two algorithms. The computational cost of the sequential clustering combination (t_s) includes the cost of the global criterion based clustering (t_g), the local criterion based information extraction (t_l), and the combination algorithm (t_c), i.e.

$$t_s = t_g + t_l + t_c \tag{19}$$

The average iteration numbers of RL and the additive linear model are about 40 and 20 separately for the clustering combination, and each iteration needs $O(N^2)$ time complexity for both of them. As the updating (14) of the linear additive model is simpler than (11), (12) of RL, the overall speed of

the additive linear model is about 5 times as fast as RL. If this fact is considered that the most of elements of W are zero, the computational cost of RL and the additive model will be lower than the above estimation. In experiments t_c is relatively very little, but t_g and t_l take more time, especially t_l. How to decrease t_l is our next research interest in the future. However, compared to the simultaneous combination method[13,14], sequential combination algorithms have low computational cost.

5 Conclusions

Clustering combination offers an effective and flexible approach to improve clustering performance by taking advantage of various clustering algorithms. But clustering combination is also a difficult problem, and we have to watch for the inconsistency between the clustering algorithms which may fail the combination. Therefore, the first task of clustering combination is to choose the clustering methods to be combined. In this paper, we choose the global criterion based clustering and the local criterion based clustering, because they complement each other, and there is not much contradiction between them. Then the next task is to choose an approach to combine the clustering algorithms. In this paper a sequential combination approach is proposed, in which an initial clustering result is first produced by the fuzzy objective function clustering algorithm, then by RL or the linear additive model the initial result is improved with the local information that is derived from the graph theoretic clustering algorithm. Compared with the simultaneous combination method[13,14], the sequential combination method is flexible, computing is fast, and algorithm realization by computer is also straightforward. The experiments illustrate that our approach is powerful and effective.

In this study, both the objective function clustering and the graph theoretic clustering in the combination use the identical features of the data set. But in some applications, we can classify the features into two classes, one for the local criterion based clustering, and the other for the global criterion based clustering. This may produce a better result than using identical features.

References

[1] Jain A K, Dubes R C. Algorithms for Clustering Data. Prentice-Hall, Englewood Cliffs, NJ, 1988.
[2] Mirkin B. Mathematical Classification and Clustering. Kluwer Academic Publishers, Dordrecht, The Netherland, 1996.
[3] Murtagh F. A survey of recent advances in hierarchical clustering algorithms. *The Computer Journal,* 1983, 26(4): 354–359.
[4] Frattale F M, Rizzi A, Panella M, Martinelli G. Scale-based approach to hierarchical fuzzy clustering. *Signal Processing,* 2000, 80: 1001–1016.
[5] Urquhart R. Graph theoretical clustering based on limited neighborhood sets. *Pattern Recognition,* 1982, 15(3): 173–187.
[6] Zahn C T. Graph-theoretic methods for detecting and describing gestalt clusters. *IEEE Trans. Computer,* 1971, 20: 68–86.
[7] Delignon Y, Marzouki A, Pieczynski W. Estimation of generalized mixtures and its application in image segmentation. *IEEE Trans. Image Processing,* 1997, 6(10): 1364–1374.
[8] Fraley C, Raftery A E. How many clusters? which clustering method? answers via model-based cluster analysis. *The Computer Journal,* 1998, 41: 578–588.
[9] Bezdek J C. Pattern Recognition With Fuzzy Objective Function Algorithms. Plenum Press, New York, 1981.
[10] Kittler J, Hatef M, Duin R, Matas J. On combining classifiers. *IEEE Trans. Pattern Analysis and Machine Intelligence,* 1998, 20(3): 226–239.
[11] Lam L, Suen C Y. Optimal combinations of pattern classifiers. *Pattern Recognition,* 1995, 16(9): 945–954.
[12] Bay S D. Nearest neighbor classification from multiple feature subsets. *Intelligent Data Analysis,* 1999, 3: 191–209.
[13] Qian Y T, Zhao R. Robust clustering based on global data distribution and local connectivity matrix. In *1997 IEEE Int. Conf. on Intelligence Processing Systems,* Beijing, China, October, 1997, pp.1629–1633.
[14] Qian Y T, Xie W, Zhao R. Robust clustering: an approach based on graph theory and objective function. *Academia Electronics,* 1998, 26(2): 91–94.
[15] Hansen P, Mladenovic N. J-means: A new local search heuristic for minimum sum of squares clustering. *Pattern Recognition,* 2001, 34: 405–413.
[16] Jaromczyk J W, Toussaint G T. Relative neighborhood graphs and their relatives. In *Proc. IEEE,* 1992, 80(9): 1502–1517.

[17] Hummel R A, Zucker S W. On the foundations of relaxation labeling processes. *IEEE Trans. Pattern Anal. Machine Intell.*, 1983, 5: 267–287.
[18] Haralick R. An interpretation for probabilistic relaxation. *Comput. Vis., Graph., Image Processing,* 1983, 22: 378–385.
[19] Peleg S. A new Probabilistic relaxation scheme. *IEEE Trans. Pattern Anal. Machine Intell.*, 1980, 2: 362–369.
[20] Geman S, Geman D. Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of image. *IEEE Trans. Pattern Anal. Machine Intell.*, 1984, 6(6): 721–741.
[21] Banerjee A, Burlina P, Alajaji F. Image segmentation and labeling using the Polya Urn model. *IEEE Trans. Image Processing,* 1999, 8(9): 1243–1253.
[22] Zucker S W. Relaxation processes for scene labeling convergence, speed and stability. *IEEE trans. Syst., Man, Cybern.*, 1978, 8: 41–48.
[23] Kosko B. Neural Networks and Fuzzy Systems: A Dynamical System Approach to Machine Intelligence. Prentice-Hall, Englewood Cliffs, NJ, 1992.
[24] Williams G. Linear Algebra with Applications. Allyn and Bacon, Inc. Massachusetts, 1984.
[25] Murphy P M, Aha D W. UCI Repository of Machine Learning Databases. http://www.ics.uci.edu/ mlearn/MLRepository.html, Irvine, CA: University of California, 1994.

QIAN Yuntao received the B.E. and M.E. degrees in automatic control from Xi'an Jiaotong University in 1989 and 1992 respectively, and his Ph.D. degree in signal processing from Xidian University in 1996. From 1996 to 1998, he was a postdoctoral fellow in Northwestern Polytechnical University. Since 1998, he has been an associate professor in Department of Computer Science, Zhejiang University. From 1999 to 2001, he was a visiting scholar to the Centre of Pattern Recognition and Machine Intelligence, Concordia University, Canada, and also to Department of Computer Science, Hong Kong Baptist University. He has published more than 20 technical papers in academic journals and conference proceedings. His present research interests include data clustering analysis, pattern recognition, image processing, wavelet theory, and neural networks.

Ching Y. Suen received his M.S. degree in engineering from the University of Hong Kong, followed by the Ph.D. degree from the University of British Columbia, Canada. In 1972, he joined the Department of Computer Science, Concordia University, Canada, and became a professor in 1979. He is the director of the Centre of Pattern Recognition and Machine Intelligence. He is the author/editor of 11 books and more than 260 papers on subjects ranging from computer vision and handwriting recognition to expert system and computational linguistics. He is the founder of a journal and an associate editor of several journals related to pattern recognition. He is a fellow of IEEE, IAPR, and the Academy of Sciences of the Royal Society of Canada.

TANG Yuanyan received his B.S. degree in electrical and computer engineering from Chongqing University, M.Eng. in electrical engineering from the Beijing University of Posts and Telecommunication, and Ph.D. in computer science from Concordia University, Canada. He is presently a professor in the Department of Computer Science, Hong Kong Baptist University. He has published more than 160 technical papers and is the author/co-author of 15 books. He is an associate editor of the International Journal of Pattern Recognition and Artificial Intelligence. His current research interests include wavelet theory and applications, pattern recognition, document processing, and artificial intelligence.