A Novel Similarity-Based Fuzzy Clustering Algorithm by Integrating PCM and Mountain Method

Vincent S. Tseng, *Member, IEEE*, and Ching-Pin Kao

***Abstract*—The fuzzy c-means (FCM) and possibilistic c-means (PCM) algorithms have been utilized in a wide variety of fields and applications. Although many methods are derived from the FCM and PCM for clustering various types of spatial data, relational clustering has received much less attention. Most fuzzy clustering methods can only process the spatial data (e.g., in Euclidean space) instead of the nonspatial data (e.g., where the Pearson's correlation coefficient is used as similarity measure). In this paper, we propose a novel clustering method, *similarity-based PCM* (SPCM), which is fitted for clustering nonspatial data without requesting users to specify the cluster number. The main idea behind the SPCM is to extend the PCM for similarity-based clustering applications by integration with the *mountain method*. The SPCM has the merit that it can automatically generate clustering results without requesting users to specify the cluster number. Through performance evaluation on real and synthetic data sets, the SPCM method is shown to perform excellently for similarity-based clustering in clustering quality, even in a noisy environment with outliers. This complements the deficiency of other fuzzy clustering methods when applied to similarity-based clustering applications.**

***Index Terms*—Fuzzy clustering, mountain method, nonspatial measure, possibilistic C-means, relational clustering, similarity-based clustering.**

I. INTRODUCTION

THE objective of cluster analysis is to partition a given set of objects into homogeneous groups based on given features, such that objects within a group are more similar to each other and more different from objects in other groups [5]. There exist two categories for clustering tasks: hard and soft clustering. In hard clustering, each data object is assigned to exactly one cluster, while in soft clustering, it is more desirable to let a data object be assigned to several clusters partially. Hence, the soft clustering is also called fuzzy clustering.

The most popular methods in fuzzy clustering are the fuzzy c-means (FCM) proposed by Bezdek *et al.* [3] and the possibilistic c-means (PCM) proposed by Krishnapuram *et al.* [17], [18]. The FCM and PCM algorithms have been utilized in a wide variety of engineering and scientific disciplines such as data mining [7], [20], pattern recognition [4], [6], image processing [1], [4], [6], [19], and bioinformatics [2], [11], [16]. Although many methods are derived from the FCM and PCM for clustering various types of spatial data, relational (similarity-based or dissimilarity-based) clustering has received much less attention. Most fuzzy clustering methods can only process the spatial data (e.g., in Euclidean space) and cannot process nonspatial data (e.g., where the Pearson's correlation coefficient is used as similarity measure). The FANNY [15], Relational fuzzy c-means (RFCM) [14], non-Euclidean RFCM (NERFCM) [13], fuzzy relational data clustering (FRC) [9], robust RFCM (R-RFCM) [9], robust NERFCM (R-NERFCM) [10], and robust-FRC (R-FRC) [9] are several relational fuzzy clustering algorithms. The original RFCM algorithm had a serious defect in that it usually failed for non-Euclidean data. The NERFCM algorithm has the advantage that it overcomes this problem. Nevertheless, the NERFCM and FRC algorithms are severely sensitive to outliers (or noises). Although the R-NERFCM and R-FRC algorithms are robust in a noisy environment with outliers, they still request the specification for the number of clusters to be generated and require good initialization. In addition, Dave and Krishnapuram [8] analyze several popular robust clustering approaches with statistical methods and provide a good overview of research related to robust fuzzy clustering methods.

In [21], Wu *et al.* proposed the self-organizing mountain method approach that incorporates the PCM technique and the idea of mountain method (MM) to perform clustering. The MM, proposed by Yager and Filev [22], searches an approximate center of each cluster by locating the maximum of the density measure. The algorithm identifies one cluster at a time and removes identified clusters by eliminating their effects to the density-like function for all the remaining points. Thus, the MM can serve as a tool to obtain the initial estimation of the cluster centers for the PCM-like methods.

In this paper, we propose a novel robust clustering method, called *similarity-based* PCM (SPCM), which is fitted for clustering nonspatial data. The main idea behind the SPCM is to extend the PCM for similarity-based clustering applications by integration with the mountain method. The input for the SPCM is a matrix that stores the degree of similarity between each pair of objects in the data set, which can be calculated via any similarity or dissimilarity measures (e.g., Euclidean distance, Pearson's correlation coefficient, etc.) depending on the application targets. Then, the SPCM can automatically generate fuzzy clustering results according to the similarity matrix without requesting users to specify the cluster number. Through performance evaluation on both of real and synthetic data sets, the SPCM is shown to perform excellently in clustering quality, even in a noisy environment with outliers. Therefore, SPCM

Manuscript received January 31, 2006; revised May 18, 2006 and July 28, 2006. This work was supported by the National Science Council, Taiwan, R.O.C., under Grants NSC 95-2221-E-006-372 and NSC 95-2422-H-006-001.

The authors are with the Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan 701, Taiwan, R.O.C. (e-mail: tsengsm@mail.ncku.edu.tw; zeno@idb.csie.ncku.edu.tw).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TFUZZ.2006.890673

can serve as a promising method for automatic similarity-based clustering applications. This complements the deficiency of other fuzzy clustering methods when applied to similarity-based clustering applications.

The rest of this paper is organized as follows: The related work on fuzzy clustering, mountain method, and correlation comparison algorithm (CCA) are described in Section II. In Section III, we present the idea and algorithm of the SPCM. In Section IV, the performance evaluation results for the SPCM are described in detail. Conclusions are drawn in Section V.

II. Related Work

A. Fuzzy Clustering Methods

The most well-known fuzzy clustering methods are probably the FCM [3] and PCM [17], [18] algorithms. Inherently, the FCM is a partitioning-based method, while the PCM is a mode-seeking method. The main disadvantage of the FCM is that it requests the specification for the number of clusters to be generated. Moreover, when the data are heavily noisy, poor clustering quality could be resulted because FCM is severely sensitive to outliers. PCM has the advantage that the membership function and the number of clusters are independent, and it is highly robust in a noisy environment with outliers. However, PCM requires good initialization and reliable scale estimate to function effectively [18].

Given n objects that are to be partitioned to c fuzzy clusters, in the PCM clustering, the task is to minimize the objective function

$$J(U,V)=\sum_{i=1}^{c}\sum_{j=1}^{n}u_{\mathrm{ij}}^{m}d_{\mathrm{ij}}^{2}+\sum_{i=1}^{c}\eta_i\sum_{j=1}^{n}(1-u_{\mathrm{ij}})^{m} \quad (1)$$

subjected to

$$u_{\mathrm{ij}}\in[0,1] \text{ for all } i \text{ and } j,$$
$$0<\sum_{j=1}^{n}u_{\mathrm{ij}}<n \text{ for all } i$$
$$\max_i u_{\mathrm{ij}}>0 \text{ for all } j. \quad (2)$$

where $U=[u_{ij}]$ is called fuzzy partition matrix, $V=(v_1,v_2,\ldots,v_i,\ldots,v_c)$ is call prototype (center) matrix, and the parameter $m\in[1,\infty]$ is called the fuzzifier. u_{ij} is the membership degree for object x_j to the ith cluster and d_{ij} is the Euclidean distance between object x_j and prototype v_i.

For the PCM method, the membership function or the degree of typicality of a data object in a cluster is assumed to be absolute. In other words, the membership degrees do not depend on that of the same data object in other clusters. Hence, each cluster is independent of the other clusters in the PCM method.

The membership degree for one object resembles the possibility that it is a member of the corresponding cluster. The membership update equation in the PCM is as follows:

$$u_{\mathrm{ik}}=\frac{1}{1+\left(\frac{d_{\mathrm{ik}}^{2}}{\eta_i}\right)^{\frac{1}{m-1}}}. \quad (3)$$

```
SPCM Method
Input: similarity matrix $S$, fuzzifier $m$, cut point $p$,
       minimal enhancement $\varepsilon$,
       minimal decreasing ratio $\delta$
Output: fuzzy partition matrix $U$

1) Estimate $r$ by CCA and obtain the mountain value for
   each object by mountain function $M^1$ (4).
2) Set cluster counter $c = 1$, $M^{1*}$ = Max(value of $M^1$).
3) Set iteration counter $l = 1$; initialize the possibilistic
   $u_c^{(0)}$; estimate $\eta_c$ using (6).
4) Choose the object with $M^{c*}$ as the initial center $v_c$.
5) Compute $u_c^{(l)}$ using re-defined membership update
   equation (5).
6) Update the similarity vector $sv_c$ by using (7).
7) Set $l = l + 1$.
8) If $\| u_c^{(l-1)} - u_c^{(l)} \| \geq \varepsilon$ then go to Step 5.
9) Set $c = c + 1$.
10) Update the mountain values using the mountain-
    elimination function $M^c$ (8).
11) Set $M^{c*}$ = Max(value of $M^c$).
12) If $M^{c*} / M^{1*} < \delta$ then terminate, else go to Step 3.
```

Fig. 1. Algorithm for SPCM method.

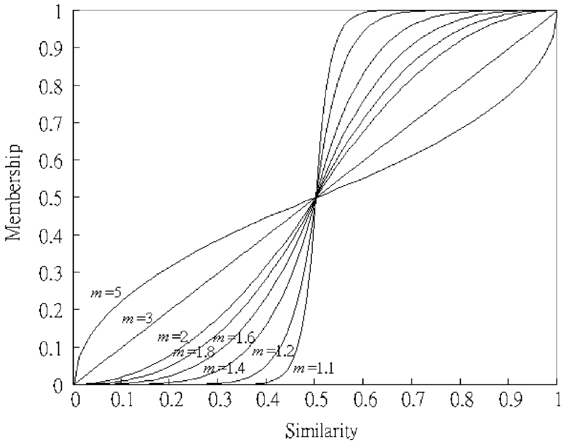

Fig. 2. SPCM membership function with $\eta_c = 0.5$ for various values of the fuzzifier parameter m.

The parameter η_i determines the distance at which the membership degree equals 0.5. η_i can be determined as follows [17], [18]:

$$\eta_i=K\frac{\sum_{k=1}^{N}u_{\mathrm{ik}}^{m}d_{\mathrm{ik}}^{2}}{\sum_{k=1}^{N}u_{\mathrm{ik}}^{m}}. \quad (4)$$

Usually K is set as one. The prototype (center) update equation for inner product distance measure is as follows:

$$v_i=\frac{\sum_{k=1}^{n}u_{\mathrm{ik}}^{m}x_k}{\sum_{k=1}^{n}u_{\mathrm{ik}}^{m}},\quad \forall i. \quad (5)$$

The PCM algorithm is listed as follows.

Step 1) Fix the number of clusters c; fix $m, 1<m<\infty$; set iteration counter $l=1$; initialize the possibilistic c-partition $U^{(0)}$; estimate η_i using (4).

TABLE I
COUNTRIES DATA (CD) DISSIMILARITY MATRIX

Country	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12
C1: Belgium	0.00	5.58	7.00	7.08	4.83	2.17	6.42	3.42	2.50	6.08	5.25	4.75
C2: Brazil	5.58	0.00	6.50	7.00	5.08	5.75	5.00	5.50	4.92	6.67	6.83	3.00
C3:China	7.00	6.50	0.00	3.83	8.17	6.67	5.58	6.42	6.25	4.25	4.50	6.08
C4: Cuba	7.08	7.00	3.83	0.00	5.83	6.92	6.00	6.42	7.33	2.67	3.75	6.67
C5: Egypt	4.83	5.08	8.17	5.83	0.00	4.92	4.67	5.00	4.50	6.00	5.75	5.00
C6: France	2.17	5.75	6.67	6.92	4.92	0.00	6.42	3.92	2.25	6.17	5.42	5.58
C7: India	6.42	5.00	5.58	6.00	4.67	6.42	0.00	6.17	6.33	6.17	6.08	4.83
C8: Israel	3.42	5.50	6.42	6.42	5.00	3.92	6.17	0.00	2.75	6.92	5.83	6.17
C9: USA	2.50	4.92	6.25	7.33	4.50	2.25	6.33	2.75	0.00	6.17	6.67	5.67
C10: USSR	6.08	6.67	4.25	2.67	6.00	6.17	6.17	6.92	6.17	0.00	3.67	6.50
C11: Yugoslavia	5.25	6.83	4.50	3.75	5.75	5.42	6.08	5.83	6.67	3.67	0.00	6.92
C12: Zaire	4.75	3.00	6.08	6.67	5.00	5.58	4.83	6.17	5.67	6.50	6.92	0.00

TABLE II
FAT-OILS DATA (FAT) SIMILARITY MATRIX

Type of Fat	F1	F2	F3	F4	F5	F6	F7	F8
F1: Linseed Oil	-	4.98	3.66	3.77	3.84	3.24	0.86	1.22
F2: Perilla Oil	4.98	-	5.70	5.88	4.70	5.30	2.78	3.08
F3: Cotton-Seed	3.66	5.70	-	7.00	6.25	6.68	4.11	4.44
F4: Sesame Oil	3.77	5.88	7.00	-	5.90	6.37	3.61	3.97
F5: Camelia	3.84	4.70	6.25	5.90	-	6.27	3.48	3.89
F6: Olive Oil	3.24	5.30	6.68	6.37	6.27	-	4.28	4.68
F7: Beef-tallow	0.86	2.78	4.11	3.61	3.48	4.28	-	6.74
F8: Lard	1.22	3.08	4.44	3.97	3.89	4.68	6.74	-

Step 2) Update the prototypes $V^{(l+1)}$ using $U^{(l)}$ and (5).
Step 3) Compute $U^{(l+1)}$ using $V^{(l+1)}$ and (3).
Step 4) $1 = l + 1$.
Step 5) If $|| U^{(l-1)} - U^{(l)}|| < \varepsilon$ then terminate, else go to Step 2).

B. Mountain Method (MM)

The MM, proposed by Yager and Filev [22], searches an approximate center of each cluster by locating the maximum of the density measure. The algorithm identifies one cluster at a time and removes identified clusters by eliminating their effects to the density-like function for all the remaining points.

Given n data points $\{x_1, \ldots, x_n\}$ in the s-dimensional space R^s. Let x_{kj} denote the jth coordinate of the kth point, where $k = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, s$. The MM discretizes the feature space to form an s-dimensional grid in the hypercube $I_1 \times \cdots \times I_s$ with nodes $N_{(i1,\ldots,is)}$, where $i_1, \ldots, i_s$ take values from the sets $[1, \ldots, r_1], \ldots, [1, \ldots, r_s]$. The intervals I_j are defined by the ranges of the coordinates x_{kj}, and each of the intervals I_j is discretized into r_j equidistant points.

The mountain function at the vertex point N_i is defined as follows:

$$M^1(Ni) = \sum_{k=1}^{n} e^{-\alpha d(xk,Ni)} \tag{6}$$

where $d\,(x_k, N_i)$ denotes the distance from the data point x_k to the grid nodes N_i and α is a positive constant. The node with maximum of the mountain function is selected as the first cluster center.

To find the next cluster center, the mountain function is revised as follows:

$$\hat{M}^k(Ni) = \hat{M}^{k-1}(Ni) - M^*_{k-1} \sum_{k=1}^{n} e^{-\beta d(N^*_{k-1},Ni)} \tag{7}$$

where M^*_{k-1} denotes the maximal value of the mountain function M_{k-1} and β is a positive constant. The revised mountain function eliminates the effects of the cluster center just identified. The searching process will be terminated when

$$\frac{M^*_1}{M^*_{k-1}} < \delta \tag{8}$$

where δ is given parameter. The mountain method algorithm is listed as follows.

Step 1) Calculate the intervals $I_j, j = (1, s)$.
Step 2) Quantize the intervals and form the grid.
Step 3) Calculate the mountain values according to (6).
Step 4) Find the cluster centers estimates and modify the mountain function according to (7) until stopping rule (8) is satisfied.

C. Correlation Comparison Algorithm (CCA)

Let density function (mountain function) J_s be the total similarity of the data point x_k to all data points with

$$Js(xk) = \sum_{j=1}^{n} f(xk)^r \tag{9}$$

TABLE III
MICROCOMPUTER DATA (COMP) SIMILARITY MATRIX

Microcomputer	**M1**	**M2**	**M3**	**M4**	**M5**	**M6**	**M7**	**M8**	**M9**	**M10**	**M11**	**M12**
M1: Apple II	-	5.27	4.40	4.32	3.29	3.42	1.14	3.68	4.31	5.02	4.56	4.09
M2: Atari 800	5.27	-	4.65	4.98	3.54	3.67	1.35	4.35	4.23	4.94	4.48	4.34
M3: Com. VIC 20	4.40	4.65	-	3.64	3.55	3.71	2.36	3.03	4.75	4.04	3.64	4.47
M4: Ex. Sorcerer	4.32	4.98	3.64	-	4.07	4.06	1.24	4.74	4.12	4.83	5.22	4.66
M5: Zenitj H8	3.29	3.54	3.55	4.07	-	5.81	1.18	4.71	2.65	3.35	3.74	5.00
M6: Zenitj H89	3.42	3.67	3.71	4.06	5.81	-	1.30	4.83	2.77	3.48	3.87	5.13
M7: Hp-85	1.14	1.35	2.36	1.24	1.18	1.30	-	0.61	1.90	1.20	1.23	2.12
M8: Horizon	3.68	4.35	3.03	4.74	4.71	4.83	0.61	-	3.06	3.77	4.16	4.02
M9: O.S. Challenger	4.31	4.23	4.75	4.12	2.65	2.77	1.90	3.06	-	5.23	4.77	3.44
M10: O.S.II Series	5.02	4.94	4.04	4.83	3.35	3.48	1.20	3.77	5.23	-	5.48	4.15
M11: TRS-80 I	4.56	4.48	3.64	5.22	3.74	3.87	1.23	4.16	4.77	5.48	-	4.61
M12: TRS-80 III	4.09	4.34	4.47	4.66	5.00	5.13	2.12	4.02	3.44	4.15	4.61	-

where $f\ (\cdot)$ is a similarity relation function. Yang and Wu [23] propose a method, CCA, to select r using a correlation comparison procedure with "$r = 5, r = 10$," "$r = 10, r = 15$," "$r = 15, r = 20$,"…. Under this method, we can easily find a good estimate of r via the density shape estimation concept, and the underestimate (small r) and overestimate (large r) are undesirable. In order to execute the correlation comparison procedure as a computer program (9) is rewritten as

$$Js(xk)rm = \sum_{j=1}^{n} f(xk)^{rm} \tag{10}$$

where $r_m = 5\,m, m = 1, 2, 3, \ldots$. The CCA algorithm is listed as follows.

Step 1) Set $m = 1$ and $\varepsilon = 0.97$.

Step 2) Calculate the correlation of the values of $J_s\ (x_k)_{rm}$ and $J_s\ (x_k)_{r(m+1)}$.

Step 3) $m = m + 1$.

Step 4) If the correlation $\geqslant \varepsilon$, then choose r_{m-1} to be the estimate of r, else go to Step 2).

III. SIMILARITY-BASED PCM (SPCM)

In this section, we present the idea and algorithm for the proposed SPCM clustering method. Before the SPCM can be applied, we have to generate a similarity matrix S based on the given data set. The matrix S stores the degree of similarity between each pair of objects in the data set, with the degrees in the range of [0, 1]. The similarity values can be obtained by using any similarity or dissimilarity measures (e.g., Euclidean distance, Pearson's correlation coefficient, etc.) depending on the applications. Then, the SPCM can cluster the data objects according to the similarity matrix S without requesting users to specify the cluster number. The algorithm for the SPCM method is as shown in Fig. 1. In the following, we will describe the algorithm in details.

A. The Redefined Mountain Function in Mountain Method

The PCM has the advantage that the membership function and the number of clusters are independent, and it is highly robust in a noisy environment with outliers. However, the PCM requires good initialization and reliable scale estimate to function effectively [18]. The MM [22] searches an approximate center of each cluster by locating the maximum of the density measure. The MM identifies one cluster at a time and removes identified clusters by eliminating their effects to the density-like function for all the remaining points. Thus, the MM can serve as a good tool for obtaining the initial estimation of the cluster centers for the PCM.

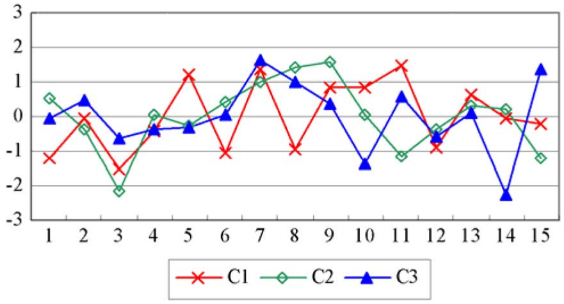

Fig. 3. Profiles of the seeds for generating GENE.

However, the SPCM does not need to conduct discretization in the feature space as the original MM does. Instead, the redefined mountain function is measured at each object x_k, where $k = 1, 2, \ldots, n$. The mountain function is redefined as follows:

$$M^1(xi) = \sum_{j=1}^{n} \exp(-(\frac{1 - S(i,j)}{(1 - \overline{S})})^r), r > 0 \tag{11}$$

where s_{ki} is the similarity between object x_k and object x_i and $\overline{s}$ denotes the sample mean of the entries of matrix S. Statistically, most links in same cluster should have their similarity greater than $\overline{s}$, except those connecting clusters and outliers (or noises). We noted that using the power parameter r in the monotone increasing function (11) is important because it can take over the effect of the normalized value 1-$\overline{s}$. Therefore, the CCA [23] is used to choose a better value of r. The CCA can estimate the approximate density shape (parameter r) of the data set using the mountain function. After processing the mountain function, the object with the maximal value of the mountain function is chosen as the initial center v_c of the cth cluster and the PCM method is applied to obtain the membership degree u_{ck} for each object x_k.

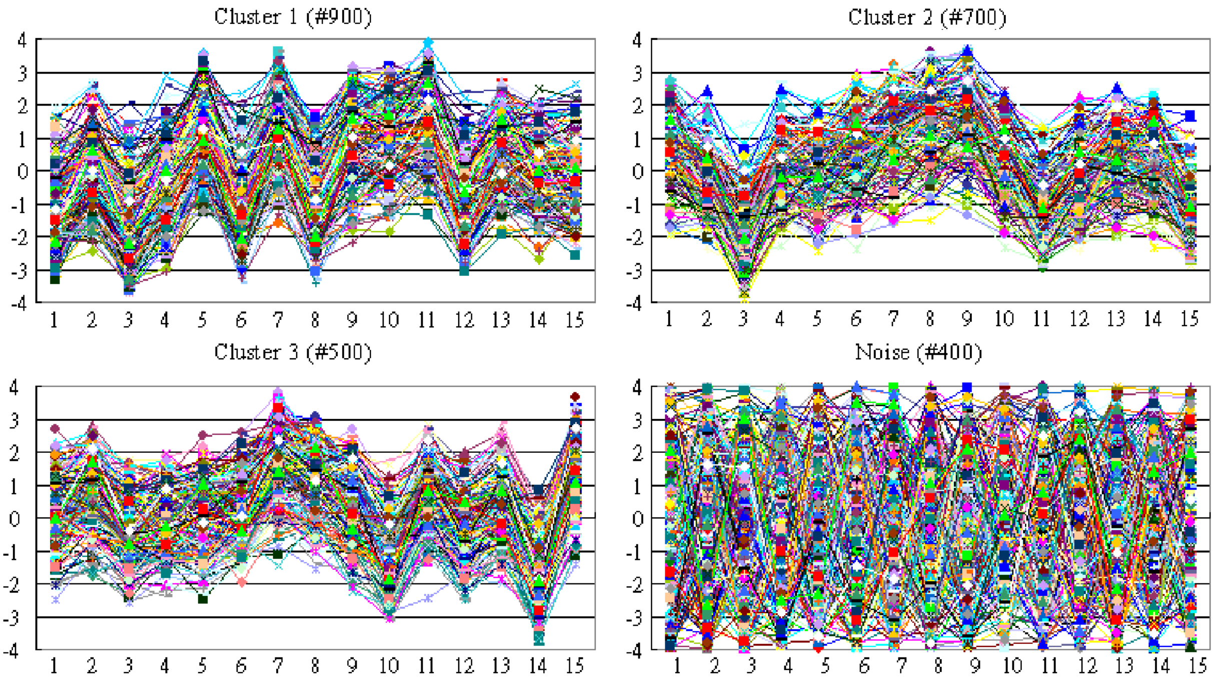

Fig. 4. Pattern profiles of each cluster in GENE.

TABLE IV
RESULTS FOR COUNTRIES DATA (CD)

Country	Mountain Value (10^{-2})				Membership			Belong to Cluster		
	M1	M2	M3	M4	C1	C2	C3	SPCM	FRC (3)	R-FRC (3)
France	4.876	0	0	0	0.971	0.093	0.246	1	1	1
USA	4.641	0	0	0	0.925	0.065	0.310	1	1	1
Belgium	4.323	0	0	0	0.885	0.085	0.364	1	1	1
Israel	1.408	0	0	0	0.640	0.093	0.204	1	1	1
Cuba	1.407	1.407	0	0	0.037	0.973	0.072	2	0	0
USSR	1.360	1.360	0	0	0.123	0.938	0.099	2	0	0
Yugoslavia	0.455	0.455	0	0	0.174	0.669	0.065	2	0	0
China	0.289	0.289	0	0	0.076	0.617	0.142	2	0	0
Zaire	0.715	0.715	0.715	0	0.218	0.095	0.960	3	2	2
Brazil	0.700	0.700	0.700	0	0.221	0.065	0.931	3	2	2
India	0.110	0.110	0.110	0.110	0.096	0.170	0.415	-	2	2
Egypt	0.208	0.208	0.208	0.208	0.350	0.199	0.382	-	2	3

B. The Redefined Membership Update Equation in PCM

The membership update equation of the original PCM cannot process the similarity-based data, and therefore has be redefined. The membership update equation is redefined as follows:

$$u_{ck} = \begin{cases} 1 - \dfrac{\left(\frac{1-svck}{1-\eta c}\right)^{\frac{2}{m-1}}}{2}, & \text{if } svck \geqslant \eta c \\ \dfrac{\left(\frac{svck}{\eta c}\right)^{\frac{2}{m-1}}}{2}, & \text{otherwise} \end{cases} \tag{12}$$

where u_{ck} is the membership degree for object x_k to the cth cluster and is subjected to (2) in the original PCM. sv_{ck} denotes the similarity between object x_k and the center v_c and is indicated in Section III-C. The parameter $m \in [1, \infty]$ is called the fuzzifier just like that in the original PCM. The parameter η_c determines the similarity at which the membership degree equals to 0.5 and is redefined by

$$\eta_c = \sqrt{K \frac{\sum_{k=1}^{N} u_{\mathrm{ck}}^m sv_{\mathrm{ck}}^2}{\sum_{k=1}^{N} u_{\mathrm{ck}}^m}}. \tag{13}$$

Usually K is set as one just like that in the original PCM. For the SPCM method, the membership degrees do not depend on that of the same data object in other clusters. Hence, each cluster is independent of the other clusters in the SPCM method. Fig. 2 shows a plot of the SPCM membership degrees resulting from (12) as a function with $\eta_c = 0.5$. A value of two for m seems to give good results in practice.

TABLE V
RESULTS FOR FAT-OILS DATA (FAT)

Type of Fat	Mountain Value (10^{-1})			Membership		Belong to Cluster			
	M1	M2	M3	C1	C2	SPCM	FRC (2)	FRC (3)	Desc. [12]
Cotton-Seed	2.065	0	0	1.000	0.395	1	0	1	1
Sesame Oil	1.692	0	0	0.959	0.310	1	0	1	1
Olive Oil	0.699	0	0	0.877	0.434	1	0	1	1
Camelia	0.211	0	0	0.714	0.294	1	0	1	1
Beef-tallow	0.576	0.576	0	0.246	0.990	2	1	2	2
Lard	0.576	0.576	0	0.288	0.990	2	1	2	2
Perilla Oil	0.037	0.037	0.037	0.480	0.186	-	0	1	1
Linseed Oil	0.001	0.001	0.001	0.198	0.024	-	0	0	1

C. The Redefined Center Update Equation in PCM

Most fuzzy clustering methods can only process the spatial data (e.g., in Euclidean space) instead of the nonspatial data (e.g., by using Pearson's correlation coefficient as similarity measure). We call the clustering methods for nonspatial data similarity-based. Unlike the PCM, SPCM is a similarity-based clustering method for nonspatial data. For the first PCM iteration of each cluster, center v_c is chosen by the redefined mountain function (11). Since the center v_c cannot be updated directly, the SPCM recomputes the similarity vector $sv_c = \{sv_{c1}, sv_{c2}, \ldots, sv_{\text{cn}}\}$ to simulate the recentering process of PCM, where sv_{ck} denotes the similarity between object x_k and the virtual center v_c of the cth cluster. The similarity vector updating equation is as follows:

$$sv_{\text{ck}} = \frac{\sum_{i=1}^{n} w_{ci} sv_{\text{ki}}}{\sum_{i=1}^{n} w_{ci}} \qquad \forall k$$

where

$$w_{ci} = \begin{cases} u_{ci}^{m} M^c(xi)^r, & \text{if } M^c(xi) \geqslant M^c_{\text{Med}} \quad \text{and} \quad u_{ci} \geqslant p \\ 0, & \text{otherwise} \end{cases} \tag{14}$$

where m denotes the fuzzifier, u_{ci} denotes the membership degree for object x_i to the cth cluster, and $M^c_{\text{Med}} =$ median of $\{M^c(xi), i = 1, \ldots, n\}$. The parameter p $(0 < p \leqslant 1)$ controls the cut point of membership degrees. The variable w_{ck} can diminish the effect of the objects in sparse regions. By experimental evaluation, a value of $p = 0.5$ is recommended to give good results.

The PCM iterations will be terminated when $\|u_c^{(l-1)} - u_c^{(l)}\| < \varepsilon$, where $u_c^{(l)} = \{u_{c1}^{(l)}, u_{c2}^{(l)}, \ldots, u_{\text{cn}}^{(l)}\}$ denotes the membership vector to the cth cluster in the lth iteration of PCM.

D. The Redefined Mountain-Elimination Function in Mountain Method

After a fuzzy cluster is formed, we may continue to search the next cluster. The main idea here is to eliminate the effects of the cluster just identified. The redefined mountain-elimination function is shown as follows:

$$M^{c+1}(xk) = \begin{cases} 0, & \text{if } uck > p \\ M^c(xk), & \text{otherwise.} \end{cases} \tag{15}$$

where M^c denotes the mountain function for the cth cluster. The parameter p $(0 < p \leqslant 1)$ controls the cut point of membership degrees, and the effects of the objects whose membership degree is greater than p will be eliminated. By experimental evaluation, a value of $p = 0.5$ is recommended to give good results.

E. The Termination Condition

In SPCM, clusters are identified one by one until the following termination condition is achieved:

$$\frac{M^{c*}}{M^{1*}} < \delta \tag{16}$$

where M^{c*} denotes the maximal value of the mountain function M^c. When all clusters have been identified, the maximal value of the redefined mountain-elimination function will descend substantially. Therefore, the process of searching the clusters will be terminated. By experimental evaluation, a value of $\delta = 1/r$ is recommended to give good results.

IV. EXPERIMENTAL EVALUATION

In order to evaluate the performance and accuracy of the SPCM method, we need some data sets in which the cluster structures are known a priori. Accordingly, we tested the SPCM method on three real data sets and one synthetic data set. All the data sets are nonspatial, so most fuzzy clustering methods are insufficient.

First, three typical data are considered from the literature. We compare the proposed SPCM method with the FRC [9] and R-FRC [9] methods in these data. The first data set is called "Countries Data" (CD) [15], [9]. Dissimilarities between 12 countries are obtained by averaging the results of a survey among political science students. The second data set is called "Fat-Oils Data" (FAT) [12], [9] and the third data set is called "Microcomputer Data" (COMP) [12], [9]. The original dissimilarity and similarity matrices for these data sets are taken from [9] and are shown in Tables I–III. The dissimilarities or similarities are converted into similarities and normalized to [0, 1] for the SPCM application. In CD, the dissimilarities are converted into similarities and normalized by $s_{ij} = (\max\{d_{ij}\} - d_{ij})/\max\{d_{ij}\}$, where d_{ij} denotes the original dissimilarity. In FAT and COMP, the similarities are normalized by $s_{ij} = s'_{ij}/7.5$ and $s_{ij} = s'_{ij}/6.5$, respectively, where s'_{ij} denotes the original similarity.

The synthetic data are correlation-inclined and generated by a data generator. The data generator can produce various kinds of

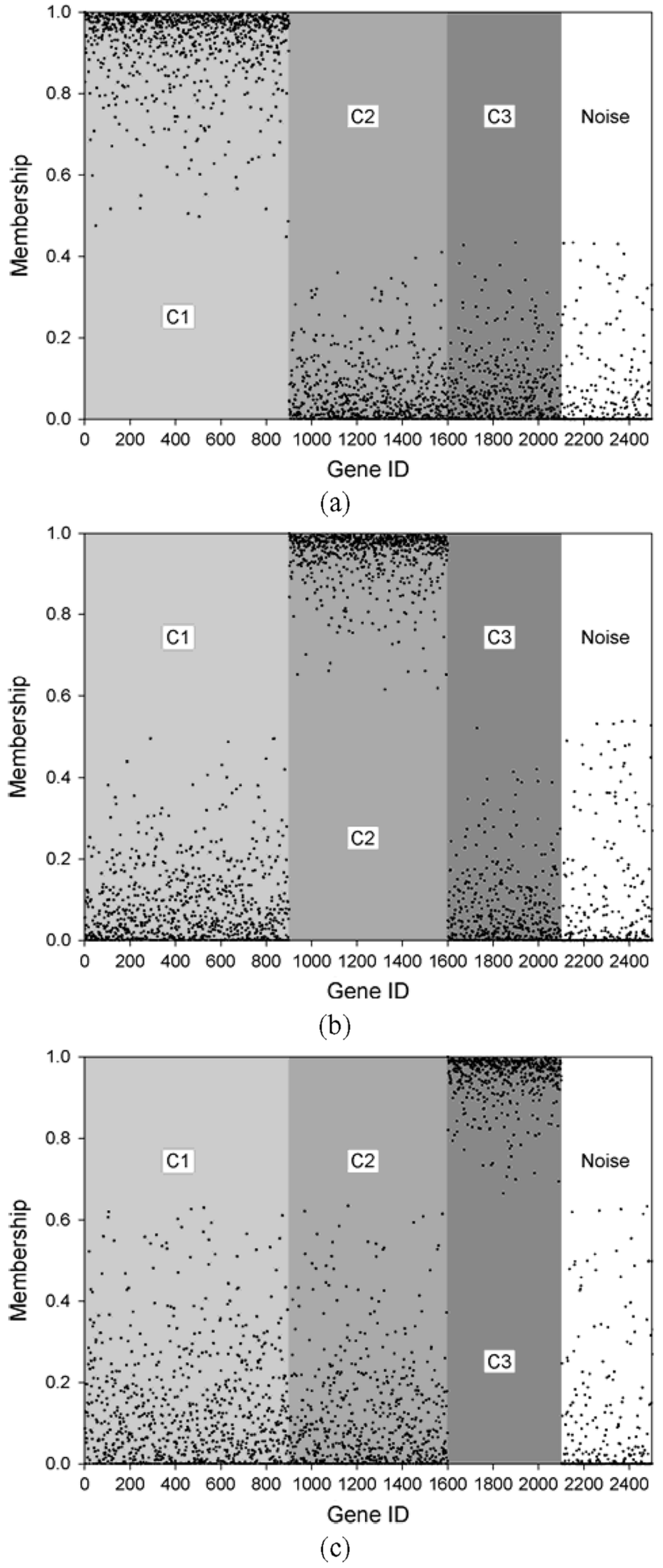

Fig. 5. Scatter plot of the gene ID and the membership degree for gene expression data. (a) Cluster #1, (b) Cluster #2, and (c) Cluster #3.

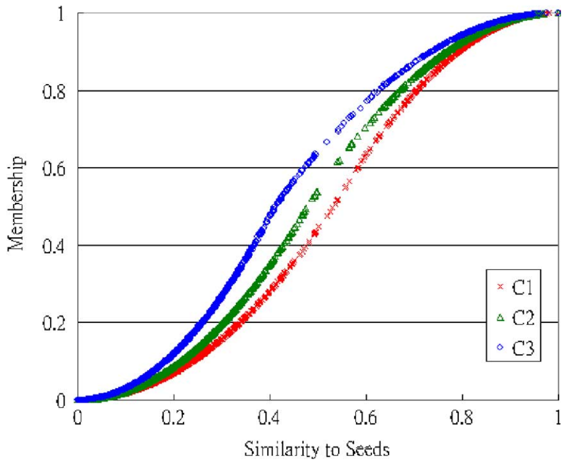

Fig. 6. Scatter plot of the similarity for genes to cluster seeds and the membership degree for GENE.

gene expression data sets with variations in terms of the number of clusters, the number of genes in each cluster, the number of noises (or outliers), etc. First, a number of seed genes are produced by the generator. All the seed genes have the same dimensions (number of conditions) with very low similarity mutually so as to ensure that the generated clusters will not overlap. Secondly, the generator produces clusters one by one. It randomly produces genes in each cluster, and all produced genes are similar to the seed gene of this cluster in the viewpoint of correlation. Thirdly, the noises are produced randomly and added into the data set. It should be noted that the noises may not fall into any cluster. In this way, all genes in the same cluster will have very high similarity and will have high dissimilarity with genes in other clusters.

We first generate the set of seeds with size three. The seeds in the set have 15 dimensions, and their correlation coefficient is less than 0.1. The profiles of the seeds are shown in Fig. 3. Then, we load the seeds into the data generator to generate a synthetic data set, called "Gene Expression Data" (GENE). The GENE contains three gene clusters of sizes 900, 700, and 500, with an additional 400 noises (or outliers). Fig. 4 shows the pattern profiles of each cluster in GENE.

A. *Typical Data Sets*

The parameter r of the mountain function (11) for CD, FAT, and COMP is estimated by CCA [23], which produces the recommended value as 10, 15, and 20, respectively. The parameters m, δ, ε, and p are set as 2, $1/r$, 0.0001, and 0.5, respectively.

The results for CD are shown in Table IV. "Mountain Value" indicates the mountain values of the each mountain function M^c. "Membership" indicates the membership degrees of each object for each cluster. The notation "-" indicates that the object is treated as an outlier by the SPCM. The notation "()" indicates the input cluster number for the FRC or R-FRC methods. It is observed that the results of the SPCM are similar to those of the FRC and R-FRC. The SPCM method produced three clusters with two outliers; United States, France, Belgium, and Israel as first cluster (developed countries); Cuba, USSR, Yugoslavia, and China as second cluster (communist countries); Zaire and Brazil as third cluster (developing countries); and Egypt and India as outliers. In [15] and [9], Egypt is also indicated as a "worst clustered" object. However, further analysis of the partition shows that Egypt and India are unlike any of the three typical clusters. Thus, Egypt and India may be classified as out-

TABLE VI
RESULTS FOR MICROCOMPUTER DATA (COMP)

Microcomputer	Mountain Value (10^{-2})			Membership		Belong to Cluster		
	M1	M2	M3	C1	C2	SPCM	FRC (4)	Desc. [12]
Zenitj H89	1.614	0	0	0.991	0.304	1	1	4
Zenitj H8	1.597	0	0	0.986	0.282	1	1	4
TRS-80 III	0.042	0	0	0.797	0.433	1	1	4
Horizon	0.009	0	0	0.705	0.357	1	1	4
Ex. Sorcerer	0.064	0.064	0	0.456	0.665	2	3	4
TRS-80 I	0.077	0.077	0	0.400	0.898	2	3	1
Apple II	0.266	0.266	0	0.312	0.698	2	3	1
Atari 800	0.083	0.083	0	0.360	0.667	2	3	1
O.S.II Series	0.290	0.290	0	0.323	0.998	2	2	1
O.S. Challenger	0.055	0.055	0	0.204	0.776	2	2	3
Com. VIC 20	0.005	0.005	0.005	0.365	0.390	-	2	3
Hp-85	0.000	0.000	0.000	0.043	0.036	-	0	2

TABLE VII
CONTINGENCY TABLE FOR GENE EXPRESSION DATA (GENE)

		Cluster Structure by SPCM			
		C1	C2	C3	Noise
Original Cluster Structure	C1	896	-	-	4
	C2	-	700	-	-
	C3	-	-	500	-
	Noise	-	5	9	386

liers. Furthermore, there is a better possibility that India belongs to cluster #3 than cluster #1 and cluster #2. Table IV also shows that the redefined mountain-elimination function successfully eliminates the effects of the cluster just identified. These results show that SPCM is a good partitioning method.

The results for FAT are shown in Table V. "Desc." indicates the description of clusters in [12]. It is observed that the results of the SPCM are similar to those of the FRC and the description of clusters. In [12], linseed oil is classified into the cluster that contains cotton-seed oil, sesame oil, etc. Moreover, linseed oil is classified into a cluster of its own by FRC (3). However, when the similarity matrix is examined, the similarity of all oils in that cluster to linseed oil and perilla oil is rather low as compared with the other members of that cluster. Furthermore, there is a better possibility that perilla oil belongs to cluster #1 than cluster #2. Again, Table V shows that the redefined mountain-elimination function successfully eliminates the effects of the cluster just identified.

The results for COMP are shown in Table VI. It is observed that partition obtained by SPCM is slightly different from the FRC and the description of clusters but is also an acceptable partition. The SPCM method produced two clusters with two outliers, while the FRC method produced three main clusters with one outlier. In [9], moreover, it is indicated that perhaps there may not be strong evidence of more than two clusters.

B. Synthetic Data Sets

In this set of experiments, because the synthetic GENE is correlation inclined, we use Pearson's correlation coefficient as similarity measure and ignore negative correlations. The parameter r of the mountain function (11) for GENE is estimated by CCA [23], which produces the recommended value as 15. The parameters m, δ, ε, and p are set as 2, $1/r$, 0.0001, and 0.5, respectively.

The SPCM method produced three clusters containing 896, 705, and 509 genes. Fig. 5 shows the scatter plot of the gene ID and the membership degree of GENE by the SPCM. In the cluster structure of synthetic GENE data, gene #1–900, #901–1600, and #1601–2100 belong to cluster #1, cluster #2, and cluster #3, respectively, and gene #2101–2500 are additional noises (or outliers). Fig. 5 shows that the SPCM detected all clusters excellently.

Fig. 6 shows the scatter plots of the similarity for genes to cluster seeds and the membership degree for GENE. In detail, the contingency table between the original cluster structure and the cluster structure detected by the SPCM is shown in Table VII, where "896" indicates the number of genes on which the original cluster structure and the cluster structure detected by the SPCM both belong to cluster #1; and "4" indicates the number of genes on which the original cluster structure belongs to cluster #1 and the cluster structure detected by the SPCM belongs to noise cluster, and so on. It is observed that the clustering result generated by the SPCM is very close to the original cluster structure. These results show that the SPCM is a good partitioning method.

V. CONCLUSION

In this paper, we propose a novel robust clustering method, called similarity-based PCM, which is fitted for clustering nonspatial data. The main idea behind the SPCM is to extend the PCM [17], [18] for similarity-based clustering applications by integration with the mountain method [22]. The SPCM has the merit that it can automatically generate clustering results without requesting users to specify the cluster number. This complements the deficiency of other existing fuzzy clustering methods when applied to similarity-based clustering applications. For example, FANNY [15], RFCM [14], NERFCM [13], and FRC [9] request the specification of the number of clusters and are severely sensitive to outliers. Although R-RFCM [9], R-NERFCM [10], and R-FRC [9] are robust in noisy environments, they request the specification of the number of clusters and require good initialization. Through performance evaluation on both of real and synthetic data sets, the SPCM is shown to perform excellently in clustering quality with various kinds of similarity measures, even in a noisy environment with outliers. Therefore, the SPCM can serve as a promising method for automatically similarity-based clustering applications.

REFERENCES

[1] M. N. Ahmed, S. M. Yamany, N. Mohamed, A. A. Farag, and T. Moriarty, "A modified fuzzy c-means algorithm for bias field estimation and segmentation of MRI data," *IEEE Trans. Med. Imag.*, vol. 21, no. 3, pp. 193–199, 2002.

[2] S. Banerjee, D. P. Mukherjee, and D. D. Majumdar, "Fuzzy c-means approach to tissue classification in multimodal medical imaging," *Inform. Sci.*, vol. 115, no. 1–4, pp. 261–279, 1999.

[3] J. C. Bezdek*, Pattern Recognition with Fuzzy Objective Function Algorithms*. New York: Plenum, 1981.

[4] J. C. Bezdek, J. Keller, R. Krishnapuram, and M. R. Pal*, Fuzzy Models and Algorithms for Pattern Recognition and Image Processing*. Norwell, MA: Kluwer Academic, 1999.

[5] M.-S. Chen, J. Han, and P. S. Yu, "Data mining: An overview from a Database Perspective," *IEEE Trans. Knowl. Data Eng.*, vol. 8, no. 6, pp. 866–883, 1996.

[6] Z. Chi, H. Yan, and T. Pahm*, Fuzzy Algorithms: With Applications to Image Processing and Pattern Recognition*. , Singapore: World Scientific, 1996.

[7] R. N. Dave, "Robust fuzzy clustering algorithms," in *Proc. IEEE Int. Conf. Fuzzy Syst.*, 1993, vol. 2, pp. 1281–1286.

[8] R. N. Dave and R. Krishnapuram, "Robust clustering mehods: A unified view," *IEEE Trans. Fuzzy Syst.*, vol. 5, no. 2, pp. 270–293, 1997.

[9] R. N. Dave and S. Sen, "Robust fuzzy clustering of relational data," *IEEE Trans. Fuzzy Syst.*, vol. 10, no. 6, pp. 713–727, 2002.

[10] J. W. Davenport and R. J. Hathaway, "Possibilistic c-means clustering for relational data," in *Proc. 1st Int. Conf. Neural, Parallel Sci. Comput.*, 1995, vol. 1, pp. 139–142.

[11] D. Dembele and P. Kastner, "Fuzzy c-means method for clustering microarray data," *Bioinformatics*, vol. 19, pp. 973–980, 2003.

[12] K. C. Gowda and E. Diday, "Symbolic clustering using a new similarity measure," *IEEE Trans. Syst., Man, Cybern.*, vol. 22, no. 2, pp. 368–378, 1992.

[13] R. J. Hathaway and J. C. Bezdek, "NERF c-means: Non-Euclidean relational fuzzy clustering," *Pattern Recognit.*, vol. 27, no. 3, pp. 429–437, 1994.

[14] R. J. Hathaway, J. W. Davenport, and J. C. Bezdek, "Relational duals of the c-Means clustering algorithms," *Pattern Recognit.*, vol. 22, no. 2, pp. 205–212, 1989.

[15] L. Kaufman and P. J. Rousseeuw*, Finding Groups in Data: An Introduction to Cluster Analysis*. New York: Wiley, 1990.

[16] S. Y. Kim, T. M. Choi, and J. S. Bae, "Fuzzy types clustering for microarray data," *Int. J. Comput. Intell.*, vol. 2, no. 1, pp. 12–15, 2005.

[17] R. Krishnapuram and J. M. Keller, "A possibilistic approach to clustering," *IEEE Trans. Fuzzy Syst.*, vol. 1, no. 2, pp. 98–110, May 1993.

[18] R. Krishnapuram and J. M. Keller, "The Possibilistic C-Means Algorithm: insights and recommandations," *IEEE Trans. Fuzzy Syst.*, vol. 4, pp. 385–393, 1996.

[19] D. L. Pham and J. L. Prince, "An adaptive fuzzy C-means algorithm for image segmentation in the presence of intensity inhomogeneities.," *Pattern Recognit. Lett.*, vol. 20, no. 1, pp. 57–68, 1999.

[20] H.-L. Shieh, Y.-K. Yang, and C.-N. Lee, "A robust fuzzy clustering approach and its application to principal component analysis," in *Proc. Int. Conf. Artif. Intell. Applicat.*, 2005, pp. 251–256.

[21] C. W. Wu, J. L. Chen, and J. H. Wang, "Self-organizing mountain method for clustering," in *2001 IEEE Int. Conf. Syst., Man, Cybern.*, 2002, vol. 4, pp. 2434–2438.

[22] R. R. Yager and D. P. Filev, "Approximate clustering via the mountain method," *IEEE Trans. Syst., Man, Cybern.*, vol. 24, no. 8, pp. 1279–1284, 1994.

[23] M. S. Yang and K. L. Wu, "A similarity-based robust clustering method," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 26, no. 4, pp. 434–448, 2004.

Vincent S. Tseng (M'99) received the Ph.D. degree in computer science from National Chiao Tung University, Taiwan, R.O.C., in 1997.

During January 1998 to July 1999, he was a Postdoctoral Research Fellow with the Computer Science Division, University of California, Berkeley. Since August 1999, he has been with the Faculty of the Department of Computer Science and Information Engineering, National Cheng Kung University, Taiwan. Since February 2004, he has also been Director of the Department of Medical Informatics, National Cheng Kung University Hospital, Taiwan. He has a wide variety of research specialties covering data mining, Web technology, and biomedical informatics. He has published a number of research papers in refereed journals and international conferences.

Dr. Tseng is a member of ACM and Phi Tau Phi. He has been a member of the Program Committee for a number of international conferences, including the IEEE International Conference on Computer-Based Medical Systems in 2006, ACM SIGKDD Workshop on Data Mining in Bioinformatics in 2003, Pacific-Asia Symposium on Knowledge Discovery and Data Mining in 2002, and International Conference on Real-Time Technology and Applications in 2001 and 2003.

Ching-Pin Kao received the B.S. degree in computer science and engineering from Yuan Ze University, Taiwan, R.O.C., in 1999 and the M.S. and Ph.D. degrees in computer science and information engineering from National Cheng Kung University, Taiwan, in 2001 and 2006, respectively.

His research interests include data mining, clustering techniques, bioinformatics, and gene expression analysis.