

Robust Interval Competitive Agglomeration Clustering Algorithm with Outliers

Jin-Tsong Jeng, Chen-Chia Chuang, Chih-Cheng Tseng, and Chang-Jung Juan

Abstract

In this study, a novel robust clustering algorithm, robust interval competitive agglomeration (RICA) clustering algorithm, is proposed to overcome the problems of the outliers, the numbers of cluster and the initialization of prototype in the fuzzy C-means (FCM) clustering algorithm for the symbolic interval-values data. In the proposed RICA clustering algorithm, the Euclidean distance measure is considered. Due to the competitive agglomeration is used, the RICA clustering algorithm can be fast converges in a few iterations and to the same optimal partition regardless of its initialization of prototype. Experimentally results show the merits and usefulness of the RICA clustering algorithm for the symbolic interval-values data with outliers.

Keywords: Symbolic interval-values data, Robust Interval Competitive Agglomeration Clustering Algorithm, Interval Fuzzy c-means clustering algorithm and Outliers.

1. Introduction

Clustering, also known as unsupervised classifications, is a process by which a data set is divided into different clusters such that elements of the same cluster are as similar as possible and elements of different clusters are as dissimilar as possible. Most existing clustering algorithms can be classified into the following two categories: hierarchical clustering algorithm and partitional clustering algorithm [1, 2]. The hierarchical clustering procedures provide a nested sequence of partitions with a graphical representation known as the dendrogram. The partitional clustering procedures generate a single partition (as opposed to a nested sequence) of the data in an attempt to recover the natural grouping present in the data. Prototype-based clustering algorithms are the most popular class of the partitional clustering algorithm. In the prototype-based clustering algorithms, each cluster is

represented by a prototype, and the sum of distances from the feature vectors to the prototypes is usually used as the objective function.

In the clustering analysis, the patterns to be grouped are usually represented as a vector of the quantitative or the qualitative measurements where each column represents a variable. Each pattern takes a single value for each variable. However, this model is too restrictive to represent complex data. In order to take into the account variability and/or the uncertainty inherent to the data, variables must assume sets of categories or intervals, possibly even with frequencies or weights. These kinds of data have been mainly studied in Symbolic Data Analysis (SDA). The aim of SDA is to provide the suitable methods (clustering, factorial techniques, decision trees, etc.) for managing aggregated data described by the multi-valued variables, where the cells of the data table contain sets of categories, intervals, or weight (probability) distributions [3, 4].

The SDA provides a number of clustering methods for the symbolic data. These methods differ in the type of the considered symbolic data, in their cluster structures and/or in the considered clustering criteria. With the hierarchical methods, an agglomerative approach has been introduced that forms composite symbolic objects using a join operator whenever mutual pairs of the symbolic objects are selected for agglomeration based on minimum dissimilarity [5] or maximum similarity [6]. In [7], authors defined generalized Minkowski metrics for mixed feature variables and presented dendrograms obtained from the application of standard linkage methods for the data sets containing the numeric and the symbolic feature values. In [8, 9], the divisive and agglomerative algorithms for the symbolic data based on the combined usage of similarity and dissimilarity measures are proposed. These proximity measures are defined on the basis of the position, span and content of symbolic data. In [10], author proposes a divisive clustering method that simultaneously furnishes a hierarchy of the symbolic data set and a monothetic characterization of each cluster in the hierarchy. In [11], a hierarchical clustering algorithm for the symbolic data based on the gravitational approach is also proposed. The agglomerative clustering algorithms based on the similarity [12] and dissimilarity functions [13] are introduced, respectively.

A number of authors have addressed the problem of

Corresponding Author: Chen-Chia Chuang is with the Department of Electrical Engineering, National Ilan University, 1, Sec. 1, Shen-Lung Road, I-Lan, Taiwan 260.
E-mail: ccchuang@niu.edu.tw

non-hierarchical (i.e. partitional) clustering algorithms for the symbolic data. In [14], a transfer algorithm is used to partition a set of symbolic objects into clusters that described by the weight distribution vectors. In [15], the classical k-means clustering algorithm is extended in order to manage data characterized by the numerical and the categorical variables. In [16], an iterative relocation algorithm is used to partition a set of symbolic objects into classes so as to minimize the sum of the description potentials of the classes. In [17], a dynamic clustering algorithm for the symbolic data is proposed. In [18], the authors proposed several clustering algorithms for the symbolic data described by interval variables, based on a clustering criterion and has thereby generalized similar approaches in the classical data analysis. In [19], authors proposed a dynamic clustering algorithm for the interval data where the class representatives are defined by an optimality criterion based on a modified Hausdorff distance. In [20], authors proposed partitioning clustering methods for the interval data based on the city-block distances, also considering the adaptive distances. In [21], an adequacy criterion based on the adaptive Hausdorff distance is introduced into the partitioning clustering algorithm for the interval-values data. Recently, the interval fuzzy c-mean (IFCM) clustering algorithm is proposed [22]. In the IFCM clustering algorithm, traditional fuzzy c-mean clustering algorithm is extended to deal with the interval-values data. Moreover, this algorithm is superior to the previous results. In [23], an interval competitive agglomeration (ICA) clustering algorithm is proposed to overcome the problems of the unknown clusters number and the initialization of prototypes in the clustering algorithm for the symbolic interval-values data.

The major drawbacks of the above approaches are the difficulty in determining the number of clusters, the sensitivity to initialization, noise and outliers. For the effects of outlier, some of robust approaches in the traditional clustering algorithms are proposed [24-26]. Similarly, some of robust approaches are also proposed in the soft-computing community [27-32]. Those robust clustering approaches can be divided into two class. In the first class, the objective function of the FCM clustering algorithm is modified to overcome the problem of outliers. These algorithms are still sensitive to initialization and other parameters used by the algorithms. The second classes of algorithms are based on robust estimators which can tolerate up to 50% noise. The standard robust methods can only be used to find a single cluster in a noisy data set. Additionally, some of these algorithms have been extended to find multiple clusters by extracting one cluster at a time. In this study, the concepts of the robust competitive agglomeration (RCA) clustering algorithm [23] is used and extended into deal with the

symbolic interval-values data.

Moreover, the advantages of the proposed clustering algorithm are also liked to the RCA clustering algorithm. The objective function of the RCA is designed so that it inherits the advantages of hierarchical clustering. Additionally, the RCA clustering algorithm starts by partitioning the data set into a large number of small clusters. As the algorithm progresses, adjacent clusters compete for the symbolic interval-values data and the clusters that lose the competition gradually become depleted and vanish. Thus, the final partition is taken to have the "optimal" number of clusters from the view of the objective function. Moreover, the final result is far less sensitive to initialization and local minima. The good properties of the RCA clustering algorithm for the crisp-values data are also shown in the proposed clustering algorithm for the symbolic interval-values data. Experiment results show the merits and the usefulness of the proposed clustering algorithm.

The organization of the rest of the paper is as follows. In Section 2, an IFCM clustering algorithm is briefly introduced. In Section 3, an ICA clustering algorithm is proposed and discussed. The simulation results are shown in the Section 4. Finally, the conclusions are summarized in the Section 5.

2. Interval fuzzy c-means (IFCM) clustering algorithm [22]

This algorithm is an extension of the standard fuzzy c-means clustering algorithm that furnishes a fuzzy partition and a prototype for each cluster by optimizing an adequacy criterion based on a suitable squared Euclidean distance between the vectors of interval-values data. An IFCM clustering algorithm is stated as follows.

Let $X = \{\bar{x}_k | k = 1, \dots, n\}$ be a set of n vectors in an n -dimensional feature space with coordinate axis labels $(x^1, \dots, x^j, \dots, x^p)$ and $U = [u_{ij}]$ is a $c \times n$ matrix called a constrained fuzzy C-partition matrix. Each pattern k is represented as vector of intervals $\bar{x}_k = (x_k^1, \dots, x_k^j, \dots, x_k^p)$ where $x_k^j = [a_k^j, b_k^j]$ with $a_k^j \leq b_k^j$. Let $G = (\bar{g}_1, \dots, \bar{g}_i, \dots, \bar{g}_c)$ represent a c -tuple of prototypes each of which characterizes one of the c clusters. The prototype \bar{g}_i can be also represented as a vector of intervals $(g_i^1, \dots, g_i^j, \dots, g_i^p)$ where $g_i^j = [\alpha_i^j, \beta_i^j]$ with $\alpha_i^j \leq \beta_i^j$.

An IFCM clustering algorithm minimizes the following objective function:

$$W^1(G, U, X) = \sum_{i=1}^c \sum_{k=1}^n (u_{ik})^2 \varphi(\bar{x}_k, \bar{g}_i)$$

$$= \sum_{i=1}^c \sum_{k=1}^n (u_{ik})^2 \left(\sum_{j=1}^p (a_k^j - \alpha_i^j)^2 + (b_k^j - \beta_i^j)^2 \right), \quad (1)$$

subject to

$$\sum_{i=1}^c u_{ik} = 1, \text{ for } k=1, \dots, n. \quad (2)$$

In (1), ϕ is the square of Euclidean distance measuring the dissimilarity between the vectors of the interval-values data and u_{ik} is the membership degree of pattern k in i th cluster.

To minimize the objective function in (1) with respect to U , the Lagrange multipliers method is applied and obtained as

$$\begin{aligned} J^1(G, U, X) \\ = \sum_{i=1}^c \sum_{k=1}^n (u_{ik})^2 \left(\sum_{j=1}^p (a_k^j - \alpha_i^j)^2 + (b_k^j - \beta_i^j)^2 \right) \\ - \sum_{k=1}^n \lambda_k \left(\sum_{i=1}^c u_{ik} - 1 \right). \end{aligned} \quad (3)$$

Then, G is fixed and solve

$$\frac{\partial J^1}{\partial u_{st}} = 2u_{st} \left(\sum_{j=1}^p (a_t^j - \alpha_s^j)^2 + (b_t^j - \beta_s^j)^2 \right) - \lambda_t = 0, \quad (4)$$

for $s=1, \dots, c$ and $t=1, \dots, n$,

to obtain an updating equation for the memberships u_{st} .

Thus, equation (4) can be rewritten as

$$u_{st} = \frac{\lambda_t}{2 \left(\sum_{j=1}^p (a_t^j - \alpha_s^j)^2 + (b_t^j - \beta_s^j)^2 \right)}, \quad (5)$$

for $s=1, \dots, c$ and $t=1, \dots, n$.

Substituting (5) into (2), λ_t is obtained as

$$\lambda_t = \frac{1}{\sum_{i=1}^c \left(1 / 2 \left(\sum_{j=1}^p (a_t^j - \alpha_i^j)^2 + (b_t^j - \beta_i^j)^2 \right) \right)}. \quad (6)$$

According to (5) and (6), the updating equation for the memberships u_{ik} is

$$u_{ik} = \left(\sum_{h=1}^c \frac{\left(\sum_{j=1}^p (a_k^j - \alpha_h^j)^2 + (b_k^j - \beta_h^j)^2 \right)^{-1}}{\left(\sum_{j=1}^p (a_k^j - \alpha_h^j)^2 + (b_k^j - \beta_h^j)^2 \right)} \right)^{-1}, \quad (7)$$

for $i=1, \dots, c, k=1, \dots, n$.

For the updating equation of G (i.e. α_i^j and β_i^j), equation (3) is minimized with respect to G . Then, we can be obtain

$$\alpha_i^j = \frac{\sum_{k=1}^n (u_{ik})^2 a_k^j}{\sum_{k=1}^n (u_{ik})^m} \quad (8)$$

and

$$\beta_i^j = \frac{\sum_{k=1}^n (u_{ik})^2 b_k^j}{\sum_{k=1}^n (u_{ik})^m}. \quad (9)$$

3. Robust Interval competitive agglomeration (RICA) clustering algorithm

To overcome the problems of the IFCM clustering algorithm (i.e. the initialization, the unknown clusters numbers and outliers), an RICA clustering algorithm is used to dealing with above problems. The proposed objective function for the RICA clustering algorithm is defined as

$$\begin{aligned} W^2(G, U, X) \\ = \sum_{i=1}^c \sum_{k=1}^n (u_{ik})^2 \rho[\phi(\bar{x}_k, \bar{g}_i)] - \eta \sum_{i=1}^c \left(\sum_{k=1}^n u_{ik} w_{ik} \right)^2, \end{aligned} \quad (10)$$

subject to

$$\sum_{i=1}^c u_{ik} = 1, \text{ for } k=1, \dots, n. \quad (11)$$

In (10), u_{ik} is the membership degree of feature (interval-values data) vector \bar{x}_k in cluster \bar{g}_i , η is balance parameter, $\rho[\cdot]$ is the robust loss function, w_{ik} is the weighting function that has a major influence on the robustness of our RICA clustering algorithm and $\phi(\bar{x}_k, \bar{g}_i)$ is the distance between the feature vector \bar{x}_k and the prototype \bar{g}_i . The objective function of the RICA clustering algorithm has two components. The first component in (10) can be regarded as a generalization of M-estimator to detect c clusters simultaneously. The global minimum of this component is reached when the number of clusters c is equal to the number of “good” sample data points. The second component in (10) is used to maximize the number of “good” points in each cluster. The weights w_{ik} represent degrees of “goodness” or typicality of point \bar{x}_k with respect to cluster i , and are used to generate robust estimates of the prototype parameters. The global minimum of this term (including the negative sign) is achieved when all points are lumped in one cluster, and all other clusters are empty. When both components are combined with a proper choice of the agglomeration parameter a , the final partition will minimize the sum of intra-cluster distances while partitioning the data set into the smallest possible number of clusters.

To minimize (10) with respect to G for fixed U and set the gradient to zero,

$$\sum_{k=1}^n (u_{ik})^2 w_{ik} \frac{\partial \phi(\bar{x}_k, \bar{g}_i)}{\partial \bar{g}_i} = 0, \quad (12)$$

where

$$w_{ik} = \frac{\partial \rho[\phi(\bar{x}_k, \bar{g}_i)]}{\partial \phi(\bar{x}_k, \bar{g}_i)}. \quad (13)$$

Further simplification of this equation depends on the loss function $\rho[\cdot]$ and the distance measure used, and will be discussed later in this section. It can be shown that the weight w in (10) is equivalent to the weight function of the W-estimator. The choice of the weight function has a major influence on the robustness of our RICA clustering algorithm, and will be discussed later in this section.

In the RICA clustering algorithm, the Euclidean distance measure is used and defined as

$$\phi(\bar{x}_k, \bar{g}_i) = \sum_{j=1}^p (a_k^j - \alpha_i^j)^2 + (b_k^j - \beta_i^j)^2. \quad (14)$$

To minimize the objective function in (10), the Lagrange multipliers method is applied and obtained as

$$\begin{aligned} J^2(G, U, X) &= \sum_{i=1}^c \sum_{k=1}^n (u_{ik})^2 \rho \left[\sum_{j=1}^p (a_k^j - \alpha_i^j)^2 + (b_k^j - \beta_i^j)^2 \right] \\ &\quad - \eta \sum_{i=1}^c \left(\sum_{k=1}^n u_{ik} w_{ik} \right)^2 - \sum_{k=1}^n \lambda_k \left(\sum_{i=1}^c u_{ik} - 1 \right). \end{aligned} \quad (15)$$

To obtain the updating equation u_{ij} that replaced by u_{st} , J^2 is minimized with respect to U for fixed G and set to zero. That is,

$$\begin{aligned} \frac{\partial J}{\partial u_{st}} &= 2u_{st} \rho \left[\sum_{j=1}^p (a_t^j - \alpha_s^j)^2 + (b_t^j - \beta_s^j)^2 \right] \\ &\quad - 2\eta \sum_{k=1}^n u_{sk} w_{sk} - \lambda_t, \end{aligned} \quad (16)$$

for $s = 1, 2, \dots, c$ and $t = 1, 2, \dots, n$, to obtain an updating equation for the membership u_{st} . The solution can be simplified considering by assuming that the membership degree values do not change significantly from one iteration to the next and by computing the term $\sum_{k=1}^n u_{sk} w_{sk}$ in (16) using the membership values from the previous iteration. With this assumption, equation (16) can be rewritten as

$$u_{st} = \frac{2\eta N_s + \lambda_t}{2\rho \left[\sum_{j=1}^p (a_t^j - \alpha_s^j)^2 + (b_t^j - \beta_s^j)^2 \right]}, \quad (17)$$

where $N_s = \sum_{k=1}^n u_{sk} w_{sk}$ is the cardinality of cluster s . Substituting (17) into (11), then

$$\sum_{i=1}^c \frac{2\eta N_i + \lambda_t}{2\rho \left[\sum_{j=1}^p (a_t^j - \alpha_i^j)^2 + (b_t^j - \beta_i^j)^2 \right]} = 1, \quad (18)$$

or

$$\begin{aligned} \lambda_t &= \frac{1 - \eta \sum_{i=1}^c \left(N_i / \rho \left[\sum_{j=1}^p (a_t^j - \alpha_i^j)^2 + (b_t^j - \beta_i^j)^2 \right] \right)}{\sum_{i=1}^c \left(1 / \rho \left[\sum_{j=1}^p (a_t^j - \alpha_i^j)^2 + (b_t^j - \beta_i^j)^2 \right] \right)}. \end{aligned} \quad (19)$$

As a result, equation (17) can be rewritten as

$$u_{st} = u_{st}^{FCM} + u_{st}^{Bias}, \quad (20)$$

where

$$u_{st}^{FCM} = \frac{\left(1 / \rho \left[\sum_{j=1}^p (a_t^j - \alpha_s^j)^2 + (b_t^j - \beta_s^j)^2 \right] \right)}{\sum_{i=1}^c \left(1 / \rho \left[\sum_{j=1}^p (a_t^j - \alpha_i^j)^2 + (b_t^j - \beta_i^j)^2 \right] \right)}, \quad (21)$$

and

$$u_{st}^{Bias} = \frac{\eta(N_s - \bar{N}_t)}{\rho \left[\sum_{j=1}^p (a_t^j - \alpha_s^j)^2 + (b_t^j - \beta_s^j)^2 \right]}. \quad (22)$$

In (22), \bar{N}_t is defined as

$$\bar{N}_t = \frac{\sum_{i=1}^c \left(1 / \rho \left[\sum_{j=1}^p (a_t^j - \alpha_i^j)^2 + (b_t^j - \beta_i^j)^2 \right] \right) N_i}{\sum_{i=1}^c \left(1 / \rho \left[\sum_{j=1}^p (a_t^j - \alpha_i^j)^2 + (b_t^j - \beta_i^j)^2 \right] \right)}, \quad (23)$$

which is simply a weighted average of the cluster cardinality.

For the updating equation of G (i.e. α_i^j and β_i^j for $i=1, 2, \dots, c$ and $j=1, 2, \dots, p$), J^2 is also minimized with respect to G for fixed U and set to zero. Then, the updating equation of α_i^j and β_i^j are obtained as

$$\alpha_i^j = \frac{\sum_{k=1}^n (u_{ik})^2 w_{ik} a_k^j}{\sum_{k=1}^n (u_{ik})^2 w_{ik}}, \quad (24)$$

and

$$\beta_i^j = \frac{\sum_{k=1}^n (u_{ik})^2 w_{ik} b_k^j}{\sum_{k=1}^n (u_{ik})^2 w_{ik}}. \quad (25)$$

In (21), it is the well-known membership term in the FCM algorithm which takes into account only the relative distances of the feature data to all clusters. The sec-

ond component in (22), u_{st}^{Bias} , is a signed bias term which depends on the difference between the cardinality of the cluster of interest and the weighted average of cardinalities from \bar{x}_t . For the clusters with cardinality higher than average value, the bias term is positive. Then, the membership value is increased. On the other hand, for low cardinality clusters, the bias term is negative. Then, the membership value is decreased. Moreover, this bias term is also inversely proportional to the distance of feature data \bar{x}_t to the cluster of the interest \bar{g}_s , which serves as an amplification factor. For example, the membership values of the feature data in spurious (low cardinality) clusters are decreased heavily when their distances to such clusters are low. This leads to a gradual reduction of the cardinality of the spurious clusters. When the cardinality of a cluster drops below a threshold, we discard the cluster, and update the number of clusters. Since the initial partition has an overspecified number of clusters, each cluster is approximated by many small clusters in the beginning. As the algorithm proceeds, the second term in (10) causes each cluster to expand and include as many data as possible. At the same time, the constraint in (11) causes adjacent clusters to compete. As a result, only a few clusters will survive, while others will shrink and eventually become extinct.

The choice of η in (10) is important matter since it reflects the importance of the second term relative to the first term. If η is too small, the second term will be neglected and the number of clusters will not be reduced. If η is too large, the first term will be neglected, and all points will be lumped into just one cluster. The value of η should be chosen such that both terms are of the same order of magnitude. As the dimensionality of feature space increases, the first term becomes larger since more components contribute to the value of the distance. Thus, to make the algorithm independent of the distance measure, η should be proportional to the ratio of the two terms. In this study, η is chosen as

$$\eta(itr) = \frac{\tau(itr)}{\sum_{i=1}^c \left[\sum_{k=1}^n u_{ik} w_{ik} \right]^2} * \quad (26)$$

$$\sum_{i=1}^c \sum_{k=1}^n (u_{ik})^2 \rho \left[\sum_{j=1}^p \left[(a_k^j - \alpha_i^j)^2 + (b_k^j - \beta_i^j)^2 \right] \right],$$

where itr is an iteration index. The best choice for τ is the exponential decay that defined as [24]

$$\tau(itr) = \tau_0 \exp(-itr/\zeta), \quad (27)$$

where τ_0 is the initial value and ζ is the time constant.

The choice of the weight function depends on the type

of clusters expected in the data set. In the regression, the errors are assumed that have a symmetric distribution about zero. Then, some robust loss function and its weighting function are used to reduce the effects of outlier [27]. However, the distance is used in the proposed approach, the symmetric distribution are not hold. In [24], authors suggest a monotonically non-increasing weight function $w(\varphi(\bar{x}_k, \bar{g}_i)): \mathcal{R}^+ \rightarrow [0,1]$ such that $w(\varphi(\bar{x}_k, \bar{g}_i)) = 0$ for $\varphi(\bar{x}_k, \bar{g}_i) > T_i + \delta S_i$ where δ is the constant, T_i and S_i are chosen as

$$T_i = \underset{\bar{x}_k \in X_i}{Med}(\varphi(\bar{x}_k, \bar{g}_i)) \text{ for } 1 \leq i \leq c, \quad (28)$$

$$S_i = \underset{\bar{x}_k \in X_i}{MAD}(\varphi(\bar{x}_k, \bar{g}_i)) \text{ for } 1 \leq i \leq c. \quad (29)$$

Choosing $w(0) = 1$, $w(T_i) = 0.5$ and $w'(0) = 0$, results in the following weight function:

$$w_{ik}(\varphi(\bar{x}_k, \bar{g}_i)) = \begin{cases} 1 - \frac{\varphi(\bar{x}_k, \bar{g}_i)^2}{2T_i^2}, & 0 \leq \varphi(\bar{x}_k, \bar{g}_i) \leq T_i \\ \frac{[\varphi(\bar{x}_k, \bar{g}_i) - (T_i + \delta S_i)]^2}{2\delta^2 S_i^2}, & T_i < \varphi(\bar{x}_k, \bar{g}_i) \leq T_i + \delta S_i \\ 0, & \varphi(\bar{x}_k, \bar{g}_i) > T_i + \delta S_i \end{cases} \quad (30)$$

The loss function $\rho_i[\cdot]$ associated with this weight function can be obtained by integrating (30). That is,

$$\rho_i[\varphi(\bar{x}_k, \bar{g}_i)] = \begin{cases} \varphi(\bar{x}_k, \bar{g}_i) - \frac{\varphi(\bar{x}_k, \bar{g}_i)^3}{6T_i^3}, & 0 \leq \varphi \leq T_i \\ \frac{[\varphi(\bar{x}_k, \bar{g}_i) - (T_i + \delta S_i)]^3}{6\delta^2 S_i^2} + \frac{5T_i + \delta S_i}{6}, & T_i < \varphi \leq T_i + \delta S_i \\ \frac{5T_i + \delta S_i}{6} + K_i, & \varphi > T_i + \delta S_i \end{cases} \quad (31)$$

where the K_i are constants used to make all ρ_i function approach the same maximum value. The constants K_i is chosen as

$$K_i = \max_{1 \leq j \leq c} \left\{ \frac{5T_j + \delta S_j}{6} \right\} - \frac{5T_i + \delta S_i}{6}, \quad (32)$$

for $1 \leq i \leq c$.

The procedure of the RICA clustering algorithms is stated as follows.

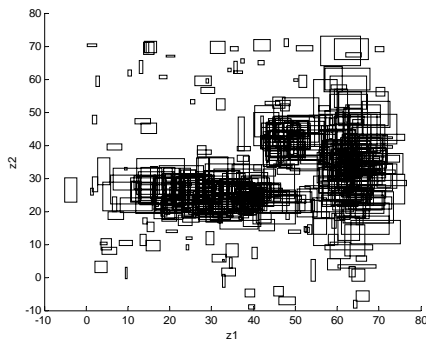


Figure 1. The interval-values data set of the ID1.

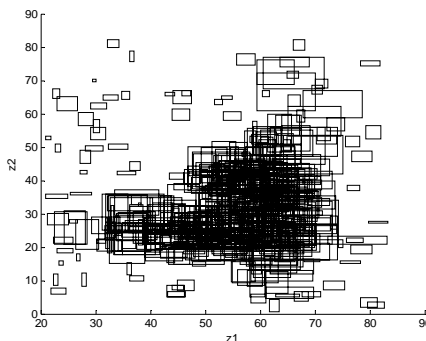


Figure 2. The interval-values data set of the ID2.

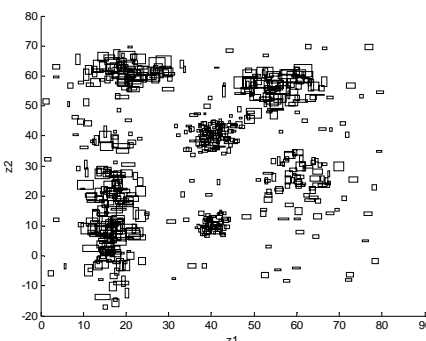


Figure 3. The interval-values data set of the ID3.

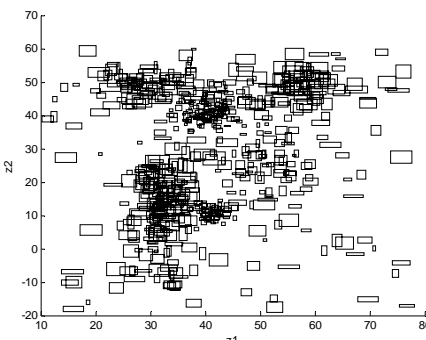


Figure 4. The interval-values data set of the ID4.

Procedures of RICA clustering algorithm

Step1. Initialization

Fix the maximum number of clusters $c = c_{\max}$.

Initialize the fuzzy c -partition matrix $U^{(0)}$.

Initialize iteration counter $itr=0$.

Step2. Compute the initial cardinalities N_i for $1 \leq i \leq c$.

Step3. Compute $\phi(\bar{x}_k, \bar{g}_i)$ for $1 \leq i \leq c$ and $1 \leq k \leq n$ in (14).

Step4. Estimate T_i and S_i by (28) and (29).

Step5. Update the weights w_{ik} by (30).

Step6. Update $\eta(itr)$ using (26) and (27).

Step7. Update the partition matrix $U^{(itr)}$ using (20).

Step8. Compute the cardinalities N_i for $1 \leq i \leq c$.

Step9. If $N_i < \varepsilon$ discard cluster \bar{g}_i .

Step10. Update the number of clusters c .

Step11. Update the centers of clusters using (24) and (25).

Step12. $itr=itr+1$.

Step13. If prototype parameters are stabilized, the procedures are terminated. Otherwise, go to Step 2.

4. Simulations

To show the usefulness of the RICA clustering algorithm for the symbolic interval-values data with outliers, some of the symbolic interval-values data sets are considered in this study. In following examples, a cluster \bar{g}_i was discarded if its cardinality N_i is less than 4.5. In (27), the initial value τ_0 and the time constant ζ are set as 5 and 10, respectively. Parameter δ is chosen as 4, respectively [24]. The initial partition matrix is randomly assigned. In this study, the data set ID1-ID4 are considered and shown in Figure 1~ Figure 4.

To illustrate the independence of the proposed algorithm from the choice of initial number of the clusters c_{\max} , the RICA clustering algorithm with different initial number of clusters are considered. For the ID1 and the ID2, the RICA clustering algorithm can be converge to the same final partition when c_{\max} are chosen as 9, 12, 15, and 18. Figures 5(a) and 5(b) show the reduction of the number of clusters in process for the RICA clustering algorithm with different c_{\max} values. For the ID3 and the ID4, the RICA clustering algorithm also converged to the same final partition when c_{\max} are chosen as 15, 20, 25, and 30. Figures 5(c) and 5(d) show the reduction of the number of clusters in process for the RICA clustering algorithm with the different c_{\max} values. These experiments illustrate the insensitivity of the RICA clustering algorithm for the interval-values data to the initial number of clusters. According the above results, the RICA clustering algorithm has fast converges in a few iterations regardless of the initial number of clusters.

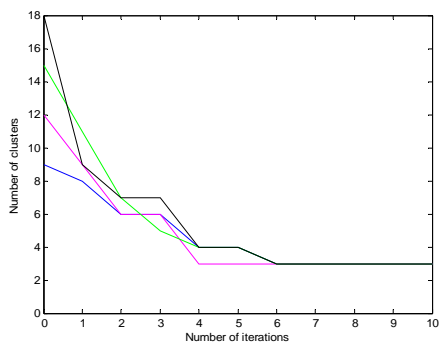


Figure 5(a). The reduction of the number of clusters for the ID1 in process is shown for the RICA clustering algorithm with the different c_{max} values.

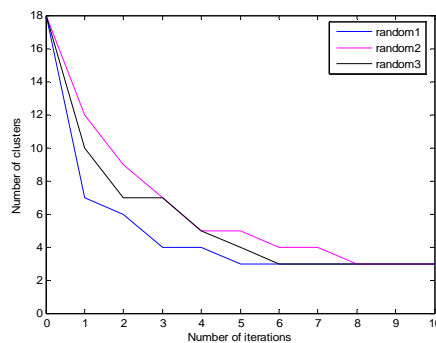


Figure 6(a). The reduction of the number of clusters for the ID1 in process is shown for the RICA clustering algorithm with the different c_{max} values.

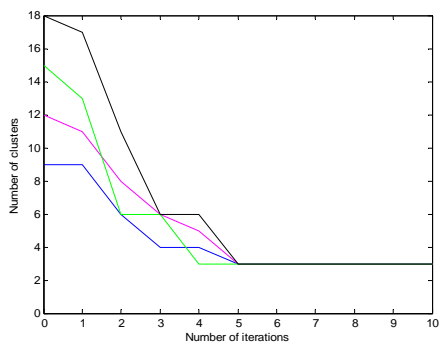


Figure 5(b). The reduction of the number of clusters for the ID2 in process is shown for the RICA clustering algorithm with the different c_{max} values.

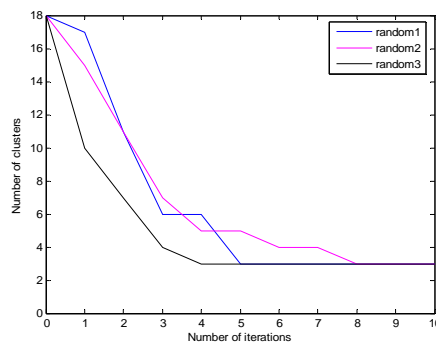


Figure 6(b). The reduction of the number of clusters for the ID2 in process is shown for the RICA clustering algorithm with the different c_{max} values.

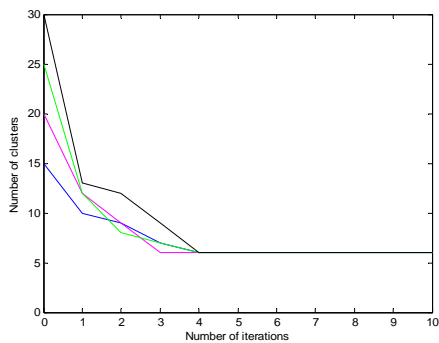


Figure 5(c). The reduction of the number of clusters for the ID3 in process is shown for the RICA clustering algorithm with the different c_{max} values.

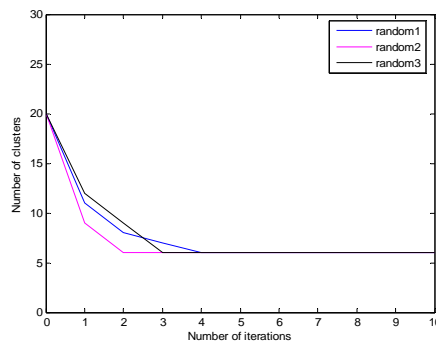


Figure 6(c). The reduction of the number of clusters for the ID3 in process is shown for the RICA clustering algorithm with the different c_{max} values.

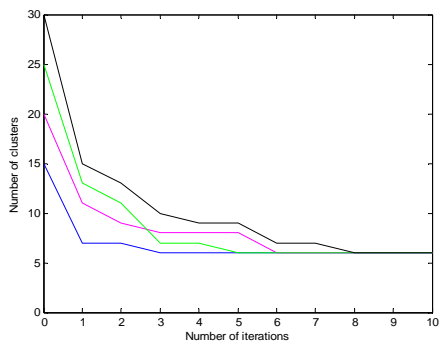


Figure 5(d). The reduction of the number of clusters for the ID4 in process is shown for the RICA clustering algorithm with the different c_{max} values.

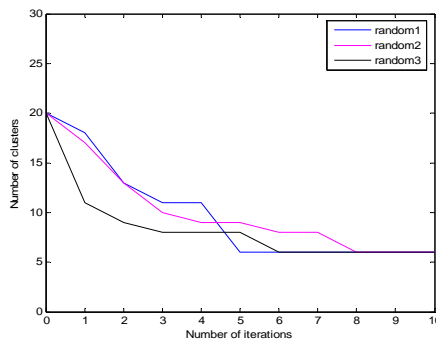


Figure 6(d). The reduction of the number of clusters for the ID4 in process is shown for the RICA clustering algorithm with the different c_{max} values.

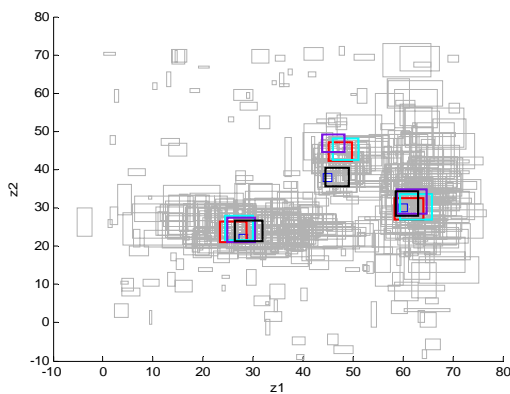


Figure 7(a). The cluster centers that obtained by the RICA algorithm (black rectangle), the partitioning clustering algorithm with city-block distances [20] (blue rectangle), the partitioning clustering algorithm with the Hausdorff distances [21] (amethyst rectangle) and the IFCM algorithm [22] (red rectangle) are shown for the ID1.

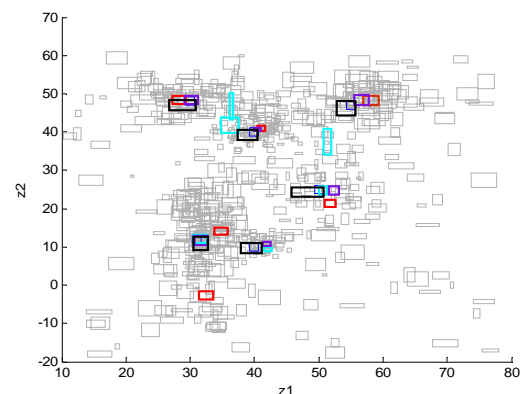


Figure 7(d). The cluster centers that obtained by the RICA algorithm (black rectangle), the partitioning clustering algorithm with city-block distances [20] (blue rectangle), the partitioning clustering algorithm with the Hausdorff distances [21] (amethyst rectangle) and the IFCM algorithm [22] (red rectangle) are shown for the ID4.

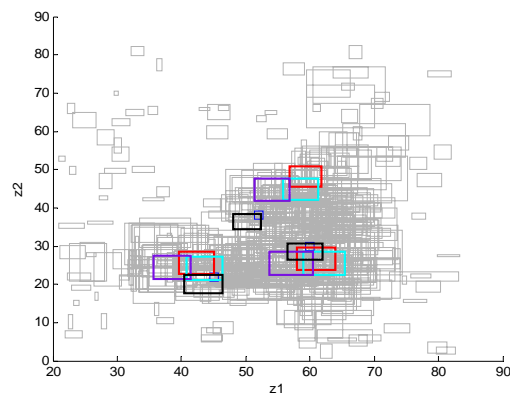


Figure 7(b). The cluster centers that obtained by the RICA algorithm (black rectangle), the partitioning clustering algorithm with city-block distances [20] (blue rectangle), the partitioning clustering algorithm with the Hausdorff distances [21] (amethyst rectangle) and the IFCM algorithm [22] (red rectangle) are shown for the ID2.

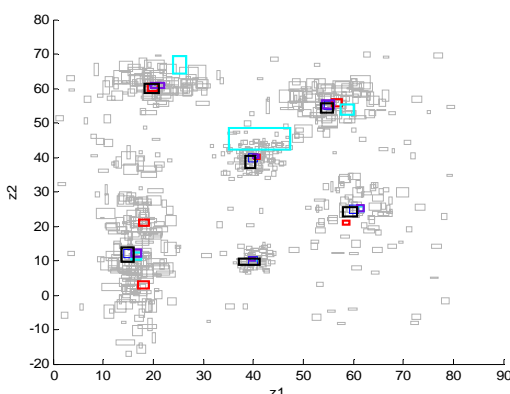


Figure 7(c). The cluster centers that obtained by the RICA algorithm (black rectangle), the partitioning clustering algorithm with city-block distances [20] (blue rectangle), the partitioning clustering algorithm with the Hausdorff distances [21] (amethyst rectangle) and the IFCM algorithm [22] (red rectangle) are shown for the ID3.

Table 1. The CR index of RICA clustering algorithm, the partitioning clustering algorithm with the city-block distances [20], the partitioning clustering algorithm with the Hausdorff distances [21] and the IFCM algorithm [22] for ID1-ID4 are tabulated.

	RICA	IFCM	City-block	Hausdorff
ID1	0.8283	0.7621	0.6638	0.6547
ID2	0.4247	0.3298	0.3485	0.3652
ID3	0.8031	0.7448	0.6869	0.7214
ID4	0.6194	0.5016	0.5421	0.5040

To illustrate the insensitivity of the RICA clustering algorithm to the initial prototypes, the above data sets are also considered. For the ID1~ID4, the initial prototypes are randomly assigned. Figures 6(a)-6(d) show the evolution of the algorithm versus the number of iterations for those data sets for the ID1, ID2, ID3 and ID4. In those figures, the RICA clustering algorithm can be converges to the same optimal partition regardless of its initialization.

Besides, the cluster centers that obtained by the RICA clustering algorithm are shown as Figures 7(a)-7(d) for the ID1~ID4, respectively. In those figures, the cluster centers for other clustering algorithms (i.e. the partitioning clustering algorithm with the city-block distances [20], the partitioning clustering algorithm with the Hausdorff distances [21] and the IFCM algorithm [22]) are also performed. Based on above figures, it is clear that the RICA clustering algorithm not affected by outliers. Additionally, the corrected Rand (*CR*) indices [22] is also used to performance index. The *CR* index measures the similarity between *a priori* hard partition and a hard partition furnished by a partitioning hard clustering algorithm or obtained from the fuzzy partition furnished by the fuzzy clustering algorithm. *CR* takes its values on

the interval $[-1, 1]$, where the value 1 indicates perfect agreement between partitions, whereas values near 0 (or negatives) correspond to cluster agreement found by chance. The CR index of RICA clustering algorithm and others algorithms are summarized as Table 1. According to above results, the RICA clustering algorithm has following advantages. First, the RICA clustering algorithm has fast converges in a few iterations regardless of the initial number of clusters. Second, the initial number of clusters c_{\max} is not predetermined. Third, the results of the RICA clustering algorithm are not affected by the initial prototypes. Fourth, the RICA clustering algorithm is better than others clustering algorithm when the outliers are existed in the symbolic interval-values data.

5. Conclusions

In this study, an RICA clustering algorithm is developed to overcome the problems of the outliers, the unknown clusters number and the initialization of prototypes in the clustering algorithm for the symbolic interval-values data. That is, the proposed RICA clustering algorithm extends the concepts of the RCA clustering algorithm that can deal with the symbolic interval-values data. Hence, the advantages of an RICA clustering algorithm are also liked to the RCA clustering algorithm and better than the IFCM clustering algorithm. Experiments with simple interval-values data sets show the merits and the usefulness of the RICA clustering algorithm. Besides, the good properties of the RCA clustering algorithm for the crisp-values data are also expressed in the RICA clustering algorithm for the interval-values data.

Acknowledgment

This work was supported by National Science Council Under Grant NSC-98-2221-E-197-020-

References

- [1] A. K. Jain, M. N. Murty and P. J. Flynn, "Data clustering: A review," *ACM Computing Surveys*, vol. 31, no. 3, pp. 264-323, 1999.
- [2] A. D. Gordon, *Classification*, CRC Press, 1999.
- [3] H. H. Bock and E. Diday, *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data*, Springer-Verlag, 2000.
- [4] L. Billard and E. Diday, "From the statistics of data to the statistics of knowledge: Symbolic data analysis," *Journal of the American Statistical Association*, vol. 98, no. 462, pp. 470-487, 2003.
- [5] K. C. Gowda and E. Diday, "Symbolic clustering using a new dissimilarity measure," *Pattern Recognition*, vol. 24, no. 6, pp. 567-578, 1991.
- [6] K. C. Gowda and E. Diday, "Symbolic clustering using a new similarity measure," *IEEE Trans. on Systems Man Cybernetic*, vol. 22, pp. 368-378, 1992.
- [7] M. Ichino and H. Yaguchi, "Generalized Minkowski metrics for mixed feature type data analysis," *IEEE Trans. on Systems Man Cybernetic*, vol. 24, no. 4, pp. 698-708, 1994.
- [8] K. C. Gowda and T. R. Ravi, "Divisive clustering of symbolic objects using the concepts of both similarity and dissimilarity," *Pattern Recognition*, vol. 28, no. 8, pp. 1277-1282, 1995.
- [9] K. C. Gowda and T. R. Ravi, "Agglomerative clustering of symbolic objects using the concepts of both similarity and dissimilarity," *Pattern Recognition Letter*, vol. 16, pp. 647-652, 1995.
- [10] M. Chavent, "A monothetic clustering method," *Pattern Recognition Letter*, vol. 19, pp. 989-996, 1998.
- [11] K. C. Gowda and T. R. Ravi, "Clustering of symbolic objects using gravitational approach," *IEEE Trans. on Systems Man Cybernetic*, vol. 29, no. 6, pp. 888-894, 1999.
- [12] D. S. Guru, B. B. Kiranagi, and P. Nagabhushan, "Multivalued type proximity measure and concept of mutual similarity value useful for clustering symbolic patterns," *Pattern Recognition Letter*, vol. 25, pp. 1203-1213, 2004.
- [13] D. S. Guru and B. B. Kiranagi, "Multivalued type dissimilarity measure and concept of mutual dissimilarity value for clustering symbolic patterns," *Pattern Recognition*, vol. 38, pp. 151-256, 2005.
- [14] E. Diday and P. Brito, *Symbolic cluster analysis: Conceptual and Numerical Analysis of Data*, Springer-Verlag, pp. 5-84, 1989.
- [15] H. Ralambondrainy, "A conceptual version of the k-means algorithm," *Pattern Recognition Letter*, vol. 16, pp. 1147-1157, 1995.
- [16] A. D. Gordon, *An interactive relocation algorithm for classifying symbolic data. In: Gaul, W. et al. (Eds.), Data Analysis: Scientific Modeling and Practical Application*, Springer-Verlag, pp. 7-23, 2000.
- [17] R. Verde, F. A. T. De Carvalho, and Y. Lechevalier, "A dynamical clustering algorithm for symbolic data. In: Tutorial on Symbolic Data Analysis," *25th Annual Conference of the Gesellschaft für Klassifikation e.V. University of Munich*, March 13, 2001.
- [18] H. -H. Bock, "Clustering algorithms and Kohonen maps for symbolic data," *Journal of the Japanese Society of Computational Statistics*, vol. 15, pp.

- 1-13, 2002.
- [19] M. Chavent and Y. Lechevallier, *Dynamical clustering algorithm of interval data: Optimization of an adequacy criterion based on Hausdorff distance*. In: Sokolowski, A., Bock, H.-H. (Eds.), *Classification, Clustering and Data Analysis*, Springer-Verlag, Heidelberg, pp. 53-59, 2002.
- [20] R. M. C. R. Souza, and F. A. T. De Carvalho, "Clustering of interval data based on city-block distances," *Pattern Recognition Letter*, vol. 25, no. 3, pp. 353-365, 2004.
- [21] F. A. T. De Carvalho, R. M. C. R. Souza, M. Chavent, and Y. Lechevallier, "Adaptive Hausdorff distances and dynamic clustering of symbolic data," *Pattern Recognition Letter*, vol. 27, no. 3, pp. 167-179, 2006.
- [22] F. A. T. De Carvalho, "Fuzzy c-means clustering methods for symbolic interval data," *Pattern Recognition Letter*, vol. 28, no. 6, pp. 423-437, 2007.
- [23] Chen-Chia Chuang, Jin-Tsong Jeng, and C. W. Tao, "Interval competitive agglomeration clustering algorithm", *Expert System with Applications*, vol. 37, no. 9, pp. 6567-6578, 2010.
- [24] H. Frigui and R. Krishnapuram, "A robust competitive clustering algorithm with applications in computer vision," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, pp. 450-465, 1999.
- [25] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*. Prentice-Hall, Englewood Cliffs, New Jersey, 1988.
- [26] R. N. Dave and R. Krishnapuram, "Robust clustering methods: A unified view," *IEEE Trans. Fuzzy System*, vol. 5, pp. 270-293, 1997.
- [27] Chen-Chia Chuang, Jin-Tsong Jeng, and C. W. Tao, "Hybrid robust approach for TSK fuzzy modeling with outliers", *Expert Systems with Applications*, vol. 36, no. 5, pp. 8925-8931, 2009.
- [28] Yih-Guang Lue, Tsu-Tian Lee, and Wei-Yen Wang, "Observer-based Adaptive Fuzzy-Neural Control for Unknown Nonlinear Dynamical Systems," *IEEE Trans. System Man and Cybernetics-Part B*, vol. 29, no. 5, pp. 583-591, October 1999.
- [29] Wei-Yen Wang, Tsu-Tian Lee, and Ching-Lang Liu and Chi-Hsu Wang, "Function Approximation Using Fuzzy Neural Networks with Robust Learning Algorithm," *IEEE Trans. System Man and Cybernetics Part B*, vol. 27, no. 4, pp. 740-747, August 1997.
- [30] Chen-Chia Chuang, "Annealing Robust Fuzzy Neural Networks for Modeling of Molecular Autoregulatory Feedback Loop Systems," *International Journal of Fuzzy Systems*, vol. 10, no. 1, pp. 11-17, 2008.
- [31] Chen-Chia Chuang, Jin-Tsong Jeng, and C. W. Tao,

"Two-Stages Support Vector Regression for Fuzzy Neural Networks with Outliers," *International Journal of Fuzzy Systems*, vol. 11, no. 1, pp. 20-28, 2009.

- [32] Chen-Chia Chuang, Shun-Feng Su, Jin-Tsong Jeng, and Chih-Ching Hsiao, "Robust Support Vector Regression Networks for Function Approximation With Outliers," *IEEE Trans. Neural Networks*, vol. 13, no. 6, pp. 1322-1330, 2002.



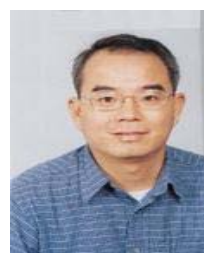
Jin-Tsong Jeng was born in Taiwan, R.O.C., in 1967. He received the B.S.E.E., M.S.E.E. and Ph. D. degrees all in Electrical Engineering from the National Taiwan University of Science and Technology, Taipei, Taiwan, in 1991, 1993, and 1997, respectively. He is currently a Professor in the Department of Computer Science and Information Engineering, National Formosa University, Huwei Jen, Yunlin, Taiwan. His primary research interests include neural networks, fuzzy system, intelligent control, support vector regression, magnetic bearing system, bio-informatics, non-holonomic control system and Microarray



Chen-Chia Chuang received the B.S. and M.S. degrees in Electrical Engineering from National Taiwan Institute of Technology, Taipei, Taiwan, in 1991 and 1993, respectively. He received Ph.D. degree in the Department Electrical Engineering at the National Taiwan University of Science and Technology, Taipei, Taiwan in 2000. He is currently a Professor with the Department of Electrical Engineering, National Ilan University, I-Lan, Taiwan. His current research interests are neural networks, data analysis, neural fuzzy systems, support vector regression and signal processing.



Chih-Cheng Tseng is currently an assistant professor of the Department of Electrical Engineering, National Ilan University, Yi-Lan, Taiwan. His research interests include the design, implementation and performance evaluation of protocols for mobile communications and wireless ad hoc/sensor networks.



Chang-Jung Juan is currently an Associate Professor with Department of Electronic Engineering, Hwa-Hsia Institute of Technology. His current research interests are intelligent control, LED, robot systems and signal processing.