International Journal of Fuzzy Logic and Intelligent Systems, vol. 6, no. 3, September 2006, pp. 210-216

Data Pattern Estimation with Movement of the Center of Gravity

Tae-Chon Ahn*, Kyung-Won Jang *, Dong-Du Shin*, Hak-Soo Kang and Yang-Woong Yoon***

***Dept. of Control and Instrumentation Engineering, Wonkwang University**
**** Dept. of Digital & Electrical Information Engineering, Vision College of Jeonju**

Abstract

In the rule based modeling, data partitioning plays crucial role because partitioned sub data set implies particular information of the given data set or system. In this paper, we present an empirical study result of the data pattern estimation to find underlying data patterns of the given data. Presented method performs crisp type clustering with given n number of data samples by means of the sequential agglomerative hierarchical nested model (SAHN). In each sequence, the average value of the sum of all inter-distance between centroid and data point. In the sequel, compute the derivation of the weighted average distance to observe a pattern distribution. For the final step, after overall clustering process is completed, weighted average distance value is applied to estimate range of the number of clusters in given dataset.
The proposed estimation method and its result are considered with the use of FCM demo data set in MATLAB fuzzy logic toolbox and Box and Jenkins's gas furnace data

Key words : SAHN (Sequential Agglomerative Hierarchical Nested) based algorithm, data pattern estimation, Clustering Algorithm

1. Introduction

In the rule-based system modeling, data-clustering algorithm has been applied to system structure identification, and its methodological efficiency has been proved by numerous previous researches and applications [9]. The benefit of this algorithm is it enables efficient rule-based model design with the relatively minimized number of rules against conventional grid partitioning.

HCM (Hard C-means Clustering) and FCM (Fuzzy C-means Clustering) is representative method to deals with efficient system identification. However, theses algorithms require a priori knowledge about data set and highly heuristic to have a satisfactory result [2]. Most of the devoted studies of clustering method are concentrated on the validity index to determine the proper number of clusters in the given data set. Xie-Beni's index [3], Fukuyama-Sugeno's index [4] and Dunn's index [5] are representative validity indexes repeatedly referred, and many validity indexes are proposed to determine the proper number of cluster from their evaluation method [6]-[10]. This is a well-known disclosed issue of the clustering method because the proper number of clusters is not known a priori.

Common aspects of cluster validation method are frequently resulted in a certain number of clusters that is called "the proper number of clusters" from the tested range of possible numbers of clusters. However, selected number of clusters may not give satisfactory result because data clustering algorithm is frequently deployed in a application dependent issues, so clustering result or its validity may evaluated by a certain object function for their own purpose or specified requirements[10]. This tendency often appears in model evaluation of the rule-based system modeling. In spite of some issues to be duly considered data-clustering algorithm is an essential method to have distribution information of the underlying data patterns in given data set.

Major interest of this study is to give a complementary choice in case of failure of validity check or minimum information instead of a 'prior knowledge' or 'the proper number of clusters' However, it is hard question to answer that how many clusters or patterns are contained in the data set. From this point of view, Mountain clustering or Subtractive Clustering should be a good choice for the question. However, common aspect of these methods requires an evaluation of all cluster center candidates and so its computation grows exponentially as the dimension is increased, and their success depends on the choice of some threshold to get the proper number of clusters [12]. Other clustering algorithm, such as HCM (Hard C-means Clustering) and FCM (Fuzzy C-means Clustering), the number of clusters is user-define parameter. For many reason, The clustering algorithm may fail to have a satisfactory result by a bad initialization or the wrong choice of algorithmic parameters [2] [12]. Therefore, in the practical application, data clustering often developed with two phases as follows: a) determine a range of possible numbers of clusters. b) Test a cluster result until satisfactory result is obtained.

In this paper, we introduce a complementary method that detecting the underlying data patterns and gives possible range of the number of clusters without less difficulties of dimensionality and heuristic threshold of validation. The

Manuscript received Apr. 12, 2006; revised Sep. 13, 2006.
This paper was supported by Wonkwang University in 2006.

presented method decrease the number of cluster from $n-1$ to 1 (*n: total number of data point*) by similarity measure between data points by means of *sequential agglomerative hierarchical nested model* [1]. In each clustering process, centroid becomes a new data point with rest of data point for next clustering process and essential indicator to estimate underlying data patterns with a presented simple parameter.

2. Clustering Algorithm and Data Pattern Estimation

2.1 SL Clustering Algorithm and data pattern estimation parameter

In this study, single linkage (SL) model of the agglomerative hierarchical nested model is applied to detect underlying data patterns of given data set. This algorithm is a graph-theoretic model that uses local connectivity criterion, in contrast with objective function model such as HCM. In the SI (System Identification) of rule-based system modeling, clustering algorithm often deals with compromise two conflicted facts that are efficiency and accuracy of identified model. Therefore, the clustering algorithm is less crucial than pattern recognition. To concentrate on our purpose, SL is applied with a pattern detection parameter is proposed here because of their structural advantages for this work. SL algorithm and proposed pattern estimation parameter is described as following steps:

1) Step 1: Similarity measure with nearest pair detection

SL algorithm performs crisp type data clustering with a given n numbers of data. In the beginning clustering process, each datum become a cluster and cluster center. With the initial number of clusters, SL calculates Euclidean distance between clusters to extract the most similar pair (most nearest) to be clustered for the next clustering process. The similarity measure with Euclidean distance forms an upper-diagonal matrix $\boldsymbol{D}$ (distance matrix) of size $n \times n$ by means of (1), where n is the cardinality of the input data set (or the total number of clusters).

a) Similarity measure; Let the given data set $i=1...n$, $j=1...m$, where n is number of data, m is dimension of data and d_{ik} is component of distance matrix $\boldsymbol{D}$. The distance matrix $\boldsymbol{D}(n \times n)$ for the similarity measure between data points is obtained from:

$$d_{ik} = \sum_{j=1}^{m} (x_{ij} - x_{kj})^2, i = 1 \cdots n, k = i+1 \cdots n \tag{1}$$

b) Find minimum entry in $\boldsymbol{D}$ and store the index of selected pair; From the result of equation (1), find the nearest data pair to be clustered and store the index of selected pair. Note that diagonal component and its below values of matrix $\boldsymbol{D}$ is not considered during the detection of the nearest pair. If there is more than one nearest pair, only one chosen among them in this study.

2) Step 2: Calculate the centroid v_s

In this step, we calculate the centroid $\boldsymbol{v}_s$ of selected data pair (or cluster) by equation (2). In the sequel, remove store the selected data pair (or cluster) from the data set and add $\boldsymbol{v}_s$ to the data set. Also, store the removed data point for the next clustering process centroid calculation.

$$v_s = \frac{1}{|c_s|} \sum_{k,\, x_k \in c_s} x_k \tag{2}$$

3) Step 3: Calculation of the pattern estimation parameter P_s and ΔP_s.

a) Average distance measure of the point-to-centorid of newly clustered data group; In equation (3), d_s denotes sum of distance of the point-to-centeroid. Equation (4) denotes average distance of the sum of distance for the point-to-centroid, where $\boldsymbol{v}_s$ is the cluster center (centroid) of the *s-th* clustering process, p_s is average distance and $|c_s|$ is cardinality of the selected data pair (or cluster) in clustering process.

$$d_s = \sum_{k,\, x_k \in c_s} \left\| x_k - v_s \right\|^2 \tag{3}$$

$$p_s = \frac{d_s}{|c_s|} \tag{4}$$

In general, when p_s are small, selected cluster in *s-th* clustering process tends to contain a small number of data point with higher similarity than subsequent clustering process. In contrast, when p_s are large, selected cluster tends to contain a large number of data point with lower similarity. However, characteristic tendency of p_s often disturbs clear variation observation of compactness because this kind of compactness indicator or similar compactness parameters shows monotonically increase when the number of clusters approaches to the 2 and vice versa [7],[8]. Therefore, to ameliorate this tendency, we deployed a weighting value to (4) as shown in equation (5) in next phase.

b) Pattern estimation parameter P_s and ΔP_s; In this phase, proposed method applied weighted average value P_s and its variation ΔP_s to observe underlying data patterns in data set. The weighted average value P_s is obtained by equation (5), and its variation ΔP_s is obtained by equation (6), respectively, where C_s is the number of clusters at the *s-th* clustering process and total process of the proposed method is *n-1*. In the sequel, go to step 1 and iterate step until the number of cluster is reached to 1.

$$P_s = \frac{1}{C_s} p_s \tag{5}$$

$$\Delta P_s = P_s - P_{s-1} \tag{6}$$

As it mentioned above, the p_s is a compactness indicator of the selected pair that is obtained from the result of $\boldsymbol{D}$ matrix in hierarchical clustering process. Therefore, overall p_s show very smooth changes with this hierarchical clustering process as far as relatively dissimilar data group is clustered. The appearance of the weighting value of equation (5) ($1/C_s$) is a reciprocal of the total number of clusters of its clustering process. This value gives a good compactness score when clustering process is close to number of data points, In contrast, when the clusters that is have a large number of data point ($1/C_s$) is clustered, gives a bad compactness score while clustering process approaches to two clusters. Therefore, P_s parameter shows more detail separation between a large number of clusters and small number of clusters in overall clustering process. In this paper, parameter P_s and ΔP_s is designed to observe variation of data patterns by clustering process in progress.

3. Numerical example

3.1 Example data set and sample run

In this section, presented method is introduced with a simple numerical example to present a detail procedure of the proposed method. The example data set is artificially composed and its scatter plot is given in figure 1. The sample data set has the distinct three clusters and no sub group (patterns) is observed in each data group. The result of P_s and ΔP_s for example data set is summarized as shown in figure 2.

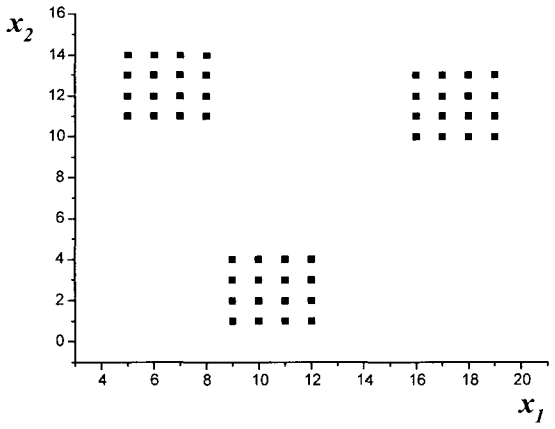
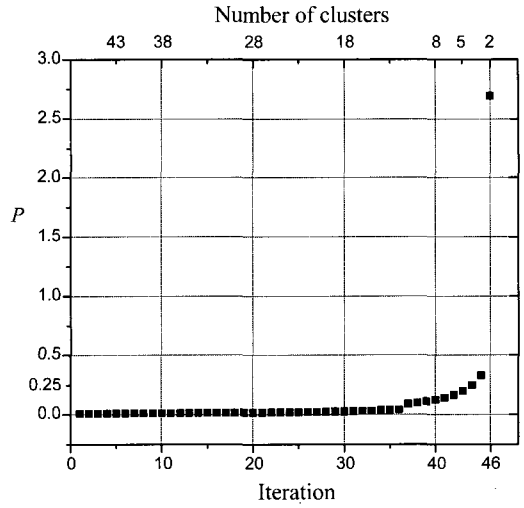
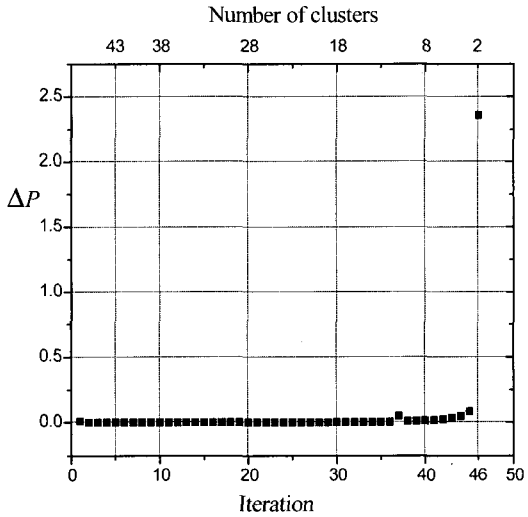


Fig. 1. Scatter plot of the example data set

The character of the p_s in equation (4) is related with compactness and sparsity of clustered data. As shown in figure 2(a), in the interval from iteration 1 to 37, P_s shows almost no changes while clustering process in progress. Therefore, by its character of P_s with clustering process of, clusters in this interval shows small number of data points with high compactness. We can estimate the number of clusters in this interval is not recommended. In contrast, in the last iteration, we can observe P_s is abruptly increased. As shown in figure 1, given data set have distinctly well separated three data groups. This data group is clustered by force of clustering process, so P_s shows a bad compactness because of the abrupt increase of the p_s by a newly formed cluster centoid. Therefore, we can assume that the proper number of clusters and a significant range of the numbers of clusters can be obtained in 37-th to 45-th iteration interval.



(a) $\mathbf{P_s}$



(b) ΔP_s

Fig. 2. Results of P_s and ΔP_s

In the 37-th to 45-th iteration, compactness of a new cluster indicates gradual increase. As it mentioned earlier in the nearest pair detection, if there is more than one nearest pair, then proposed method choose only one among them. Therefore, not selected minimum pairs transferred to next clustering process, and then one of minimum pairs from them will be selected after

previous clustering process ended. At this point, selected pair will have a same p_s value in its clustering process. In spite of same p_s, this value will have different P_s because of weighting value by equation (5).

But it is a negligible difference. However, when clusters have enough amount of data point, this difference will be gradually increase as shown in 37-th to 45-th iteration interval. Note that example data set is artificially composed to have uniform distance in their data group. Unfortunately, this kind of data distribution does not exist in the practical application. However, we still need more precise parameter whether this tendency is occurred by weighting value or by their nature data distribution. Therefore, in this paper, another estimation parameter is used for more detail observation as shown in equation (6). The role of ΔP_s is to observe relative variation of the P_s. If ΔP_s indicates a moderate slop, then we can assume that it is resulted from weighting value. On the other hand, if ΔP_s shows an exponential slope, then we can assume that it caused by their sparsity.

3.2. Estimation of underlying data patterns

As a final stage ΔP_{mean} is applied to observe underlying data patterns of given data set.

1). Determination of maximum number of clusters:

To determine a maximum number clusters ΔP_{mean} is applied. The role of ΔP_{mean} is to eliminate unnecessary cluster candidate from $\Delta \boldsymbol{P}$ in figure 2(b). As it mentioned earlier, a significant range of the numbers of clusters can be obtained in 37-th to 45-th iteration interval. Criterion of this phase is the center of the gravity (average of $\Delta \boldsymbol{P}$) as shown in equation (7). The role of ΔP_{mean} (center of the gravity) is it can give reference point of the balance of power between the range of ΔP_s that shows almost no change and the remainder range of ΔP_s that shows an obvious change. According to equation (7), disregards ΔP_s values if the ΔP_s does not exceed ΔP_{aean} from $\Delta \boldsymbol{P}$.

$$\Delta P_{mean} = \frac{\sum_{s=1}^{n-2} \Delta P_s}{Number\_of\_iteration} \tag{7}$$

2). Determination of minimum number of clusters:

With result of phase (a), repeat the center of gravity calculation with the rest of result of iteration. In the sequel, disregard ΔP_s values if the ΔP_s exceed ΔP_{mean} from the rest of $\Delta \boldsymbol{P}$. Therefore, as shown in figure (3), a possible range of numbers of clusters is determined within 1.2213~0.058535 of the $\Delta \boldsymbol{P}$. The selected number of cluster is 3. In this example, the last result of P_s and ΔP_s is abandoned. When the number of cluster is reached two, P_s abruptly increased because of the distribution of given sample data set, so *n-1 th* result has extremely bad score than *n-2 th* result. Therefore, this value can perturb to have balanced range by their ΔP_s. If given data set have distinct 2 clusters, ΔP_s will be abruptly increased when number of clusters reached 1. Therefore, range of a denominator in equation (7) needs a careful consideration.

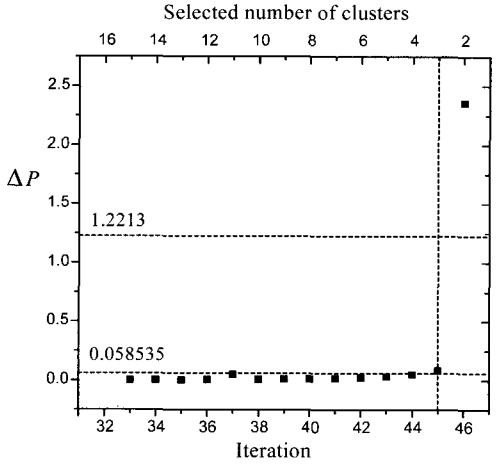
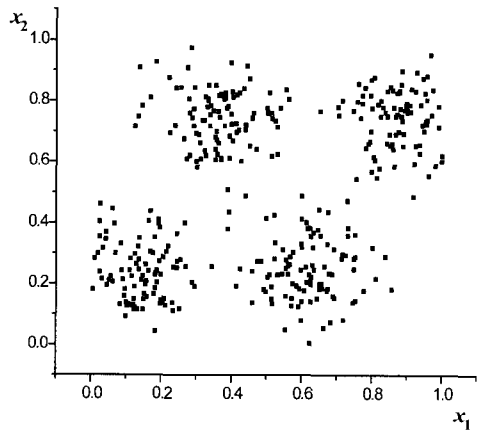


Fig. 3. Selected number of cluster for example data

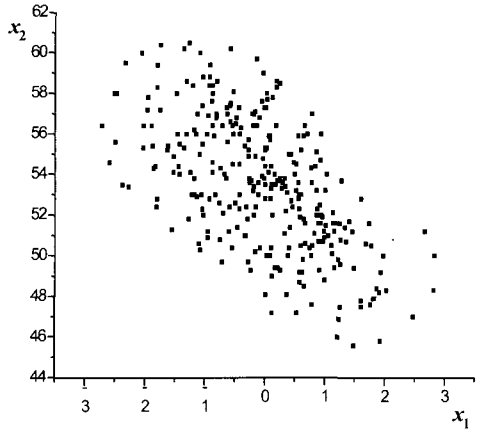
4. Numerical Experiments

In this experiment, to observe performance of the proposed method, two sample data sets are applied. One is extracted from FCM demo of the MATLAB fuzzy logic toolbox, and the other is Box and Jenkins's gas furnace data that is well-known time series data set in rule-based system modeling[1][14].

Matlab demo data show relatively distinct separation but it is clear enough. Result of gas furnace data is compared with a previous survey [13].

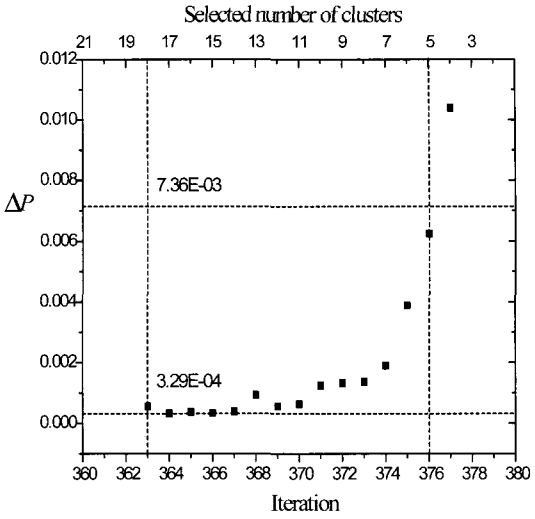


(a) matlab demo data



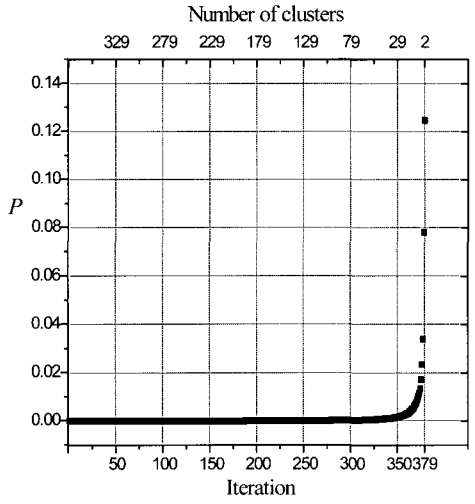
(b) gas furnace data(u(t-4),y(t1))

Fig. 4. Scatter plot of sample data set and Box and Jenkin`s gas furnace data(u(t-4),y(t-1))

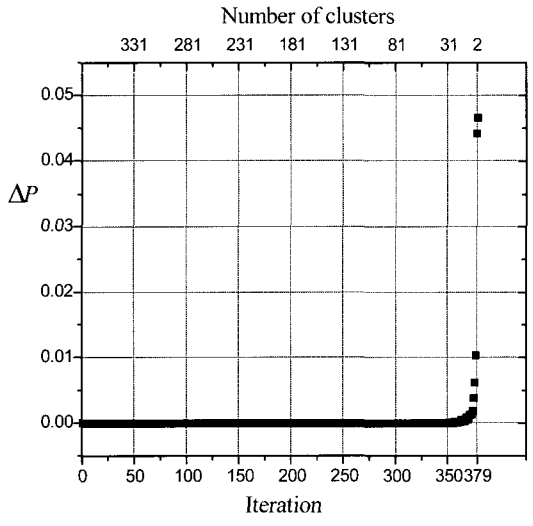


(c) Selected range

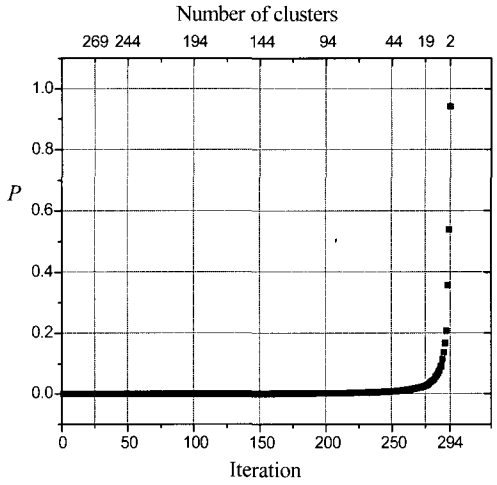
Fig. 5. Simulation result of sample data set in figure 4(a)



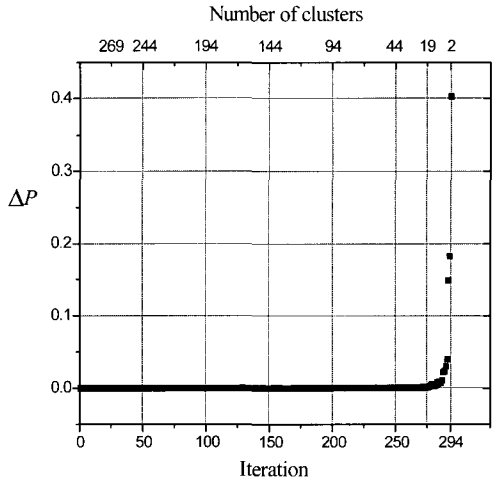
(a) P



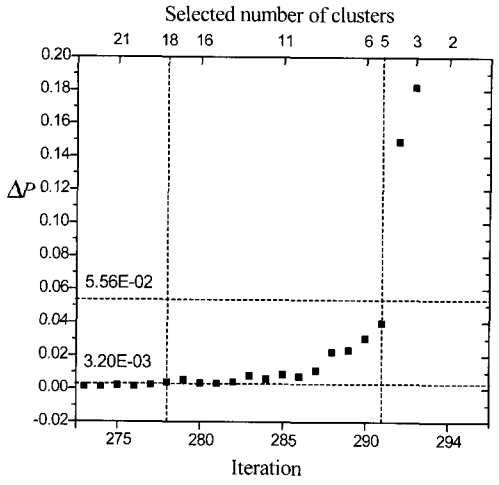
(b) ΔP



(a) P



(b) ΔP



(c) Selected range

Fig. 6. Simulation result of gas furnace data

In the figure 5(a), intuitional measure of the number of clusters is 5 but this data set also has several sub groups in data. As a consequence of simulation, possible range of the number of clusters is from 5 to 18. In the figure 6 (c) shows selected number of clusters. Possible range of the number of clusters is from 5 to 18. This result includes same number of clusters in joo's [13] study, which is 5 clusters.

5. Discussion and concluding remarks

In this paper, an empirical study of the data pattern estimation method is presented and its numerical experiment is carried out. Purpose of this method is detecting possible number of data patterns as it mentioned earlier. As it shown in simulation results, when data set have distinct number of clusters and shows clear separation, proposed method can detect exact number of clusters. On the other hand, if there is a underlying data patterns, proposed method indicate a range of numbers clusters by inherent characteristic of the ΔP_{mean}(average or center of gravity), but ΔP_{mean} can be too much dependent to under clustered values of ΔP_s. However, we believe that Ps and ΔP_s shows a crucial information of the underlying data patterns. In the future research, more precise criterion to determine the adequate range estimation of the number of clusters will be considered.

References

[1]. A. Jane and R. Dubes, *'Algorithm for Clustering Data'*. Prentice-hall, NJ., 1998.

[2]. J-S. R. Jang, C. T. Sun and E. Mizutani, *'Neuro-Fuzzy and Soft Computin'*, Prentice-hall, NJ., 1997.

[3]. X. L. Xie and G. A. Beni, "Validity Measure for Fuzzy Clustering", *IEEE Trans. on Pattern Analysis and Machine Intelligence.*, Vol. 3, No. 8, pp. 7-31, 1991.

[4]. Y. Fukuyama and M. Sugeno, "A New Method of Choosing the Number of Clusters for the Fuzzy C-means Method", *Proc. of 5th Fuzzy Systems Symposium.*, pp. 247-250, 1989.

[5]. J. C. Dunn, "A Fuzzy Relative of the ISODATA Process and its Use in Detecting Compact Well-Sperated Clusters", *J. Cybern.*, Vol 3, pp. 32-57, 1973.

[6]. I. Gath and A. B. Geva, "Unsupervised Optimal Fuzzy Clurstering", *IEEE Trans. on Pattern Alalysis and Machine Intelligence*, Vol. 11, No. 7, pp.773-781, 1989

[7]. N. R. Pal and J. C. Bezdek, "On Cluster Validity for the Fuzzy C-means model", *IEEE Trans. on Fuzzy Systems*, Vol. 3, No. 3, pp. 370-379, 1995.

[8]. S. H. Kwon, "Cluster Validity Index for Fuzzy Clustering", *IEE Electronic Letters.* Vol. 34., No. 22, pp. 2176-2177, 1998.

[9]. A. O. Boudraa, "Dynamic Estimation of Number of Clusters in Data Set", *IEE Electronic Letters*, Vol. 35, No. 19, pp. 1606-1607, 1999.

[10]. H. Sun, S. Wang and Q. Jiang, "A New Validation Index for Determining the Number of Clusters in a Data Set", *Proc. IJCNN '01, Neural Networks*, Vol. 3, pp. 31582-1857, 2001.

[11]. M. Sugeno and T. Yasukawa, "A Fuzzy Logic Based Approach to Qualitative Modeling", *IEEE Trans. on Fuzzy Systems*, Vol. 1, No. 1, pp.7-31, 1993.

[12]. N. R. Pal, J. C. Bezdek and T. A. Runkler, "Some Issues in System Identification using Clustering", *Int. Conference on Neural Networks*, Vol. 4, pp. 2524-2529, 1997.

[13]. Y. H. Joo, H. S. Hwang, K. B. Kim and K. B. Woo, "Linguistic Model Identification for Fuzzy System", *IEE Electronic Letters*, Vol. 31, No. 4, pp. 330-331, 1995.

[14]. G. E. P. Box and G. M Jenkins, *'Time Series Analysis-Forecasting and Control'*. Holden-Day Inc., 1976.

Tae-Chon Ahn

was born in Inchon, Korea on October 11, 1955. He received the B.S., M.S. and Ph.D. degrees in Electrical Engineering from the Yonsei University, Korea in 1978, 1980 and 1986, respectively. From 1987 to 1988, he was a visiting scholar in the department of Control System Engineering at Upscale University, Sweden. From 1996 to 1997, he was a visiting professor in the department of electrical and computer engineering at Georgia Institute of Technology, USA. Since 1981, he has been with the School of Electrical and Electronics Engineering at Wonkwang

University, Korea, where he is currently a Professor. He is engaged in research on digital granular control, real-time data measuring and processing, FPGA design, intelligent systems and control of synchronous machines.
Phone : +82-63-850-6344
Fax : +82-63-853-2196
E-mail : tcahn@wonkwang.ac.kr

Kyung-Won Jang was born in Korea in 1974. He received the B.S. and M.S. degrees in control and instrumentation engineering from Wonkwang University, Iksan, Korea, in 2001 and 2003, respectively. He is currently working toward the Ph. D. degree in control and instrumentation engineering at Wonkwang University. His current research deals with the model based controller design, pattern recognition, granular computing and its application.
Phone : +82-63-850-6344
Fax : +82-63-853-2196
E-mail : jaang@wonkwang.ac.kr

Dong-Du Shin was born in Korea in 1979. He received the B.S. degree in in control and instrumentation engineering from Wonkwang University, Iksan, Korea, in 2005. He is currently working toward the M. S. degree in control and instrumentation engineering at Wonkwang University. He is engaged in research on embedded systems and microprocessor systems.
Phone : +82-63-850-6344
Fax : +82-63-853-2196
E-mail : shindu@wonkwang.ac.kr

Yang-Woong Yoon was born in Jeonju, Korea, on September 25, 1941. He received the B.S., M.S., Ph. D. degrees in Electrical Engineering of Chonbuk National University, Korea in 1967, 1974 and 1985, respectively. Since 1979, he has been with the School of Electrical and Electronics Engineering at Wonkwang University, Korea, where he is currently a Professor. He is engaged in research on Electric machinery and control of synchronous machines.
Phone : +82-63-850-6732
Fax : +82-63-853-2196
E-mail : ywyoon@wonkwang.ac.kr

Hak-Soo Kang was born in Iksan, Korea. He received the B.S and M.S. degrees in Electrical Engineering of Wonkwang University, Korea in 1984 and 1986, respectively. Since 1987, he has been with the Dept. of Digital & Electrical Information Engineering, Vision College of Jeonju, where he is currently a Professor. He is engaged in research on Electric machinery, control of synchronous machines and microprocessor.
Phone : +82-10-4690-7698
Fax : +82-63-853-2196
E-mail : hskang@wonkwang.ac.kr