World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:1, No:11, 2007

Multi-Agent Systems for Intelligent Clustering

Jung-Eun Park, Kyung-Whan Oh

Abstract**—** Intelligent systems are required in order to quickly and accurately analyze enormous quantities of data in the Internet environment. In intelligent systems, information extracting processes can be divided into supervised learning and unsupervised learning. This paper investigates intelligent clustering by unsupervised learning. Intelligent clustering is the clustering system which determines the clustering model for data analysis and evaluates results by itself. This system can make a clustering model more rapidly, objectively and accurately than an analyzer. The methodology for the automatic clustering intelligent system is a multi-agent system that comprises a clustering agent and a cluster performance evaluation agent. An agent exchanges information about clusters with another agent and the system determines the optimal cluster number through this information. Experiments using data sets in the UCI Machine Repository are performed in order to prove the validity of the system.

Keywords**—** Intelligent Clustering, Multi-Agent System, PCA, SOM, VC(Variance Criterion)

I. INTRODUCTION

IN the Internet based analysis system, the need for an agent based model for analysis is discussed. In particular, intelligent systems can perform more intelligence data mining and can find pertinent patterns from vast quantities of data. Clustering, one of the methods of data mining, plays a role in binding with similarly object in Internet transactions(e.g., users, web documents)[1]. Most events occur at real time in the Internet environment. This means that the analyzer has difficulty performing clustering analysis and analyzing the results continuously. For this reason, an intelligent agent that takes on most of the analyzer's tasks and that undertakes automatic clustering is required. In this study, the approach to an automatic clustering intelligent system is a multi-agent system that comprises a clustering agent and a cluster performance evaluation agent. The agent exchanges information about clusters with other agents and the whole system determines the optimal cluster number using this information. The cluster performance evaluation of the proposed system used data from the UCI Machine Repository[2] and artificially created data (artificial synthetic data) was also used for the performance evaluation. Studies related to the proposed system are discussed in section 2. Section 3 introduces the integrated design and process for the proposed system. In section 4, experiments and results relating to the proposed methods are introduced. Finally, section 5 sets out the conclusions of this study and recommended areas for future study.

This research was performed for the Intelligent Robotics Development Program, one of the 21st Century Frontier R&D Programs funded by the Ministry of Commerce, Industry and Energy of Korea.

J. P. Author is with the department of Computer Science, Sogang University, Seoul, 121-742, Korea(South) (corresponding author to provide phone: 82-2-703-7626; fax: 82-2-704-8273; e-mail: fayemint@empal.com).

K. O. Author received the B.S. degree in Mathematics from Sogang University, Seoul, 121-742, Korea(South), in 1978, and the M.S. and Ph.D. degrees in Computer Science from Florida State University, Tallahassee, FL 32306, USA, in 1985 and 1988, respectively. He is currently with the department of Computer Science at Sogang University, Seoul, 121-742, Korea(South), where he is a Professor (e-mail:kwoh@sogang.ac.kr).

II. RELATED WORKS

A. Principal Component Analysis

1) Principal Components

Principal component analysis(PCA) is concerned with explaining the variance-covariance structure through a few linear combinations of the original variables. Its general objectives are data reduction and interpretation.[3] Algebraically, principal components are particular linear combinations of the random variables $X_1, X_2, \ldots, X_p$.[4] Geometrically, these linear combinations represent the selection of a new coordinate system obtained by rotating the original system with $X_1, X_2, \ldots, X_p$ as the coordinate axes. The new axes represent the directions with maximum variability and provide a simpler and more parsimonious description of the covariance structure.[5] As is noted below, principal components depend solely on the covariance matrix Σ with eigenvalues $\lambda_1 \geq \lambda_2 \cdots \geq \lambda_p \geq 0$. Consider the linear combinations

$$\begin{aligned} Y_1 &= l'_1 X = l_{11}X_1 = l_{21}X + \cdots + l_{p1}X_p \\ Y_2 &= l'_2 X = l_{12}X_1 = l_{22}X + \cdots + l_{p2}X_p \\ &\vdots \\ Y_p &= l'_p X = l_{1p}X_1 = l_{2p}X + \cdots + l_{pp}X_p \end{aligned} \quad (1)$$

Then, using the linear combinations,

$$Var(Y_i) = l'_i \Sigma l_i \quad i = 1, 2, \ldots, p \quad (2)$$

$$Cov(Y_i, Y_k) = l'_i \Sigma l_k \quad i, k = 1, 2, \ldots, p \quad (3)$$

The principal components are those uncorrelated linear combinations Y1, Y2, …, Yp whose variances in Equation (2) are as possible. The first principal component is the linear

combination with maximum variance; that is, the linear combination maximizes Var(Yi) = l'i Σli. It is clear that Var(Yi) = l'i Σli can be increased by multiplying any l1 by a particular constant. To eliminate this indeterminacy, it is convenient to restrict the discussion to coefficient vectors of unit length. Therefore, following Equations (4) to (6) can be states:

First principal component = linear combination l'_1X that maximizes

$$Var(l'_1X) \quad subject\ to \quad l'_1l_1 = 1 \tag{4}$$

Second principal component = linear combination l'_2X that maximizes

$$\begin{aligned} &Var(l'_2X) \quad subject\ to \quad l'_2l_2 = 1 \\ &and \quad Cov(l'_1X, l'_2X) = 0 \end{aligned} \tag{5}$$

ith principal component = linear combination l'_iX that maximizes

$$\begin{aligned} &Var(l'_2X) \quad subject\ to \quad l'_2l_2 = 1 \\ &and \quad Cov(l'_1X, l'_2X) = 0 \end{aligned} \tag{6}$$

2) Criterion for determining the number of principal components

The criterion for determining the number of principal components can be decided by a contribution degree about total variance. This determines the minimum number of components that account for over a particular(e.g., 80~90%) fraction of whole variance. The size of the eigen-values can be used. This criterion means that the eigen-value that is ascribed to the principal components must be at least more than one. This is known as Kaiser's criterion.[6]

B. Self-Organizing Feature Maps

1) Kohonen Networks

The Self-Organizing Map (SOM) developed by professor Kohonen can be categorized in the competitive learning network.[7] It is based on unsupervised learning, and provides a topology preserving mapping from high dimensional space to map units. Map units, or neurons, usually form a two-dimensional lattice. Thus, the mapping process involves mapping from high dimensional space onto a two-dimensional plane. The characteristic of topology preservation means that the mapping preserves the relative distance between the points. Points that are close to each other in the input space are mapped to nearby map units in the SOM. Thus, the SOM can serve as a cluster analyzing tool for high-dimensional data. The SOM also has capability to generalize data. This generalization capability means that the network can recognize or characterize inputs never encountered before. A new input is assimilated with the map unit it is mapped to. The SOM is a two-dimensional array of neurons as characterized by Equation (7):

$$M = \{m_1, K, m_{p\times q}\} \tag{7}$$

One neuron is a vector called the codebook vector as set out in Equation (8):

$$m_i = \{m_{i1}, K, m_{in}\} \tag{8}$$

This has the same dimension as the input vectors (n-dimensional). The neurons are connected to adjacent neurons by a neighborhood relation.[8]

2) Self-Organizing Feature Maps Algorithm

The Kohonen SOM algorithm can be expressed in the steps set out below.[9]

- ***Step 1*** : Initialization:
  Choose random values for the initial weights.
- ***Step 2*** : Winner finding:
  Find the winner neuron j^* at time k, using the minimum-distance criterion:

$$j^* = \arg\min_j \| x(k) - w_j \| \quad , \quad j = 1, K, N^2$$

where $x(k)$ represents the kth input pattern and $\|.\|$ indicates the Euclidean norm.

- ***Step 3*** : Weight updating:
  Adjust the weights of the winner and its neighbors, using the following rule:

$$w_j(k+1) = \begin{cases} w_j(k) + \eta(k)(x(k) - w_j(k)) & if \quad j \in N_{j^*}(k) \\ w_j(k) & o.w. \end{cases}$$

where, $\eta(k)$ is a positive constant and $Nj^*(x)$ is the neighborhood set of the winner neuron j^* at time k.

The above steps are *repeated* until satisfaction of the given conditions.

3) A dimension decision problem of feature map

A Kohonen neural network has problems as well as advantages. One of the problems is that a dimension of a feature map must be decided subjectively. If a dimension grows large gradually, the cluster is increased.[10] On the other hand, if a dimension grows small gradually, the cluster is decreased. A solution to this problem of the SOM involves the use of Principal Component Analysis. This is set out in the following section for a dimension decision problem of the feature map.[11]

III. Integrated Design of Proposed System

A. Clustering using Multi-Agent system

Clustering plays a role in binding similar objects in Internet transactions. Most events occur at real time in the Internet environment. Thus, the analyzer has difficulty taking part in clustering and in analyzing results continuously. For this reason, an intelligent agent that takes on most of the analyzer's task is required and this agent undertakes automatic clustering.[12] A multi-agent system for intelligent automatic clustering is proposed here. This system is composed of a clustering agent and a performance evaluation agent. Optimal cluster results are obtained by two agents which automatically exchange information.[13]

1) Clustering Agent

The clustering agent performing automatic clustering used the self-organization feature map as the clustering algorithm. This is because a self-organization feature map enables very fast clustering and is suitable for the clustering of real time Internet data. The optimal dimension was determined by using principal component analysis among the multivariable statistical methods in order to solve the problem that the number of optimal cluster grows large if the dimension of the feature map grows large in the self-organizing feature map. The number of a general cluster is determined by means of a scatter diagram which uses principal components of 2 or 3 by a principal component analysis of original data. This is used to form the dimension of the feature map as set out in Equation (9):

$$FS_{SOM} = (CS_{PCA}) \times (CS_{PCA}) \tag{9}$$

In Equation (9), FS_{SOM} is the feature map dimension number of SOM; CS_{PCA} is the cluster number decided by the scatter diagram which has the principal component as an axis. Equation (9) was heuristically obtained through a lot of experimentation for the purposes of this paper.

2) Performance Evaluation Agent

A clustering performance evaluation agent takes charge of performance evaluation of the clustering results. If the evaluation reveals poor performance, the performance evaluation agent requires the clustering agent to perform a better cluster analysis. Performance evaluation of the clustering results is performed through the Variance Criterion *(VC)* proposed in this paper. This criterion decides the best result as the minimum value in variance for continuous variables. Generally, objects belonging in one cluster are very similar within the cluster and are different with objects belonging in other clusters; this is the basis for using this criterion. Moreover, there is a penalty for restricting the number of clusters in the *VC*. This means that analysis with excessive number of clusters is meaningless. The *VC* consists of a continuous cluster variable and a penalty is given when the number of clusters increases. The *VC* measurement proposed this paper is defined as follows:

$$VC_M = \sum_{i=1}^{M} v_i / M + 0.1 \times M \tag{10}$$

In Equation (10), M is the number of clusters; v_i is average value of the standard derivation for *ith* cluster. $0.1 * M$ in the second term is the penalty according to the number of clusters. This means that the smaller the value of the equation, the better the quality of the cluster.

3) MAS-IC System

The Multi-Agent System for Intelligent Clustering (MAS-IC), including a clustering agent and a clustering performance evaluation agent, has system architecture set out in Figure 1.

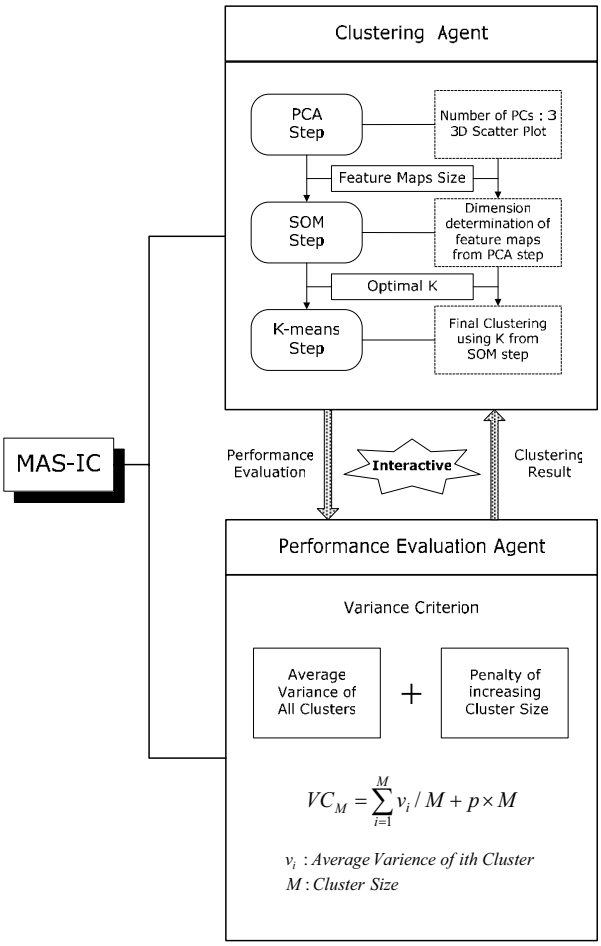

Fig. 1 System Architecture of MAS-IC

As for the MAS-IC, a clustering agent and a clustering performance evaluation agent are charged to each other with the interactive structure set out in Figure 1. The clustering agent carries out only clustering and the performance evaluation agent performs only performance evaluation.

B. Necessity of automatic clustering

In this paper, intelligent clustering is performed as shown in Figure 2. First, the scatter diagram for the entire data set is drawn by PCA. In this time, three main principal components are used. The reason for drawing 3-D scatter diagrams using three components is that 3-D is the limit for visible scatter diagrams. The overall cluster of the data is determined by this scatter diagram which also determines the dimension of the feature map in SOM. In this paper, the dimension of the feature map in SOM is determined as *(cluster*cluster)* by the scatter diagram. This determination was made heuristically through many experiments for the purposes of this study. By using SOM, the unsupervised learning process for determining the optimal cluster number is performed. After the optimal cluster number is determined, the number is used as initial cluster number *(K)* for K-means. A clustering agent takes charge of the above steps. Next, a cluster evaluation agent estimates the quality of the cluster using the variance criterion.

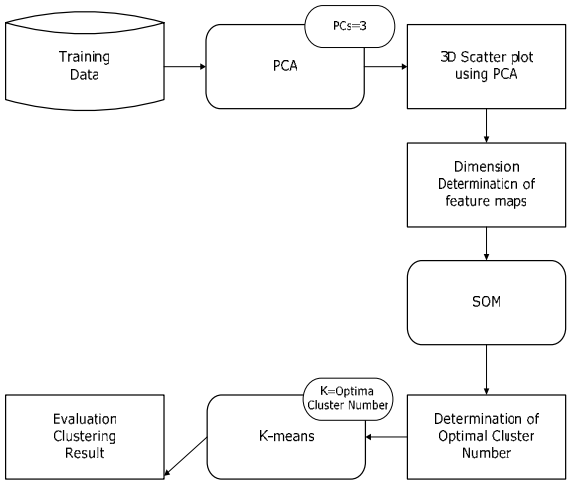

Fig. 2 Automatic Clustering Procedure using MAS-IC

IV. Results of Experiments and Discussion

A. Experiments and results using artificial synthetic data

Artificial synthetic data was created for the experiments of the proposed method by using five input variables. The following is a pseudo-code for the creation of the artificial synthetic data used in the experiments.

```
DO G=1 TO 5;
IF      G=1 THEN DO; X1=1;  X2=4;  END;
ELSE IF G=2 THEN DO; X1=1;  X2=1;  END;
ELSE IF G=3 THEN DO; X1=6;  X2=1;  END;
ELSE IF G=4 THEN DO; X1=15; X2=5;  END;
ELSE             DO; X1=3;  X2=20; END;
DO N=1 TO 20;
         X3=X1+RANNOR(12345);
         X4=X2+RANNOR(12345);
         X5=RANUNI(12345);
         OUTPUT;
         END;
END;
```

A summary of the artificial synthetic data created by the above pseudo-code is set out in Table 1.

TABLE I
SUMMARIZED INFORMATION

	Min	*Max*	*Mean*	*SD*
X_1	1	15	5.2	5.2570
X_2	1	20	6.2	7.1181
X_3	0.7540	16.4736	5.1760	5.2357
X_4	-1.3980	21.7720	6.0633	7.2751
X_5	0.0024	0.9962	0.4847	0.2802

Table 1 shows the Min, Max, Mean, and SD for about five input variables. The results of a PCA carried out using this data are shown in Table 2. The eigen-value and accumulative explanative information were obtained for about three principal components. Table 2 shows that three principal components explained about 99.78% of all data.[5]

TABLE II
PCA RESULT OF ARTIFICIAL SYNTHETIC DATA

Component	*eigenvalue*	*accumulation*
1	2.1504	0.5376
2	1.8251	0.9939
3	0.0159	0.9978

Figure 3 shows a 3-dimensional Scatter Diagram for three principal components which use the PCA results of Table 2 concerning the artificial synthetic data which had five input variables.

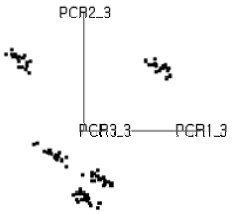

Fig. 3 3D Scatter Diagram of 3 principal components with Artificial synthetic data

Figure 3 indicates that five clusters form for all the data. Therefore, the feature map in Figure 4 is decided by SOM (5*5) using the results in Figure 3. Figure 4 shows that the clustering performance results of SOM have a (5*5) feature map dimension.

Dark nodes which are colored darker than others contain relatively more objects. This confirms that a lot of objects are gathering compared to data in the other four nodes. Therefore, as for the clustering performance evaluation agent, it was decided that for this experiment the optimal cluster number was 4. Table 3 sets out the K-means clustering results for four clusters. The variance and homogeneity of each cluster can be determined from this table.

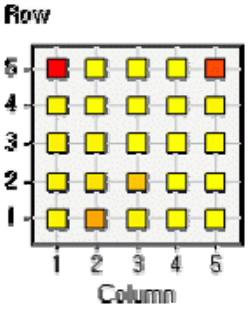

Fig. 4 (5*5) SOM learning results of Artificial Synthetic data

TABLE III
K-MEANS(K=4) RESULT OF ARTIFICIAL SYNTHETIC DATA

Cluster	X_1	X_2	X_3	X_4	X_5	*Avg.*
1	0.000	0.000	0.918	1.279	0.286	0.4966
2	0.000	0.000	1.072	0.721	0.277	0.4140
3	0.000	0.000	0.914	0.945	0.278	0.4274
4	0.000	0.506	0.852	1.939	0.268	0.7130

Four variances for each cluster are calculated on each of the five input variables, and the standard derivation of each cluster for five input variables was calculated with the last. The *VC* value, the clustering performance evaluation measurement, can be calculated for these values and is as follows:

$$VC_4 = \frac{(0.4966 + 0.4140 + 0.4274 + 0.7130)}{4} + 0.1 \times 4 = 0.9128 \quad (11)$$

The *VC* values of each cluster in Artificial Synthetic data are shown in Table 4.

TABLE IV
CLUSTER NUMBERS AND VC VALUES

Cluster numbers	*VC values*
3	1.0139
4	0.9128
5	0.9343
6	1.0076
7	1.1013

Table 4 clearly indicates that the VC value of the four clusters decided by the proposed system is the smallest.

B. Experiments and results using Iris data

The Iris Plants Database consists of 150 learning data sets. There are four input variables - X_1 (sepal length in cm), X_2 (sepal width in cm), X_3 (petal length in cm) and X_4 (petal width in cm) - and these variables are decided on a kind of an iris. Table 5 shows the simple characteristics of these data.[2]

TABLE V
SUMMARIZED INFORMATION OF IRIS DATA

	Min	*Max*	*Mean*	*SD*
X_1	4.30	7.90	5.84	0.83
X_2	2.00	4.40	3.05	0.43
X_3	1.00	6.90	3.76	1.76
X_4	0.10	2.50	1.20	0.76

First of all, the results obtained by a principal component analysis of the Iris data, and the four input variables, are shown in Table 6. This shows that three principal components explained about 99.48% of all data. Therefore, there is hardly any loss of information along a dimension reduction.

TABLE VI
PRINCIPAL COMPONENT ANALYSIS OF IRIS DATA

Component	*eigenvalue*	*Accumulation(%)*
1	2.910818	72.77
2	0.921221	95.80
3	0.47353	99.48

Figure 5 shows a 3-dimensional Scatter Diagram for all data using three principal components. Three clusters can be observed. Therefore, the feature map in Figure 6 is decided by SOM (3*3) using the results in Figure 5.

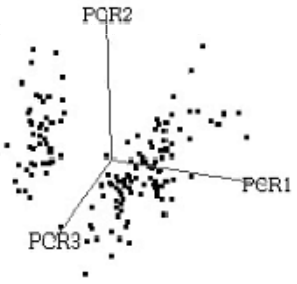

Fig. 5 3D Scatter Diagram of 3 principal components with Iris data

The clustering result of SOM to have a (3*3) feature map dimension is shown in Figure 6.

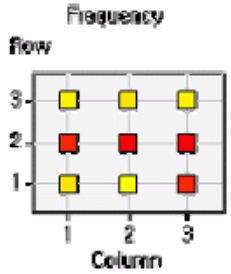

Fig. 6 (3*3) SOM learning results of Iris data

Dark color nodes among the nine nodes of Figure 6 are the

nodes where clusters formed. According to the figure, three clusters were formed. Table 7 sets out the K-means results that clustered with the Iris data with a k value of 3.

TABLE VII
K-MEANS(K=3) RESULTS OF IRIS DATA

Cluster	*X1*	*X2*	*X3*	*X4*	*Avg.*
1	0.50	0.29	0.46	0.24	0.37
2	0.35	0.38	0.17	0.11	0.25
3	0.48	0.30	0.54	0.30	0.41

Table 7 shows the variance value of each input variable about each cluster, and the last column (Avg.) shows the standard derivation of each cluster. The VC value about three clusters can be calculated from this table as set out below.

$$VC_3 = (0.37 + 0.25 + 0.41)/3 + 0.1 \times 3 = 0.64 \quad (12)$$

The *VC* values for each set of clustering results calculated from the results set out in these tables are displayed in Table 8.

TABLE VIII
THE VC VALUES OF EACH CLUSTERING RESULTS

Cluster numbers	*VC values*
2	0.7
3	0.64
4	0.69
5	0.7

It is shown in Table 8 that the *VC* value of the three cluster decided by the proposed system is the smallest. Therefore, the cluster results obtained by the proposed clustering system are shown to be optimal.

V. CONCLUSION AND FUTURE WORK

The number of optimal clusters is decided by principal component analysis and by a self-organizing feature map in a Multi-Agent System for Intelligent Clustering proposed in this paper. The optimal cluster uses a K-means algorithm. This K value is based on the results of the SOM derived from the PCA. In this paper, a variance criterion is proposed and applied in the performance evaluation of cluster results. The analysis process was performed with two agents. It is shown that the clustering method proposed in this paper can obtain better homogeneity than previously proposed clustering methods. If a multi-agent system for intelligent clustering as designed in this paper is dynamically applied to a web mining process, it is submitted that it will show outstanding performance.

In the future, the development of an intelligent agent for prediction will make it possible to establish an intelligent multi-agent system that includes supervised learning and unsupervised learning.

REFERENCES

[1] M. Ankerst, M. Breunig, H. P. Kriegel, J. Sander, "Optics: Ordering points to identify the clustering structure" In *Proc. ACM SIGMOD Int. Conf. on Management of Data (SIGMOD'99)*, 1999, pp. 49–60.
[2] http://www.ics.uci.edu/~mlearn/MLRepository.html
[3] Lindsay I Smith, "A tutorial on Principal Components Analysis", February, 2002 [Online] Available: http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf
[4] I.T. Jolliffe, 2002, Principal component analysis 2nd ed., Springer
[5] A. R. Johnson, D. W. Wichern, Applied Multivariate Statistical Analysis, 356~357, 1992, Prentice Hall,
[6] M. Emre Celebi, Y. Alp Aslandogan, "Content Based Image Retrieval Incorporating Models of Human Perception", in *IEEE International Conference on Information Technology, Coding and Computing*, Las Vegas, NV, April 2004.
[7] T. Kohonen, Self-Organizing Maps, 1995, Springer
[8] J. Vesanto, J. Himberg, " SOM Toolbox For Matlab 5 ", *Espoo 2000*, 2000
[9] S. H. Jun, J. Yang, K. W. Oh, "Automatic Determination of Cluster Size Using Machine Learning Algorithms", *SSGRR 2002*, 2002
[10] J. Friedman, "On bias, variance, 0/1-loss and the curse of dimensionality", *Data Mining and Knowledge Discovery*, 1:55–77, 1997.
[11] J. Hollmén, "Process Modeling Using the Self-Organizing Map", 1996, [Online] Available: http://citeseer.ist.psu.edu/458515.html
[12] G. Karypis, H. Han, V. Kumar, "Chameleon: A hierarchical clustering algorithm using dynamic modeling", IEEE Computer, 32(8):68–75, August 1999
[13] J.E. Park, S.H. Jun, K.W. Oh., " Intelligent Data Mining Agents for Automatic Clustering", in *Conf. KIISS2002 Conf.*, Seoul(Korea), 2002, pp. 370–376

Jung-Eun Park was born in Seoul, Korea, in 1978. She received the B.S. degree in computer science from SungKongHoe University, Seoul, Korea in 2001 and the M.S. degree in computer science from Sogang University, Seoul, Korea in 2003. She has been a Ph.D. candidate in the Department of Computer Science, Sogang University, Seoul, Korea.

Her published articles are as follow :

[1] "Intelligent data mining agents for automatic clustering", in *Conf. KIISS2002 Conf.*, Seoul(Korea), 2002, pp. 370–376

[2] "Discretization of continuous-valued attributes considering data distribution", Journal of Fuzzy Logic and Intelligent Systems, vol. 013, n.004, pp.391–396, 1225–1127, Aug. 2003.

[3] "Global search strategy using enhanced bacteria chemotaxis algorithm", in *Conf. KISS2005 Conf.*, Seoul(Korea), Nov. 2005, VOL. 32 NO. 02, pp. 0790–0792.

Her research interests include Evolutionary Algorithm, Intelligent Robot, Reasoning about Uncertainty, Data Mining.

Ms.Park is now a regular member of KISS(The Korea Information Science Society).