Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.18576/amis/110328

Intelligent Customer Segmentation System Using Hybrid of Artificial Immune Network and Particle Swarm Optimization Algorithm

R. J. Kuo[1,*] *and S. S. Chen*[2]

[1] Department of Industrial Management, National Taiwan University of Science and Technology, No. 43, Section 4, Kee-Lung Road, Taipei, Taiwan, ROC
[2] Tsmc Solid State Lighting Co., Ltd. No. 9, Li-Hsin 4th Road, Hsinchu Science Park, Hsinchu, Taiwan, ROC

Received: 2 Mar. 2017, Revised: 13 Apr. 2017, Accepted: 19 Apr. 2017
Published online: 1 May 2017

Abstract: This study attempts to propose a novel algorithm, hybrid of artificial immune network (aiNet) and particle swarm K-means optimization (PSKO)(aiNet-PSKO) algorithm, for cluster analysis. In order to verify the proposed methods, four benchmark data sets, Iris, Glass, Wine, and Breast Cancer, are first employed. The computational results indicate that aiNet-PSKO algorithm outperforms artificial immune system and particle swarm optimization related algorithms. Thereafter, these methods are further applied to the transaction database for an internet florist. The results also show that the aiNet-PSKO algorithm has the lowest sum of Euclidean distance value. The results can be used to make the marketing strategies for different clusters in order to provide different products or services customers prefer.

Keywords: Cluster analysis, Artificial immune network, Particle Swarm Optimization Algorithm, K-meansalgorithm, RFM analysis.

1 Introduction

Due to fast growth of information technology, a plenty of data are stored and processed by the enterprises. Therefore, how to effectively apply these data is a very important issue for the enterprises. Basically, data mining techniques can help us analyze the large database and extract some interesting or hidden patterns. Especially the cluster analysis, it is able to group the data into some clusters. Cluster analysis is a widely used technique, whose goal is to partition a set of patterns into disjoint and homogeneous clusters [29]. Basically, clustering algorithm groups the data with high similarity altogether and so minimizes the similarity among each cluster. It allows us to find how data spreads as well as their features. Through the process of reducing the complexity, clustering algorithm also allows us to analyze the hidden commercial implication under the clustered data and would be beneficial for business decision making. However, a challenge in data clustering is to determine the optimal number of clusters in the data set. Therefore, a clustering technique or model which can dynamically determine the number of clusters for a data set and which is computationally inexpensive will have an added advantage [11]. On the other hand, evolutionary algorithms (EAs) are heuristic and stochastic search procedures based on the mechanics of natural selection, genetics, and evolution, which allow them to find the global solution for a given problem [37]. Some of such algorithms are evolutionary strategy, like genetic algorithm (GA), particle swarm optimization (PSO), artificial immune system (AIS), etc. Therefore, this study considers the use of a kind of AISs, artificial immune network (aiNet) for cluster analysis. Therefore, this study attempts to propose a novel clustering algorithm which is an integration of aiNet and PSO and K-means (aiNet-PSKO). In order to testify the proposed aiNet-PSKO algorithm, four benchmark data sets, Iris, Wine, Glass, and Breast Cancer, were employed. The computational results are compared with those of AIS-based algorithms and PSO-based algorithms. Additionally, the Internet is well known of high degree of interaction with customers. It has become a new and effective channel for the enterprises [21]. Most

* Corresponding author e-mail: rjkuo@mail.ntust.edu.tw

enterprises hope to preserve their competitive advantage in the electronic commerce environment through running a well-established customer relationship management (CRM). Furthermore, to extract valuable information that contain commercial implication among the huge and complex transactional data collected from the Internet, the cluster analysis algorithm could be considered to apply where the result could support the decision making for business. Therefore, the proposed aiNet-PSKO algorithm was then applied to a real-world problem considering customer transactional database from a cyber store in Taiwan. The purpose is to apply the proposed aiNet-PSKO algorithm on the transaction database and carry out clustering according to customer values and attributes. The results can be applied as a CRM analysis for this enterprise to provide the appropriate recommendation for products and service. The remainder of this study is organized as follows. Section 2 briefly presents the necessary background, while the proposed aiNet-PSKO algorithm is proposed in Section 3. Sections 4 and 5 discuss the experiment analysis results using four benchmark data sets and the model evaluation results for a real-world customer segmentation problem, respectively. Finally, the concluding remarks are made in Section 6.

2 Literature Survey

This section will briefly present the literature survey regarding cluster analysis, artificial immune network (aiNet), and hybrid of aiNet and particle swarm optimization.

2.1 Cluster Analysis

Cluster analysis which is a very important technique in data mining can partition data into a certain number of clusters. A cluster is described by considering internal homogeneity and external differentiation, i.e., patterns in the same cluster should be similar to each other, while patterns in different clusters should not [43]. Basically, clustering algorithms can be divided into two categories: hierarchical and partitional clustering ([16]). Hierarchical clustering algorithms organize data into a hierarchical structure according to the proximity matrix. The benefit is that the results of hierarchical clustering are usually depicted by a binary tree or den diagram. Hierarchical clustering algorithms can be further classified as agglomerative and divisive methods.Many hierarchical clustering algorithms have been presented including CURE [12], BIRCH [46], ROCK [13], CHAMELEON [19], DIANA [20], AUTOCLUST [7]and AMOEBA [6]. Unlike to hierarchical clustering which produces a successive level of clusters by iterative fusions or divisions, partitional clustering assigns a set of objects into clusters with no hierarchical structure. These algorithms intend to minimize certain criteria, like square error function, and can therefore be treated as optimization problems. Partitional clustering aims to optimize cluster centroids as well as the number of clusters [15]. Partitional clustering algorithms are mainly classified into supervised and unsupervised clustering algorithms. The main difference is that supervised clustering algorithms need to pre-determine the number of clusters while the other does not. Since the algorithm proposed in this study belongs to supervised clustering algorithm, it will be discussed with more details. The most widely used supervised clustering algorithm is the K-means algorithm [9].Latter and Bezdek [2] developed a fuzzy version of the K-means algorithm, called the fuzzy C-means algorithm. Unlike K-means algorithm, the object can belong to all of the clusters with a certain degree of membership. Other modifications of fuzzy C-means algorithms are the possibilistic-means clustering algorithm [23], fuzzy c-shells [3], and hierarchical unsupervised fuzzy clustering [10]. In addition, evolutional algorithms are also employed to improve the performance of K-means. Clustering can be regarded as a category of optimization problems that uses evolutional algorithms, like GA, PSO or ant colony optimization (ACO) algorithms. Many techniques support this method, such as the genetically guided [14], the genetic K-means [22], Tabu search [1], the ant colony clustering [24], and the particle swarm K-means optimization (PSKO) algorithm [25].

2.2 Artificial Immune Network

In 1974, Jerne proposed the first mathematical model in artificial immune system, which initiated subsequent researches and discussions. [32] suggested that Artificial Immune System (AIS) is a novel intelligent algorithm method, inspired from the biological immune system. The main principle of AIS is to imitate artificial immune system, and AIS can acquire evolutionary learning capability by learning the biological protection principle. The Opt-aiNET derives from aiNet, which uses clonal selection and affinity maturation in the immune theory and the immune network theory to establish the network model as an important application model of the artificial immune system. Its advantages include noise acceptability, unsupervised learning and self-organization. The applications include data processing, optimization learning and fault diagnosis. [42] suggested that the optimal artificial immune network model is an adaptive process based on an immune principle to perform search and optimization. The procedures in the model include the fitness value evaluation made by antibody for objective function, clonal expansion antibody maturation and suppression. There are some characteristics for opt-aiNet suggested by [39].The population size will grow or reduce with

iteration. Clone produces similar antibodies by copying populations. The number of copies in each population is the same. Besides, the variation of cloned antibodies is related to the fitness value of the original populations. Basically, a better fitness value indicates smaller variation. The excessively similar antibodies are suppressed. The affinity of antibodies is smaller than preset threshold s. The average fitness value evaluation aims to explore development space. If there is a deviation from the last generation, the exploration can be continued, and the suppression phase can be performed until no deviation occurs. New cells will be added after suppression to prevent solution from falling on the regional optimal solution. [4] proposed the adaptive radius immune algorithm (ARIA), through which the most density information of the compressed data can be retained. The simulation experiment showed that the clustering effect of the method is better than aiNet clustering algorithm. [41] used aiNet model in the more complex file clustering field. Based on the immune network and mutation principle, aiNet algorithm has good clustering result, and better performance than hierarchical clustering method and K-means clustering method. [40] applied aiNet to cluster analysis for customer online shopping. The method is compared with hierarchical, fuzzy c-means and spectral analysis methods. The experiment has demonstrated aiNet clustering analysis yields better result. [17] combined AIS with SDM (Sparse Distributed Memories) to propose a model used for processing a mass of dynamic data. The model can follow up dynamic data set and memorize past clustering results. [38] suggested that AIS algorithm can be used for network mining of clustering of dynamic data strings, which are clicked on continuously in student course grouping. [44] applied clonal selection in basic artificial immune system to data clustering, suggested an easier algorithm, and used four simulation data sets and two reference data sets to test data clustering. The experimental result indicated that the data sets can be clustered correctly. Based on anti-spam technology, [45] proposed a new e-mail service system using artificial immune clustering method to filter all messages. [27] suggested self-adaptive capability of multiple-clone clustering algorithm based on the biological immune and clonal selection principle. The data simulation experiment can prove that the algorithm proposed is a rational and effective clustering method and can yield better result as compared to K-means. Based on immune control and immune selection mechanism in the biological immune system, [35] suggested a new data mining methodimmune Dominance Clonal Multi-objective Clustering algorithm (IDCMC). The parent population antibody is divided into three sub-antibody sets. Different measurement methods and evolution selection strategies are used. Through simulation experiment, this algorithm has better clustering performance than those of GA and K-means. [5] suggested that hybrid of artificial immune system and ant algorithm can be used for air-conditioner market segments of 3*C* shopping. The experiment proved that the hybrid method has better clustering result than the SOM clustering method. [26] applied artificial immune network to cluster analysis. The results indicated that it is superior to PSO and AIS related methods.

2.3 Hybrid of AIS and PSO

[33] suggested that PSO with global detectability and unique artificial immune theory can improve global performance and avoid early convergence. [47] and [8] indicated that hybrid of AIS and PSO can improve search capability, and can increase the rate of convergence and effectively prevent local optimal solution. In regard to hybrid of AIS and PSO, [33,47,8,34] performed PSO evolution first and then followed by AIS. [27] performed AIS first and then followed by PSO evolution.

3 Methodology

This section will propose a novel clustering algorithm based on hybrid of artificial immune network and particle swarm optimization. In artificial immune network and particle swarm K-means optimization algorithm (aiNet-PSKO), antibody is expressed by centroid of each cluster. For example, if data are divided into 3 groups, then K is equal to 3. Xi is the cluster center vector where $i = 1, 2, and 3$. Based on the different data dimensions, expression of the cluster centroid vector is shown in Figure 1. The aiNet-PSKO developed by this study uses

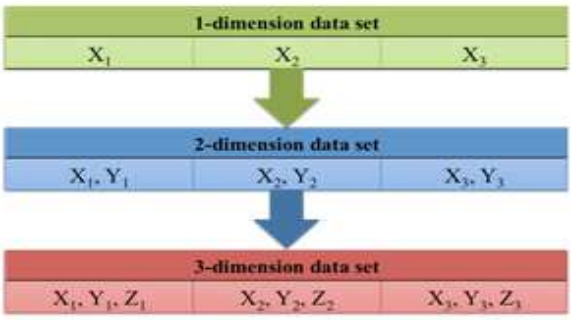

Fig. 1: Real value coding method of aiNet-PSKO algorithm.

aiNetK algorithm as the basic clustering model. In the process of aiNetK evolution, PSO evolution is added. Figure 2 shows flowchart of aiNet-PSKO algorithm and is described as follows.

Step 1:Set up related parameters

In aiNet-PSKO algorithm, the parameters include number of iterations, number of memory cells, M,

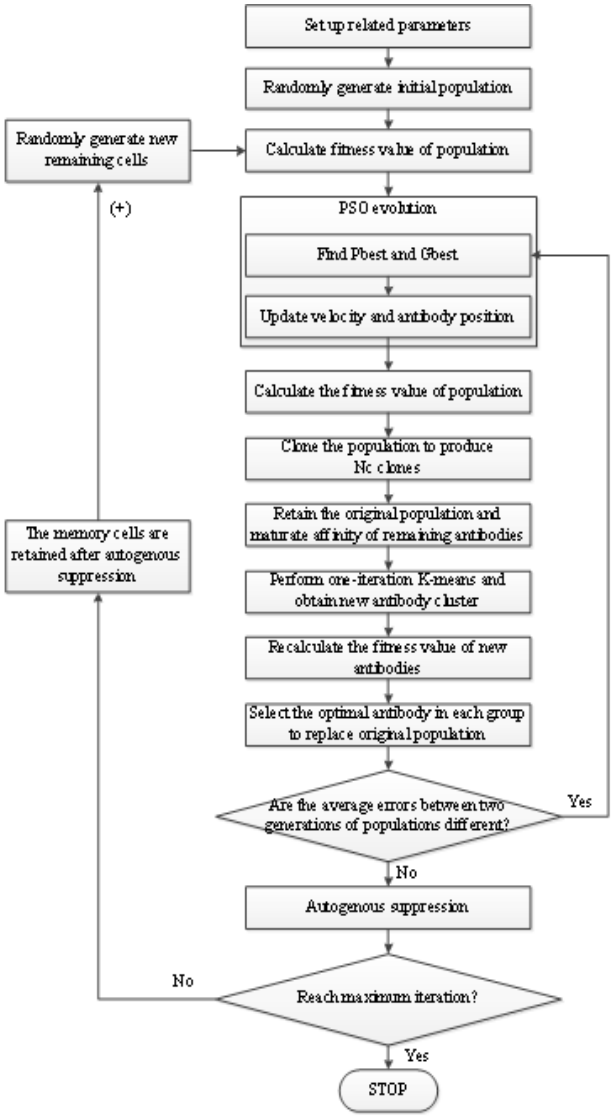

Fig. 2: Flowchart of aiNet-PSKO algorithm.

number of remaining cells, R, number of clones, Nc, optimal parameter of PSO (weight w, cl, and c2), error threshold between two generations and autogenous suppression threshold σ_s.

Step 2: Randomly generate initial population

At the beginning of iteration, M memory cells and R remaining cells are generated at random, and merged into initial population set P. Their relational expression is $P = M + R$. The population set P will gradually tend towards the optimal solution with evaluation in the subsequent steps. At the beginning of aiNet-PSKO algorithm, number of clusters K should be given. Each antibody consists of centroid vector of K clusters as follows:

$$P_{id} = (X_{i1},,X_{ij},,X_{ik}), i \in sizeof\,parent\,population set P and d \in sizeof\,cluster number K, \quad (1)$$

where P_{id} is the i^{th} parent population set.

Step 3: Calculate fitness value of population

This study uses Euclidean distance to calculate fitness value, and the calculation steps are as follows:

(i) Calculate Euclidean distance from each data X in each population to each cluster centroid as follows:

$$d(X_i, C_{ij}) = \|X_i - C_{ij}\| \quad (2)$$

(ii) In each population, each data X is assigned to the nearest cluster centroid and belongs to the corresponding cluster.

(iii) Calculate the fitness value of each antibody $(Ab)_i$ as follows:

$$(Ab)_i = \frac{1}{(1+D_i)}, 0 \le (Ab)_i \le 1; i \in P \quad and \quad (3)$$

$$D_i = SED_i = \sum_{j=1}^{K} \sum_{\forall X_i \in n_{ij}} \|X_i - C_{ij}\|, \quad (4)$$

where D_i: SED (Sum of Euclidean Distance) of the i^{th} antibody.

X_i: vector of i^{th} data.

n_{ij}: total quantity of data which belongs to the j^{th} cluster in the i^{th} antibody.

C_{ij}: Cluster centroid vector of the j^{th} cluster in the i^{th} antibody.

$\|X_i - C_{ij}\|$: Euclidean distance between each data vector and a centroid vector of each cluster, namely, Euclidean distance between data point X in the j^{th} cluster and its cluster centroid C_{ij}.

Step 4: Perform one-phase PSO evolution

One-phase PSO evolution steps are as follows:

(i) Find Pbest and Gbest.

(ii) Use the following two equations to update velocity and move antibody position:

$$\mathbf{V}_i(t+1) = w \times \mathbf{V}_i(t) + r_1 c_1 (\mathbf{x}_{pBest} - \mathbf{x}_i(t)) + r_2 c_2 (\mathbf{x}_{gBest} - \mathbf{x}_i(t)) \quad (5)$$

$$\mathbf{x}_i(t+1) = \mathbf{x}_i(t) + \mathbf{V}_i(t+1) \quad (6)$$

where $\mathbf{x}_i(t)$: position of the ith particle in the tth generation.

$V_i(t+1)$: velocity of the ith particle in the tth generation.

$\mathbf{x}_{pBest}$: position of optimal solution of each particle in historical generations.

$\mathbf{x}_{gBest}$: position of optimal solution of all particles.

w: parameter that controls impact of velocity of last iteration generation.

c_1, c_2: set acceleration constant between 1 and 3.

r_1, r_2: random value with uniform distribution between 0 and 1.

Step 5:Calculate the fitness value of population
In this phase, calculate the fitness value of population. The calculation in this phase is the same as that in the Step 3.

Step 6:Clone the population to produce Nc clones
N_c is set parameter. This study employed Taguchi parameter design to obtain the value in the subsequent sections. This step clones each original population to produce N_c clones. Thus, antibodies will have $P * N_c$ copies.

Step 7:Retain the original population and maturate affinity of remaining antibodies.
In this part, the original population needs to be retained first and then affinity of remaining antibodies can be maturated to prevent the antibodies after variation from being worse than the parameters before variation. The variation steps are as follows:

(i) Calculate mutation rate. The mutation rate is related to the fitness value. The mutation rate of the antibodies with higher affinity is smaller. The equation is presented by following equation:

$$Maturaterate = \alpha_i = \frac{1}{\rho} exp^{-(Ab^*)_i}; \tag{7}$$

where $(Ab^*)_i = \frac{(Ab)_i}{(Ab)_{max}}$, $(Ab^*)_i$ is a normalized affinity value, ρ is constant, and $i \in (P * N_c) - P$.

(ii) Maturate affinity using following equation:

$$c^{'} = c + \alpha i * N(0,1), \tag{8}$$

where $c^{'}$: antibody serial after affinity maturation.
$c = (c_1, , c_2, , c_x)$ stands for cloned antibody serial.
αi: mutation rate.
$N(0,1)$: standard normal distribution with the average value 0 and standard deviation 1.

Step 8:Perform one-iteration K-means and obtain new antibody cluster
Calculation steps of one-iteration K-means are as follows:

(i) Calculate Euclidean distance from each data X in each antibody to centroid of each cluster as follows:

$$d(X_i, C_{ij}) = \|X_i - C_{ij}\|; \tag{9}$$

(ii) In each antibody, each data X is assigned to the nearest cluster centroid and belongs to the cluster.

(iii) Calculate new cluster centroid vector in each antibody and obtain new antibody cluster as follows:

$$C_{ij}^{new} = \frac{1}{n_{ij}} \sum_{\forall X_i \in n_{ij}} X_i, \tag{10}$$

where C_{ij}^{new}: new centroid vector of the j-th cluster in the i^{th} antibody. $n_i j$: total quanity of data of the j-th cluster in the i^{th} antibody.

Step 9:Recalculate the fitness value of new antibodies.
The cluster centroids in the antibodies may change after affinity maturation and one-iteration K-means. Therefore, the fitness value of the new antibody should be calculated before selection of optimal antibody. The calcuation of the fitness value is the same as Step 3.

Step 10:Select the optimal antibody in each group to replace original population.
In this step, there are P groups of new antibody clusters, and each cluster has N_c clones. In order to obtain optimal antibody of each group, the antibodies in each group are prioritized in terms of the fitness value in selection. After prioritization, the first antibody (with the highest affinity) in each group is selected to replace original population P. At this time, the optimal antibody selected from each group may be the worse antibody in other groups.

Step 11:Determine whether average errors between two generations of populations have difference.

Identify whether average errors betweeb two generations of populations have difference. If the difference value (absolute value) is greater than error threshold, this implies that there is difference between two generations and the population has not been searched completely. It is necessary to return to Step 4. The threshold value should be pre-determined. This study employs Taguchi parameter design to obtain the value in the subsequent section. The calculation of population error is shown as follows:

$$Averagepopulationerror = Populationerror = \sum_{i}^{p} \frac{SED_i}{n}, \tag{11}$$

where n stands for number of population in this generation.

Step 12:Autogenous suppression The autogenous suppression aims to expand the search scope by deleting excessively similar antibodies and prevent regional optimal solution. This study arrays the optimal populations selected in the last step in descending order, and then calculate the Euclidean distance between pairwise antibodies, as shown in Eq. (12). If the distance is smaller than the autogenous suppression threshold s, the antibody with lower affinity is deleted, and the better antibody is retained

Table 1: Benchmark data sets.

Data Set	Cases	Feature Dimensions	Clusters
Iris	150	4	3
Wine	178	13	3
Glass	214	9	6
Breast Cancer	683	9	2

and used as memory cell M; memory cell M is multiplied by d% which is number of the remaining cells R. Thus, the number of memory cells and remaining cells in each iteration is different with autogenous suppression. The autogenous suppression s and d% are also parameters needing determination. This study employs Taguchi parameter design to obtain the values in the subsequent sections.

$$d(C_i, C_j) = \|C_i - C_j\|; \tag{12}$$

Step 13:Determine whether the number of iterations is statisfied or not.
If the preset iteration number is reached, then stop; Otherwise, go back to Step 3.

4 Simulational Results

This section will employ the benchmark data sets to verify the proposed clustering algorithm.

4.1 Data Set Collection

The four benchmark data sets used in this study are from the website of Department of Information and Computer Science at University of California. The benchmark data sets include Iris, Wine, Glass and Breast Cancer which are used to testify efficiency of the method proposed in this study. This study summarizes the benchmark data sets based on the data quantity, dimension and number of clusters, as shown in Table 1.

4.2 Taguchi Design of Experiments

For designing aiNet-PSKO algorithm, it is necessary to determine the related parameters including: size of population cell (N), multiples of clone (Nc), percentage (d) of new cells, suppression threshold (belta), initial weight (wei) and learning factors (c1 and c2). These 8 parameters are used as design factors of the Taguchi experimental design. Each factor has three levels. This study refers to the parameters in the past literatures. Table 2 shows parameter setting of the literatures and an optimal parameter combination from several tests in this study. It is found that the Nc in the literature is the same. Thus, this study slightly adjusts Nc in order to obtain better parameter combination. This section uses the Iris data set as the object of the experimental combination analysis. The orthogonal array $L_{27}(8^3)$ is used for experiment and software package Minitab is used for Taguchi experimental design. SED (sum of Euclidean distance) is applied as the criterion. A smaller SED indicates better clustering result. Thus, the experiment belongs to small-the-better. S/N ratio is signal to noise ratio, and is used to evaluate system quality stability in the Taguchi design of experiments. A higher S/N ratio indicates smaller noise, and better generated parameter quality. To sum up, the parameter setting of the proposed aiNet-PSKO algorithm and PSO, AIS, ai-Net, PSKO, AISK and aiNetK algorithm are shown in Table 3.

4.3 Experimental Results and Verification

4.3.1 Algorithm convergence

This study presents the convergence of different algorithms in the four benchmark data sets, as shown in Figures 3 to 6. Based on the SED convergence map, ai-Net can evaluate average fitness value error and suppress excessively similar antibodies. Thus, convergence of the four data sets is quicker than the AIS and the PSO. Through local search of K-means algorithm with quick convergence, the proposed aiNet-PSKO algorithm in the four data sets has faster and better convergence than other algorithms.

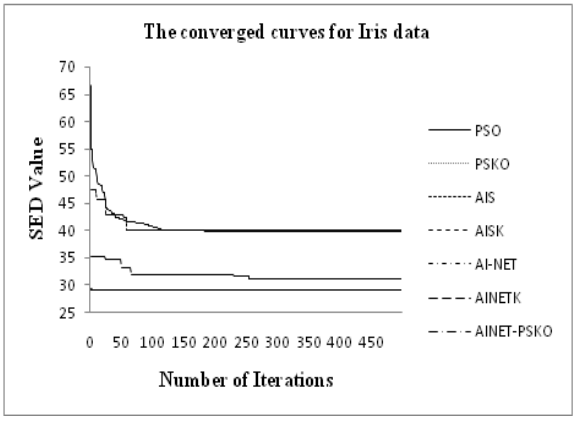

Fig. 3: SED convergence curves of Iris Data Set using different algorithms.

4.3.2 Algorithm clustering results

(1)Average value and standard deviation of SED.
This study tests each algorithm in each benchmark data

Table 2: Factor levels.

Level	N	Nc	d	Suppression	beta	wei	c1	c2
1	20	9	0.4	0.2	0.001	0.72	1.47	1.47
2	10	10	0.5	0.01	0.01	0.95	0.5	0.5
3	30	11	0.9	0.001	0.005	0.5	2.6	2.6

Table 3: Parameter setup for each algorithm.

Parameter setting	of the algorithms
Number of iterations	200
Algorithms	AIS, AISK
Size of population cells	$N = 30$
Size of memory cells	$M = 20$
Number of selected antibodies	$n = 20$
Clone multiple	$f = 2$
Constant parameter ρ	$\rho = 5$
Percentage of new cells	$d = 0.2$
Algorithms	ai-Net, aiNetK, aiNet-PSKO
Size of parent cells	$N = 30$
Clone multiple	$Nc = 11$
Percentage of new cells	$d = 0.9$
Suppression threshold	$Suppression = 0.01$
beta	$beta = 0.005$
Algorithms	PSO, PSKO, aiNet-PSKO
Number of particles	$P = 30$
Initial weight	$wei = 0.5$
Learning factor c1	$c1 = 1.47$
Learning factor c2	$c2 = 2.6$

Table 4: Average Clustering Results of SED for 30 Run of each algorithm

Algorithm	Iris	Wine	Glass	Breast Cancer
PSO	33.9485 ± 2.9754	116.1199 ± 4.5403	85.4623 ± 3.4951	413.1796 ± 30.2675
PSKO	29.1729 ± 0.0103	$88.6952 \pm 7.11E-14$	46.4 ± 1.4795	327.789 ± 0.0819
AIS	38.7704 ± 1.6063	141.782 ± 2.6937	98.6811 ± 3.1429	450.5403 ± 21.0532
AISK	29.1546 ± 0.0033	88.7101 ± 0.0057	45.177 ± 0.8148	327.7334 ± 0.0082
ai-Net	31.464 ± 0.4499	120.2075 ± 7.2344	81.1545 ± 6.4916	352.565 ± 5.4716
aiNetK	29.1501 ± 0.0034	$88.6952 \pm 7.10543E-14$	43.7845 ± 0.0436	327.7278 ± 0.0037
aiNet-PSKO	29.1478 ± 0.0042	$88.6952 \pm 7.10543E-14$	43.7165 ± 0.0314	327.7248 ± 0.0012

set 30 times and took the average value as the testing result, as shown in Table 4. As seen, in the most benchmark data sets, the SED and standard deviation calculated by the aiNet-PSKO algorithm are both the lowest.

(2)Computational time.
For computational time, since aiNet has the characteristic of evaluating average fitness value error, the time needed by the proposed aiNet-PSKO is longer than those of other algorithms. The computational time for each algorithm is listed in Table 5.

Table 5: Computational time (sec) of each algorithm for each benchmark data set.

Algorithm	Iris	Wine	Glass	Breast Cancer
PSO	1	2	3.4	3.7
PSKO	1.3	3	5.7	6.5
AIS	1	2	3.5	4
AISK	1.6	3.7	6.7	8
ai-Net	12.6	94.6	24.8	34.1
aiNetK	40.7	270.667	39.6	770
aiNet-PSKO	51.1	146.33	199.4	939

(3)Required number of iterations for convergence.
To compare rapid convergence capability of the algorithms, the average value of SED of each benchmark data set is used as threshold value. Each algorithm is stopped and the iteration number is recorded when the threshold value is reached. Contrarily, if the algorithm does not reach the threshold value when the iteration exceeds 2000 times, convergence is regarded to fail. This study tests each algorithm 10 times in each benchmark data set, and the average iteration number of convergence is shown in Table 6. As shown, PSO and AIS have fast

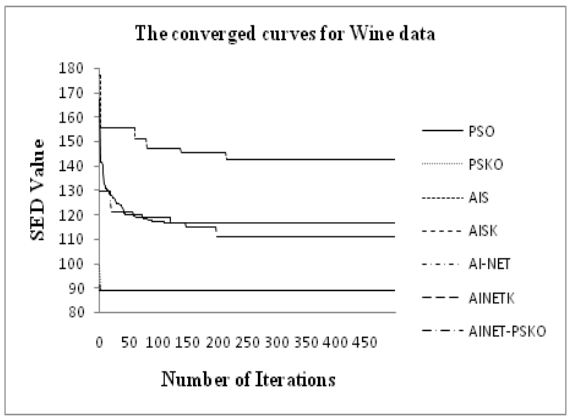

Fig. 4: SED convergence curves of Wine Data Set using different algorithms.

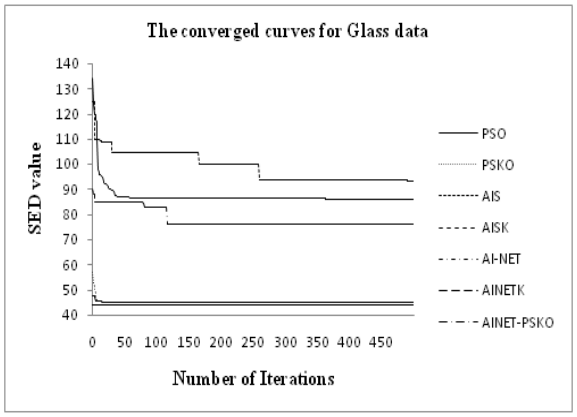

Fig. 5: SED convergence curves of Glass Data Set using different algorithms.

Table 6: Average convergence time of each algorithm for each benchmark data set.

Algorithm	Iris	Wine	Glass	Breast Cancer
PSO	1065	2000	2000	1775.3
PSKO	1.4	1	1.5	1.4
AIS	2000	2000	2000	2000
AISK	1	1	1	1
ai-Net	211.7	2000	2000	87.3
aiNetK	1	1	1	1
aiNet-PSKO	1	1	1	1

computing speed but worse search capability, and thus, most of data sets cannot reach threshold value. The aiNet-PSKO needs longer computational time but has better searching capability. It can reach convergence threshold needing minimum number of iterations.

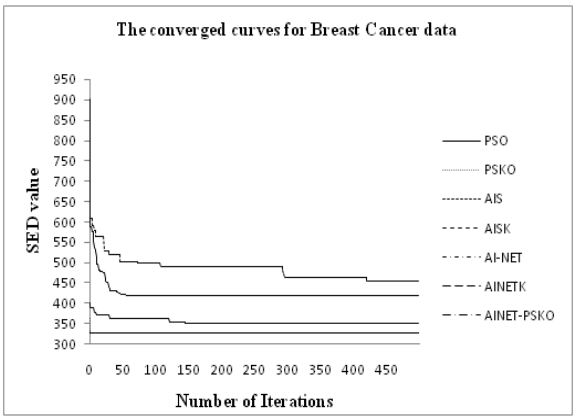

Fig. 6: SED convergence curves of Breast Cancer Data Set using different algorithms.

Table 7: Accuracy of clustering result of each algorithm for each benchmark data set.

Algorithm	Iris	Wine	Glass	Breast Cancer
PSO	0.6667	0.6124	0.9159	0.9004
PSKO	0.8867	0.9494	0.986	0.9649
AIS	0.7067	0.6011	0.771	0.8683
AISK	0.86	0.9494	0.986	0.9693
ai-Net	0.86	0.6011	0.9626	0.9327
aiNetK	0.8933	0.9551	0.9907	0.9722
aiNet-PSKO	0.8933	0.9551	0.9907	0.9722

(4)Accuracy.

The accuracy of the clustering result of each algorithm for the four benchmark data sets is shown in Table 7. It indicates that the accuracy of proposed aiNet-PSKO for the four data sets is the highest among other algorithms.

4.3.3 Statistical test

Through confidence level of 95% ($\alpha = 0.05$), it is verified whether difference between aiNet-PSKO and other six algorithms is significant or not. $\mu_{aiNet-PSKO}$ is average SED value of aiNet-PSKO, μ_{PSO} is average SED value of PSO, μ_{PSKO} is average SED value of PSKO, μ_{AIS} is average SED value of AIS, μ_{AISK} is average SED value of AISK, μ_{ai-Net} is average SED value of ai-Net, and μ_{aiNetK} is average SED value of aiNetK. Then, pairwise comparison is verified: The test statistic (Z_0) is as follows:

$$\frac{(\bar{x_1} - \bar{x_2}) - (\mu_1 - \mu_2)}{\sqrt{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})}} \tag{13}$$

The left tail test of single-tailed test was conducted. If the test statistic value $(Z_0) < -Z_\alpha$, then H_0 is rejected.

Table 8: Comparison of the proposed algorithm with other algorithms for four data sets.

aiNet-PSKO vs. Other Algorithms	aiNet-PSKO vs. PSO	aiNet-PSKO vs. PSKO	aiNet-PSKO vs. AIS	aiNet-PSKO vs. AISK	aiNet-PSKO vs. ai-Net	aiNet-PSKO vs. aiNetK
Iris(Z_0)	-8.8373	-12.3594	-32.8114	-6.973	-28.1969	-2.3313
Test result	Reject H_0	Reject H_0	Reject H_0	Reject H_0	Reject H_0	Reject H_0
Wine(Z_0)	-33.084	0	-107.944	-14.3177	-23.8582	0
Test result	Reject H_0	non-Reject H_0	Reject H_0	Reject H_0	Reject H_0	non-Reject H_0
Glass(Z_0)	-65.4179	-9.9323	-95.7837	-9.8105	-31.5876	-6.9319
Test result	Reject H_0	Reject H_0	Reject H_0	Reject H_0	Reject H_0	Reject H_0
Breast Cancer(Z_0)	-15.464	-4.293	-31.9518	-5.6839	-24.8657	-4.2244
Test result	Reject H_0	Reject H_0	Reject H_0	Reject H_0	Reject H_0	Reject H_0

Table 8 shows test results of other algorithms in the four data sets by using aiNet-PSKO.

Based on the statistic test, in Wine data set aiNet-PSKO has no enough evidence to prove it is better than aiNetK and PSKO. In most data sets, aiNet-PSKO has enough evidence to show its average SED value is superior to those of other algorithms. Based on the above results, the computational time of the proposed methods is longer than other algorithms, and the average computational time of the aiNet-PSKO is the longest since aiNet-PSKO integrates the three algorithms. In addition, its algorithm structure is more complex than other algorithms, thus resulting in longest computational time. However, aiNet-PSKO has faster convergence. Further, aiNet-PSKO can be converged to minimum average SED value. The velocity of convergence, convergence result and accuracy of clustering result of aiNet-PSKO are better than those of other algorithms.

5 Customer Segmentation Problem

The empirical case in this section belongs to the unknown number of clusters. Thus, this study employs two-phase clustering analysis methods. In the first phase, the Adaptive Resonance Theory 2 (ART2) neural network is used to automatically find number of the clusters. In the second phase, PSO, PSKO, AIS, AISK, ai-Net, aiNetK and aiNet-PSKO algorithms are compared to find the optimal clustering results. The framework is illustrated in Figure 7. The case company is a business-to-consumer cyber flower shop located in Taiwan. The overall transaction data for the cyber flower shop is from 2006/1/1 to 2006/12/31. The data are calculated and pre-processed through the RFM (recency, frequency, and monetary) rule first ([36,32]). Thus, there are three input attributes for each piece of data. Then the proposed clustering method is applied to segment the customers. The goal is to categorize customers with similar characteristics into one cluster, so the enterprise can develop corresponding marketing strategy based on each customer cluster.

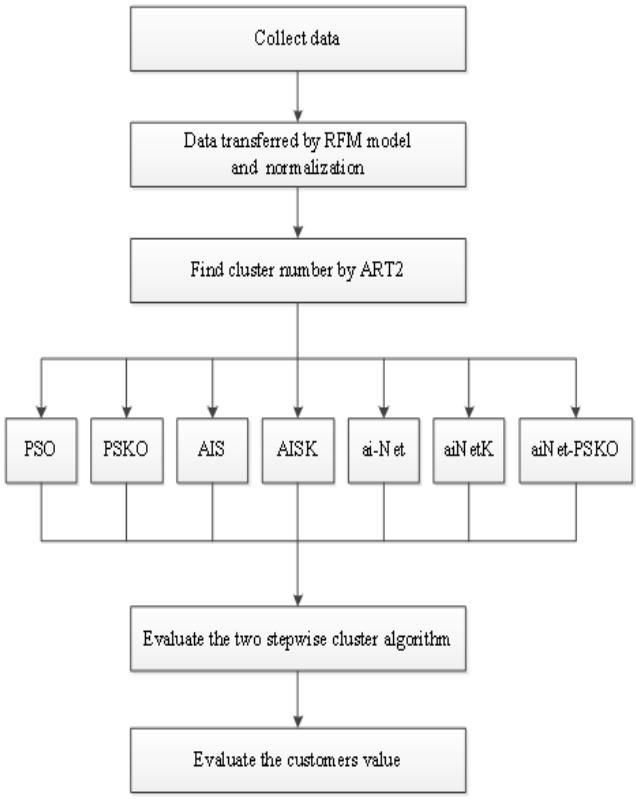

Fig. 7: The research framework.

5.1 Data Preprocessing

[36] noted that enterprises should understand recency, frequency, and monetary (RFM) variables to determine suitable marketing strategy. The RFM analysis technique identifies customer behavior according to three variables as follows [18]: Recency (CR) of the last purchase: It refers to the interval between the time that the latest purchase happened and the present. Frequency (CF) of purchases: It refers to the number of transactions in a particular period. Monetary value (CM) of a purchase: It

refers to the amount of a particular order. Basically, customer lifetime value or loyalty can be evaluated in terms of RFM variables by integrating the rate of each cluster [32]. The RFM rule is generally acknowledged as the most popular customer value analytical method at present [28].

5.2 Determination of Cluster Number Using ART2 Neural Network

This study employs Matlab 7.0 to code ART2 algorithm. The algorithm mainly uses threshold value to determine number of clusters, and the threshold value has significant impact on the clustering results. Wilk's Lambda is used as the indicator to measure number of clusters [30]. Wilk's Lambda is defined as percentage of within-group variation in total variation, as shown in Eq. (14). Wilk's Lambda is often used in multivariate analysis of variance to determine cluster number. It is assumed that Wilk's Lambda in different clusters increases sharply, and cluster number before variation is regarded as the optimal cluster number.

$$Wilk^{'}sLambda = \frac{SS_within}{SS_total} \qquad (14)$$

Given different threshold value , this study uses Wilks Lambda calculated from ART2 neural network and the relation between Wilks Lambda and the cluster number, as shown in Table 9 and Figure 8. When the cluster number is reduced from 4 to 3, Wilk's Lambda increases sharply. Thus, this study regarded the cluster number of 4 before variation as the optimal cluster number.

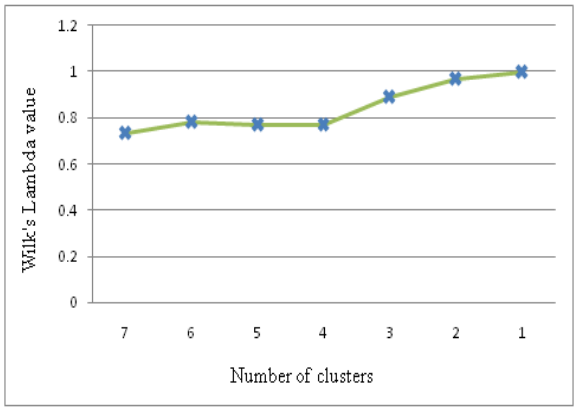

Fig. 8: Relation between Wilk's Lambda and cluster number.

5.3 Comparison of Clustering Results

After determining the cluster number in the first phase, this study compares SED convergence map, SED average value and standard deviation of PSO, PSKO, AIS, AISK, ai-Net, aiNetK and aiNet-PSKO. As shown in Figure 9, aiNet-PSKO has fast velocity of convergence in RFM data set. In RFM data set, this study tests each clustering algorithm for 30 times. The average SED value and standard deviation of these tests are listed in Table 10. As shown, ART2+aiNet-PSKO has the minimum SED value, and this represents aiNet-PSKO can yield optimal clustering result in the empirical experiment.

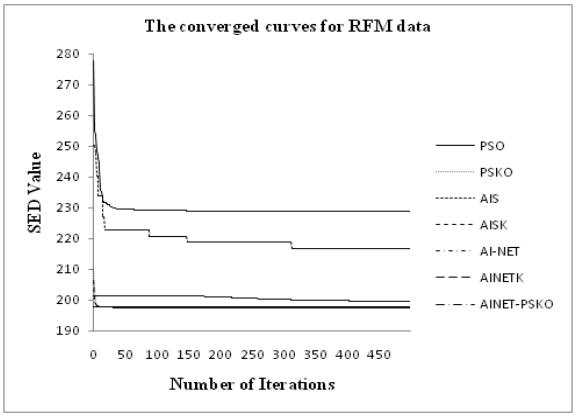

Fig. 9: SED convergence curves of RFM Data Set using different algorithms.

5.4 Statistical Test

Through confidence level of 95% ($\alpha = 0.05$), it is verified whether difference between aiNet-PSKO and other six algorithms (including ai-NetK) is significant or not. $\mu_{aiNet-PSKO}$ is average SED value of aiNet-PSKO, μ_{PSO} is average SED value of PSO, μ_{PSKO} is average SED value of PSKO, μ_{AIS} is average SED value of AIS, μ_{AISK} is average SED value of AISK, μ_{ai-Net} is average SED value of ai-Net, and μ_{aiNetK} is average SED value of aiNetK. Then, pairwise comparison is verified. The test statistic (Z_0) is as Eq. 14. The left tail test of single-tailed test was conducted. If the test statistic value $(Z_0) < -Z$, H_0 is rejected. Table 11 shows test results of other algorithms in the four data sets using aiNet-PSKO. Based on the statistic test, in RFM data set the aiNet-PSKO has enough evidence to show its average SED value is superior to that of other algorithms. Based on the above results, this study employed ART2+aiNet-PSKO as clustering analysis tool in the empirical analysis.

6 Conclusions

This study has proposed a novel cluster analysis technique, aiNet-PSKO algorithm. It owns both the merits

Table 9: Wilk's Lambda of different threshold value ρ.

Threshold value (rho)	clusters	Wilk's Lambda	Difference value
0.994	7	0.734141	0.04815
0.993	6	0.782291	-0.01224
0.992	5	0.770051	0.000566
0.99	4	0.769485	0.121022
0.975	3	0.890507	0.076817
0.96	2	0.967324	0.032676
0.92	1	1	

Table 10: Clustering result of each algorithm.

Clustering analysis method	ART2 + PSO	ART2 + PSKO	ART2 + AIS	ART2 + AISK	ART2 + ai-Net	ART2 + aiNetK	ART2 + aiNet-PSKO
SED	207.14387.5693	197.79760.0413	215.42444.3978	197.74180.028	200.37120.7359	197.73310.0447	197.67680.0609

Table 11: Test of other algorithms in the RFM data set by using aiNet-PSKO.

aiNet-PSKO vs. Other Algorithms	aiNet-PSKO vs. PSO	aiNet-PSKO vs. PSKO	aiNet-PSKO vs. AIS	aiNet-PSKO vs. AISK	aiNet-PSKO vs. ai-Net	aiNet-PSKO vs. aiNetK
RFM(Z_0)	-6.8502	-8.9918	-22.1016	-5.3115	-19.9858	-4.082
Test result	Reject H_0	Reject H_0	Reject H_0	Reject H_0	Reject H_0	Reject H_0

of aiNet and PSO algorithms. This can improve the clustering performance. The computational results of four benchmark data sets reveal that it has the lowest SSE value compared to other clustering algorithms. Besides, the results of a real-world problem for customer clustering also illustrate that aiNet-PSKO algorithm is superior to other clustering algorithms. The clustering results can be applied to design the corresponding marketing strategies for different clusters. This can enhance the sales performance for the case company. In the future, the current idea can be applied to propose the automatic clustering algorithm which does not need the number of clusters in advance.

Acknowledgements

This study was financially supported by the National Science Council of the Taiwanese Government, under contract number NSC99-2410-H-011-009-MY3. This support is greatly appreciated.

References

[1] 4 Al-Sultan, K., A Tabu search approach to the clustering problem, Pattern Recognition, Vol. 28, No. 9, pp.1443-1451, 1995.

[2] Bezdek, J., A convergence theorem for the fuzzy ISO-DATA clustering algorithm, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2, pp.1-8, 1980.

[3] Bezdek, J. and Hathaway, R., Numerical convergence and interpretation of the fuzzy c-shells clustering algorithms, IEEE Transactions Neural Network, Vol. 3, No. 5, pp.787-793, 1992.

[4] Bezerra, G. B., Barra, T. V., De Castro, L.N. and Von Zuben, F. J.,Adaptive radius immune algorithm for data clustering, Lecture Notes in Computer Science, Vol. 3627, pp.290-303, 2005.

[5] Chiu, C. Y., Kuo, I. T. and Lin, C. H., Applying artificial immune system and ant algorithm in air-conditioner market segmentation, Expert Systems with Applications, Vol. 36, No. 3, pp.4437-4442, 2009.

[6] Estivill-Castro, V. and Lee, I., AMOEBA: hierarchical clustering based on spatial proximity using delaunay diagram, Proceedings 9th International Spatial Data Handling (SDH2000), pp.10-12, 2000a.

[7] Estivill-Castro, V. and Lee, I., AUTOCLUST: Automatic clustering via boundary extraction for massive point data sets, Proceedings 5th InternationalConference Geo-Computation, pp.23-25, 2000b.

[8] Farzaneh, A., Alireza, M. and Ashkan, R.K., A novel binary particle swarm optimization method using artificial immune system, Proceedings of the International Conference on Computer as a Tool, pp.217-220, 2005.

[9] Forgy, E., Clustering analysis of multivariate data: efficiency versus interpretability of classification, Biometrics, Vol. 21, pp.768-769, 1965.

[10] Geva, A.B, Hierarchical unsupervised fuzzy clustering, IEEE Transactions on Fuzzy Systems, Vol. 7, No. 6, pp.723-733, 1999.

[11] Graaff, A. J. and Engelbrecht, A. P., Clustering data in stationary environments with a local network neighborhood artificial immune system, International Journal of Machine Learning and Cybernetics, Vol. 3, No. 1, pp.1-26, 2012.

[12] Guha, S., Rastogi, R. and Shim, K., CURE: an efficient clustering algorithm for large databases, Proceedings ACM SIGMOD International Conference Management of Data, pp.73-84, 1998.

[13] Guha, S., Rastogi, R. and Shim, K., ROCK: a robust clustering algorithm for categorical attributes, Information Systems, Vol. 25, No. 5, pp.345-366, 2000.
[14] Hall, L., zyurt, I., and Bezdek, J., Clustering with a genetically optimized approach, IEEE Transactions on Evolutionary Computation, Vol. 3, No. 2, pp.103112, 1999.
[15] Hamerly, G. and Elkan, C, Learning the K in K-means, Proceedings of 7th Annual Conference on Neural Information Processing Systems, 2003.
[16] Han, J. and Kamber, M., Data Mining: Concepts and Techniques, 2nd Edition, Morgan Kaufmann, 2006.
[17] Hart, E. , Ross, P., Exploiting the analogy between the immune system and sparse distributed memories, Genetic Programming and Evolvable Machines, Vol. 4, No. 4, pp.333-358, 2003.
[18] Hosseini, S.M.S., Maleki, A. and Gholamian, M. R., Cluster analysis using data mining approach to develop CRM methodology to assess the customer loyalty, Expert Systems with Applications, Vol. 37, No. 7, pp.5259-5264, 2010.
[19] Karypis, G., Han, E. and Kumar, V., Chameleon: hierarchical clustering using dynamic modeling, IEEE Computer, Vol. 32, No. 8, pp.68-75, 1999.
[20] Kaufman, L. and Rousseeuw, P., Finding groups in data: an introduction to cluster analysis, Wiley, 1990.
[21] Kiang, M. Y., Raghu, T. S. and Shang, K. H. M., Marketing on the Internet Who can benefit from an online marketing approach? Decision Support Systems, Vol. 27, pp.383-393,2000.
[22] Krishna, K. and Murty, M. N., Genetic K-means algorithm, IEEE Transactions on Systems Man, and Cybernetics, Vol. 29, No. 3, pp.433-439, 1999.
[23] Krishnapuram, R. and Keller, J., A possibilistic approach to clustering, IEEE Transactions Fuzzy Systems, Vol. 1, No. 2, pp.98-110, 1993.
[24] Kuo, R. J., Wang, H. S., Hu, T. L. and Chou, S. H., Application of ant K-means on Clustering Analysis in Data Mining, International Journal of Computers and Mathematics with Applications, Vol. 50, pp.1709-1724, 2005.
[25] Kuo, R. J., Wang, M. J. and Huang, T. W., An application of particle swarm optimization algorithm to clustering analysis, Journal of Soft Computing, Vol. 15, No. 3, pp.533-542, 2011.
[26] Kuo, R.J., Chen, S.S., Cheng, W.C., and Tsai, C. Y., Integration of artificial immune network and K-means for cluster analysis, Knowledge and Information Systems, Vol. 40, pp.541557, 2014.
[27] Li, X. Y., Xu, H. L. and Cheng, Z. G., One immune simplex particle swarm optimization and its application, Proceedings of the 4th International Conference on Natural Computation, pp.331-335, 2008.
[28] Liang, Y. H., Integration of data mining technologies to analyze customer value for the automotive maintenance industry, Expert Systems with Applications, Vol. 37, No. 12, pp.7489-7496, 2010.
[29] Liao, X. F., Hu, L. T. and Jin, H., Energy optimization schemes in cluster with virtual machines, Cluster Computing, Vol. 13, pp.113-126, 2010.
[30] Lin, C. R. and Chen, M. S., A robust and efficient clustering algorithm based on cohesion self-merging, Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.582-587, 2002.
[31] Liu, D. R. and Shih, Y. Y., Hybrid approaches to product recommendation based on customer lifetime value and purchase preferences, Journal of Systems and Software, Vol. 77, No. 2, pp.181-191, 2005.
[32] Liu, F., Wang, Q. and Gao, X., Survey of artificial immune system, Proceedings of the 1th International Symposium on Systems and Control in Aerospace and Astronautics, pp.19-21, 2006.
[33] Lu, H., A particle swarm optimization based on immune mechanism, Proceedings of the International Joint Conference on Computational Sciences and Optimization, pp.670-673, 2009.
[34] Luo, Y. and Che, X., Chaos immune particle swarm optimization algorithm with hybrid discrete variables and its application to mechanical optimization, Proceedings of the IEEE International Conference on Intelligent Information Technology Application Workshops, pp.190-193, 2009.
[35] Ma, W., Jiao, L. and Gong, M., Immunodominance and clonal selection inspired multiobjective clustering, Progress in Natural Science, Vol. 19, No. 6, pp.751-758, 2009.
[36] Marcus, C., A practical yet meaningful approach to customer segmentation, Journal of Consumer Marketing, Vol. 15, No. 5, pp.494-504, 1998.
[37] Mu, Y. and Sheng, A., Evolutionary diagonal recurrent neural network with improved hybrid EP-PSO algorithm and its identification application, International Journal of Innovative Computing, Information and Control, Vol. 5, No. 3, pp.1615-1624, 2009.
[38] Nasraoui, O., Rojas, C. and Cardona, C.,A framework for mining evolving trends in Web data streams using dynamic learning and retrospective validation, Computer Networks, Vol. 50, No. 10, pp.1488-1512, 2006.
[39] Pasti, R. and Castro, L.N.D., An immune and a gradient-based method to train multi-layer perceptron neural networks, Proceedings of the International Joint Conference on Neural Networks, pp.2075-2082, 2006.
[40] Sotiropoulos, D. N., Tsihrintzis, G. A., Savvopoulos, A. and Virvou, M., Artificial immune system-based customer data clustering in an e-shopping application, Lecture Notes in Computer Science, Vol. 4251, pp.960-967, 2006.
[41] Tang, N. and Vemuri, V., An artificial immune system approach to document clustering, Proceedings of the ACM Symposium on Applied Computing, Vol. 2, pp.918-922, 2005.
[42] Timmis, J. and Edmonds, C., A comment on Opt-AiNET: An immune network algorithm for optimization, Lecture Notes in Computer Science, Vol. 3102, pp.308-317, 2004.
[43] Xu, R. and Wunsch, D., Survey of clustering algorithms, IEEE Transactions on Neural Networks, Vol. 16, Issue 3, pp.645-678, 2005.
[44] Younsi, R. , Wang, W., A new artificial immune system algorithm for clustering, Lecture Notes in Computer Science, Vol.3177, pp.58-64, 2004.
[45] Yue, X., Abraham, A., Chi, Z. X., Hao, Y. Y. and Mo, H., Artificial immune system inspired behavior-based anti-spam filter, Soft Computing, Vol. 11, No. 8, pp.729-740, 2007.
[46] Zhang, T., Ramakrishnan, R. and Livny, M., BIRCH: an efficient data clustering method for very large databases, Proceedings of the ACM SIGMOD Conference Management of Data, pp.103-114, 1996.
[47] Zhang, X., Ma, T. and Han, X., Optimizing Fixed Shelf Order-Picking for AS/RS Based on Immune Particle

Swarm Optimization Algorithm, Proceedings of the IEEE International Conference on Automation and Logistics, No. 8, pp.2824-2829, 2007.

R. J. Kuo received the M.S. degree in Industrial and Manufacturing Systems Engineering from Iowa State University, Ames, IA, in 1990 and the Ph.D. degree in Industrial and Management Systems Engineering from the Pennsylvania State University, University Park, PA, in 1994. Currently, he is the Distinguished Professor in the Department of Industrial Management at National Taiwan University of Science and Technology, Taiwan. His research interests include architecture issues of computational intelligence and their applications to data mining, electronic business, logistics and supply chain management and decision support systems.

S. S. Chen received the M.S. degree in Industrial Management from National Taiwan University of Science and Technology, Taiwan. His research interests include architecture issues of computational intelligence and their applications to data mining and operations management.