Available online at www.sciencedirect.com

SCIENCE @ DIRECT®

ELSEVIER Computers and Mathematics with Applications 50 (2005) 1709–1724

An International Journal
computers & mathematics
with applications

www.elsevier.com/locate/camwa

Application of Ant K-Means on Clustering Analysis

R. J. Kuo, H. S. Wang, Tung-Lai Hu and S. H. Chou
Department of Industrial Engineering and Management
National Taipei University of Technology
Taipei, Taiwan 106, R.O.C.
rjkuo@ntut.edu.tw

(Received April 2005; accepted May 2005)

Abstract—This paper intends to propose a novel clustering method, ant K-means (AK) algorithm. AK algorithm modifies the K-means as locating the objects in a cluster with the probability, which is updated by the pheromone, while the rule of updating pheromone is according to total within cluster variance (TWCV). The computational results showed that it is better than the other two methods, self-organizing feature map (SOM) followed by K-means method and SOM followed by genetic K-means algorithm via 243 data sets generated by Monte Carlo simulation. To further testify this novel method, the questionnaire survey data for the plasma television market segmentation is employed. The results also indicated that the proposed method is the best among these three methods based on TWCV. © 2005 Elsevier Ltd. All rights reserved.

Keywords—Data mining, Clustering analysis, Ant colony optimization.

1. INTRODUCTION

Data mining, extracting interesting and valuable information such as trends, features, or patterns from hidden predictive data, is the multidisciplinary field that is at the intersection of statistics, machine learning, database management, and data visualization, to provide a new perspective on data analysis.

Clustering analysis, which is the subject of active research in several fields such as statistics, pattern recognition, machine learning, and data mining, is to partition a given set of data or objects into clusters (or called groups, classes). It also has been applied in a large variety of applications, for example, image segmentation, object and character recognition, document retrieval, etc. Some of the clustering techniques, which support these applications, are shown in the next section.

Ant colony optimization (ACO) is a recently proposed metaheuristic approach for solving hard combinatorial optimization problems [1]. The ant is able to communicate with each other according to its chemical trail called "pheromone". The characteristics of an ant colony include positive feedback, distributed computation, and use of a constructive greedy heuristic [2].

This study is partially supported by National Science Council of Taiwan, ROC under Contract NSC92-2416-H-027-002. Her support has been appreciated.

0898-1221/05/$ - see front matter © 2005 Elsevier Ltd. All rights reserved. Typeset by $\mathcal{A}\mathcal{M}\mathcal{S}$-TEX
doi:10.1016/j.camwa.2005.05.009

Due to ACO's promising results, this study proposes a novel method, ant K-means (AK) algorithm, to solve the clustering problem. The proposed method needs the initial number of corresponding clusters and centroids. The computational performance of the proposed method is compared with the other two methods, self-organizing feature map (SOM) followed by K-means method [3,4] and SOM followed by genetic K-means algorithm [5], via 243 data sets generated by Monte Carlo simulation [6,7]. To further testify its performance, the questionnaire survey data for the plasma television market segmentation is employed. The results also illustrated that the proposed method has the smallest total within clustering variance among three methods.

The rest of the paper is organized as follows. Section 2 presents some general background for data mining, clustering analysis, and ant colony optimization, while the proposed method is explained in Section 3. Both the simulation and real-world problem results are illustrated in Sections 4 and 5, respectively. Finally, Section 6 makes the concluding remarks.

2. LITERATURE REVIEW

2.1. Data Mining

Fayyad, Piatetsky-Shapiro and Smyth [8] had defined the knowledge discovery in databases (KDD) as a nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data [9]. By the term process shows that KDD is made up of several steps, which involve selection, preprocessing, transformation, data mining, and interpretation/evaluation. In [10,11], they show that data mining is a multidisciplinary field that is at the intersection of statistics, machine learning, database management, and data visualization, to provide a new perspective on data analysis. Fayyad [9] stated the following.

> "Data mining is a step in the KDD process consisting of applying computational techniques that, under acceptable computational efficiency limitations, produce a particular enumeration of patterns (or models) over the data."

He further explained that the data mining is concerned with the algorithmic means by which patterns are extracted and enumerated from data. Therefore, we may say that the data mining is the process of analyzing the transformed or observational data sets to find unsuspected relationships and to summarize the data in novel techniques such as neural network or genetic algorithm for unknown patterns or models to data owner.

2.2. Clustering Analysis

The goal of clustering analysis is to group similar objects together. There are many methods being applied in clustering analysis, like hierarchical clustering, partition-based clustering, density-based clustering, and artificial intelligence-based clustering. In this section, the artificial intelligence-based clustering which includes artificial neural networks (ANN) and genetic algorithm (GA) is illustrated. The other approaches are introduced in survey research [12–14].

2.2.1. Applications of ANN in clustering analysis

The most widely applied unsupervised learning scheme is Kohonen's feature maps (also called self-organizing feature map, SOM) [15]. The SOM typically has two layers. The input layer is fully connected to the output layer that is a two-dimensional layer. The output layer nodes each measure the Euclidean distance of its weights to the incoming input values. The output nodes with the smallest distance called as winning nodes and the weights of them are adjusted to be closer to the values of the input nodes. The other widely applied unsupervised learning scheme is adaptive resonance theory (ART). The basic features of the ART architecture are shown as follows. Patterns of activity that develop over the nodes in the two layers of the attentional subsystem trances because they exist only in association with a single application of an input vector. The weights associated with the bottom-up and top-down connections between F_1 and F_2

traces because they encode information that remains a part of the network for an extended period. In [16], Carpenter *et al.* presented the adaptive resonance theory 2 (ART 2). ART 1 and ART 2 differ in the nature of their input patterns. ART 1 requires that the input vectors be binary. ART 2 is suitable for analog patterns. Both ART 1 and ART 2 have an attentional subsystem and an orienting subsystem. The attentional subsystem of each architecture consists of two layers, F_1 and F_2. The orienting subsystem of each network performs the identical function. The basic differential equations that govern the activities of the individual processing in ART 2, Carpenter and Grossberg have had to split the F_1 layer into a number of sublayers containing both feed-forward and feedback connections [16,17].

2.2.2. Applications of GA in clustering analysis

In [18], Selim and Ismail proved that the conventional statistics methods, like K-means algorithm, are easy to find a local minimum. It is necessary to develop a more robust method for clustering analysis. Maulik and Bandyopadhyay [19] proposed a GA-based method to solve the clustering problem and experiments on synthetic and real life data sets to evaluate the performance. The results showed that the GA-based method may improve the final output of K-means. According to [20], the researcher proposed a novel approach called genetic K-means algorithm (GKA) for clustering analysis. It defines a biased mutation operator specific to clustering called distance-based-mutation. Using finite Markov chain theory, it proved that the GKA converges to the best known optimum.

2.2.3. Two-stage clustering analysis

Kuo *et al.* proposed a two-stage method which integrates both the SOM and K-means (called S + K). The results indicated that the proposed method is much better than only using SOM or K-means [3]. In [5,21], Kuo modified Krishna's GKA and used SOM's solution as the initial solution for modified GKA. The results showed that it is better than previously published method, SOM + K-means. Kuo *et al.* [5] proposed SOM + GA-Based clustering method (called S + G) on clustering analysis, and S + G gets better computational performance than SOM and S + K.

2.3. Ant Colony System

Dorigo *et al.* [22] applied the ant system to the well-known traveling salesman problem (TSP). The corresponding algorithm is shown in Figure 1. Let a set of n cities, the TSP is the problem of finding a minimal length closed tour that visits each city once. The edge (or arc) between City i and City j is called d_{ij}; in the case of Euclidean TSP, d_{ij} is the Euclidean distance between City i and City j, i.e., $d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$. An example of the TSP is given by a graph (N, E), where N is the set of cities and E is the set of edges between cities.

Let m ants walk on the n cities. Each ant randomly set in n cities, and moves to next city with the *probability* (called *random-proportional rule*), given by equation (1), which ant k in City i

```
Procedure Ant System algorithm for TSP
    Set parameters, initialize pheromone trails
    While (termination iteration not met) Do
        Do
            Determinate next city with probability $p_k(i,j)$ given by Equation (1)
        While (each ant complete a tour)
        Updating the pheromone on each walked edge by Equation (2)
    End
    Summarize all data
End AS algorithm for TSP
```

Figure 1. The ant system algorithm for TSP.

chooses to move to the City j with the probability $p_k(i,j)$, i.e.,

$$p_k(i,j) = \begin{cases} \dfrac{[\tau(i,j)]^\alpha \cdot [\eta(i,j)]^\beta}{\sum\limits_{s \in \text{allowed}_k} [\tau(i,s)]^\alpha \cdot [\eta(i,s)]^\beta}, & \text{if } j \in \text{allowed}_k, \\ 0, & \text{otherwise,} \end{cases} \tag{1}$$

where τ is the pheromone, η (or called *visibility*) is the inverse of the distance between i and j, i.e., $\eta = 1/d_{ij}$, allowed$_k$ is the set of cities not visited by ant k. The parameter α and β control the relative importance of pheromone trail versus visibility.

When each ant passing by the whole cities (a complete tour), the pheromone will be laid on the edge which is walked by that ant. This is important mechanism to update the pheromone on the edge (also called global updating rule). Once all ants have built their tours, pheromone is updated on all edges according to

$$\tau(i,j) \leftarrow (1-\rho) \cdot \tau(i,j) + \sum_{k=1}^{m} \Delta\tau_k(i,j), \tag{2}$$

where

$$\Delta\tau_k(i,j) = \begin{cases} \dfrac{Q}{L_k}, & \text{if } (i,j) \in \text{ tour donc by ant } k, \\ 0, & \text{otherwise,} \end{cases} \tag{3}$$

$0 < \rho < 1$ is a pheromone decay parameter, Q is a constant, and L_k is the length of the tour performed by ant k, and m is the number of ants. This updating rule is called *ant cycle* algorithm. Dorigo *et al.* [22] also bought up other two algorithms for updating rule, which is called *ant-density* model and *ant-quantity* model, as follows,

$$\Delta\tau_k(i,j) = \begin{cases} Q, & \text{if } (i,j) \in \text{ tour done by ant } k, \\ 0, & \text{otherwise,} \end{cases} \tag{4}$$

$$\Delta\tau_k(i,j) = \begin{cases} \dfrac{Q}{d_{ij}}, & \text{if } (i,j) \in \text{ tour done by ant } k, \\ 0, & \text{otherwise,} \end{cases} \tag{5}$$

Equation (4) is the updating rule of *ant-density* model, where Q is the quantity of trail which is left on edge(i,j) every time an ant goes from i to j; in equation (5) called *ant-quantity* model, where an ant going from I to j leaves a quantity Q/d_{ij} of trail on edge (i,j) every time it goes from i to j. They differ from ant-cycle, where both of these rules allow each ant lays its trail at each step, without waiting for the end of the tour.

2.4. Ant in Clustering Analysis

Applying ant colony system in clustering analysis is still a very novel research area. Tsai and his colleagues [23] employed the ant system with differently favorable strategy for data clustering. It is named ant colony optimization with different favor (ACODF). ACODF algorithm has the following desirable strategies. It first uses differently favorable ants to solve the clustering problem. Then, the proposed ant colony system adopts simulated annealing concept for ants to decreasingly visit the amount of cities and get the local optimal solutions. Finally, it utilizes tournament selection strategy to choose a path. Every ant only needs to visit few cities instead of all of cities. Thus, the ant will reduce visiting the cities every iterations. After several iterations, the trail intensity close between nodes of trails will be increased. On the other hand, the trail intensity far between nodes of trails will be decreased. Therefore, ants will favor to visit the closer nodes and then reinforcing the trail with their own pheromone. Finally, a number of clusters will be built.

In [24], they applied the ant colony system (ACS) for clustering problem. Based on ACS, it treats the data (objects or elements) as the ants. Thus, each ant has different properties. Basically, the process of data clustering is the process of ant looking for food.

3. METHODOLOGY

This section intends to present the proposed method called Ant K-means algorithm. In this method, it is necessary to provide the number of clusters just like the conventional K-means algorithm. Both the random or predetermined initial start point can be applied. In the current study, the number of clusters and the centriod of each cluster are generated from ant system-based clustering algorithm (ASCA) which is developed by the authors. In order to validate clustering analysis solutions, Monte Carlo framework had been used in many literatures. One of the main advantages is that the researchers can use the analytical data with a known structure [25].

Following are the three methods which will be evaluated by using the data sets generated from the Monte Carlo method.

Use the SOM to determine the numbers of clusters and the initial points, then employ the K-means algorithm to find the final solution (called S + K in this paper) [3].

Use the SOM to determine the numbers of clusters and then employs genetic K-means to find the final solution (called S + G in this paper) [5].

Use ant system-based clustering algorithm to determine the numbers of clusters and the initial points, then employs proposed ant K-means to find the final solution (called ASCA + AK in this paper).

3.1. Monte Carlo Study

In this study, the data sets are generated by Monte Carlo method proposed in [6,7], for validating the feasibility of the clustering methods. By Monte Carlo method, the solutions of the clusters are known. Thus, the misclassify rate and the within cluster variance can be got for validating the solutions of the clustering methods. The factors for experiment design are

1. the number of clusters of the data sets,
2. the number of the data sets,
3. the density of the data sets, and
4. the error perturbation of the data sets.

Therefore, this experiment is the four-way factorial design, and each factor is at three levels arranged in this factorial experiment (shown in Table 1).

Table 1. Factors and levels of this experiment.

Factors \ Levels	I	II	III
The number of clusters	3	5	7
The number of dimensions	6	8	10
The level of density	Equal	10%	60%
The level of error perturbation	No	Low	High

Through the $3 \times 3 \times 3 \times 3$ full factorial design with three replications, 243 data sets are generated and each data set contains 120 data points.

3.2. Algorithm

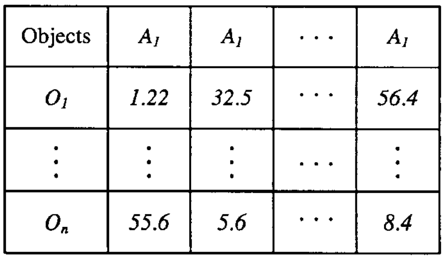
This subsection intends to present the proposed method, ant K-means. Since the initial solution is from ant system-based clustering algorithm, it will be introduced first.

3.2.1. Definitions and notations

The following terms and notations are used throughout this study.

Let

$$E = \{O_1, O_2, \dots O_n\}$$



Objects	A_1	A_1	$\cdots$	A_1
O_1	1.22	32.5	$\cdots$	56.4
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$
O_n	55.6	5.6	$\cdots$	8.4

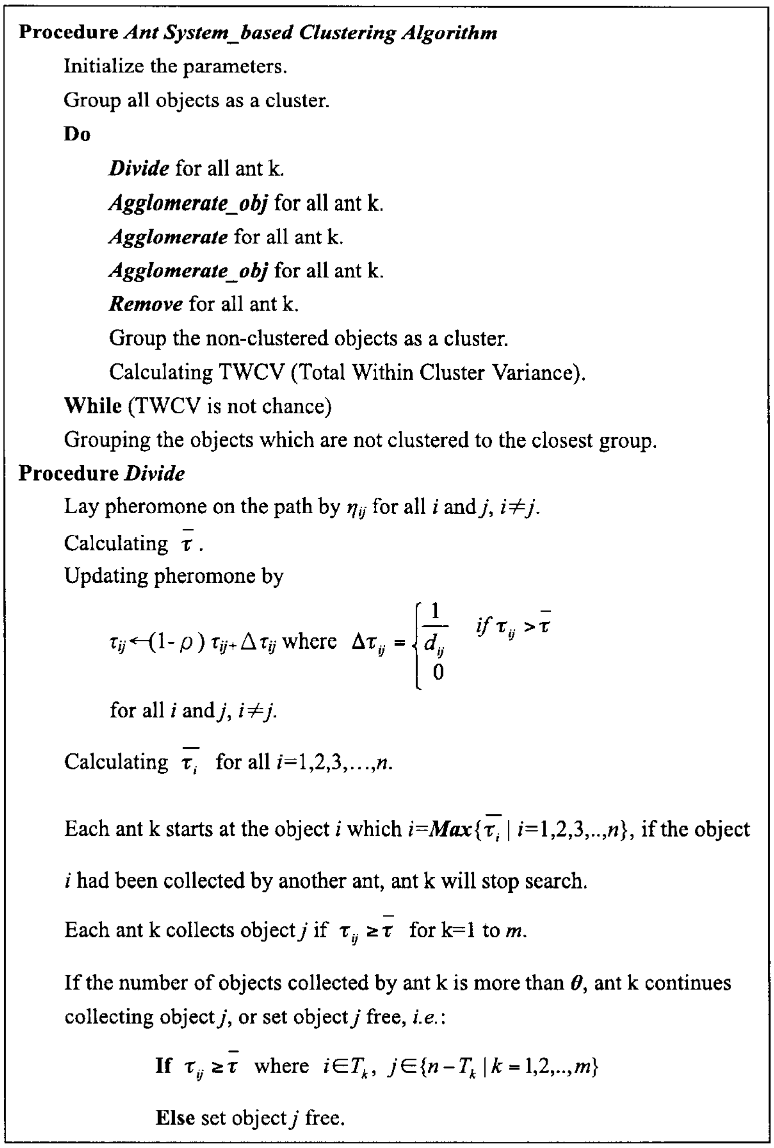
Figure 2. The format of the data set.

be the set of n data or objects, where O is the objects (or data, item) collected from the database. Each object has k attributes, where $k > 0$. This is shown in Figure 2.

- α: the relative importance of the trail,
$$\alpha \geq 0.$$
- β: the relative importance of the visibility,
$$\beta \geq 0.$$
- ρ: the pheromone decay parameter,
$$0 < \rho < 1.$$
- Q: a constant.
- n: number of objects.
- m: number of ants.
- nc: number of clusters.
- T is the set includes used objects. The maximal number recorded by T array will be n, i.e.,
$$T = \{O_a, O_b, \dots, O_t\},$$
where $a, b, \dots, t$ are the points that ant has been.
- T_k: the set T is performed by ant k.
- $O_{\text{center}}(T)$: the object which is the center of all objects in T, i.e.,
$$O_{\text{center}}(T) = \frac{1}{n_T} \sum_{O_i \in T} O_i, \tag{6}$$
where n_T is the number of objects in T.
- TWCV: total within cluster variance, i.e.,
$$\sum_{k=1}^{nc} \sum_{i \in k} (O_i, O_{\text{center}}(T_k))^2. \tag{7}$$

3.2.2. Ant system-based clustering algorithm (ASCA)

The algorithm of ASCA is including four subprocedures, that is, *divide*, *agglomerate_obj*, *agglomerate*, and *remove*. Following is the subscribing of procedures of ASCA. First, initialize the parameters and group all the objects as a cluster. Then, the subprocedure divide will divide the cluster into several subclusters and some object which does not belong to any subclusters through the consistency of the pheromone and some criterion. After divide, the agglomerate_obj



```
Procedure Ant System_based Clustering Algorithm
    Initialize the parameters.
    Group all objects as a cluster.
    Do
        Divide for all ant k.
        Agglomerate_obj for all ant k.
        Agglomerate for all ant k.
        Agglomerate_obj for all ant k.
        Remove for all ant k.
        Group the non-clustered objects as a cluster.
        Calculating TWCV (Total Within Cluster Variance).
    While (TWCV is not chance)
    Grouping the objects which are not clustered to the closest group.
Procedure Divide
    Lay pheromone on the path by $\eta_{ij}$ for all $i$ and $j$, $i \neq j$.
    Calculating $\bar{\tau}$.
    Updating pheromone by
        $\tau_{ij} \leftarrow (1-\rho)\,\tau_{ij} + \Delta\tau_{ij}$ where $\Delta\tau_{ij} = \begin{cases} \frac{1}{d_{ij}} & if\ \tau_{ij} > \bar{\tau} \\ 0 \end{cases}$
        for all $i$ and $j$, $i \neq j$.
    Calculating $\overline{\tau_i}$ for all $i$=1,2,3,…,$n$.

    Each ant k starts at the object $i$ which $i$=Max{$\overline{\tau_i}$ | $i$=1,2,3,..,$n$}, if the object
    $i$ had been collected by another ant, ant k will stop search.
    Each ant k collects object $j$ if $\tau_{ij} \geq \bar{\tau}$ for k=1 to $m$.
    If the number of objects collected by ant k is more than $\theta$, ant k continues
    collecting object $j$, or set object $j$ free, i.e.:
            If $\tau_{ij} \geq \bar{\tau}$ where $i \in T_k$, $j \in \{n - T_k \mid k = 1,2,..,m\}$
            Else set object $j$ free.
```

Figure 3. The procedure of ant system based clustering algorithm.

is the next step at this algorithm in order to agglomerate the objects into the suitable subcluster. Fourth, agglomerate is the subprocedure to merge the similar two subclusters into a cluster. Then, run agglomerate_obj again. Sixth, after agglomerating the similar object into the suitable subcluster, the remove subprocedure tries to remove the unsimilar from subcluster. Calculate the total within cluster variance (TWCV). If TWCV is not changed, grouping the nonclustered objects to the closest cluster, and stop the procedure. Otherwise, repeat the subprocedure divide, agglomerate_obj, agglomerate, agglomerate_obj, remove, round and round until TWCV is not changed. The detail algorithm of ASCA is introduced in [25] as shown in Figure 3.

3.2.3. Ant K-means clustering algorithm (AK)

The proposed method of this paper is ant K-means algorithm (AK). AK modifies the K-means as locating the objects in a cluster with the probability which is modified by the pheromone. The rule of updating pheromone is according to total within variance. The process is as following. The first step is initializing the parameters including the number of clusters and its centroid. Then,

```
Procedure Agglomerate_obj
    Let $C$ be the collector which collects $O_j$ to $T_k$ with the following equation.
    Let $C = \phi$.
    If $O_j$ satisfied with the following equation i.e.,
        $D(O_{center}(T_k), O_j) \geq D_{mean}(T_k) + 3Dev(T_k)$  where $j \in \{n - T_k \mid k = 1,2,\ldots,m\}$,
        add $O_j$ to $C$.
    Else $C = \phi$.
    If $(C \neq \phi)$
        If $(C < 2)$ Assign $O_j$ to $T_k$.
        Else $(C \geq 2)$ Assign $O_j$ to $T_k$ if the distance of $O_{center}(T_k)$ and $O_j$ is minimum.
    Else continue.
Procedure Agglomerate
    Do
        Let $C$ be the collector which collects the $T_k$ satisfied with the following
        equation.
        Let $C = \phi$.
        If $T_k$ satisfied with the following equation i.e.,
        $D(O_{center}(T_{k_i}), O_{center}(T_{k_j})) \leq D_{mean}(T_{k_i}) + D_{mean}(T_{k_j}) - \xi\, Dev(T_{k_i}) - \xi\, Dev(T_{k_j})$
        for all $k_i$ and $k_j$, $k_i \neq k_j$ $i$ ,$j$=1,2,3,...,$m$, add $T_k$ into $C$.
        Else $C = \phi$.
        If $(C \neq \phi)$
            If $(C < 2)$ Agglomerate $T_{k_i}$ and $T_{k_j}$ as a cluster.
            Else $(C \geq 2)$ Agglomerate $T_{k_i}$ and $T_{k_j}$ as a cluster if the distance of
                $O_{center}(T_{k_i})$ and $O_{center}(T_{k_j})$ is minimum.
        Else $(C = \phi)$ continue.
    While (there is no more cluster could be agglomerated)
Procedure Remove
    Remove object $j$ from Ant k if $D(O_j, O_{center}(T_k)) > 2Dev(T_k)$, where
    $j \in T_k$ and $T_k = 1,2,3,\ldots,m$.
```

Figure 3. (cont.)

lay equal pheromone on each path. Third, each ant k chooses the centroid to move with P, i.e.,

$$P_{ij}^k = \frac{\tau_{ij}^{\alpha} \eta_{ij}^{\beta}}{\sum_{c}^{nc} \tau_{ic}^{\alpha} \eta_{ic}^{\beta}}, \tag{8}$$

where i is the start point, j is the end point (centroid) which ant k chooses to move, c is the centroid and nc is the number of centroids. Therefore, if the value of P_{ij} is bigger than others, ant k will move from point i to point j, i.e., object i belongs to centroid j. Fourth, update the pheromone by

$$\tau_{ij} \leftarrow \tau_{ij} + \frac{Q}{\mathrm{TWCV}},$$

where Q is the constant, TWCV is the total within cluster variance. Then, calculate $O_{\text{center}}(T_k)$, where $k = 1, 2, 3, \ldots, nc$. After that, calculate TWCV. If TWCV is changed, go back to third

```
Procedure Perturbation
    Each Ant starts at random object and chooses the centroid randomly of cluster
    to move for all Ant k.
    Calculating $O_{center}(T_k)$ where $k = 1,2,...,nc$ and TWCV.
Procedure Ant_K-means
    Input the number of clusters and the corresponding centroids, and set the
    parameter $\alpha$, $\beta$, $\rho$, number of iterations and ants.
    Lay equal pheromone on each path.
    While (the number of iterations is not reached)
        Do
            Updating pheromone by $\tau_{ij} \leftarrow \tau_{ij} + \frac{Q}{TWCV}$.
            Each Ant k chooses the centroid to move with P, i.e.,
                $P = \frac{\tau_{kc}^{\alpha} \cdot \eta_{kc}^{\beta}}{\sum_{i=1}^{nc} (\tau_{ki}^{\alpha} \cdot \eta_{ki}^{\beta})}$.
            Calculate $O_{center}(T_k)$ where $k = 1,2,...,nc$
            Calculate TWCV (Total Within Cluster Variance).
        While (TWCV is not changed)
        If TWCV is smaller than the smallest TWCV, replace it.
        Perturbation.
```

Figure 4. The procedure of ant K-means.

step; otherwise, if TWCV is smaller than the smallest TWCV, replace it. The next step is to run the procedure perturbation in order to leap from the local minimal solution. If the number of iterations is not reached, go back to third step; otherwise, stop this algorithm. Figure 4 shows the procedure of ant K-means algorithm.

3.3. Evaluation of Three Clustering Methods

The number of misclassification rates of three clustering methods is compared with respect to their sensitivity and robustness. There are five hypotheses for evaluating the performance of three clustering methods.

Hypothesis 1. The number of misclassifications does not differ across three clustering methods.

Hypothesis 2. The number of misclassifications does not differ across the number of clusters in data set.

Hypothesis 3. The number of misclassifications does not differ across the number of dimensions of each observation.

Hypothesis 4. The number of misclassifications does not differ across the levels of error.

Hypothesis 5. The number of misclassifications does not differ across the levels of density.

4. SIMULATION RESULTS

This section will depict the results for the data obtained from the Monte Carlo simulation.

4.1. Verification of Random Number Generator

The simulation data sets were implemented (described in Section 3.1) in order to verify the efficiency of the three clustering tools. The random number generator must correspond to normal and uniform distribution. First, 1000 data points were generated via the random number generator, which grouped them into 20 segments. The chi-square test showed that these data fit the normal and uniform distribution. The testing results show that the random numbers generated by the generator fit a normal and uniform distribution. Thus, it is reasonable to accept the reliability of the random number generator.

4.2. The Result of ASCA + AK

Table 2 shows the result of the misclassification rates and total within cluster variance for three methods. That is, the misclassification rates and total within cluster variance of ASCA+AK are smaller than S + K in three, five, and seven clusters data sets, and smaller than S + G in three and seven clusters.

Table 2. Factors and levels of this experiment.

	Misclassification rate			TWCV		
	ASCA + AK	S + G	S + K	ASCA + AK	S + G	S + K
3 clusters	6.8271	7.1481	8.9259	46112.3232	46398.0610	8.9259
5 clusters	0.6543	0.3951	3.0370	47223.4444	46699.7432	3.0370
7 clusters	0.1358	0.4321	2.8148	45923.5556	49830.2254	2.8148

4.3. The Comparison of Three Clustering Methods

To examine the performance of four clustering methods, SPSS 10.0 is used for multivariate analysis of variance. The mean misclassification rate for three replications is used as the dependent variable for multivariate analysis of variance. Table 3 shows the results of multivariate analysis of variance for four methods. In Table 3, dimension, density level, and error level are significantly different at the $\alpha = 0.05$ level for each method. Density level × error level is significantly different in ASCA + AK and S + G. The mean misclassification rates under different levels of factors for each method are listed in Table 4. All factors of ASCA + AK are smaller than that of S + K, and the factors of ASCA + AK, which are three and seven clusters, eight and ten dimensions, equal density level and all of error level, are smaller than S + G.

Table 3. The format of the data set.

*The mean difference is significant at the $\alpha = 0.05$ level.

Factors \ Methods	ASCA + AK	S + G	S + K
Cluster Number	.734	.322	.181
Dimension	.000*	.000*	.000*
Density Level	.000*	.000*	.004*
Error Level	.000*	.000*	.000*
Cluster Number × Dimension	.015	.456	.214
Cluster Number × Density Level	.002*	.052	.603
Cluster Number × Error Level	.006	.452	.721
Dimension × Density Level	.000*	.085	.112
Dimension × Error Level	.012	.753	.428
Density Level × Error Level	.000*	.003*	.206
Cluster Number × Dimension × Density Level	.000*	.010*	.596
Cluster Number × Dimension × Error Level	.886	.999	.765
Cluster Number × Density Level × Error Level	.242	.437	.243
Dimension × Density Level × Error Level	.022	.988	.545
Cluster × Dimension × Density Level × Error Level	.487	.224	.762

4.4. The Analysis of Experiments

The main effect of each factor is discussed in this subsection. There are some further discussions as follows according to the results in Section 4.2.

Table 4. The procedure of ant system-based clustering algorithm.

	Level	ASCA + AK	S + G	S + K
Total Average		2.539	2.659	4.926
Cluster Number	3	6.827	7.148	8.926
	5	0.654	0.395	3.037
	7	0.136	0.432	2.815
Dimension	6	3.802	3.457	5.519
	8	1.889	2.099	4.667
	10	1.926	2.420	4.593
Density Level	Equal	2.407	3.074	5.000
	10%	2.420	2.235	5.914
	60%	2.790	2.667	3.864
Error Level	Free	0.111	0.321	11.198
	Low	1.260	1.358	2.580
	High	6.247	6.296	1.000

Table 5. The ANOVA for three clustering methods.

*The mean difference is significant at the $\alpha = 0.05$ level.

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0	P-Value
Methods	879.070	2	439.535	5.156	.006
Error	61883.695	726	85.239		
Total	62762.765	728			

Table 6. The simulation results of three methods.

*The mean difference is significant at the $\alpha = 0.05$ level.

(I) Methods	(J) Methods	Mean Difference(I-J)	P-value
ASCA + AK	S + K	−2.3868*	.018*
	S + G	−.1193	.990
S + K	ASCA + AK	2.3868*	.018*
	S + G	2.2675*	.026*
S + G	ASCA + AK	.1193	.990
	S + K	−2.2675*	.026*

Table 7. The multivariate analysis of variance for three methods.

Methods	ASCA + AK	S + G	S + K
Average total within cluster variance	46420.77	47643.67	53056.86

Hypothesis 1. The number of misclassifications does not differ across four clustering methods.

According to Table 5, these three clustering methods have significant difference. Further, Table 6 shows the Scheffé's multiple comparison test is employed for these three methods, and the number of misclassifications does not differ across ASCA + AK and S + G, means these two methods are not difference at mean misclassification. However, Table 7 shows that the average total within cluster variance of ASCA + AK is smaller than those of ASCA and S + G. Therefore, it could be said that the performance of ASCA + AK is the best method among these three clustering analysis methods.

Hypothesis 2. The number of misclassifications does not differ across the number of clusters in the data set.

According to Table 3, it is shown that the number of clusters does not affects cluster recovery of three methods significantly at $\alpha = 0.05$ level.

Hypothesis 3. The number of misclassification does not differ across the number of dimensions of each observation.

In Table 3, the number of dimensions affect cluster recovery of these three methods significantly at $\alpha = 0.05$ level.

Hypothesis 4. The number of misclassification does not differ across the levels of density.

Shown as Table 3, the density level affects the number of misclassification recovery of these three methods significantly at $\alpha = 0.05$ level.

Hypothesis 5. The number of misclassifications does not differ across the levels of error.

The error factor affects cluster recovery of three methods is as significant as the number of clusters at $\alpha = 0.05$ level in Table 3.

5. CASE STUDY AND DISCUSSION

The proposed method, ASCA + AK, is excellent for clustering analysis as shown in Section 4. To further check on this proposed method, a useful comparison of three methods applies the real case data. Recently, the price of plasma TV is gradually decreasing, the factories have to rethink the strategy of the price of plasma TV in such a low-price strain. Therefore, this study intends to apply the proposed method to address the suggestions of the sale promotion to be aimed at the different clusters via the questionnaire [26].

5.1. Clustering Analysis

Catherine and Paul [27] stated that the result will be better without rotation/standardization of factor score in clustering analysis. Therefore, this study takes the original data with 15 attributes as criterions in clustering analysis for the strategy of the sale promotion in each cluster.

As mentioned in Section 4, there is no significantly difference between S + G and ASCA + AK via the simulation data. However, in Table 7, the average total within cluster variance of ASCA + AK is smaller than that of S + G. Thus, the performance of ASCA+AK is better than that of S + G. Then, testifying the sensitivity and robustness of the proposed method, ASCA + AK, is illustrated in the following subsection by using the real-world problem.

5.2. Parameters Setup

According to [22], there are several combinations for determining parameters setup as applying ant colony system algorithm. They are $\alpha = \{0, 0.5, 1, 2, 5\}$, $\beta = \{0, 1, 2, 5\}$, $\rho = \{0.3, 0.5, 0.7, 0.99, 0.999\}$, and $Q = \{1, 100, 10000\}$. Thus, there are 300 combinations for parameters ($5 \times 4 \times 5 \times 3$). After coding the proposed method using Visual C++, the results showed that as $\alpha = 0.5$, $\beta = 1$, $\rho = 0.9$, and $Q = 1$, the proposed method has the smallest total within cluster variance, 7530.55. This result will be employed to make the comparison with other two existing methods.

5.3. Evaluation of Two Methods

After determining the number of clusters, three, through ASCA, AK is employed to cluster 354 samples. In Table 8, ASCA + AK, which determines the number of clusters and its centroid by ASCA and then employ the AK to find the final solutions, has the best efficiency compared with S + G, which is also the two-stage method. Thus, ASCA + AK is employed as the clustering tool for this case study.

Table 8. The mean misclassification rates under different factors for three methods.

	ASCA + AK	S + G	S + K
TWCV	7530.55	7702.49	7753.64

5.4. Factors Named

According to the result of the clustering by ASCA + AK, the sorting of the average importance level of each item in different clusters are listed in Table 9.

Table 9. The sorting of the average importance level of each item in different clusters.

Item #	Cluster 1	Item #	Cluster 2	Item #	Cluster 3
6	6.13	11	6.20	14	4.83
14	5.89	14	5.80	11	4.67
13	5.52	6	5.72	6	4.53
11	5.46	10	5.58	15	4.46
12	5.42	12	5.56	13	4.25
15	5.38	13	5.37	5	4.10
9	5.27	9	5.33	9	4.00
4	5.13	15	5.26	12	3.99
10	4.88	4	4.54	10	3.87
2	4.60	8	4.49	4	3.65
1	4.57	5	4.46	7	3.49
3	4.57	7	2.93	8	3.28
8	4.44	1	2.60	3	2.62
5	4.10	3	2.34	1	2.60
7	3.55	2	2.23	2	2.46

a. Cluster 1: The top 5 of Cluster 1 are
- "the higher of the discount of the production price, the higher the buying intention,"
- "the more positive in evaluation of the brand, the higher the buying intention,"
- "the evaluation is good in this plasma TV (Pioneer),"
- "if the price is lower in other advertisement than this advertisement, it will influence the buying intention," and
- "the quality of this brand (Pioneer) is good."

The above items get the higher score than others in Cluster 1. Therefore, Cluster 1 is named as "cluster of discount and brand consideration."

b. Cluster 2: The top 5 of Cluster 2 are
- "if the price is lower in other advertisement than this advertisement, it will influence the buying intention,"
- "the more positive in evaluation of the brand, the higher the buying intention,"
- "the higher of the discount of the production price, the higher the buying intention,"
- "buying a high quality plasma TV by paying a reasonable price according to the advertisement," and
- "the quality of this brand (Pioneer) is good."

The above items get the higher score than others in Cluster 2. Thus, Cluster 2 can be named as "cluster of discount considered, but easily influenced by external environment."

c. Cluster 3: The top 5 of Cluster 3 are
- "the more positive in evaluation of the brand, the higher the buying intention,"
- "if the price is lower in other advertisement than this advertisement, it will influence the buying intention,"
- "the higher of the discount of the production price, the higher the buying intention,"
- "if I want to buy the plasma TV, this brand of plasma TV will be considerable," and
- "the evaluation is good in this plasma TV (Pioneer)."

The above items get the higher score than others in Cluster 3. Therefore, Cluster 2 is called "cluster of high identification in brand."

5.5. The Mean Difference Analysis for Each Factor

After three clusters are generated by the proposed method, ASCA + AK, the mean difference of clusters for each factor are concerned. Table 10 shows the mean differences of clusters of 15 factors are all significant, further, in order to know the different factors have difference or not for each cluster, Scheffé's multiple comparison test are employed as shown in Table 11. There are 82.22% factors are significantly different between each cluster. Therefore, ASCA + AK can help this research find quite good clusters.

Table 10. The mean difference of clusters for each factor.

*The mean difference is significant at $\alpha = 0.025$ level.

Factor	Item_1	Item_2	Item_3	Item_4	Item_5	Item_6	Item_7	Item_8
P-Value	0.00*	0.00*	0.00*	0.00*	0.00*	0.00*	0.00*	0.00*
Factor	Item_9	Item_10	Item_11	Item_12	Item_13	Item_14	Item_15	
P-Value	0.00*	0.00*	0.00*	0.00*	0. 00*	0.00*	0.00*	

Table 11. The Scheffé's multiple comparison test for each cluster.

Factors \ Methods	1-2	1-3	2-3
Item_1	0.00*	0.00*	1.00
Item_2	0.00*	0.00*	0.18*
Item_3	0.00*	0.00*	0.10
Item_4	0.00*	0.00*	0.00*
Item_5	0.00*	0.00*	0.00*
Item_6	0.00*	0.00*	0.00*
Item_7	0.01*	0.94	0.00*
Item_8	0.00*	0.00*	0.00*
Item_9	0.96	0.00*	0.00*
Item_10	0.00*	0.00*	0.00*
Item_11	0.00*	0.00*	0.00*
Item_12	0.62	0.00*	0.00*
Item_13	0.52	0.00*	0.00*
Item_14	0.84	0.00*	0.00*
Item_15	0.73	0.00*	0.00*

5.6. Suggestions

In case study, this research accord to analyze the customer's buying will and the production's discount, the suggestions of marketing strategy for three clusters are represented as follows.

Cluster 1. Discount and brand consideration.

The customers belonging to this cluster think that the discount is the most important consideration. Sometimes, they will have higher buying intention as the product is discounted. They also pay much attention to the brand because this group of customers thinks that good brand always has good quality. Therefore, the company has to maintain the quality of product and tries to enhance the branding value.

Cluster 2. Discount considered, but easily influenced by the external environment.

The customers belonging to this cluster also consider that the discount is important, but they are easily influenced by the external environment. For example, if there are two plasma TVs, one is the well-known brand and the other is a new brand but its plasma TV has a quite good quality even as good as the well-known brand and cheaper than it. The customer belonging to this cluster will have high intention to buy the second plasma TV. Therefore, no doubt, the higher quality and lower price sale promotion is the only way to sell the product to these customers in this cluster.

Cluster 3. High identification in brand.

The customers in this cluster think that the brand is everything. They will not rashly buy the product which is the new brand, even if this product is cheaper. That is, these customers are called the cluster of high identification in brand. Therefore, the company has to continue maintaining these customers' loyalty to the brand. Also, of course, the lower price sale promotion will raise the intention of buying.

6. CONCLUSIONS

This study has demonstrated that the proposed clustering method, AK. The only problem for AK is that the number of clusters is required. In practice, it is also very difficult to determine the number of clusters. Thus, ASCA proposed by the authors is able to handle this problem. In Section 4, the experimental result for the data generated by Monte Carlo shows that ASCA + AK is better than S + G as well as S + K. In addition, in Section 5, the result of clustering analysis from real-world problem also illustrates that ASCA + AK is better than S + G based on average total within cluster variance. According to the above results, the proposed method, ant K-means algorithm (AK), which needs the number of clusters and the initial points, is a robust clustering method. It can be applied to many different kinds of clustering problems or combined with some other data mining techniques for getting more promising results for industries.

REFERENCES

1. M. Dorigo and T. Stützle, *The Ant Colony Optimization Metaheuristic: Algorithms, Applications, and Advances*, Technical Report IRIDIA, (2000).
2. D. Corne, M. Dorigo and F. Glover, *New Ideas in Optimization*, McGraw-Hill, (1999).
3. R.J. Kuo, L.M. Ho and C.M. Hu, Integration of self-organizing feature map and K-means algorithm for market segmentation, *International Journal of Computers and Operations Research* **29**, 1475–1493, (2002).
4. R.J. Kuo, K. Chang and S.Y. Chien, Integration of self-organizing feature map and genetic algorithm based clustering method for market segmentation, *Journal of Organizational Computing and Electronic Commerce* **14** (1), 43–60, (2004).
5. R.J. Kuo, C.L. Liao and C. Tu, Integration of ART2 neural network and genetic K-means algorithm for analyzing web browsing paths in electronic commerce, *Decision Support Systems* **40** (2), 355–374, (2005).
6. G.W. Milligan, An Examination of the Effect of Six Types of Error perturbation on fifteen Cluster Algorithms, *Psychometrika* **45** (3), 325–342, (1980).
7. G.W. Milligan, An Algorithm for generating Artificial Test Clusters, *Psychometrika* **50** (1), 123-127, (1985).
8. U. Fayyad, G. Piatetsky-Shapiro and P. Smyth, From Data Mining to Knowledge Discovery in Database, *American Association for Artificial Intelligence*, 37–54, (August 1996).
9. U. Fayyad, Data Mining and Knowledge Discovery in Databases: Implications for Scientific Databases, *Scientific and Statistical Database Management, 1997. Proceedings, Ninth International Conference*, 2–11, (1997).
10. F.H. Daniels and M. Holsheimer, Methodological and Practical Aspects of Data Mining, *Information and Management* **37**, 271–281, (2000).
11. I.K. Sethi, *Data Mining: An Introduction, Data Mining For Design and Manufacturing*, Kluwer Academic Publishers, (2001).
12. A.K. Jain, M.N. Murty and P.J. Flynn, Data clustering: A review, *ACM Computing Surveys* **31** (3), (September 1999).
13. I.H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, Morgan Kaufmann Publishers, (2000).
14. P. Berkhin, Survey of Clustering Data Mining Techniques, *Accrue Software, Inc.* `http://www.accrue.com/products/researchpapers.html`, (2002).
15. T. Kohonen, Self-organizing maps: Optimization approaches, In *Artificial Neural Networks*, (Edited by K. Makisara, O. Simula and J. Kangas), pp. 981–990, Elsevier, Amsterdam, The Netherlands, (1991).

16. G.A. Carpenter and S. Grossberg, ART2: Self-Organization of Stable Category Recognition Codes for Analog Input Pattern, *Applied Optics* **26**, 4919–4930, (1987).
17. J.A. Freeman and D.M. Skapura, *Neural Networks: Algorithms, Applications, and Programming Techniques*, Addison-Wesley Publishing Company, Inc., (1992).
18. S.Z. Selim and M.A. Ismail, K-means-type algorithms: A generalized convergence theorem and characterization of local optimality, *IEEE Trans. on Pattern Analysis and Machine Intelligence* **6** (1), 81–87, (1984).
19. H. Maulik and S. Bandyopadhyay, Genetic Algorithm-Based Clustering Technique, *Pattern Recognition* **33**, 1455–1465, (2000).
20. K. Krishna and M. Murty, Genetic K-Means Algorithm, *IEEE Transactions on Systems, Man, and Cybernetics—Part B: Cybernetics* **29** (3), 433–439, (1999).
21. R.J. Kuo and W.J. Chung, Integration of self-organizing map and genetic K-means algorithm for data mining, *Proceedings of 30th International Conference of Computer and Industrial Engineering, Tinos Island, Agean Sea, Greece, June 29–July 2, 2002*, (2002).
22. M. Dorigo, V. Maniezzo and A. Colorni, The Ant System: Optimization by a colony of cooperating agents, *IEEE Transactions on Systems, Man and Cybernetics—Part B* **26** (1), 1–13, (1996).
23. C.F. Tsai, H.C. Wu, and C.W. Tsai, A new clustering approach for data mining in large databases, *Proceedings of the international Symposium on Parallel Architectures, Algorithms and Networks (ISPAN'02), IEEE Computer Society*, 1087–4089, (2002).
24. X.B. Yang, J.G. Sun and D. Huang, A new clustering method based on ant colony algorithm, *Proceedings of the 4th World Congress on Intelligent Control and Automation* **2222–2226**, (June 2002).
25. R.J. Kuo, C.L. Cha, S.H. Chou, C.W. Shih and C.Y. Chiu, Integration of ant algorithm and case based reasoning for knowledge management, *Proceedings of International Conference on IJIE, Nov. 10–12, 2003, Las Vegas, NV, U.S.A., (in CD-R)*, (2003).
26. Y.N. Ying, *The Study of Price Promotion on Brand Evaluation and Purchase Intention—An Empirical Investigation of the PDP Industry*, Master Thesis, (in Chinese), National Taipei University of Technology, (2003).
27. C.M. Schaffer and P.E. Green, Cluster-based market segmentation: Some of further comparisons of alternative approaches, *Journal of the Market Research Society* **40** (2), 155–163, (April 1998).