ELSEVIER

European Journal of Operational Research 135 (2001) 413–427

EUROPEAN JOURNAL OF OPERATIONAL RESEARCH

www.elsevier.com/locate/dsw

Theory and Methodology

Genetic clustering algorithms

Yu-Chiun Chiou [a], Lawrence W. Lan [b,*]

[a] *Aviation and Maritime Management Department, Chang Jung Christian University, Tainan 711, Taiwan, ROC*
[b] *Institute of Traffic and Transportation, National Chiao Tung University, 4F, 114 Sec. 1, Chung-Hsiao W. Rd., Taipei 10012, Taiwan, ROC*

Received 15 November 1999; accepted 20 November 2000

Abstract

This study employs genetic algorithms to solve clustering problems. Three models, SICM, STCM, CSPM, are developed according to different coding/decoding techniques. The effectiveness and efficiency of these models under varying problem sizes are analyzed in comparison to a conventional statistics clustering method (the agglomerative hierarchical clustering method). The results for small scale problems (10–50 objects) indicate that CSPM is the most effective but least efficient method, STCM is second most effective and efficient, SICM is least effective because of its long chromosome. The results for medium-to-large scale problems (50–200 objects) indicate that CSPM is still the most effective method. Furthermore, we have applied CSPM to solve an exemplified *p*-Median problem. The good results demonstrate that CSPM is usefully applicable.

Keywords: Genetic algorithms; Clustering; *p*-Median problem

1. Introduction

Clustering, so-called set partitioning, is a basic and widely applied methodology. Application fields include statistics, mathematical programming (such as location selecting, network partitioning, routing, scheduling and assignment problems, etc.) and computer science (including pattern recognition, learning theory, image processing and computer graphics, etc.). Clustering is mainly to group all objects into several mutually exclusive clusters in order to achieve the maximum or minimum of an objective function. Clustering is rapidly becoming computationally intractable as problem scale increases, because of the combinatorial character of the method. Brucker [7] and Welch [34] proved that, for specific objective functions, clustering becomes an NP-hard problem when the number of clusters exceeds 3. Even the best algorithms developed for some specific objective functions, exhibit complexities of $O(N^3 \log N)$ or $O(N^3)$ [15], leaving much room for improvement. The heuristic algorithms for clustering can be divided into four categories:

* Corresponding author. Tel.: +886-2-2311-0094; fax: +886-2-2331-2160.
E-mail address: lwlan@cc.nctu.edu.tw http://www.itt.nctu.edu.tw (L.W. Lan).

0377-2217/01/$ - see front matter
PII: S0377-2217(00)00320-9

conventional statistics clustering, mathematical programming, network programming, and genetic algorithms (GAs). The algorithms for conventional statistic clustering [3,14,18,25,33] include agglomerative hierarchical clustering method and *K*-means. The algorithms for mathematical programming [8,11,17,27,28,30–32] range from dynamic programming, Lagrangian relaxation, linear relaxation, column generation, branch-and-price and Lipschitz continuous. The algorithms for the network programming [2,12] include graph theoretic relaxations and network relaxation. The algorithms for GAs are rapidly developed recently [4,6,10,19,20,22–24,26] including group-numbers encoding method (e.g. binary code, Boolean matching code), group-separators encoding method and evolution program method.

While the aforementioned studies have proposed ways to solve clustering problems, two main research gaps still remain. First, the number of clusters must be subjectively determined in advance. This number cannot be determined simultaneously by the model. Therefore, the above studies involve a complex procedure that exhaustively compares all the optimum clustering for every given number of clusters, then determines the number of clusters of best objective value. Exceptions for this gap are the studies of Lozano et al. [22] and Lunchian et al. [25] in which they only solved the optimal number of clusters without developing an explicit algorithm for the assignment problem. Second, most of the non-GAs based algorithms are limited in applications. They are proposed under a specific form of the objective function such as convex function, or proposed by the assumption that the feasible set is a convex hull, or proposed with the help of additional information such as the gradient of the objective function.

GAs, first proposed by Holland [16], are general-purpose search algorithms that have the characteristics of stochastic search, multi-points search, direct search and parallel search. The related articles have proved the effectiveness and efficiency of GAs in application to the combinatorial optimization problems [5,9,13,21,29]. Directly using the fitness to evaluate the chromosomes, GAs can be applied to various objective functions without a need for additional information in the search. This study attempts to develop coding/decoding techniques for GAs to solve simultaneously the optimal number of clusters and the optimal clustering result in comparison to the conventional statistic clustering method (the agglomerative hierarchical clustering method).

2. Mathematical model of clustering

The mathematical model of clustering of given number (m) of clusters is

$$[\mathrm{CA}_m]$$

$$\text{Max } F(X) \tag{1}$$

subject to

$$\sum_j X_{ij} = 1 \quad \text{all } i, \tag{2}$$

$$\sum_j X_{jj} = m, \tag{3}$$

$$X_{ij} \leqslant X_{jj} \quad \text{all } i, j, \tag{4}$$

$$X_{ij} = \{0, 1\} \quad \text{all } i, j, \tag{5}$$

where $X_{ij} = 1$ denotes that ith object is assigned to jth cluster, $X_{ij} = 0$ otherwise, $i, j = \{1, 2, \ldots, N\}$, N is the number of objects, m is the number of clusters, $F(X)$ is the objective function. In the application field of statistics, $F(X)$ can be generally defined as [30]:

$$F(X)^d = \min_{P_m \in \pi_m} \|\{d(S_1), \ldots, d(S_m)\}\|_p, \tag{6}$$

$$F(X)^r = \max_{P_m \in \pi_m} \big\|\{r(S_i, S_j),\ 1 \leqslant i \leqslant j \leqslant m\}\big\|_p, \tag{7}$$

$$F(X)^s = \max_{P_m \in \pi_m} \|\{s(S_1), \ldots, s(S_m)\}\|_p, \tag{8}$$

where S_i is ith cluster, $i = 1, \ldots, m$. P_m is a clustering result of m-clustering, $P_m = \{S_1, \ldots, S_m\}$. π_m is a set of all possible clustering results of m-clustering. $F(X)^d$ is the diameter function of clustering and $d(S_j) = \max_{O_k, O_l \in S_j} d_{kl}$ is the diameter of jth cluster; $F(X)^r$ is the distance function of clustering and $r(S_i, S_j) = \min_{O_k \in S_i, O_l \in S_j} d_{kl}$ is the distance or discrimination between ith and jth clusters; $F(X)^s$ is the split function of clustering and $s(S_j) = \min_{O_k \in S_j, O_l \notin S_j} d_{kl}$ is shortest distance between jth

cluster with other clusters. $\|\cdot\|_p$ represents the l_p-norm. $\{\cdot\}$ represents the measure of a vector of diameter or distance. O_k, O_l are kth and lth objects, respectively.

The total number of feasible solutions of $[CA_m]$ is $|\pi_m| = (1/m!)\sum_{j=0}^{m}(-1)^{m-j}\binom{m}{j}j^N$ [1]. There are a total of 511 feasible solutions for 10 objects ($N = 10$) to be divided into 2 clusters ($m = 2$). There are a total of 42,525 feasible solutions for 10 objects to be divided into 5 clusters.

If the number of clusters (m) is not given exogenously, then $[CA_m]$, without the constraint (3), becomes [CA], that is:

[CA]

$$\text{Max } F(X) \tag{9}$$

subject to

$$\sum_j X_{ij} = 1 \quad \text{all } i, \tag{10}$$

$$X_{ij} \leqslant X_{jj} \quad \text{all } i, j, \tag{11}$$

$$X_{ij} = \{0, 1\} \quad \text{all } i, j, \tag{12}$$

As to the number m of feasible solutions of [CA] is $|\pi| = \sum_{m=1}^{N}|\pi_m|$. There are 52 feasible solutions at $N = 5{,}113{,}608$ at $N = 10$ and 1.99×10^7 at $N = 15$. Obviously, the complexity of [CA] is exponential to the problem size.

3. Genetic algorithms

Genetic algorithms are general-purpose search algorithms that use principles inspired by natural population genetics to evolve solutions to problems. The basic idea of genetic algorithms is to maintain a population of chromosomes that represent candidate solutions. A chromosome is composed of a series of genes that represent decision variables or parameters. Each member of the population is evaluated and assigned a measure of its fitness as a solution. There are three genetic operators: selection, crossover and mutation [13]:

The first operator – selection – is to assign the reproduction possibilities to chromosomes based on their fitness. A Monte Carlo wheel is often employed. That is, the higher fitness a chromosome is, the more possible it is selected.

The second operator – crossover – is to combine the features of two parent structures to form two offsprings. The simplest way to make a crossover is to swap a corresponding segment of the parents. One-point crossover, two-point crossover and uniform crossover are often employed.

The last operator – mutation – is to alter one or more genes of offsprings with a very low probability to avoid being trapped in a local optimum. The resulting offspring is then evaluated and inserted back into the population. This process continues until a predetermined criterion (e.g. maximum number of generations, minimum value of fitness improved between two adjacent generations or certain mature rate) is reached.

4. Problem formulation

The effectiveness and efficiency of GAs vary with various coding/decoding techniques. This study proposes three coding/decoding techniques for GAs to solve clustering problems. They are the simultaneously clustering method (SICM), the stepwise clustering method (STCM) and the cluster seed points method (CSPM). Then, these models are compared with a conventional statistics clustering model – the agglomerative hierarchical clustering method (AHCM). The details of these four models are described as follows.

4.1. Agglomerative hierarchical clustering method (AHCM)

AHCM involves a series of successive merges. Initially, there are as many clusters as objects. These initial groups are merged according to their degree of improvement in the objective values. Eventually, all subgroups are fused into a single cluster [33]. The following are the steps in AHCM for grouping N objects in a maximizing problem for example:

Step 0. Start with N clusters, each containing a single object, that is, $S_i^{(1)} = \{O_i\}$, $i = 1, \ldots, N$. An

$N \times N$ symmetric matrix increments of an objective function $MF = \{\Delta F_{ij},\ i,j = 1,\ldots,N,\ i \neq j\}$, where ΔF_{ij} represents the incremental of objective value in case that ith cluster and jth cluster are fused into a single cluster. Let $k = 1$.

Step 1. If $\Delta F_{vu} = \max\{\Delta F_{ij},\ i,j = 1,\ldots,N-k,\ i \neq j\}$ and $v > u$, then let $S_u^{(k+1)} = S_u^{(k)} \cup S_v^{(k)}$, $S_1^{(k+1)} = S_1^{(k)},\ldots,S_{u-1}^{(k+1)} = S_{u-1}^{(k)},\ S_{u+1}^{(k+1)} = S_{u+1}^{(k)},\ldots,S_v^{(k+1)} = S_{v+1}^{(k)},\ldots,S_{N-v-1}^{(k+1)} = S_{N-v}^{(k)}$ and $S_{N-k}^{(k+1)} = \Phi$. Calculate the objective value $F(X)^{(k)}$ of the partition. Let $k = k+1$.

Step 2. Repeat Step 1 until $k = N-1$. $F(X)^* = \max\{F(X)^{(k)},\ k = 1,\ldots,N-1\}$.

4.2. Simultaneously clustering method (SICM)

If there must be at least two objects to form a cluster, the maximal number of clusters, K, is equal to $[N/2]$ ([] is the Gauss sign). Then, there are altogether $N \times K$ decision variables of problem [CA], as shown in Table 1. If there are more than two variables with value of 1 in the same column, it means that they mutually form a cluster. However, if X_{ik} is encoded as a gene, it will cause the length of chromosome be too long (for instance, the number of decision variables for 10 objects is 50, for 20 objects is 200, and for 60 objects is 1800) and will result in an insufficiency of computer memory. In addition, it will be difficult to handle the constraints if one and only one variable equals 1 and else equals 0 in the same row.

In order to deal with the problem, SICM uses a coding/decoding technique to replace each row of the decision variable matrix with a shorter gene string. Take 17 objects for example, there are at most 8 partition sets ($8 = [17/2]$). Because each object is likely to be assigned to any set, these clusters require three genes to represent them, as shown in Table 2.

Table 1
Relationship between N objects and K clusters

Object	Clusters					
	1	2	...	k	...	K
1	X_{11}	X_{12}		X_{1k}		X_{1K}
2	X_{21}	X_{22}		X_{2k}		X_{2K}
⋮						
i	X_{i1}	X_{i2}		X_{ik}		X_{ik}
⋮						
N	X_{N1}	X_{N2}		X_{Nk}		X_{NK}

Table 2
Matching rules for gene strings, integers, and clusters

Gene strings	Integers	Clusters
000	0	1
001	1	2
010	2	3
011	3	4
100	4	5
101	5	6
110	6	7
111	7	8

Replacement of each row with these three genes not only curtails the length of chromosomes (four genes can represent the problem of 33 objects, five genes can represent the problem of 65 objects) but also avoids the problem that an object might be assigned to several clusters or be unassigned. Table 3 illustrates a feasible clustering result for these 17 objects. The chromosome of Table 3 is composed of 51 genes (000001011000101101010000001000 00001101100000010101O). Every three genes of the chromosome are then decoded into an integer of 0–7 sequentially, representing the cluster into which each object is assigned according to the matching rules stated in Table 2. After being decoded, the chromosome represents that five clusters are formed. The clusters consist of $6, 2, 2, 3, 3$ objects, respectively.

4.3. Stepwise clustering method (STCM)

STCM successively solves the optimal binary clustering of a cluster until the objective value cannot be further improved. An initial single cluster containing all objects is divided into two subgroups such that the objective function is optimized at this stage. Through each binary clustering process, a cluster is divided into two subgroups. A cluster is called fathomed when it cannot be further binary clustered to improve the objective value. This concept is similar to that of

Table 3
Relationship between objects and clusters (17 objects for example)

Objects	Clusters								Encoding
	1	2	3	4	5	6	7	8	
1	1	0	0	0	0	0	0	0	000
2	0	1	0	0	0	0	0	0	001
3	0	0	0	1	0	0	0	0	011
4	1	0	0	0	0	0	0	0	000
5	0	0	0	0	0	1	0	0	101
6	0	0	0	0	0	1	0	0	101
7	0	0	1	0	0	0	0	0	010
8	1	0	0	0	0	0	0	0	000
9	0	1	0	0	0	0	0	0	001
10	1	0	0	0	0	0	0	0	000
11	1	0	0	0	0	0	0	0	000
12	0	0	0	1	0	0	0	0	011
13	0	0	0	1	0	0	0	0	011
14	1	0	0	0	0	0	0	0	000
15	1	0	0	0	0	0	0	0	000
16	0	0	0	0	0	1	0	0	101
17	0	0	1	0	0	0	0	0	010
Subtotal	6	2	2	3	0	3	0	0	

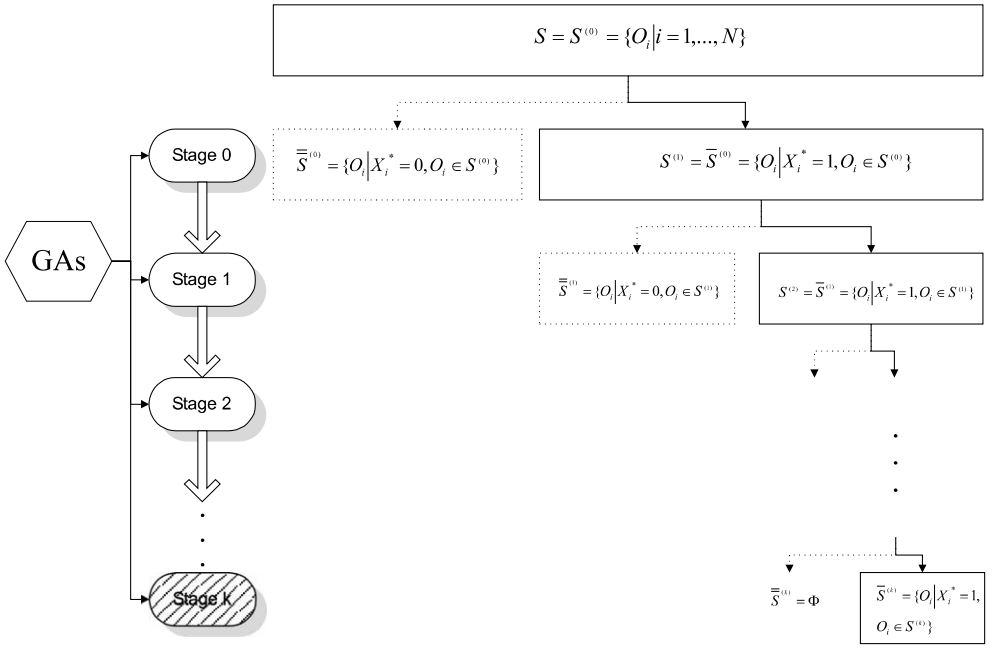

Fig. 1. Framework of STCM.

branch-and-bound. When all clusters are fathomed, STCM has attained the optimal clustering. The framework of the model is depicted by Fig. 1. The following are the algorithms for STCM under the depth first principal, which fathoms individual branch one at a time.

Step 0. Let $S^{(0)}$ stand for the cluster containing all objects. The problem of optimally dividing $S^{(0)}$ into two subgroups, namely $\bar{S}^{(0)}$ and $\bar{\bar{S}}^{(0)}$, can be formulated as the following 0–1 mathematical programming:

$\mathrm{MP}^{(0)}$

$$\text{Max } F(X)^{(0)} \tag{13}$$

$$\text{subject to} \quad X_i = \{0,1\} \quad i = 1,\ldots,|S^{(0)}|, \tag{14}$$

where $X_i = 1$ denotes that ith object of $S^{(0)}$ is grouped in the cluster $\bar{S}^{(0)}$, $X_i = 0$ denotes that ith object of $S^{(0)}$ is grouped in the cluster $\bar{\bar{S}}^{(0)}$. X_i is encoded as the gene of chromosomes (the length of chromosomes is $|S^{(0)}|$), then GAs are employed to solve $MP^{(0)}$ by maximizing $F(X)^{(0)}$ to attain the optimal binary clustering: $\bar{S}^{(0)} = \{O_i | X_i^* = 1,\ O_i \in S^{(0)}\}$ and $\bar{\bar{S}}^{(0)} = \{O_i | X_i^* = 0,\ O_i \in S^{(0)}\}$.

Step 1. Let $S^{(1)} = \bar{S}^{(0)}$ and renumber the objects of $S^{(1)}$. Formulate the optimal binary clustering problem of $S^{(1)}$ as $\mathrm{MP}^{(1)}$, which is also solved by GAs. $F(X)^{(1)^*}$ is the objective value of optimal binary clustering of $S^{(1)}$ under the assumption that the other cluster ($\bar{\bar{S}}^{(0)}$) remaining unchanged. The clustering result is: $\bar{S}^{(1)} = \{O_i | X_i^* = 1,\ O_i \in S^{(1)}\}$ and $\bar{\bar{S}}^{(1)} = \{O_i | X_i^* = 0,\ O_i \in S^{(1)}\}$. Three clusters are formed, they are: $\bar{\bar{S}}^{(0)}$, $\bar{S}^{(1)}$ and $\bar{\bar{S}}^{(1)}$. $F(X)^{(1)^*}$ is the objective value of these three clusters.

Step 2. Let $S^{(i)} = \bar{S}^{(i-1)}$ and solve $\mathrm{MP}^{(i)}$ by GAs.

Step 3. Repeat Step 2 until $\bar{\bar{S}}^{(k)} = \Phi$, then this branch is fathomed. There is a total of $k+1$ clusters, that is $\bar{\bar{S}}^{(0)}, \bar{\bar{S}}^{(1)}, \ldots, \bar{\bar{S}}^{(k-1)}$ and $\bar{S}^{(k)}$. $\bar{S}^{(k)}$ can no longer be divided in the following steps. $F(X)^{(k)*}$ is the optimal objective value of these $k+1$ clusters.

Step 4. Choose one of the remaining branches to be binarily clustered by repeating Steps 2 and 3 until it is fathomed.

Step 5. If all branches are fathomed, then stop. The clusters formed are the optimal clustering result of STCM. Otherwise, go to Step 4.

In comparison to SICM, the coding/decoding of STCM are much simpler because the length of chromosome can be largely curtailed and can be further reduced in the evolutions of optimization stages. Consider N objects for instance, let $|S^{(0)}| = N$ denote N objects in set $S^{(0)}$, the length of the chromosome at the stage 0 is N. If $|\bar{\bar{S}}^{(0)}| = L_0$, the length of the chromosome at the stage 1 is $N - L_0$. If $|\bar{\bar{S}}^{(1)}| = L_1$, the length of chromosome at the stage 2 can be further shortened as $N - L_0 - L_1$, and so forth.

4.4. Cluster seed points method (CSPM)

CSPM first employs GAs to select the most suitable cluster seeds from all objects, then assigns the rest of the objects to each cluster according to their similarity to the cluster seed or to their degree of improvement of the objective function. The number of cluster seeds represents the number of clusters and the characteristics of these cluster seeds determine the clustering result. The framework of CSPM is depicted by Fig. 2.

The following are the steps of the assignment algorithm in Fig. 2.

Step 0. Let $k = 1$ and S be a set of all objects, that is, $S = \{O_1, \ldots, O_N\}$. CP_m is a set of cluster seeds, that is, $CP_m = \{c_1, \ldots, c_m\}$. NP is a set of non-cluster seeds, that is, $NP = S - CP_m \cdot S_j^{(0)} = \{c_j\}$, $j = 1, \ldots, m$.

Step 1. Let O_k denote kth object of NP. If $F(S_1^{(k)}, \ldots, S_{j-1}^{(k)},\ S_j^{(k)} \cup \{O_k\},\ S_{j+1}^{(k)}, \ldots, S_m^{(k)}) = \mathrm{Max}_i\{F(S_1^{(k)}, \ldots,\ S_{i+1}^{(k)}, S_i^{(k)} \cup \{O_k\},\ S_{i+1}^{(k)}, \ldots, S_m^{(k)}\}$, then O_k is assigned to jth cluster.

Step 2. Let $S_j^{(k)} = S_j^{(k-1)} \cup \{O_k\}$ and $k = k+1$. If $k < N - m + 1$, return to Step 1, otherwise terminate.

Once the clustering result, P_m, is obtained, the objective value $F(X)$ which represents the fitness of this set of cluster seeds is also determined. However, CSPM employs GAs to search for the optimal cluster seeds by encoding variables X_i as genes to represent related objects, where $X_i = 1$ denotes that the ith object is chosen as a cluster seed, and $X_i = 0$ otherwise.

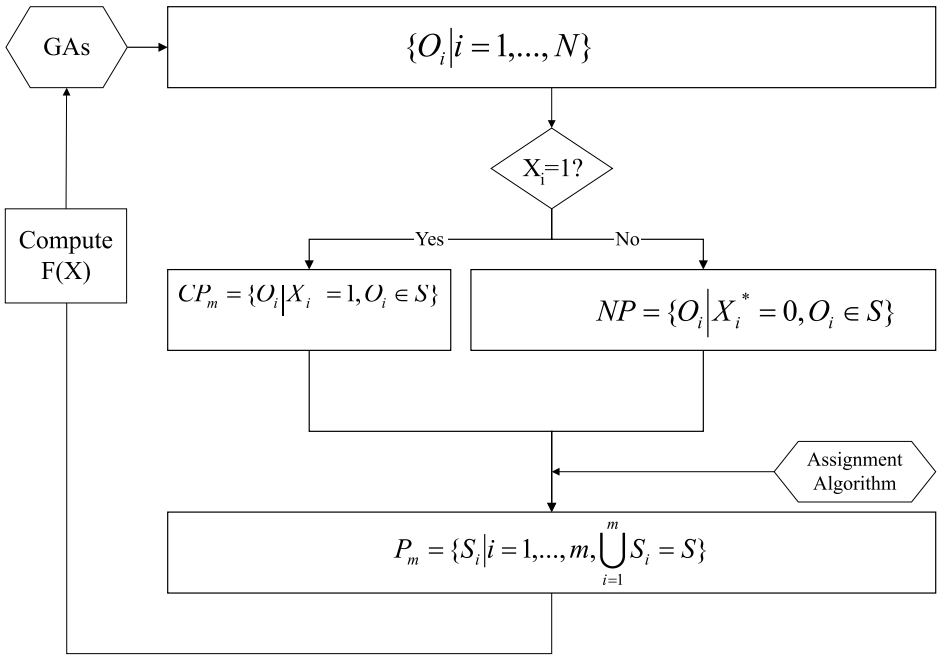

Fig. 2. Framework of CSPM.

5. Computational experiments

5.1. Experimental design

A random number generator is using to generate 200 two-dimensional objects (a_i and $b_i, i = 1,\ldots,200$) as shown in Fig. 3. All objects with coordinate of (a_i, b_i) are uniformly distributed within the square formed by four corners of which coordinates are $(0,0)$, $(0,1)$, $(1,0)$ and $(1,1)$, respectively. In order to analyze the effectiveness and efficiency of four models under varying problem sizes, we choose the first 50 objects merely in testing of small problems (10–50 objects) and use all the objects to be involved in testing of medium-to-large problems (50–200 objects).

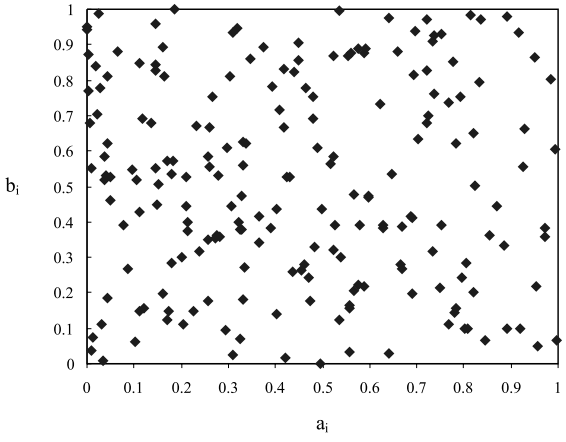

Fig. 3. Distribution of 200 two-dimensional objects.

Generally, minimizing the sum of squared errors is chosen as the objective function for clustering problems. However, it is only applicable to problems in which the number of clusters is specified. Employing this objective function to solve the optimal number of clusters will optimally result in N clusters such that the sum of squared errors is 0. Nevertheless, in this study, the objective function is to maximize the ratio of between-clusters variability to within-clusters variability, to determine simultaneously the optimal number of clusters and the optimal result of clustering [15].

$$F = \frac{N-m}{m-1} \times \frac{\sum_{k=1}^{m} n_k \left[\left(\bar{a}_k - \bar{\bar{a}}\right)^2 + \left(\bar{b}_k - \bar{\bar{b}}\right)^2 \right]}{\sum_{k=1}^{m} \sum_{i=1}^{n_k} \left[(a_i - \bar{a}_k)^2 + (b_i - \bar{b}_k)^2 \right]}, \quad (15)$$

where $\bar{a}_k = (1/n_k)\sum_{i=1}^{n_k} a_i$, $\bar{b}_k = (1/n_k)\sum_{i=1}^{n_k} b_i$, $\bar{\bar{a}} = (1/N)\sum_{i=1}^{N} a_i$, $\bar{\bar{b}} = (1/N)\sum_{i=1}^{N} b_i \cdot n_k$ is the number of objects in the kth cluster, that is, $n_k = |S_k|$. In order to avoid clusters with only one object, a penalty of $J \times M$ (J is the number of one-object clusters; M is a big number) is added to the objective value.

The mechanism of genetic algorithms in these three models is tested according to the following: population of each generation = 100, roulette wheel selecting, two points crossover at a rate of 1.00 and gene mutation at a rate of 0.01. The stopping rule is preset as mature rate reaching 80%. That is, GAs will continue to evolve until there are over 80% chromosomes with the same fitness in an epoch. Due to the stochastic characteristics of GAs, the following empirical comparison of our proposed methods are analyzed using hypothesis test on the results obtained from 30 different executions.

5.2. The results

5.2.1. Small scale problems ($N \leqq 50$)

Table 4 and Fig. 4 summarize the objective values solved by AHCM, SICM, STCM and CSPM under various problem sizes. The objective values solved by SICM are all significantly (at 5% level of significance) inferior to those of AHCM, showing the ineffectiveness of SICM. At $N = 10$, 20 and 30, the objective values solved by STCM are not significantly different from those of AHCM, but at $N = 40$ and 50 STCM shows more effective than AHCM. The objective values solved by CSPM are not significantly different from those of AHCM at $N = 10$ and 20, but CSPM demonstrates more effective than AHCM at $N = 30$, 40 and 50. For further comparison between STCM and CSPM, we test the null hypothesis $H_0 : \bar{F}_2 = \bar{F}_3$ against alternative hypothesis $H_1 : \bar{F}_2 < \bar{F}_3$ at $N = 10$, 20, 30, 40 and 50, respectively. Their corresponding Z values are 0.00, 0.69, 13.45, 16.70 and 11.17, indicating that CSPM is significantly more effective than STCM at $N = 30$, 40 and 50. The testing results show that CSPM is the most effective method, followed by STCM, and SICM is the least effective due to its long chromosome (five times that of CSPM) which leads to the need to search a larger feasible space.

Table 4
Effectiveness of four methods ($N \leqq 50$)[a]

Number of objects	AHCM	SICM		STCM		CSPM	
	F	$\bar{F}_1$ ($\bar{F}_1/F$)	δF_1 (Z_1)	$\bar{F}_2$ ($\bar{F}_2/F$)	δF_2 (Z_2)	$\bar{F}_3$ ($\bar{F}_3/F$)	δF_3 (Z_3)
10	13.06	12.45	0.73	13.06	0.00	13.06	0.00
		(0.95)	(−4.63*)	(1.00)	(0.00)	(1.00)	(0.00)
20	26.95	11.38	4.49	26.77	0.69	26.87	0.36
		(0.42)	(−18.97*)	(0.99)	(−1.43)	(1.00)	(−1.28)
30	28.70	13.83	3.83	28.82	2.81	36.40	1.28
		(0.48)	(−21.25*)	(1.00)	(0.22)	(1.27)	(32.90*)
40	37.10	11.55	4.44	38.72	2.88	48.99	1.75
		(0.31)	(−31.55*)	(1.04)	(3.08*)	(1.32)	(37.20*)
50	38.24	14.81	4.50	51.18	4.47	61.01	1.81
		(0.39)	(−28.53*)	(1.34)	(15.86*)	(1.60)	(68.98*)

[a] (1) F stands for the objective value of AHCM. (2) $\bar{F}_i$ and δF_i represent the means and standard deviations of objective values solved by SICM, STCM and CSPM with 30 different executions, $i = 1, 2, 3$. (3) Effectiveness index of the ith method is defined as $\bar{F}_i/F$. (4) $Z_i = (\bar{F}_i - F)/(\delta F_i/\sqrt{30})$ which follows Normal distribution (0,1). (5)* denotes that the null hypothesis test ($H_0 : \bar{F}_i = F$) is rejected at the 5% level of significance.

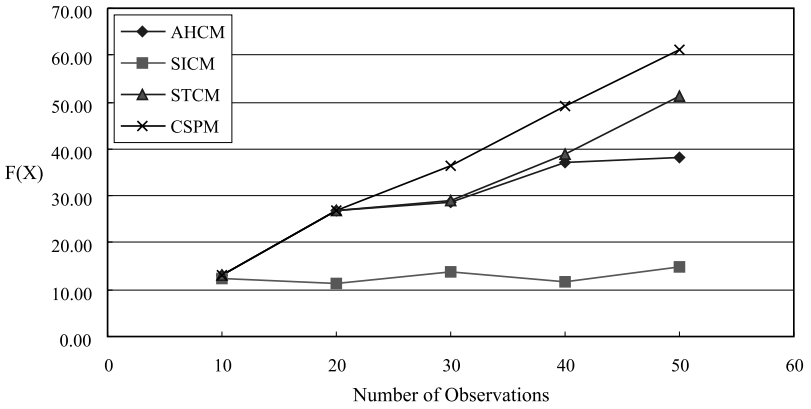

Fig. 4. Effectiveness of four methods ($N \leqq 50$).

Table 5
Efficiencies of four methods ($N \leqq 50$)[a]

Number of objects	AHCM	SICM		STCM		CSPM	
	SS	$\overline{SS}_1$ ($SS/\overline{SS}_1$)	δSS_1 (Z_1)	$\overline{SS}_2$ ($SS/\overline{SS}_2$)	δSS_2 (Z_2)	$\overline{SS}_3$ ($SS/\overline{SS}_3$)	δSS_3 (Z_3)
10	172	1987	1193	2917	418	11,372	1724
		(0.09)	(9.06)	(0.06)	(36.08)	(0.02)	(36.08)
20	1347	7383	2506	7533	1020	133,743	42,927
		(0.18)	(16.08)	(0.18)	(17.06)	(0.01)	(17.06)
30	4521	12,963	3538	14,323	2018	511,915	220,156
		(0.35)	(20.02)	(0.32)	(12.74)	(0.01)	(12.74)
40	10,696	33,020	10,250	20,970	1776	1,257,755	454,935
		(0.32)	(17.62)	(0.51)	(15.14)	(0.01)	(15.14)
50	20,871	68,033	17,141	28,080	2143	2,226,590	633,469
		(0.31)	(21.73)	(0.74)	(19.25)	(0.01)	(19.25)

[a] (1) SS stands for the number of solutions searched by AHCM. (2) $\overline{SS}_i$ and δSS_i represent the means and standard deviations of the number of solutions searched by SICM, STCM and CSPM with 30 different executions, $i = 1, 2, 3$. (3) Efficiency index of the ith method is defined as $SS/\overline{SS}_i$. (4) $Z_i = (\overline{SS}_i - SS)/(\delta SS_i/\sqrt{30})$ which follows Normal distribution (0,1). (5) * denotes that the null hypothesis test (H_0: $\overline{SS}_i = SS$) is rejected at the 5% level of significance.

Table 5 summarizes the number of solutions searched until the "optimal solution" being obtained by four methods under various problem sizes. Obviously, AHCM, which has the least number of solutions searched, is the most efficient method. Further comparison between SICM and STCM is made by a two-tailed test of $H_0: \overline{SS}_1 = \overline{SS}_2$ against $H_1: \overline{SS}_1 \neq \overline{SS}_2$. The Z values at $N = 10$, 20, 30, 40 and 50 are 4.03, 0.30, 1.83, −6.34 and −12.67, respectively, implicitly showing that SICM is more efficient than STCM at $N = 10$, not significantly different from STCM at $N = 20$ and 30, and less efficient than STCM at $N = 40$ and 50. Similarly, the results of two-tailed tests of $H_0: \overline{SS}_1 = \overline{SS}_3$ against $H_1: \overline{SS}_1 \neq \overline{SS}_3$ and $H_0: \overline{SS}_2 = \overline{SS}_3$ against $H_1: \overline{SS}_2 \neq \overline{SS}_3$ have shown that CSPM is the least efficient method.

Fig. 5 shows an optimal clustering result of CSPM at $N = 50$. The figure shows that, by maximizing the ratio of between-clusters variability to within-cluster variability, these 50 objects have been divided into 11 clusters. Each cluster contains 2–8 objects. Since the objects in the same clusters are adjacent to each other and no object is obviously grouped into a wrong cluster, the result of clustering appears to be good.

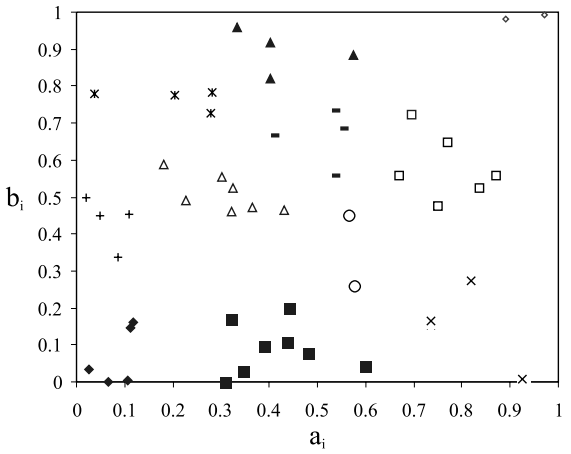

Fig. 5. Optimal clustering result of CSPM ($N = 50$).

5.2.2. Medium-to-large scale problems ($50 \leqq N \leqq 200$)

If SICM were applied to larger scale problems, the length of the chromosome would be too long, saturating the computer memory. Thus we only analyze the clustering results of AHCM, STCM and CSPM, illustrated in Fig. 6. It shows obviously that CSPM is the most effective, especially for the larger scale problems. AHCM is slightly more effective than STCM. It shows that STCM is more effective than AHCM for small scale problems ($N \leqq 50$), but less effective than AHCM for medium-to-large scale problems ($50 \leqq N \leqq 200$).

6. Applications

CSPM can be applied to problems of p-Median because of not only its effectiveness, but also its search procedure (cluster seed points are chosen first, and the rest of the objects are assigned later). A *p*-Median problem is chosen as an example to examine the applicability of this method.

6.1. Problem statement

There are 25 districts uniformly distributed in a square area. Each link that connects two adjacent districts is 1 km long. We consider setting up several public facilities, such as hospitals, schools or fire departments in some districts to serve all districts. More facilities represent a higher total set-up cost. However, if facilities are insufficient or set up in the wrong place, the total cost of inconvenience across all districts using the facilities will increase. Thus the problem of optimal locations

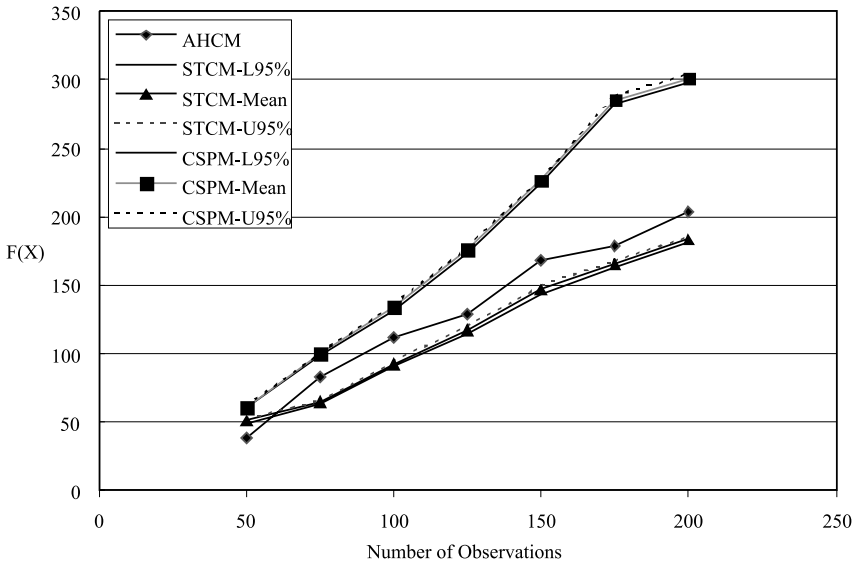

Fig. 6. Means and 95% upper/lower confidence intervals of STCM and CSPM $50 \leqq N \leqq 200$.

and serviceable areas of facilities can be formulated as follows:

[SC]

$$\text{Min } Z = \alpha \sum_j X_{jj} + \beta \sum_i \sum_j (|a_i - a_j| + |b_i - b_j|) X_{ij} \tag{16}$$

subject to

$$\sum_j X_{ij} = 1 \quad \text{all } i, \tag{17}$$

$$X_{ij} \leqslant X_{jj} \quad \text{all } i,j, \quad i = 1,\ldots,25, \tag{18}$$

$$X_{ij} = \{0,1\} \quad \text{all } i,j, \quad i = 1,\ldots,25, \; j = 1,\ldots,25, \tag{19}$$

where Z represents the total cost function, α is the set-up cost of a public facilities (dollars/facility), β is the unit distance cost for a district to access its public facility (dollars/kilometer), (a_i, b_i) is the coordinate of ith district, $X_{jj} = 1$ denotes that jth district has a public facility, $X_{jj} = 0$ otherwise. $X_{ij} = 1$ denotes that ith district uses the public facility located at jth district, $X_{ij} = 0$ otherwise. The mechanism of GAs follows the set-ups in Section 5.1.

6.2. The results

Without a loss of generality, let $\beta = 1$ dollars/km. Figs. 7–14 show the optimal designs of CSPM according to different set-up costs (α), ranging from 0.5 to 15 dollars/facility, respectively. In these figures, a circle denotes a district, a shadowed circle indicates that a public facility has been

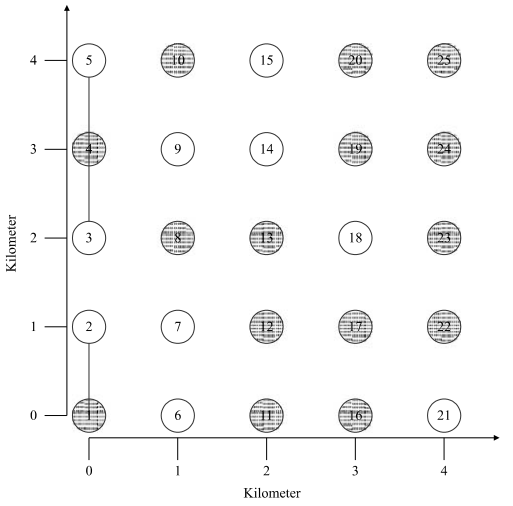

Fig. 8. The optimal design at $\alpha = 1$.

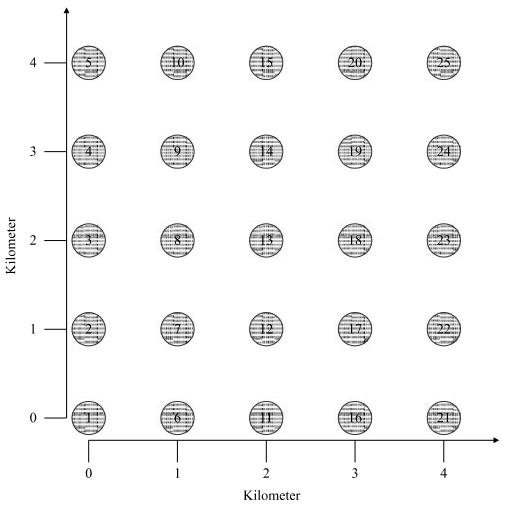

Fig. 7. The optimal design at $\alpha = 0.5$.

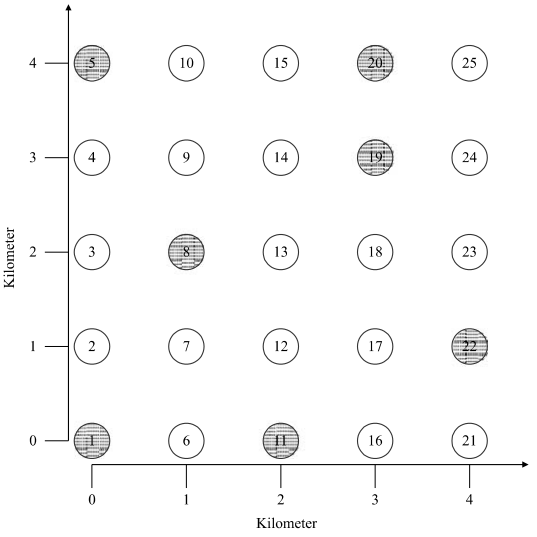

Fig. 9. The optimal design at $\alpha = 2$.

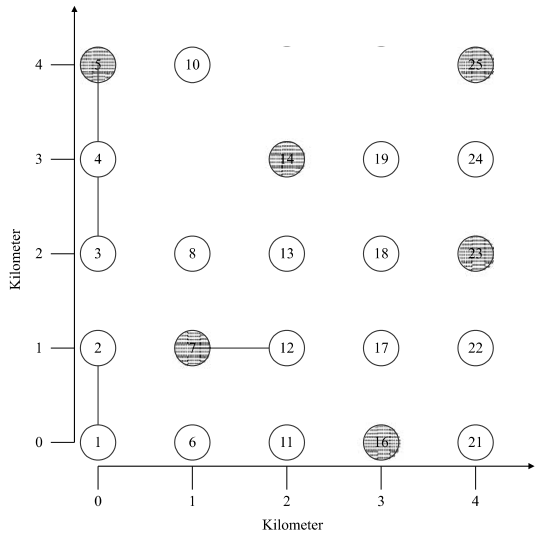

Fig. 10. The optimal design at $\alpha = 3$.

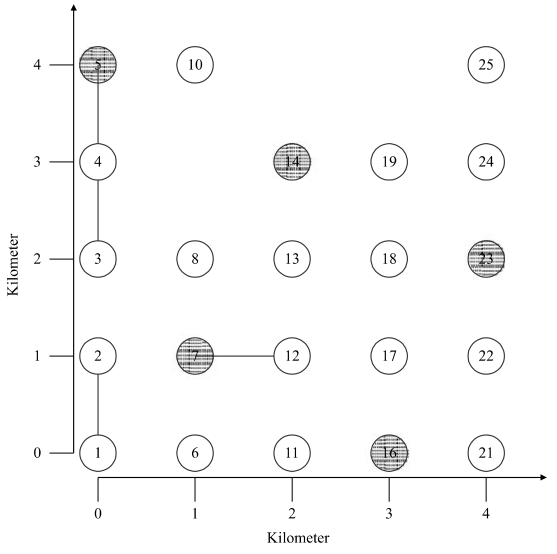

Fig. 11. The optimal design at $\alpha = 4$.

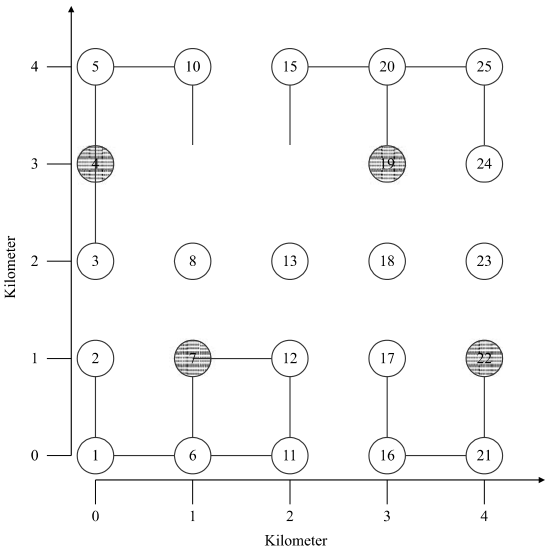

Fig. 12. The optimal design at $\alpha = 5$.

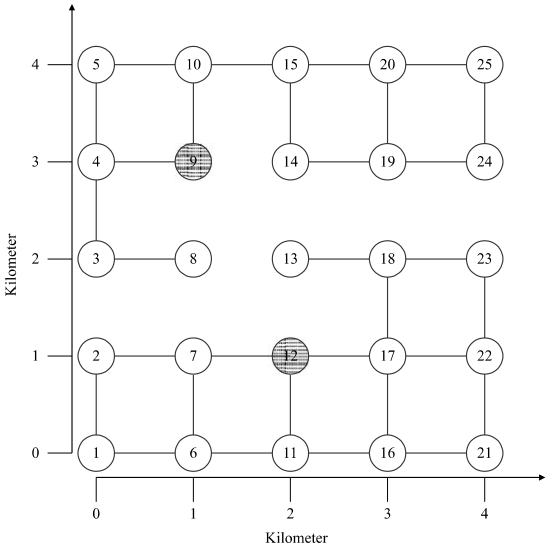

Fig. 13. The optimal design at $\alpha = 10$.

set up in that district, and the districts connected with links are all served by the same public facility. At $\alpha = 0.5$ dollars/facility, the optimal number of facilities is 25, that is, each district has one public facility and the total distance cost is 0 (Fig. 7). At $\alpha = 3$ dollars/facility, the optimal number of facilities is 6 and total distance cost is 22 dollars (Fig. 10). At $\alpha = 15$ dollars/facility, the optimal number of facilities is only one and the total distance cost reaches 60 dollars (Fig. 14). Fig. 15 further illustrates that while the set-up cost of a facility increases, the optimal number of facilities decreases and the total distance cost increases.

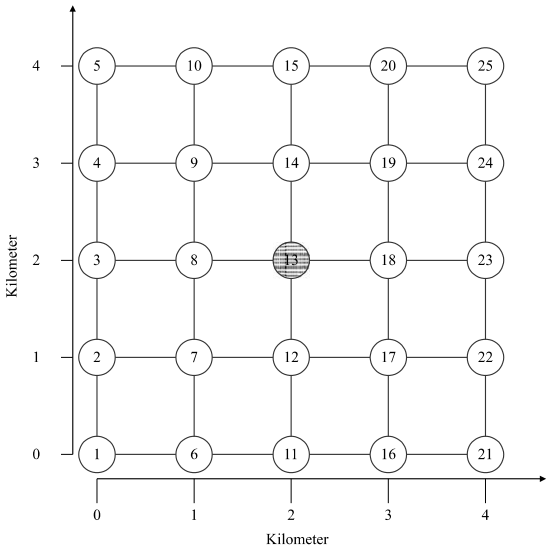

Fig. 14. The optimal design at $\alpha = 15$.

6.3. Concluding remarks

This study discusses the effectiveness and efficiency of solving clustering problems by employing GAs. Varying the techniques of coding/decoding, we proposed SICM, STCM and CSPM models and tested them with 200 two-dimensional objects. The results from small scale problems (10–50 objects) show that CSPM is most effective but least efficient, STCM is second most effective and efficient, and SICM is least effective because of its long chromosome required. The results from medium-to-large problems (50–200 objects) indicate that CSPM is still the most effective method. AHCM is slightly more effective than STCM. Both results show that CSPM can solve clustering problems effectively. CSPM can be easily applied to problems of p-Median or location selection because of its effectiveness and its search procedure (cluster seed points is chosen first, and the rest of the objects are assigned later). CSPM is highly applicable as evidenced by the reasonable results of the application to an exemplified p-Median problem.

Since the search space for CSPM is proportional to 2^N (search space for $N = 10$ is $2^{10} = 1024$, $N = 100$ is $2^{100} = 1.27 \times 10^{30}$), the larger the scale of the problem is, the less effectiveness of CSRM would be anticipated. Future studies can examine the feasibility of employing a hybrid GA model (combining other heuristic algorithms, such as simulated annealing) to further enhance its effectiveness and efficiency. The clustering problems solved by GAs using personal computers, however, involves the storage of a certain population of chromosomes, which is likely to cause an insufficiency of computer memory as the problem scale gets larger (say $N > 200$). Therefore, reducing the storage requirements is worthy of further investigation. Future studies can also be conducted by comparing the effectiveness and efficiency of our proposed methods with the GA based algorithms mentioned in the references.

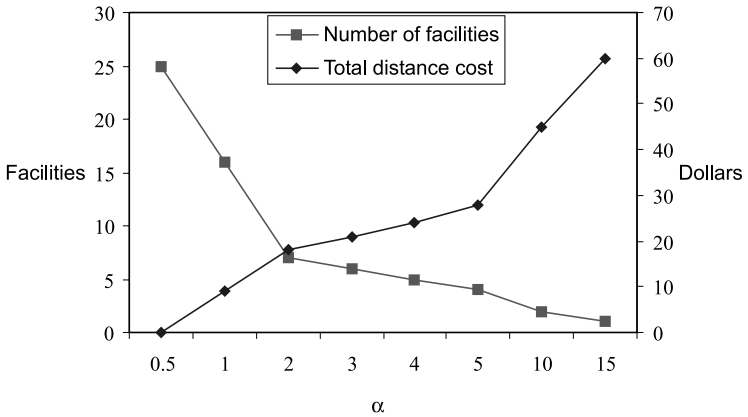

Fig. 15. Optimal number of facilities and total distance costs vs. facility set-up cost.

Acknowledgements

The authors are greatly indebted to two referees for their constructive comments. This paper is partially sponsored by the National Science Council of the Republic of China under contract number NSC88-2211-E-009-019.

References

[1] M. Abramowitz, I.A. Stegun (Eds.), Handbook of Mathematical Functions, National Bureau of Standards, Applied Mathematical Series 55, U.S. Department of Commerce, 1964.
[2] A.I. Ali, H. Thiagarajan, A network relaxation based enumeration algorithm for set partitioning, European Journal of Operational Research 38 (1989) 76–85.
[3] M.R. Anderberg, Cluster Analysis for Applications, Academic Press, New York, 1973.
[4] C. Alippi, R. Cucchiara, Cluster partitioning in image analysis classification: A genetic algorithm approach, in: Proceedings of the 1992 IEEE International Conference on Computer Systems and Software Engineering, 1992, pp. 423–427.
[5] H. Aytug, G. Koehler, J.L. Snowdon, Genetic learning of dynamic scheduling within a simulation environment, Computers and Operations Research 21 (1994) 909–925.
[6] J.N. Bhuyan, V.V. Raghavan, V.K. Elayavalli, Genetic algorithms with an ordered representation, in: Proceedings of the Fourth International Conference on Genetic Algorithms, 1991, pp. 408–415.
[7] P. Brucker, On the complexity of clustering problems, in: M. Beckmenn, H.P. Kunzi (Eds.), Optimization and Operations Research, Lecture Notes in Economics and Mathematical Systems, vol. 157, Springer, Berlin, 1978, pp. 45–54.
[8] D.G. Cattrysse, M. Salomon, L.N. Van Wassenhove, A set partitioning heuristic for the generalized assignment problem, European Journal of Operational Research 72 (1994) 167–174.
[9] D.G. Conway, M.A. Venkataramanan, Genetic search and dynamic facility layout problem, Computers and Operations Research 21 (1994) 955–960.
[10] R. Cucchiara, Analysis and comparison of different genetic models for the clustering problem in image analysis, in: Proceedings of the International Conference on Artificial Neural Nets and Genetic Algorithms, Innsbruck, Austria, 1993, pp. 423–427.
[11] J. Etcheberry, The set covering problem: A new implicit enumeration algorithm, Operations Research 25 (1977) 760–772.
[12] E. El-Darzi, G. Mitra, Graph theoretic relaxations of set covering and set partitioning problems, European Journal of Operational Research 87 (1995) 109–121.
[13] D. Goldberg, Genetic Algorithms in Search Optimization and Machine Learning, Addison-Wesley, Reading, MA, 1989.
[14] P. Hansen, M. Delattre, Complete-link cluster analysis by graph coloring, Journal of the American Statistical Association 73 (1978) 397–403.
[15] P. Hansen, B. Jaumard, Minimum sum of diameters clustering, Journal of Classification 4 (1987) 215–226.
[16] J.H. Holland, Adaptation in Nature and Artificial Systems, University of Michigan Press, Ann Arbor, MI, 1975.
[17] R.E. Jensen, A dynamic programming algorithm for cluster analysis, Operations Research 12 (1969) 1034–1057.
[18] R.A. Johnson, D.W. Wichern, Applied Multivariate Statistical Analysis, Second ed., Prentice-Hall, New York, 1988.
[19] D.R. Jones, M.A. Beltramo, Solving partitioning problems with genetic algorithms, in: Proceedings of the Fourth International Conference on Genetic Algorithms, 1991, pp. 442–449.
[20] F. Kettaf, J. Asselin de Beauville, Genetic and fuzzy based clustering, in: Proceedings of the Fifth Conference on International Federation of Classification Societies, 1996, pp. 100–103.
[21] F.T. Lin, C.Y. Kao, C.C. Hsu, Applying the genetic approach to simulated annealing in solving some NP-hard problems, IEEE Transactions on Systems Man and Cybernetics 23 (1993) 1752–1767.
[22] J.A. Lozano, P. Larranga, M. Grana, Partitional cluster analysis with genetic algorithm: Searching for the number of clusters, Data Science, Classification and Related Methods (1998) 117–125.
[23] C.B. Lucasius, A.D. Dane, G. Kateman, On *k*-medoid clustering of large data sets with the aid of a genetic algorithm: Background, feasibility, and comparison, Analytica Chimica Acta 282 (1993) 647–669.
[24] S. Lunchian, H. Lunchian, M. Petriuc, Evolutionary automated classification, in: Proceedings of the First IEEE Conference on Evolutionary Computation, 1994, pp. 585–589.
[25] J. MacQueen, Some methods for classification and analysis of multivariate observations, in: Proceedings of Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, 1967, pp. 281–297.
[26] I.R. Moraczawski, W. Borkowski, A. Kierzek, Clustering geobotanical data with the use of a genetic algorithm, Coenoses 10 (1995) 17–28.
[27] E.C. Moshe, C.A. Tovey, J.C. Ammons, Circuit partitioning via set partitioning and column generation, Operations Research 44 (1996) 65–76.
[28] J.M. Mulvey, H.P. Crowder, Cluster analysis: An application of Lagrangian relaxation, Management Science 25 (1979) 329–340.
[29] A.L. Nordstrom, S. Tufekci, A genetic algorithm for the talent scheduling problem, Computers and Operations Research 21 (1994) 941–954.
[30] J. Pinte'r, G. Pesti, Set partition by globally optimized cluster seed points, European Journal of Operational Research 51 (1991) 127–135.

[31] M. Savelsbergh, A branch-and-price algorithm for the generalized assignment problem, Operations Research 45 (1997) 831–841.
[32] M.A. Trick, A linear relaxation heuristic for the generalized assignment problem, Naval Research Logistic 39 (1992) 137–152.
[33] J.H. Ward, Hierarchical grouping to optimize an objective function, Journal of the American Statistical Association 58 (1963) 236–244.
[34] J.W. Welch, Algorithmic complexity: Three NP-hard problems in computational statistics, Journal of Statistical Computation and Simulation 15 (1983) 17–25.