



ELSEVIER

Expert Systems with Applications 27 (2004) 265–276

Expert Systems  
with Applications

[www.elsevier.com/locate/eswa](http://www.elsevier.com/locate/eswa)

# A purchase-based market segmentation methodology

C.-Y. Tsai\*, C.-C. Chiu

*Industrial Engineering and Management Department, Yuan-Ze University, 135 Yuan-Tung Rd., ChungLi, 320, Taoyuan, Taiwan, ROC*

## Abstract

Market segmentation is critical for a good marketing and customer relationship management program. Traditionally, a marketer segments a market using general variables such as customer demographics and lifestyle. However, several problems have been identified and make the segmentation result unreliable. This paper develops a novel market segmentation methodology based on product specific variables such as purchased items and the associative monetary expenses from the transactional history of customers to resolve these problems. A purchase-based similarity measure, clustering algorithm, and clustering quality function are defined in this paper. A genetic algorithm approach is adopted to ensure that customers in the same cluster have the closest purchase patterns. After completing segmentation, a designated RFM model is used to analyze the relative profitability of each customer cluster. The findings from a practical marketing implementation study will also be discussed.

© 2004 Elsevier Ltd. All rights reserved.

**Keywords:** Market segmentation; Purchase behavior; Clustering; Genetic algorithm; RFM analysis

## 1. Introduction

The mass marketing approach cannot satisfy the needs of diverse customers today. This diversity should be satisfied using segmentation that divides markets into customer clusters with similar needs and/or characteristics that are likely to exhibit similar purchasing behaviors (Dibb & Simkin, 1996). Segmentation theory proposes that groups of customers with similar needs and purchasing behaviors are likely to demonstrate a more homogeneous response to marketing programs that target specific consumer groups. With proper market segmentation, enterprises can arrange the right products, services and resources to a target customer cluster and build a close relationship with them. Market segmentation has consequently been regarded as one of the most critical elements in achieving successful modern marketing and customer relationship management (CRM) (Berson, Smith, & Thearling, 2000).

A critical issue to successful market segmentation is the selection of the segmentation variables. Segmentation variables can be broadly classified into general variables and product specific variables (Wedel & Kamakura, 1997). The general variables include the customer demographics and lifestyles. The product specific variables involve

customer purchasing behaviors and intentions. Many researches have devoted themselves to using general variables to partition customers because the variables are intuitive and easy to operate (Beane & Ennis, 1987; Hammond et al., 1996). In the study by Natter (1999), an artificial neural network clustering method that incorporated both clusters and segment discriminant analysis was proposed. The relationship between the consumer demographics was estimated using this method. Chou et al. (2000) used customer demographics to identify prospective customers without conducting designed marketing campaigns. This provided an intuitive measure to guide in the selection of marketing targets. Kuo et al. (2002) introduced a two-stage method that combined self-organizing feature maps with the K-means algorithm. The self-organizing feature maps first determined the number of clusters and the starting point. The K-means method was then employed to find the final solution.

Market segmentation based on general variables is more intuitive and easier to conduct than product specific variables. However, the assumption that customers with similar demographics and lifestyles will exhibit similar purchasing behavior is doubtful. Today, customers can easily locate abundant product information from various marketing channels. To present uniqueness and identity, each customer pursues personalized products and services

\* Tel.: +886-34638800512; fax: +886-34638907.

E-mail address: [cytsai@saturn.yzu.edu.tw](mailto:cytsai@saturn.yzu.edu.tw) (C.-Y. Tsai).

even within groups with similar demographics and lifestyles. This makes their purchase patterns difficult to determine using only general variables. Another problem is that most general variables are viewed as private property by many individuals. Data collection for those variables could be difficult and time-consuming. Although the data can be obtained, this data varies as time goes by. For example, occupation, income, and marriage status data collected now might not be valid two years later if no continuous revision is performed. These problems make market segmentation using general variables questionable (Drozdenko & Drake, 2002).

This study developed a novel market segmentation methodology based on product specific variables such as items purchased and the associative monetary transactional history of customers. This identifies groups of customers with similar purchasing behaviors with a more homogeneous response to marketing programs. A genetic algorithm (GA) approach is developed in this methodology that increases the clustering quality. This ensures that customers in the same cluster have the closet similar purchase patterns. The remainder of this paper is organized as follows. Section 2 examines the framework of the proposed purchase-based segmentation methodology including data preparation, a similarity measure, a clustering algorithm and a clustering quality function. In Section 3, an initial cluster centers generation process using a heuristic genetic algorithm is developed to improve the clustering quality. A designated RFM model is then proposed to analyze the relative profitability of each customer cluster in Section 4. Section 5 provides a practical marketing implementation method to demonstrate the benefit of the proposed methodology. A summary and conclusion are presented in Section 6.

## 2. A purchase-based market segmentation methodology

This section introduces a novel market segmentation methodology based on the purchase behaviors of customers. The core of the methodology includes data preparation, a purchase-based similarity measure, clustering algorithm and a clustering quality function.

### 2.1. Data preparation

The purpose of data preparation is to integrate, select and transform the data from one or more databases into the data required for the proposed methodology. Let  $I$  be the set of all items provided by an enterprise, and  $T^0$  be the transaction database. In  $T^0$  a transaction  $t^0$  consists of at least one row of data that records a customer ID, transaction time, an item, quantity, monetary expense and so on. To observe customer purchasing behavior, we need to retrieve his/her purchased items and the aggregated monetary expenses for these items over a period. Let  $id_i$

be the customer ID of a customer  $c_i$ ,  $itemset_i = \{i_{ia} | i_{ia} \in I\}$  be the set of items purchased by  $c_i$ . Let  $moneyset_i = \{m_{ia} | a = 1, \dots, \|itemset_i\|\}$  be the set of aggregated monetary expenses for purchased items where  $m_{ia}$  is the aggregated monetary expense for item  $i_{ia}$  and  $\|A\|$  is the number of members of a set  $A$ . Therefore, an aggregated record that describes the purchase behavior of a customer  $c_i$  can be represented as  $t_i^c \equiv (id_i, itemset_i, moneyset_i)$  and stored in the cumulative transaction database  $T^c$ . A data preparation example is shown in Fig. 1.

### 2.2. The purchase-based similarity measure

A simple matching coefficient and Jaccard's coefficient are two popular similarity measures (Manning & Schutze, 1999; Romesburg, 1984) that can be used to evaluate the similarity between two customers  $c_i$  and  $c_j$  based on their respective  $itemset_i$  and  $itemset_j$ . The two measures accumulate the number of matching and non-matching items between two customers and calculate their similarity. Although the measures are easy to use with low

The transaction database  $T^0$

Customer ID	transaction time	product item	quantity	expense
101	371	milk	2	150
101	371	bread	6	100
101	371	cookie	7	300
102	371	shoes	1	950
102	371	hat	1	1500
103	373	soap	1	240
103	373	bag	2	160
103	373	soda	8	100
101	376	cookie	4	240
101	376	beer	15	480
103	377	table	1	600
103	377	soda	20	250

Retain attributes of Customer ID, product item, and expense

Customer ID	product item	expense
101	milk	150
101	bread	100
101	cookie	300
102	shoes	950
102	hat	1500
103	soap	240
103	bag	160
103	soda	100
101	cookie	240
101	beer	480
103	table	600
103	soda	250

Transform to desired data format

The cumulative transaction database  $T^c$

Customer ID	product item / aggregated expense
101	milk / 150 , bread / 100 , cookie / 540 , beer / 480
102	shoes / 950 , hat / 1500
103	soap / 240 , bag / 160 , soda / 350 , table / 600

Fig. 1. A data preparation example.

computational cost, they are not appropriate to our case. First, most of the time a customer purchases only a small subset of thousands upon thousands of items provided by an enterprise (Agrawal, Imielinski, & Swami, 1993). This makes the similarity between any two customers very small so that the discriminative ability is not significant. Second, the importance between any two items should not be equivalent. A co-purchase association for any two items should be included in the similarity measure when evaluating the importance between any two items. For example, if item A is often co-purchased with item B while less with item C, the co-purchase association between A and B should be stronger than that between A and C. Ignoring these associations and treating all items equally creates a similarity bias. Third, the importance of each customer should also be different based on his/her contribution to the profit for an enterprise. It is necessary to include the profitability of each customer when evaluating customer purchase behavior similarity. However, the previous measures do not consider the importance.

A purchase-based similarity measure is developed in this section to fulfill these conditions. The similarity measure considers the co-purchase association between two items and the profitability of each customer. An intimacy measure is defined to include the co-purchase association. The intimacy measure is inspired by the *support* concept in the association rule (Agrawal & Srikant, 1994) to improve its discriminability. Let  $\text{Supp}(\{i_i, i_j\})$  be the proportion of transactions containing the itemset  $\{i_i, i_j\}$  to all transactions in  $T^0$  and represented as:

$$\text{Supp}(\{i_i, i_j\}) = \frac{\|\{t^0 \in T^0 | t^0 \text{ contains } \{i_i, i_j\}\}\|}{\|T^0\|} \quad (1)$$

where  $i_i, i_j \in I$ . However, the support value could be very low if an itemset contains rarely co-purchased item(s) (Mannila, 1998). To reduce the impact of the rare item problem, the intimacy measure of an itemset  $\{i_i, i_j\}$  is defined as:

$$\text{Int}(\{i_i, i_j\}) = \frac{\text{Supp}(\{i_i, i_j\})}{\text{Supp}(i_i) + \text{Supp}(i_j) - \text{Supp}(\{i_i, i_j\})} \quad (2)$$

$\text{Int}(\{i_i, i_j\})$  is ranged from 0 to 1. If  $i_i = i_j$ ,  $\text{Int}(\{i_i, i_j\}) = 1$ . After knowing the intimacy of any two items, the purchase-based similarity measure can be evaluated as follows. Let the aggregated record  $t_i^c$  for customer  $c_i$  be  $(\text{id}_i, \text{itemset}_i, \text{moneyset}_i)$  where  $\text{itemset}_i = \{i_{i1}, i_{i2}, \dots, i_{is}\}$  and  $\text{moneyset}_i = \{m_{i1}, m_{i2}, \dots, m_{is}\}$  and the aggregated record  $t_j^c$  for customer  $c_j$  be  $(\text{id}_j, \text{itemset}_j, \text{moneyset}_j)$  where  $\text{itemset}_j = \{i_{j1}, i_{j2}, \dots, i_{jt}\}$  and  $\text{moneyset}_j = \{m_{j1}, m_{j2}, \dots, m_{jt}\}$ . The purchase-based

similarity measure is defined as:

$$\text{Sim}(c_i, c_j) = \frac{\sum_{a=1}^s \sum_{b=1}^t [m_{ia} \times m_{jb} \times \text{Int}(\{i_{ia}, i_{jb}\})]}{\sum_{a=1}^s \sum_{b=1}^t [m_{ia}^i \times m_{jb}^j]} \quad (3)$$

### 2.3. The purchase-based segmentation algorithm

A purchase-based segmentation (PBS) algorithm is developed based on the similarity measure in Eq. (3) to perform market segmentation. In this algorithm, users need to specify the number of customer clusters,  $K$ . The  $K$  value can be subjectively determined according to the marketing program objective or objectively evaluated using our suggested mechanism. The suggestion mechanism will be introduced in Section 2.4. After the  $K$  value has been determined, the  $K$  initial cluster centers are selected from the aggregated  $T^c$  records. The selection procedure can be a random selection process or a heuristic selection process. The heuristic process genetic algorithm that improves the clustering quality will be introduced in Section 3.

Let  $G = \{c^n | n = 1, \dots, K\}$  be the set of  $K$  cluster centers and  $c^n$  be the cluster center of the  $n$ th cluster  $G^n$  where  $c^n \in T^c$ . Therefore,  $(T^c - G) = \{c_i | i = 1, \dots, \|T^c - G\|\}$  is the set of remaining customers that were not selected as cluster centers. That is,  $c_i \in T^c$  and  $c_i \notin G$ . After the similarities between all cluster centers  $c^n$  and a remaining customer  $c_i$  are evaluated using  $\text{Sim}(c^n, c_i)$  of Eq. (3), customer  $c_i$  will be assigned to the  $n$ th cluster  $G^n$  if the similarity between  $c_i$  and  $c^n$  is maximum. This can be expressed as:

$$\text{Max}_{1 \leq n \leq K} \{\text{Sim}(c_i, c^n)\} \text{ where } c_i \in (T^c - G) \quad (4)$$

After each remaining customer is assigned to a proper cluster, the next step is to recalculate the new cluster center for each cluster. Typically, a customer is assigned to a new cluster center if the sum of the similarities between him and the other customers in the same cluster is maximum and the sum of the similarities between him and the other cluster centers is minimum. To satisfy both requirements, a priority measure is developed to evaluate the chance for a customer being assigned to a new cluster center. Suppose that  $c_i$  and  $c_j$  are two customers in cluster  $G^n$ , and the cluster center in  $G^n$  is  $c^n$ . The priority of customer  $c_i$  can be defined as:

$$\text{Pio}(c_i) = \sum_{c_j \in G^n, j \neq i} \text{Sim}(c_i, c_j) / \sum_{c^m \in G, m \neq n} \text{Sim}(c_i, c^m) \quad (5)$$

where  $c^m$  is the center of cluster  $G^m$ ,  $\sum_{c_j \in G^n, j \neq i} \text{Sim}(c_i, c_j)$  represents the sum of the similarities between  $c_i$  and other customers in the same cluster  $G^n$ , and  $\sum_{c^m \in G, m \neq n} \text{Sim}(c_i, c^m)$  represents the sum of the similarities between  $c_i$  and other cluster centers except for  $G^n$ . For all customers in cluster

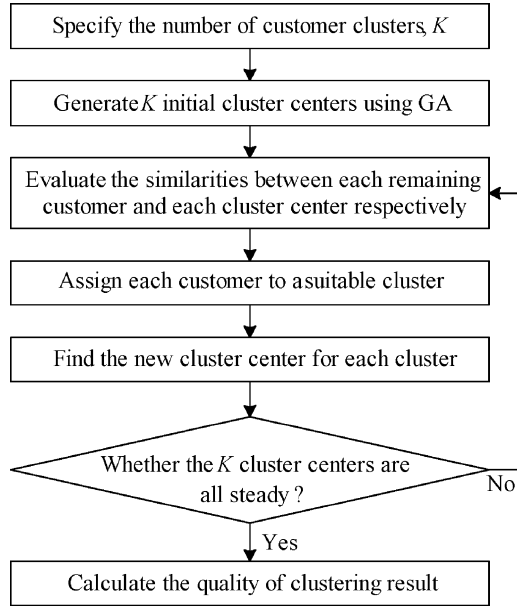


Fig. 2. The framework of the PBS algorithm.

$G^n$ , a customer with the largest priority measure is selected as the new cluster center. The process can be represented as:

$$c^n \equiv \arg \max_{c_i \in G^n} \{ \text{pio}(c_i) \} \quad (6)$$

After recalculating the new cluster center for each cluster, this algorithm accomplishes one iteration. The algorithm goes back to the beginning, sequentially executes Eqs. (3)–(6) until no cluster center been changed. Now, customers having similar purchase behaviors are clustered in the same cluster. The procedure for the proposed PBS algorithm is summarized in Fig. 2.

#### 2.4. The clustering quality function

The objective of the PBS algorithm is to maximize the sum of the similarities between a cluster center and all customers in the same cluster, and to minimize the sum of the similarities between two cluster centers in different clusters. Therefore, the quality of the clustering result with  $K$  clusters can be defined as Eq. (7):

$$\rho(K) = \frac{1}{K} \sum_{n=1}^K \left( \min_{1 \leq m \leq K, m \neq n} \left\{ \frac{\eta_n + \eta_m}{\delta_{nm}} \right\} \right) \quad (7)$$

$$\eta_n = \frac{1}{\|G^n\|} \sum_{c_i \in G^n} \text{Sim}(c_i, c^n) \quad (8)$$

$$\eta_m = \frac{1}{\|G^m\|} \sum_{c_j \in G^m} \text{Sim}(c_j, c^m) \quad (9)$$

$$\delta_{nm} = \text{Sim}(c^n, c^m) \quad (10)$$

Eq. (8) defines  $\eta_n$  as the average of similarities between cluster center  $c^n$  and all customers in cluster  $G^n$ . Eq. (9) states that  $\eta_m$  is the average of the similarities between

cluster center  $c^m$  and all customers in cluster  $G^m$ . Eq. (10) defines  $\delta_{nm}$  as the similarity between  $c^n$  and  $c^m$ .

With the clustering quality defined in Eq. (7), we can determine a suggested value for  $\hat{K}$  ranging between the lower boundary  $s$  and the higher boundary  $t$ . This process can be represented as:

$$\hat{K} \equiv \arg \max_{s \leq K \leq t} \{ \rho(K) \} \quad (11)$$

Using Eq. (11), an optimal value for  $K$  can be objectively determined for market segmentation.

### 3. Cluster center initialization using genetic algorithm

The initial cluster centers can be selected from the cumulative transaction database  $T^c$  through a random selection process. However, random selection often causes the clustering quality to fall into local optimization (Bradley & Fayyad, 1998). Meila and Heckerman (1998) suggested performing a large number of runs with random initial cluster centers and choosing the best one as the clustering result. Dimitriadou et al. (1999) used a voting approach in each run to combine the present clustering result with the prior clustering result to produce a better result. Although these researches did enhance the clustering quality, they are not appropriate to our case because it is time-consuming to run several clustering processes in a large dataset.

To avoid this problem, a heuristic selection process using a genetic algorithm (GA) (Holland, 1975) is developed in this section to generate better initial cluster centers resulting in a more stable clustering quality. The GA is a computational abstraction of biological evolution used to solve optimization problems through a series of genetic operations such as reproduction, crossover and mutation on a population of chromosomes (Goldberg, 1989).

#### 3.1. Chromosome encoding

Typically, a chromosome can be used to represent a candidate solution to a problem where each gene in the chromosome represents a parameter of the candidate solution. In this study, a chromosome is regarded as a set of  $K$  initial cluster centers and each gene is a cluster center. Specifically, a chromosome  $f_i$  can be represented as  $f_i = [y_1, \dots, y_j, \dots, y_K]$  where  $y_j$  is the  $j$ th gene and  $K$  is total number of genes. A real-value encoding scheme is suitable to represent a gene because each customer has a unique customer ID. Fig. 3 illustrates a chromosome encoding example.

#### 3.2. Population initialization

Let  $P^e$  be the  $e$ th-generation genetic operation population where  $0 \leq e \leq E$  and  $E$  is the maximal number of generations to terminate GA. The number of chromosomes

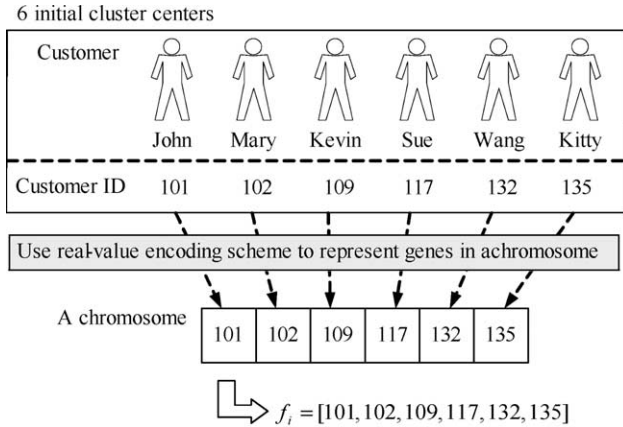


Fig. 3. A chromosome encoding example.

is fixed for all generations. Therefore, the  $e$ th-generation population can be represented as  $P^e = \{f_i | i = 1, \dots, L\}$  where  $f_i$  is the  $i$ th chromosome and  $L$  is the number of chromosomes in a population. Note that  $L$  is an even integer number specified by users for executing genetic operations.

### 3.3. Fitness value of a chromosome

The fitness value of a chromosome evaluates whether a chromosome is suitable for survival or not. The clustering

quality in Eq. (7) is used as the fitness function in this practice. Therefore, the formula for calculating the fitness value of a chromosome  $f_i$  is given as:

$$\text{fitness}(f_i) = \rho(K) \text{ where } f_i = [y_1, \dots, y_j, \dots, y_K] \quad (12)$$

According to Eq. (12), all chromosomes in  $P^e$  are equally divided into  $P_{\text{good}}^e$  and  $P_{\text{bad}}^e$  based on individual fitness values.  $P_{\text{good}}^e$  is the class of chromosomes with higher fitness values, and  $P_{\text{bad}}^e$  is the class of chromosomes with lower fitness values. That is,  $\|P_{\text{good}}^e\| = \|P_{\text{bad}}^e\| = I/2$ .

### 3.4. Reproduction

The purpose of reproduction is to eliminate chromosomes with lower fitness from the population and duplicate chromosomes with higher fitness in the population. This operation selects offspring chromosomes from better parent chromosomes. Therefore, all chromosomes in  $P_{\text{bad}}^e$  are eliminated and all chromosomes in  $P_{\text{good}}^e$  are retained. The chromosomes in  $P_{\text{good}}^e$  are then selected and duplicated to fill the spaces left by the chromosomes in  $P_{\text{bad}}^e$ . In  $P_{\text{good}}^e$  the probability that chromosome selection will depend on its fitness value. The higher the fitness value a chromosome has, the higher the probability that chromosome has for selection. The formula for calculating

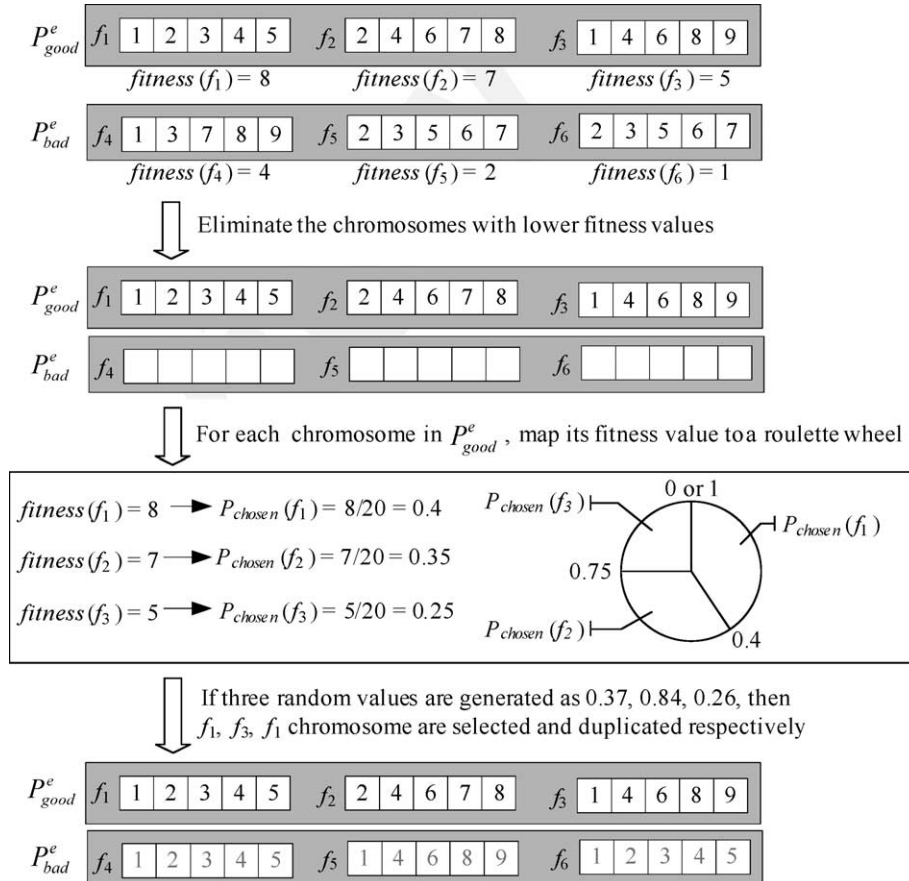


Fig. 4. A GA reproduction operation.



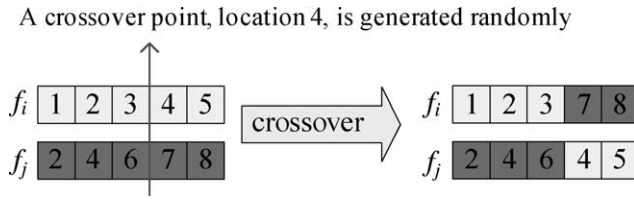


Fig. 5. A GA crossover operation.

the selection probability is given below:

$$\text{Prob}(f_i) = \text{fitness}(f_i) / \sum_{f_i \in P_{\text{good}}^e} \text{fitness}(f_i) \quad (13)$$

The reproduction operation ensures that all chromosomes in an offspring population are generated from excellent parent population chromosomes. An example of reproduction is illustrated in Fig. 4.

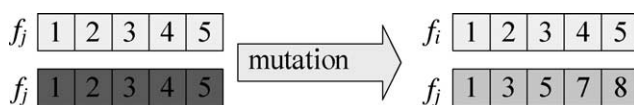
### 3.5. Crossover and mutation

After reproduction, crossover and mutation operations are initiated to generate unexpected offspring chromosomes. A chromosome  $f_i$  from  $P_{\text{good}}^e$  and a chromosome  $f_j$  from  $P_{\text{bad}}^e$  are chosen to form an operation unit. If the genes from  $f_i$  and  $f_j$  are not totally equal, the crossover operation is adopted on the offspring from  $f_i$  and  $f_j$ . To do this, a single crossover point is randomly selected. For  $f_i$  and  $f_j$ , the genes in front of the crossover point are retained but the genes after the crossover point are swapped one by one. An example of the crossover operation is illustrated in Fig. 5. If the genes from  $f_i$  and  $f_j$  are all the same, the mutation operation is adopted for their offspring chromosomes. In this case, a new chromosome is generated afresh to replace either  $f_i$  or  $f_j$ . This selection is made randomly. An example of the mutation operation is depicted in Fig. 6.

The GA process used to generate the initial cluster centers is summarized in Fig. 7. After completing the GA operations, the best  $K$  initial cluster center is then determined for the PBS algorithm.

## 4. A RFM model for profitability evaluation

This section introduces a designated RFM model to analyze the relative profitability for each customer cluster from the segmentation result after executing the proposed PBS algorithm. With this model, an enterprise can quickly



Generate a new chromosome,  $f_j$ , to replace the original one

Fig. 6. A GA mutation operation.

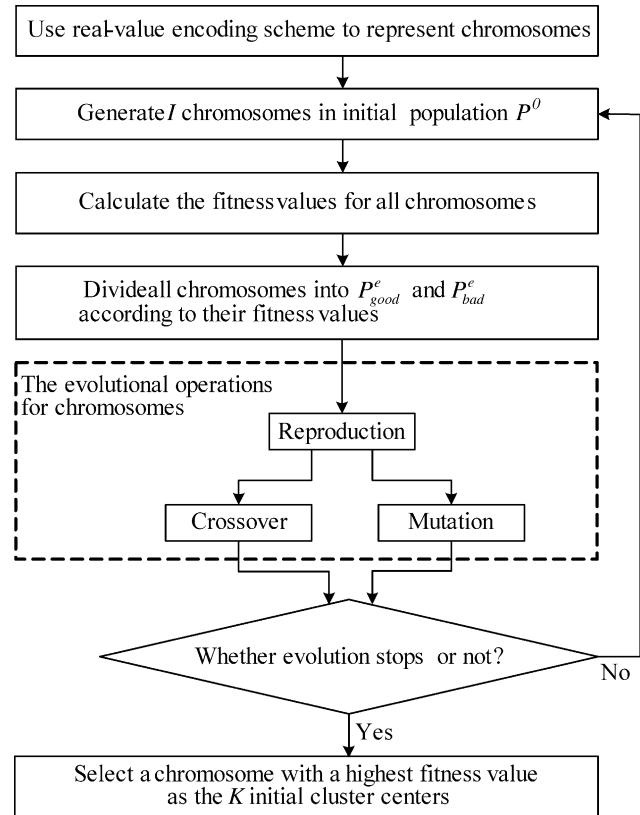


Fig. 7. The proposed GA selection process to generate initial cluster centers.

find the target clusters and adjust its marketing programs and business initiatives to provide the right products, services and resources to the target clusters. The RFM model measures the customer value based on Recency (R), Frequency (F), and Monetary (M) criteria (Hughes, 1994). Recency measures the interval between the most recent transaction time and the analyzing time. Frequency measures the purchase frequency within a specified period. Monetary measures the total monetary expenditure within a specified period. Based on this scheme, the value of a customer  $c_i$  can be represented as:

$$V(c_i) = W^R \times R(c_i) + W^F \times F(c_i) + W^M \times M(c_i) \quad (14)$$

where  $R(c_i)$ ,  $F(c_i)$ , and  $M(c_i)$  represent the scores for customer  $c_i$  in terms of the  $R$ ,  $F$ , and  $M$  criteria, respectively.  $W^R$ ,  $W^F$ , and  $W^M$  represent the importance weights for the  $R$ ,  $F$ , and  $M$  criteria, respectively. In addition,  $W^R + W^F + W^M = 1$ .

The scores can vary depending on the types of applications and scoring approaches (Hughes, 1994; Stone, 1995). The scores retrieved from the original transaction database are treated with z-score normalization before calculating the value of a customer. Therefore, the  $R(c_i)$ ,  $F(c_i)$ , and  $M(c_i)$  scores can be redefined

as follows:

$$R(c_i) = \frac{O_i^R - \mu^R}{\sigma^R}; \quad F(c_i) = \frac{O_i^F - \mu^F}{\sigma^F};$$

$$M(c_i) = \frac{O_i^M - \mu^M}{\sigma^M} \quad (15)$$

where  $O_i^R, O_i^F$ , and  $O_i^M$  represent the original values for a customer  $c_i$  derived from  $T^0$  according to the definition of  $R, F$ , and  $M$ .  $\mu^R, \mu^F$ , and  $\mu^M$  represent the averages for the  $O_i^R, O_i^F$ , and  $O_i^M$  values for all customers.  $\sigma^R, \sigma^F$ , and  $\sigma^M$  represent the standard deviations of the  $O_i^R, O_i^F$ , and  $O_i^M$  values for all customers.

The profitability of the  $n$ th customer cluster  $G^n$  can be acquired by calculating the average for all customer values in the cluster. This can be defined as Eq. (16):

$$V(G^n) = W^R \times R(G^n) + W^F \times F(G^n) + W^M \times M(G^n) \quad (16)$$

$$R(G^n) = \frac{\sum_{c_i \in G^n} R(c_i)}{\|G^n\|}; \quad F(G^n) = \frac{\sum_{c_i \in G^n} F(c_i)}{\|G^n\|};$$

$$M(G^n) = \frac{\sum_{c_i \in G^n} M(c_i)}{\|G^n\|} \quad (17)$$

where  $R(G^n), F(G^n), M(G^n)$  represent the scores for the  $n$ th cluster  $G^n$  in terms of  $R, F$ , and  $M$ , respectively. After the profitability for all clusters is known, the clusters are ranked and the most important one is identified. This is helpful for an enterprise to offer customized products and services to target specific customer clusters.

## 5. Demonstration

To demonstrate the performance of the proposed market segmentation methodology, we use the purchase data from the sales department of a retail store as an example. There were 9729 transactions generated jointly by

4223 customers in a transaction database containing 1560 items.

### 5.1. Validation for the PBS algorithm

The PBS algorithm was developed to generate clusters in which customers with similar purchasing behaviors would be located together. That is, the items purchased by customers in the same cluster should be similar. Although the items purchased in each cluster could be distinct, the purchase patterns of a customer cluster can be sketched by calculating the frequency for each item purchased by the customers in the cluster. In fact, the frequency at which an item has been purchased in one-itemset can be called ‘support,’ similar to the measure defined in Eq. (1). That is, the support for item  $i_i$  purchased by customers in cluster  $G^n$  can be evaluated as  $\|\{t^D \in D | t^D \text{ contains } i_i\}\|/\|D\|$  where  $D$  is a set of all transactions made by customers in the cluster  $G^n$  and  $t^D$  is a transaction in  $D$ .

To validate this idea, the K-means algorithm (MacQueen, 1967), using traditional customer demographics as the segmentation variables, was utilized in comparison with the PBS algorithm. In this experiment, the number of clusters was set at 30 for both algorithms. In each cluster, the average for the top five highest supports, the average of supports for all products, and the standard support deviation for all products were evaluated. The results for a K-means algorithm and the PBS algorithm are summarized in Tables 1 and 2, respectively.

To validate whether the two algorithms are significantly different in these three aspects, as described in Tables 1 and 2, one-way ANOVA tests were adopted. Table 3 illustrates the details from the test results. It was found that the average for the top five highest supports evaluated using the PBS algorithm was significantly larger than that using the K-means algorithm. Therefore, we claim that the PBS algorithm can generate clusters in which the customers tend to purchase similar products. That is, those customers

Table 1  
The purchase pattern result using the K-means clustering algorithm

Cluster id	Top five highest supports for all products						Average of supports for all products	Standard deviation of supports for all products
	1	2	3	4	5	Average		
01	0.0236	0.0189	0.0189	0.0189	0.0142	0.0189	0.0063	0.0025
02	0.0208	0.0156	0.0156	0.0156	0.0156	0.0172	0.0063	0.0027
03	0.0138	0.0138	0.0138	0.0138	0.0138	0.0138	0.0047	0.0025
04	0.0175	0.0175	0.0175	0.0175	0.0175	0.0175	0.0056	0.0028
05	0.0259	0.0185	0.0185	0.0148	0.0148	0.0185	0.0054	0.0025
06	0.0131	0.0114	0.0114	0.0098	0.0098	0.0111	0.0035	0.0018
07	0.0294	0.0221	0.0221	0.0221	0.0221	0.0236	0.0084	0.0033
08	0.0107	0.0107	0.0107	0.0107	0.0092	0.0104	0.0035	0.0017
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
28	0.0209	0.0167	0.0167	0.0167	0.0167	0.0175	0.0054	0.0026
29	0.0244	0.0244	0.0244	0.0244	0.0244	0.0244	0.0094	0.0035
30	0.0158	0.0158	0.0158	0.0158	0.0126	0.0152	0.0044	0.0023

Table 2  
The purchase pattern result using the PBS algorithm

Cluster id	Top five highest supports of all products						Average of supports of all products	Standard deviation of supports of all products
	1	2	3	4	5	Average		
01	0.0934	0.0588	0.484	0.450	0.381	0.0567	0.0052	0.0058
02	0.1019	0.0906	0.0340	0.0302	0.0302	0.0574	0.0060	0.0070
03	0.0799	0.0523	0.0303	0.0275	0.0275	0.0435	0.0050	0.0051
04	0.0403	0.0361	0.0361	0.0297	0.0297	0.0344	0.0040	0.0045
05	0.0985	0.0606	0.0492	0.0455	0.0379	0.0583	0.0061	0.0028
06	0.0909	0.0455	0.0455	0.0420	0.0350	0.0518	0.0052	0.0064
07	0.0466	0.0443	0.0373	0.0350	0.0326	0.0392	0.0042	0.0055
08	0.0463	0.0379	0.0379	0.0358	0.0295	0.0375	0.0039	0.0070
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
28	0.0591	0.0540	0.0411	0.0411	0.0360	0.0463	0.0047	0.0039
29	0.1412	0.0734	0.0621	0.0565	0.0452	0.0759	0.0080	0.0043
30	0.0606	0.0404	0.0379	0.0303	0.0303	0.0400	0.0048	0.0041

Table 3  
Validation using one-way ANOVA tests

		Sum of squares	Df	Mean square	F value	P value	$\alpha = 5\%$
average of top 5 highest supports	Variations between groups	$1.938 \times 10^{-2}$	1	$1.938 \times 10^{-2}$	245.882	0.000	$P < \alpha$
	Variations within groups	$4.572 \times 10^{-3}$	58	$7.883 \times 10^{-5}$			
	total variations	$2.395 \times 10^{-2}$	59				
average of supports for all products	Variations between groups	$1.402 \times 10^{-7}$	1	$1.402 \times 10^{-7}$	0.073	0.788	$P > \alpha$
	Variations within groups	$1.113 \times 10^{-4}$	58	$1.918 \times 10^{-6}$			
	total variations	$1.114 \times 10^{-4}$	59				
Standard deviation of supports for all products	Variations between groups	$1.176 \times 10^{-4}$	1	$1.176 \times 10^{-6}$	103.242	0.000	$P < \alpha$
	Variations within groups	$6.607 \times 10^{-5}$	58	$1.139 \times 10^{-6}$			
	total variations	$1.837 \times 10^{-4}$	59				

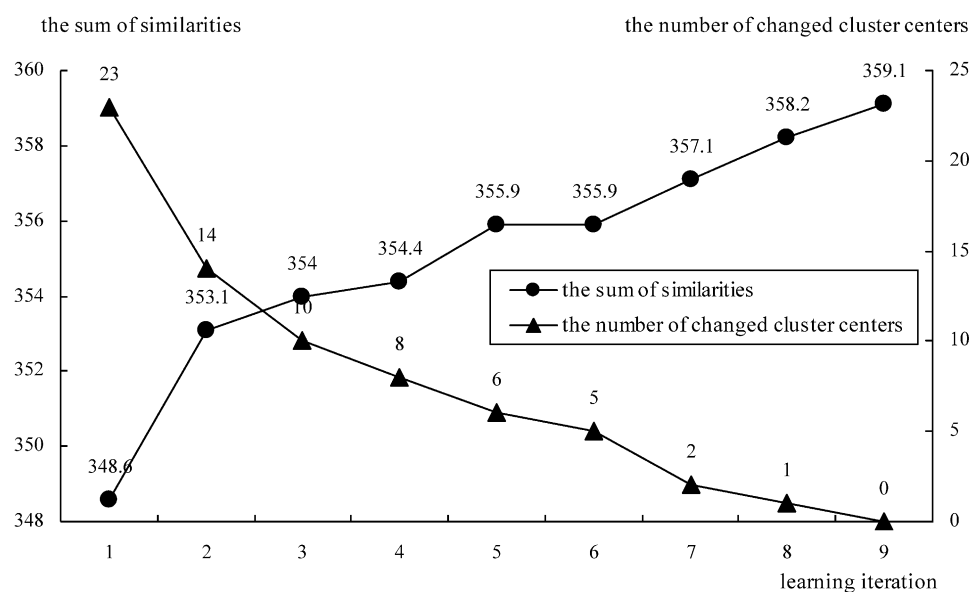


Fig. 8. Convergence in learning iterations.



that have the closest purchase behaviors. In addition, the PBS algorithm result was also significantly larger than the K-means algorithm result in the standard support deviation for all products. That means that the PBS algorithm can generate a cluster in which items have significantly different purchase attraction.

The PBS algorithm can converge quickly after a few learning iterations. To show the convergence, the sum of similarities between each customer and his/her cluster center and the number of changed cluster centers were observed. As shown in Fig. 8, when the number of iterations increased, the sum of the similarities between each customer and his/her cluster center increased, and the number of changed cluster centers decreased. It was found that after nine iterations the algorithm reached its stopping criteria and completed the calculation.

### 5.2. GA performance evaluation

As mentioned in Section 3, the initial cluster centers generated using a random selection process tended to make the clustering quality fall into local optimization. In this paper, a heuristic genetic algorithm (GA) selection process was developed to generate better initial cluster centers, allowing the clustering quality to be improved. To show the benefit, a series of experiments were conducted to compare the GA and random selection process performance. We tested cases involving initial cluster centers ranging from 30 to 80. The clustering quality defined in Eq. (7) was used to judge the clustering performance. As shown in Fig. 9, GA selection generated better clustering results than random selection in all cases. Fig. 10 shows the computation time for executing the PBS algorithm using the two selection processes. We observed that the computation time using the GA selection process was about two times longer than the process using random selection. In addition, the time increased slowly when the number of clusters grew linearly. Note that when a random selection process was adopted, the computation time decreased after the number

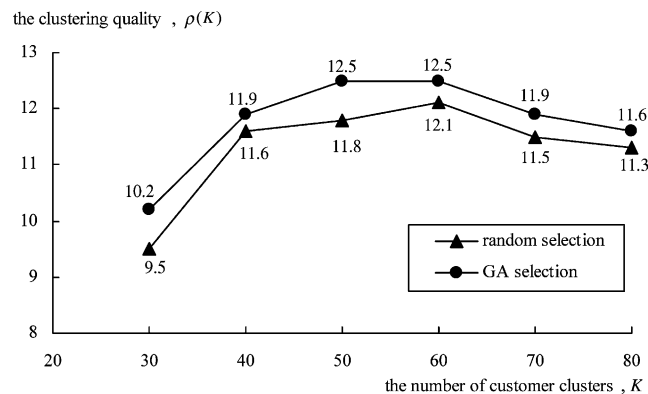


Fig. 9. The clustering quality result using GA and random selection processes.

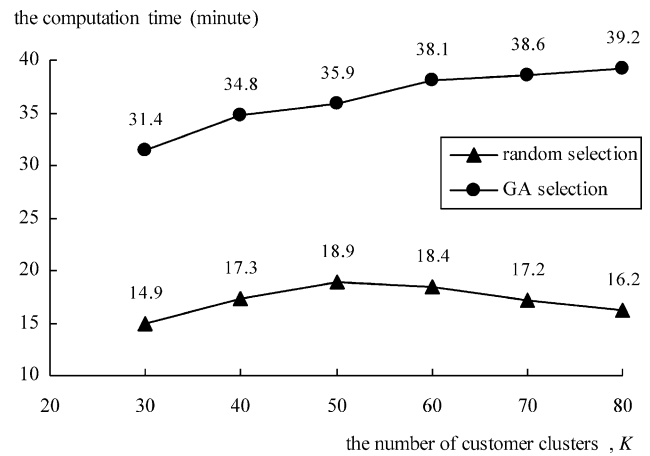


Fig. 10. The computation time using GA and random selection processes.

of initial centers grew to more than 50. The reason is that some clusters contained few customers after the number of clusters rose above 50.

Several parameters should be adjusted appropriately in the GA. Among them, the number of chromosomes in a population  $L$  and the maximum number of generations  $E$  are critical. To understand how these two parameters affect the clustering quality, the following experiments were conducted. Let the number of customer clusters be 30. In Fig. 11, the vertical axis represents the clustering quality and the horizontal axis represents the maximum number of generations from 5 to 35. Each curve in the figure represents a different number of chromosomes in a population, i.e. 10–60. We found that when the maximum number of generations grew, the clustering quality also increased for all curves. However, when the maximum number of

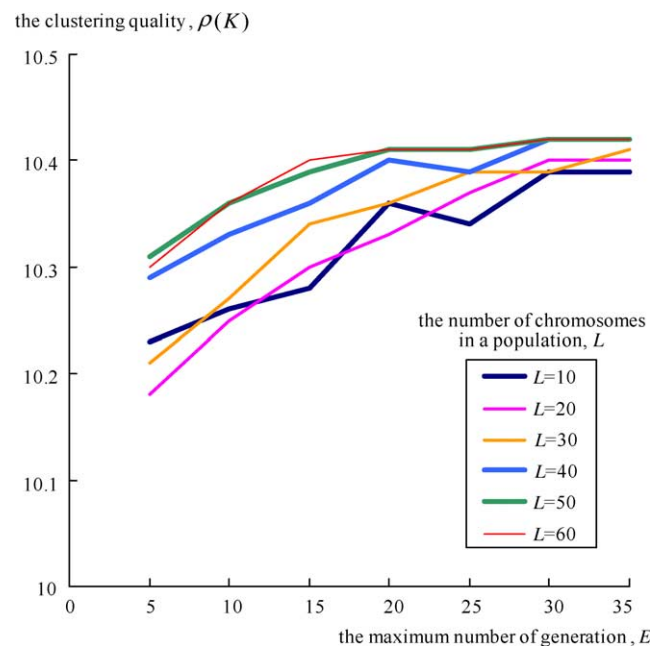


Fig. 11. The clustering quality result with different GA settings.

Table 4  
The RFM analysis result

Cluster id	$R(G^n)$	$F(G^n)$	$M(G^n)$	$V(G^n)$	Rank	Cluster id	$R(G^n)$	$F(G^n)$	$M(G^n)$	$V(G^n)$	Rank
1	-0.020	0.090	0.023	0.041	22	16	-0.065	0.252	0.197	0.166	11
2	-0.055	0.310	0.259	0.216	5	17	-0.055	-0.005	-0.095	-0.050	28
3	-0.080	0.221	0.108	0.115	15	18	-0.195	0.034	-0.041	-0.041	27
4	-0.167	-0.015	-0.082	-0.071	30	19	0.081	0.151	0.047	0.095	16
5	-0.027	0.238	0.215	0.175	10	20	-0.011	0.130	0.053	0.070	18
6	-0.056	0.092	0.010	0.029	24	21	0.081	0.264	0.132	0.174	10
7	0.040	0.306	0.201	0.210	6	22	0.065	0.103	-0.040	0.038	23
8	0.118	0.322	0.248	0.251	4	23	-0.003	0.183	0.170	0.140	14
9	0.126	0.351	0.266	0.272	3	24	0.068	0.213	0.142	0.155	12
10	-0.052	0.109	0.028	0.044	21	25	-0.098	0.149	0.108	0.083	17
11	-0.149	-0.002	-0.072	-0.059	29	26	-0.125	0.042	-0.072	-0.036	25
12	0.049	0.246	0.180	0.180	9	27	0.071	0.225	0.249	0.203	7
13	0.233	0.528	0.401	0.418	1	28	-0.104	0.256	0.156	0.144	13
14	0.214	0.366	0.285	0.303	2	29	-0.050	0.149	0.046	0.067	19
15	-0.178	0.314	0.264	0.195	8	30	0.041	0.108	0.035	0.065	20

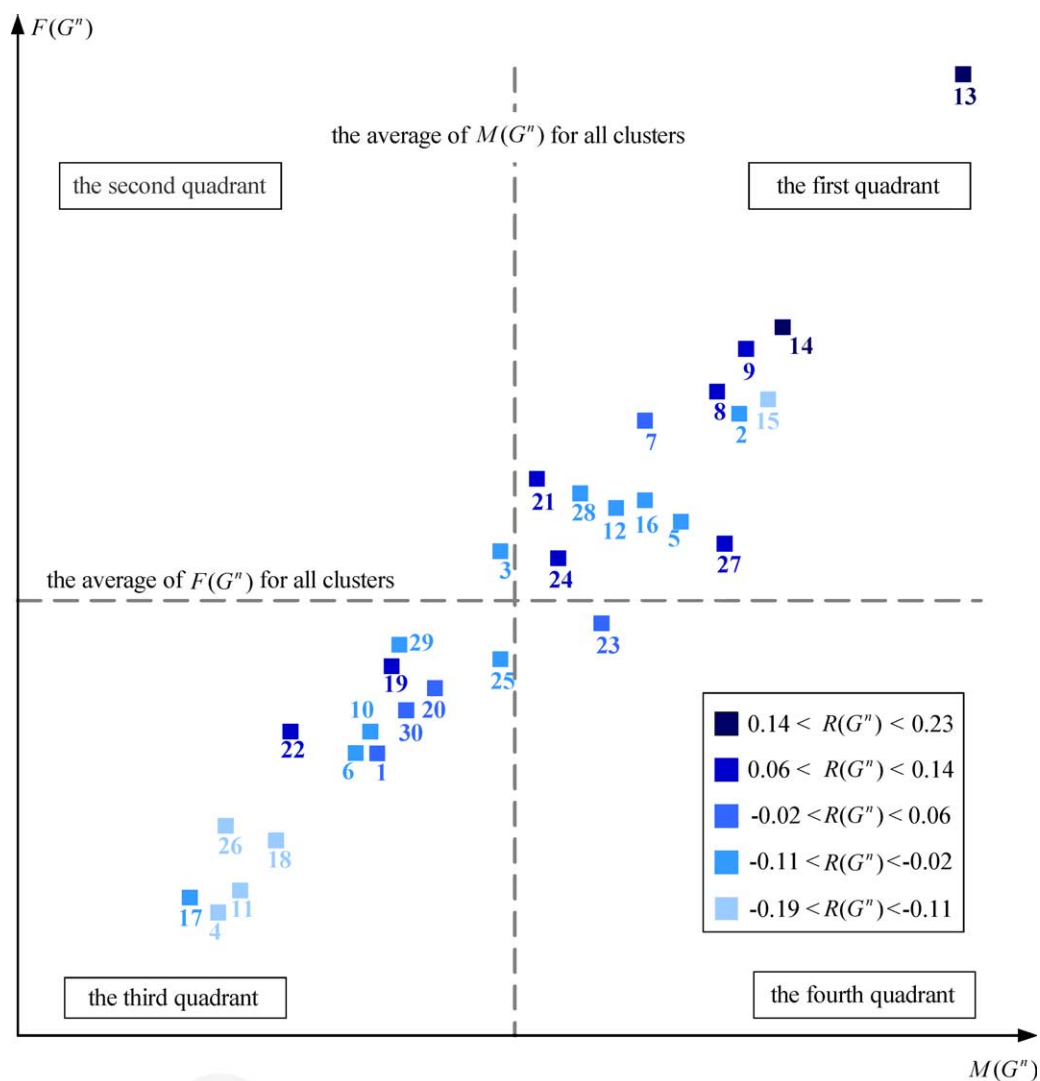


Fig. 12. The two-dimensional RFM model graph.

generations became greater than 25, the clustering quality increased slowly. When the number of generations became greater than 35, no significant difference in the clustering quality occurred for all curves. An interesting finding was that the clustering quality did not linearly conform to the number of chromosomes.

### 5.3. RFM profitability analysis

The RFM model was used to analyze the relative profitability of each customer cluster. To demonstrate the proposed RFM model performance, an analysis example using the clustering result generated in Section 5.2 is introduced. The weights for the  $R$ ,  $F$ ,  $M$  criteria were set as  $W^R = 0.2$ ,  $W^F = 0.4$ , and  $W^M = 0.4$  in this case. After a series of calculations using Eqs. (14)–(17), the RFM profitability analysis result is summarized in Table 4. To provide a clear view for marketing programs, a 2-dimensional RFM model graph is depicted in Fig. 12. A customer cluster  $G^n$  is located at position  $(x, y)$  with a gray color level where  $x$  represents the  $M(G^n)$  value,  $y$  represents the  $F(G^n)$  value, and the gray level represents the  $R(G^n)$  value. The graph is further partitioned into four quadrants so that marketers can easily locate interesting clusters. The horizontal line represents the average  $F(G^n)$  values for all clusters and the vertical line represents the average  $M(G^n)$  values for all clusters. The clusters in the first quadrant have higher  $M(G^n)$  and  $F(G^n)$  values, while the clusters in the third quadrant have lower  $M(G^n)$  and  $F(G^n)$  values.

From Fig. 12, we find that most clusters belong to the first and third quadrants except that the third and 23rd clusters are in the second and fourth quadrants, respectively. In the first quadrant, the 9th, 13th, and 14th clusters can be identified as the most profitable groups. Among them the 13th cluster is the most valuable group, because the customers in this group spend large amounts of money (high  $M(c_i)$  values), purchase frequently (high  $F(c_i)$  values) and have contributed recently (high  $R(c_i)$  values). Note that although the 15th cluster is located in the first quadrant, its  $R(G^n)$  value is relatively low compared to the others in the same quadrant. Marketers should try to understand why the customers in this cluster have not purchased recently and solve this problem before losing them. In addition, marketers may allocate fewer resources to the 4th, 11th, 18th, and 26th clusters or adjust their marketing programs and business incentives to encourage them to purchase, because these clusters are located in the left-bottom corner of the third quadrant.

## 6. Conclusions

The mass marketing approach cannot satisfy diverse customer needs today. This diversity should be exploited using market segmentation that divides the market into

customer clusters with similar needs, characteristics and purchasing behaviors. This paper introduced a novel purchase-based market segmentation methodology. This methodology was developed based on product specific variables such as the purchased items and associated monetary expenses from transactional customer histories. This allows groups of customers with similar purchasing behaviors, providing a more homogeneous response to marketing programs. To ensure that customers in the same cluster have the closest purchase patterns, a heuristic genetic algorithm (GA) was embedded into our clustering algorithm. The GA performance was examined using a series of experiments. These experiments showed that GA adoption generated much better clustering quality within an acceptable computation time. After segmentation, a designated RFM model was introduced to analyze the relative profitability of each customer cluster. The RFM profitability analysis highlighted more marketing opportunities and helps marketers to revise their marketing strategies. In the future, we will enhance the proposed methodology using fuzzy set theory to refine the uncertain relationship between the customers and clusters.

## Acknowledgements

This work was partially supported by the National Science Council of Taiwan No. NSC 91-2213-E155-052.

## References

- Agrawal, R., Imielinski, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. *Proceedings of the ACM SIGMOD international conference on management of data* (pp. 207–216).
- Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules. *Proceedings of the 20th international conference on very large databases* (pp. 487–499).
- Beane, T. P., & Ennis, D. M. (1987). Market segmentation: a review. *European Journal of Marketing*, 21(5), 20–42.
- Berson, A., Smith, S., & Thearling, K. (2000). *Building data mining applications for CRM*. New York: McGraw-Hill.
- Bradley, P. S., & Fayyad, U. M. (1998). Refining initial points for K-means clustering. *Proceedings of the 15th international conference on machine learning* (pp. 91–99).
- Chou, P. B., Grossman, E., Gunopulos, D., & Kamesam, P. (2000). Identifying prospective customers. *Proceedings of the sixth international conference on knowledge discovery and data mining* (pp. 447–456).
- Dibb, S., & Simkin, L. (1996). *The market segmentation workbook: target marketing for marketing managers*. Routledge, London.
- Dimitriadou, E., Weingessel, A., & Hornik, K. (1999). Voting in clustering and finding the number of clusters. *International Congress on Computational Intelligence: Methods and Applications*, 291–296.
- Drozdenko, R. G., & Drake, P. D. (2002). *Optimal database marketing: Strategy, development, and data mining*. London: Sage.

- Goldberg, D. E. (1989). *Genetic algorithms in search, optimization and machine learning*. Reading, MA: Addison-Wesley.
- Hammond, K., Ehrenberg, A. S. C., & Goodhardt, G. J. (1996). Market segmentation for competitive brands. *European Journal of Marketing*, 30(12), 39–49.
- Holland, J. H. (1975). *Adaptation in natural and artificial systems*. Ann Arbor, MI: The University of Michigan Press.
- Hughes, A. M. (1994). *Strategic database marketing: The masterplan for starting and managing a profitable, customer-based marketing program*. Probus Pub Co.
- Kuo, R. J., Ho, L. M., & Hu, C. M. (2002). Integration of self-organizing feature map and K-means algorithm for market segmentation. *Computers and Operations Research*, 29(11), 1475–1493.
- MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Conference on Mathematical Statistics and Probability*, 1, 281–297.
- Mannila, H. (1998). Database methods for data mining. *Proceedings of the fourth international conference on knowledge discovery and data mining*, New York.
- Manning, C. D., & Schutze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.
- Meila, M., & Heckerman, D. (1998). An experimental comparison of several clustering and initialization methods. *Proceedings of the 14th conference on uncertainty in artificial intelligence* (pp. 386–395).
- Natter, M. (1999). Conditional market segmentation by neural networks: a Monte-Carlo study. *Journal of Retailing and Consumer Services*, 6(4), 237–248.
- Romesburg, H. C. (1984). *Clustering analysis for researchers*. Belmont: Lifetime Learning Publications.
- Stone, B. (1995). *Successful direct marketing method*. Lincolnwood, IL: NTC Business Books.
- Wedel, S., & Kamakura, W. (1997). *Market segmentation: Conceptual and methodological foundations*. Boston: Kluwer.