A Market Segmentation System for Consumer Electronics Industry Using Particle Swarm Optimization and Honey Bee Mating Optimization

Chui-Yu Chiu[1], I-Ting Kuo and Po-Chia Chen

Department of Industrial Engineering and Management, NTUT, Taiwan.

Abstract. The use of information technologies in various business areas is emerging in recent years. With the development of information technology, how to find useful information existed in vast data has become an important issue. The most broadly discussed technique is data mining, which has been successfully applied to many fields and analytic tools. Clustering analysis which tries to segment data into homogeneous clusters is one of the most useful technologies in data mining methods. Market segmentation is among the important issue of most companies. Market segmentation relies on the data clustering in a huge data set. In this study, we propose a clustering system which integrated particle swarm optimization and honey bee mating optimization methods. Simulations for a benchmark test functions show that our proposed method possesses better ability to find the global optimum than other well-known clustering algorithms. The results show that system through PSHBMO can effectively find the global optimum solution, and extend the application of market segmentation to solve the RFM model.

Keywords. Market segmentation, Particle swarm optimization, Honey bee mating optimization.

1 Introduction

Recently, the progress of information technology has transformed the way of marketing and information management in companies. With the large number of data from customer behavior, it has become possible to realize the consumer insight by the variety of data mining techniques and CRM tools. In general, market segmentation is the most important issue for companies to recognize different group of customers, who have some similar characteristics explaining different customer value.

Market segmentation divides a market into distinct subsets of buyers, each has some similar attributes. It's most important variable is purchasing behavior. In order to describe customer's purchasing behavior, the RFM analytic model

[1] Corresponding Author : Department of Industrial Engineering and Management, National Taipei University of Technology, 1, Section 3, Chung-Hsiao East Road, Taipei 106, Taiwan, ROC; Tel.: +886 2 2771 2171x2365; fax: +886 2 2731 7168. E-mail address: cychiu@ntut.edu.tw (C. Y. Chiu).

proposed by Hughes [1], is usually applied. It is a model that tells important customers from large data by three attributes, i.e., interval of customer purchasing, frequency and amount of money. The definitions of RFM model are summarized as follows: (1) R-Recency of the last purchase; (2) F-Frequency of the purchases; (3) M-Monetary value of the purchases.

Data clustering is a useful technique for the discovery of some knowledge from a dataset. Clustering is the process of grouping a set of abstract or physical objects into classes of similar objects. The purpose of clustering analysis is to find the difference among each groups and the similarity in the same group. Aldenderfer and Blashfield [2] concluded five basic steps that characterized all clustering studies. Each of these steps is essential to clustering as follows. (1)Selection of a sample to be clustered; (2) Definition of a set of variables on which to measure the entities in the sample: (3) Consumption of the similarities among the entities; (4) Use of clustering method to create groups of similar entities, and (5) Validation of the resulting cluster solution.Clustering techniques are relatively popular for market segmentation due to its short computation time and easy accommodation. For examples, Shin et al. [3] used three clustering algorithms, including K-means, FCM and SOM, to find properly graded stock market trading brokerage commission rates based on the 3-month long total trades of two different transaction modes and concluded that FCM is the most robust approach. Huang et al. [4] used support vector clustering for market segmentation by the drink company. The results can be seen that the proposed method can outperform to the k-means and the SOFM methods. Liao et al. [5] proposed the apriori algorithm and clustering analysis as methodologies for data mining. The results was illustrated what functionalities best fit the consumers' needs and wants for life insurance products by extracting specific knowledge patterns and rules from consumers and their demand chain.

Particle swarm optimization (PSO) is an evolutionary computation technique developed by Kenney and Eberhart [6]. The method has been developed through a simulation of simplified social models. Each particle also has its own coordinates and velocity to change the direction of flying in predefined search domain with *D*-dimensional. All particles fly through the search domain by following the direction of current optimal particle. He et al. improved the standard PSO with passive congregation (PSOPC), which can improve the convergence rate and accuracy of the SPSO efficiently [7]. Liu [8] proposed CPSO (chaotic particle swarm optimization) algorithm. It applies PSO to perform global exploration and chaotic local search to perform local search on the solutions produced in the global exploration process. Over the years, many successful applications of PSO, to image registration [9], have been reported. PSO algorithm is a powerful optimization technique for solving multimodal continuous optimization problems [10].

Afshar [11] brought up a swarming relation in society model assuming a polygynous colony called honey bee mating optimization (HBMO). It studies the behavior of social insects and uses their models to solve the optimum problems. Each bee performs sequences of actions according to genetic, environmental, and social regulation. Result of each action itself became a portion of the environment and greatly influences the subsequent actions of both single bee and many drones. The marriage process represents one type of action that was difficult to study

because the queens mate during their mating-flight far from the nest. Fathian and Amiri [12] applied honeybee mating optimization in clustering problems. It compared HBMO means with other heuristics algorithm in clustering, such as GA, SA, TS, and ACO, by implementing them on several well-known datasets. It was found that the HBMO algorithm works than the best one.

2. Methodology

In most customer relationship management problems, one may notice that no tool for data mining is perfect because there are many uncertain variables. In this study, we propose an integrating particle swarm optimization with honey bee mating optimization (PSHBMO) market segmentation system based on the structure of decision support system to solve the clustering problem.

The proposed method was divided into four phases: (1) Data pre-processing (2) Using particle swarm optimization to search initial solution (3) Using honey bee mating optimization to search the best solution (4) Comparing clustering performance evaluation, as follows:

2.1 Data pre-processing

Before feeding data into the clustering algorithms, database variables should be normalized to eliminate scale effects. Normalization entails relatively minor additional computations during application of a solution to new data, which must also be normalized. For some attributes whose preferences are monotonically increasing, such as frequency and population et al., a simple positively linear normal function is shown in Eq. (3.1). However, exhibits a monotonically decreasing preference. An inverse function is shown in Eq. (3.2) is then applied.

$$NormalizedValue = (log_{Xi} - log_{Xmin}) / (log_{Xmax} - log_{Xmin}) \quad (3.1)$$

$$NormalizedValue = 1-(log_{Xi} - log_{Xmin}) / (log_{Xmax} - log_{Xmin}) \quad (3.2)$$

2.2 Using particle swarm optimization to search initial solution

This step utilizes particle swarm optimization (PSO) to decide the initial vector of cluster center. A single particle represents the k cluster centroid vectors. That is, each particle X_{id} has its vectors V_{id} that be constructed as follows:

$$X_{id} = (z_{i1},\ldots\ldots,z_{ik}) \quad (3.3)$$

where z_{ij} refers to the k-th cluster centroid vector of the i-th particle

(1) Initialize each particle to contain k randomly selected cluster centroids.

(2) Calculate the Euclidean distance between all of data to cluster centroids, and assign for the minimum distance.

(3) Calculate the fitness using equation follow:

$$J_e = \frac{\sum_{j=1}^{k} \left[\sum_{\forall x \in n_{ij}} d(x, z_{ij}) / \left|n_{ij}\right|\right]}{k} \quad (3.4)$$

(4) Update the global best, local best positions and cluster centroids (X_{id} and V_{id}) as follow, and recalculate back to step (2):

$$V_{id}^{new} = W \times V_{id}^{old} + c_1 \times rand_1 \times (P_{id} - X_{id}) + c_2 \times rand_2 (P_{gd} - X_{id}) \quad (3.5)$$

$$X_{id}^{new} = X_{id}^{old} + V_{id}^{new} \quad (3.6)$$

(5) Proceed until meet epochs equal to a parameter. Keep the global best solution and locate best solution.

2.3 Using honey bee mating optimization to search the best solution.

Having got the initial solution of PSO, this step was searching the optimization solution. This algorithm is constructed with the following five main stages:

(1) Initialize mutation rate, cross rate and each parameter. Let the global best solution be queen's chromosome and location best solution be drone's chromosome.

(2) Use simulate annealing and roulette wheel selection to select the set of drones from the list for the creation of broods. After each transition in space, the queen's speed and energy decays according to the following equations:

$$S(t+1)=\alpha(t)\times S(t) \quad (3.7)$$

where α is a factor[0,1] and is the amount of speed reduction after each transition.

(3) Create new set of broods by crossover the drone's genotypes with the queens.

(4) Calculate the Euclidean distance between all of data to cluster centroids, and assign for the minimum distance.

(5) Calculate the fitness using equation as formulate 3.4. Determining that is the new best solution better than the previous one.

(6) Randomly mutate chromosome and replace weaker queens by fitter broods.

(7) Proceed until meet epochs equal to a parameter.

2.4 Comparing clustering performance evaluation.

This algorithm tries to minimize the error function – Mean Square Error (MSE). The main purpose is to compare the quality of the respective cluster, where quality is measured according to the following three criteria:

(1) The mean square error as defined in Eq.(3.8).

$$\text{MSE}= \sum_{i=1}^{n} \sum_{j=1, x_i \in c_j}^{k} \left| x_i - m_j \right|^2 \quad (3.8)$$

where x_i (i=1,2,…n) is a data set X with n objects,k is the number of clusters,m_j is the centroid of cluster C_j (j=1,2,…k).

(2) The intra-cluster distances: the distance between data vectors within a center, where the objective is to minimize the intra-cluster distances. It is defined in Eq. (3.9).

$$\text{intra}-\text{clusterdistance}=\sum_{i=1}^{n}\sum_{j=i+1}^{n} d\left(x_i,x_j\right) \tag{3.9}$$

where x_i (i=1,2,…n) is a data set X with n objects, x_j (j=1,2,…n) is a data set X with n objects.

(3) The inter-cluster distances: the distance between the centroids of the clusters, where the objective is to maximize the distance between clusters. It is defined in Eq. (3.10).

$$\text{inter}-\text{clusterdistance}=\sum_{i=1}^{k}\sum_{j=i+1}^{k} d\left(m_i,m_j\right) \tag{3.10}$$

where m_i is the centroid of cluster C_i (i=1,2,…k), m_j is the centroid of cluster C_j (j=1,2,…k),k is the number of clusters.

This research considered inter-cluster and intra-cluster distances at the same time to make sure that the latter ensures compact clusters with little deviation from the cluster centroids, while the former ensures larger separation between the different clusters. In order to get the value which maximize the distance between clusters and minimize the intra-cluster distances, this study according to the notion of Intra−cluster Distance and Inter−cluster Distance.

3 Evaluation of proposed model on data sets

We presented a set of experiments to show the performance of the PSHBMO algorithm. The experiment was conducted on a Pentium 3.40 GHz, 512 MB RAM computer and coded with Borland C++ 6.0 Builder software.

3.1 Experiment-IRIS

A well-know database, the Iris Plant, is utilized to test the performance of our proposed method. Iris plant is a database with 4 numeric attributes, 3 classes and 150 instances. We compared the PSHBMO algorithm with PSO+K-means and SOM+K-means by Chiu [13]. Finally, we provided the clustering results and used three criteria to evaluate the quality of the results as in following table.

Table1 summarizes the results obtained from the clustering algorithm for the problem above. The values reported are averaged 30 simulations, for which standard deviations indicates the range of values where the algorithms converge. If the algorithm could cluster data with a lower MSE value, the similarity within a segment increases. For the problem, the PSHBMO algorithm had a smallest average MSE. When considering Intra −cluster Distance and Inter −cluster Distance, the PSHBMO algorithm also had a smallest value comparing with PSO +k-means and SOM +k-means.

Table 1. Comparison of three methods of proposed system

Algorithm	MSE	Inter-cluster Distance	Intra-cluster Distance	$\frac{\text{Intra - cluster Distance}}{\text{Inter - cluster Distance}}$
PSO+k-means	0.218	0.281	0.803	0.350068
SOM+k-means	0.224	0.290	0.792	0.366445
PSHBMO	0.194	0.264	0.804	0.329403

3.2 Case study-RFM model

For the case study, a real-world database is collected from Podak Co., an authorized agent of Panasonic that provides passive and active electronic components for consumer electronics, telecommunications, computers etc. The period of the business transaction data is from 2003/1/1 to 2006/6/15.

In this study, we employ a two-stage clustering suggested by Punj and Stewart [14]. In the first stage, PSHBMO is used to cluster the normalized data set into different groups. The result is shown in Table2. It is found that the intra-cluster distance /inter-cluster distance is the lowest at Group=9 and the distance is relatively decreasing flatly when the number of clusters is more than nine. Therefore, it is implies that the best number of clusters should be nine.

Table 2. Clustering Result Form Group=3 to Group=10

Index	Group=3	Group=4	Group=5	Group=6	Group=7	Group=8	Group=9	Group=10
Intra-cluster distance	0.305664	0.305267	0.274965	0.265611	0.248516	0.223385	0.218951	0.239033
Inter-cluster distance	0.585759	0.612778	0.580098	0.598582	0.578599	0.571645	0.575373	0.591121
$\frac{\text{Intra – cluster}}{\text{Inter – cluster}}$	0.521826	0.49817	0.473997	0.443734	0.429513	0.390776	0.380537	0.404371

In the second stage, we compared the results of the PSHBMO, SOM+*k*-means and PSO+*k*-means algorithms showed the best number of groups is six. The result is presented in Table 3.

Table 3. Comparison of three methods

Algorithm	MSE	Inter-cluster Distance	Intra-cluster Distance	$\frac{\text{Intra - cluster Distance}}{\text{Inter - cluster Distance}}$
PSO+k-means	0.185	0.238	0.580	0.410162
SOM+k-means	0.190	0.246	0.574	0.429780
PSHBMO	0.168	0.242	0.603	0.401326

As a result, PSHBMO posseses lowest MSE and $\frac{\text{Intra - cluster Distance}}{\text{Inter - cluster Distance}}$. PSHBMO is better than PSO+ *k*-menas and SOM +*k*-means. The clustering results of PSHBMO is further utilized for make marketing strategies; Table 4 shows the clustering results for customers.

Table 4. Clustering Results of PSHBMO

Cluster	Customer Counts	Recency (Avg.)	Frequency (Avg.)	Monetary (Avg.)	RFM Status
1	2	253.5000	1773.0000	89,582,024.00	R↑F↑↑M↑↑
2	10	16.0000	440.3000	4,420,767.00	R↓↓F↑M↑
3	9	152.8889	194.8889	10,984,008.00	R↓F↑M↑
4	12	12.6667	8.5000	183,222.40	R↓↓F↓↓M↓
5	16	118.0000	23.5000	93,436.13	R↓F↓M↓↓
6	29	331.1034	2.7931	15,980.83	R↑F↓↓M↓↓
7	16	355.0625	10.8125	314,885.10	R↑F↓M↓
8	12	1.4167	402.2500	11,089,204.00	R↓↓F↑M↑
9	14	20.7143	74.7143	491,144.30	R↓↓F↓M↓
Total Avg.		163.94167	135.9000		3,928,079.2833

In table 4, the sign ↑ denotes that the value was greater than an average, and the sign ↑↑ denotes that the value was the much greater than an average. the sign ↓ denotes that the value was smaller than an average, and the sign ↓↓ denotes that the value was the much smaller than an average. Cluster 2, Cluster 3 and Cluster 8 which has R↓(↓)F↑M↑ can be considered as loyal ones who frequently deal with and make a large purchase. Cluster 4, Cluster 5 and Cluster 9 who has R↓F↓M↓ was probably a new customer who recently dealt with. Cluster 1 who has R↑F↑↑M↑↑ is promising one who might be promoted to the loyal customer. Cluster 6 and Cluster 7 who have R↑F↓M↓ is likely to be vulnerable customers who have not dealt with for a long time.

Among the nine clusters, Cluster 3 is selected as a target customer segment with the first priority, followed by Cluster 6. It is because that the effect to these target segments might become potentially greater than the effect to others from the RFM point of view.

4 Conclusion

Data mining is the process of posting various queries and extracting useful information, patterns, and trends often previously unknown from large quantities of data possibly stored in databases. Database technology has been used with great success in traditional business data processing. Through this study, we built that an effective and accurate market segmentation system based on intelligent clustering methods integrated particle swarm optimization and honey bee mating

optimization methods. We used three clustering algorithms to solve the Iris dataset. For the problem, the PSHBMO algorithm had a smallest value of MSE and Intra−cluster Distance/Inter−cluster Distance which compared with PSO+k-means and SOM+k-means. In the business database, we decided RFM variables are to be used for analyzing. First, we ran PSHBMO for this normalized data to find which group had the lowest of the intra-cluster distance /inter-cluster distance at nine groups. Then, the result found that Cluster 3 was a target customer segment. With the above clustering information, the proposed method can help marketers to develop proper tactics for their customers.

In order to segment customers precisely, many enterprises have installed customer clustering system. However, most of customer clustering systems in operation are lack of an accuracy performance evaluation analysis based on logical inference or just provide a rough clustering method. An improper clustering result usually leads to a conclusion that can't meet firms and customers' requests and expectation. Consequently, this study proposes a system using the technique of data mining to serve as decision support system by finding useful information hided out.

Furthermore, the procedure proposed in this study can be also applied to other industries. According to the clustering analysis, enterprises of different industry can design various marketing strategies and advertisement models to aim at different group of customers with common features to achieve the needs of customers, enhance the strength with customers for company, satisfy their demands, increase customer loyalty and finally obtain better profits.

5 References

[1] Hughes AM. Strategic Database Marketing: The masterplan for starting and managing a profitable customer-based marketing program. Chicago: Probus Publishing. 1994

[2] Aldenderfer MS, Blashfield RK. Cluster Analysis: Quatitative applications in the social sciences. Beverly Hills: Sage Publication, 1984.

[3] Shin HW, Sohn SY. Segmentation of stock trading customers according to potential value. Expert Systems with Applications. Vol. 27, pp.27-33, 2004.

[4] Huang, J. H., Tzeng GH, Ong CS. Marketing segmentation using support vector clustering. Expert Systems with Applications. Vol.32, pp.313-317,2007.

[5] Liao SH, Chen YN, Tseng YY, Mining demand chain knowledge of life insurance market for new product development. Expert Systems with Applications. Vol. 36, pp.9422–9437,2009.

[6] Kennedy J, Eberhart RC. Particle swarm optimization. International Joint Conference,Neural Networks, Vol. 4, pp. 1942-1948, 1995.

[7] He S, Wu QH, Wen JY, Saunders JR, Paton RC. A particle swarm optimizer with passive congregation. Biosystem,78:135-47,2004.

[8] Liu B, Wang L, Jin YH, Tang F, Huang DX. Improved particle swarm optimization combined with chaos. Chaos, Solitons &Fractals,25:1261-71,2005.

[9] Yin PY. Particle swarm optimization for point pattern matching. J. Visual Commun. Image Representation 17, 143-162,2006.

[10] Paterlini S, Krink T. Differential evolution and particle swarm optimisation in partitional clustering. Computational Statistics & Data Analysis 50 (5): 1220–1247, 2006.
[11] Afshar A, Bozorg O. Honey-bee mating optimization (HBMO) algorithm for optimal reservoir operation. Journal of the Franklin Institute, Vol.344, pp.452-462 , 2007.
[12] Fathian M, Amiri B. Application of honey-bee mating optimization algorithm on clustering, Applied Mathematics and Computation. Vol.199, pp.1502-1513, 2007.
[13] Chiu CY, Chen YF, Kuo IT, Ku HC. An Intelligent Market Segmentation System Using K-Means and Particle Swarm Optimization. Expert Systems with Applications, 2008.
[14] Punj G, Stewart DW. Cluster Analysis in Marketing Research: Review and Suggestions for Application. Journal of Marketing Research, Vol. 20, pp. 134-148, 1983.