Taller Latino Iberoamericano de Investigación de Operaciones
"La IO aplicada a la solución de problemas regionales y de administración pública"

A heuristic approach to a market segmentation problem with multiple attributes

Diana L. Huerta-Muñoz[1], Roger Z. Ríos Mercado[1], Rubén Ruiz[2]

[1]Graduate Program in Systems Engineering
Universidad Autónoma de Nuevo León
San Nicolás de los Garza, Nuevo León, México

[2]Department of Applied Statistics,
Operations Research and Quality
Universitat Politècnica de València
Valencia, Spain

Resumen: Este trabajo trata un problema real de segmentacion de clientes de una empresa distribuidora de bebidas. Dado un conjunto de clientes, los cuales comparten atributos geográficos y comerciales (volumen de compra que el cliente demanda de un determinado tipo de producto llamado SKU, tipo de contrato y tipo de establecimiento), la empresa desea particionar dicho conjunto en segmentos según determinados requerimientos. Se desea que los clientes asignados a un mismo segmento tengan atributos muy similares. La razón principal de obtener una partición con estas características es debido a que la empresa requiere aplicar diferentes estrategias de mercadotecnica para cada uno de los segmentos establecidos. Además se desea obtener segmentos compactos, es decir, los clientes que forman un mismo segmento deben estar relativamente cercanos unos de otros. En este trabajo proponemos un modelo matemático y una metodología de solución basada en una Búsqueda Local Iterativa Voraz. Resultados computacionales son mostrados en este trabajo.

Abstract: This paper addresses a real-world customer segmentation problem from a beverage distribution firm. Given a set of customers, which share geographical and marketing attributes (volume that customers demand from a specific type of product called SKU, type of contract, and type of store), the firm wants to partition this set into segments according to certain requirements. It is desired that customers allocated to the same segment must have very similar attributes. The main reason to get a partition with these features is because the firm wants to try different product marketing strategies. In addition, the firm wishes compact segments, that is, customers within a segment must be relatively close to each other. In this work, we propose a mathematical model and a solution metodology based on an Iterated Greedy Local Seach in a variable neighborhood enviroment. Computational results are presented.

Keywords: Customer Segmentation, Metaheuristics, Variable Neighborhood Search, *k*-means, Iterated Local Search.

Introduction

Market segmentation is a process that involves dividing the total market for a product or service in several smaller groups that are internally homogeneous. The essence of segmentation is to know what consumer needs. Market segmentation has been widely studied in different forms [1-4].

In this paper we are studying a real-world case from a beverage distribution firm. We are interested on finding partitions of a set of customers with respect to four specific attributes. In this work we proposed a mathematical model and developed an iterated greedy local search heuristic composed by one construction phase and one destruction phase to solve this problem. To the best of our knowledge, the particular market segmentation problem presented in this work has not been addressed before.

Problem Description

We are working on a case study with real-world data. Given a set of customers V who share geographical and marketing attributes (the total volume a_{is} that a customer i demands of a specific product type s (SKU), the type of contract c, and the type of store e), the company wants to partition this set into p segments according to certain requirements. It is desired that customers allocated to the same segment must have very similar attributes. In addition, the firm wishes compact segments.

The following requirements are given by the firm: (a) if the types of contract (store) of two different customers are the same, the dissimilarity of these customers with respect to that attribute is equal to zero and equal to one otherwise; (b) the Euclidean distance is used to measure the dispersion between two different customers; (c) the way to measure dissimilarities between two customers with respect to the volume of product demand is based on the squared root of sum of squared differences of the average volumes for each pairwise of customers.

Mathematical Model

Given the features of each of the attributes and the requirements described above, we propose a combinatorial optimization model to represent the problem faced by the company.

Parameters

V: set of customers; K: set of segments, ($|K|=p$); d_{ij}: euclidean distance between customers i and j; S: set of SKU types; $A = (a_{is})$: matrix that relates customers with respect to SKU type, where a_{is} represents the volume that customer i demands from a certain type of SKU s; C: set of contract; c_i: type of contract of customer i ; E: set of strores; e_i: type of store of customer i.

Computed Parameters

q_{ij}^{SKU}: dissimilarity between customers i and j with respect to the SKU attribute; h_{ij}: Dissimilarity between customers i and j with respect to the type of contract, where $h_{ij}=0$ if $c_i=c_j$ and $h_{ij}=1$ otherwise; g_{ij}: dissimilarity between customers i and j with respect to the type of establishment, where $g_{ij}=0$ if $e_i=e_j$ and $g_{ij}=1$ otherwise.

Let Π be the collection of all p-partitions of V. The problem consist of finding a p-partition $X=(X_1,\ldots, X_p)$ of V as follows,

$$\min_{X\in\Pi} f(X) = \alpha_1 f_{disp}(X) + \alpha_2 f_{sku}(X) + \alpha_3 f_{cont}(X) + \alpha_4 f_{est}(X), \quad (1)$$

where the functions that measures dispersion, SKU volume, type of contract, and type of store are given by:

$$f_{disp}(X) = \sum_{k\in K} \sum_{i<j\in X_k} d_{ij}, \quad (2)$$

$$f_{sku}(X) = \sum_{k\in K} \sum_{i<j\in X_k} q_{ij}^{SKU}, \quad (3)$$

$$f_{cont}(X) = \sum_{k\in K} \sum_{i<j\in X_k} h_{ij}, \quad (4)$$

$$f_{estab}(X) = \sum_{k\in K} \sum_{i<j\in X_k} g_{ij}. \quad (5)$$

The equation (2) represents the dissimilarity with respect to dispersion between customers, (3) represents the dissimilarity between customers with respect to the SKU attribute where $q_{ij}^{SKU} = \sqrt{\sum_{s\in S}\left(\frac{a_{is}}{a_i^T} - \frac{a_{js}}{a_j^T}\right)^2}$ and a_i^T is the total volume of the customer i for all SKUs. The equations (4) and (5) represents the dissimilarity with respect to the type of contract and the type of store of the customers that composed the partition X.

Proposed Heuristics

We developed a heuristic method called IGACS (Iterated Greedy Algorithm for Customer Segmentation). Iterated greedy heuristics have been successfully applied to other combinatorial optimization problems [5]. This algorithm is composed of two main phases: construction and destruction. **Figure 1** shows the outline of this procedure.

In the first step we obtain an initial partition using a popular technique for clustering problems called p-means algorithm [6]. This is typically applied to problems where the "distance" between units is given by geographic measures. In our case, the "distance" between units considers the weighted sum of dissimilarity function of the four attributes.

In the p-means algorithm, p objects are selected to represent the initial centroids, the others n-p objects are assigned to its closest centroid, and iteratively recalculates those centroids based on the most centered customer (with respect to our concept on distance) of each segment.

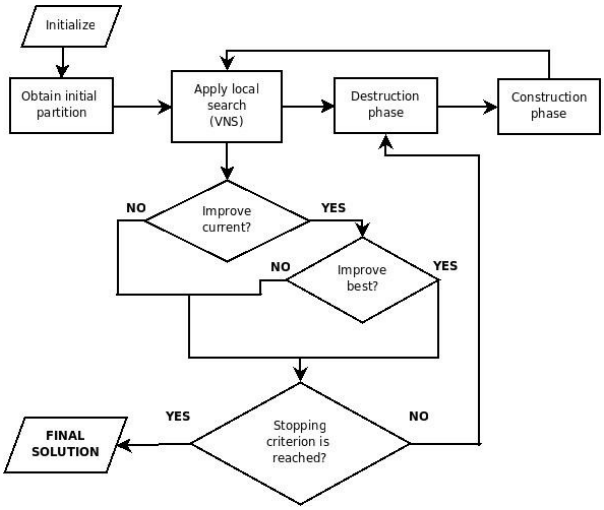

Fig. 1 Steps of the proposed IGACS.

It is well-known that the *p*-means algorithm strongly depends on the initial configuration of centroids. So in addition, we propose a Greedy Randomized Adaptive Search Procedure (GRASP) [7] based on a heuristic for the *p*-dispersion problem [8,9] for attempting to find a "disperse" initial configuration of centroids selected from a Restricted List of Candidates (LRC) built using a merit function based on a quality parameter called β. The goal of introducing a GRASP-construction is to find a good configuration of initial centers and obtains better quality solutions.

In the second step, we implemented a local search procedure to improve the initial partition found by the modified *p*-means algorithm. This local search, based on a variable neighborhood search procedure (VNS) [10], is composed of two simple local searches:

a) ***LS1:*** it consists of inserting one customer i from one segment $X_{t(i)}$ to another segment X_k, $i \in V$, $k \in K$ that reduces the dissimilarity of the partition,

b) ***LS2:*** swap two customers i and j, $i,j \in V$, from different segments.

Figure 2 and **Figure 3** show the moves of insertion and interchange of customers applied in the VNS, where $\varphi(i,k)$ and $\varphi(i,j)$ represent the functions that measures the benefit of the movement. If function $\varphi(i,k)$ is positive the insertion move is performed and the same for the function $\varphi(i,j)$ in the LS2.

In the next step (step 3), the best partition found by the VNS is destroyed by randomly removing d elements from the partition. After that, the solution is built (step 4) by adding one element (one by one) to the current partition, where each element is added if the dissimilarity increases as little as possible. Then the VNS is applied and the solution is updated if there is an improvement. After that, the destruction and construction phases (step 3 and 4) are applied again. This procedure finishes when a stopping criterion is reached.

$$\phi(i,k) = \sum_{j \in X_{t(i)}} f_{ij} - \sum_{q \in X_k} f_{iq};$$

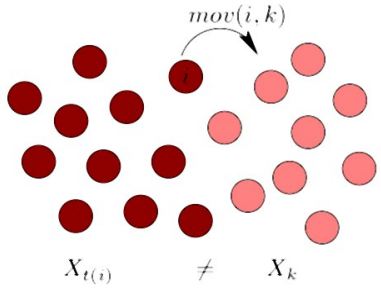

Fig. 2 Insertion move (LS1).

$$\phi(i,j) = \left(\sum_{q \in X_{t(i)}} f_{iq} - \sum_{r \in X_{t(j)}} f_{ir} \right) + \left(\sum_{r \in X_{t(j)}} f_{jr} - \sum_{q \in X_{t(i)}} f_{jq} \right)$$

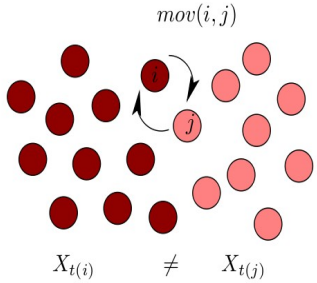

Fig. 3 Swap move (LS2).

Computational Results

The IGACS was implemented in C++ using an Ubuntu-Linux 9.04 operative system. The experiments are carried out on a Dell server with Intel Core(TM) 2 Quad 1 processor and 2.4 Ghz. CPU with 3.2 Gb. RAM. For the experiments we used two real-world instances, Inst1 and Inst2. The Inst1 instance is composed by n=8566 customers, 12 types of contracts, 32 types of stores and 233 SKUs. The Inst2 instance is about n=17332 customers, 12 types of contracts, 32 types of stores and 201 SKUs. **Figure 4** and **Figure 5** show the geographic location of the customers of instances Inst1 and Inst2 respectively. The number of segments is fixed to p={30, 40} for Inst1 and p={60, 70} for Inst2, and the weights of the alphas are fixed to $\alpha 1 = \alpha 2 = \alpha 3 = \alpha 4 =$ 0.25 for both instances.

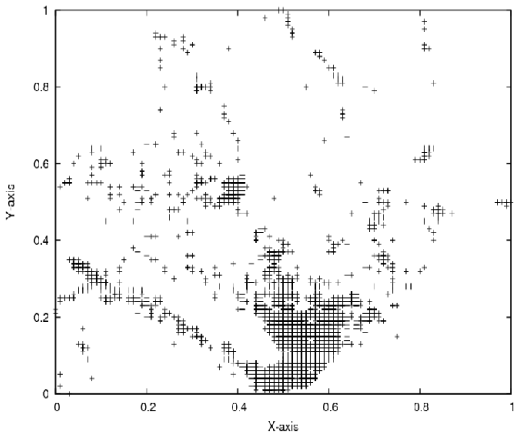

Fig. 4 **Geographic location of customers of the instance Inst1.**

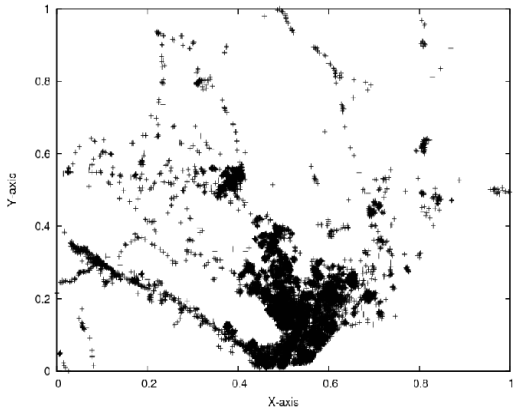

Fig. 5 **Geographic location of customers of the instance Inst2.**

For the evaluation of the GRASP procedure the p-means algorithm, used to get an initial partition, is executed 50 times. We tried different values of the quality parameter β ={0, 0.2, 0.4, 0.6, 0.8, 1}. When β = 0, the selection of centroids is totally greedy (centroids are the most disperse) and when β = 1 the selection is completely random. **Table 1** shows the objective function value of the best partition found after applying the p-means algorithm using different values of the quality parameter β to select the initial configuration of centroids.

Table 1: Comparison between the objective function value of the partitions found by the p-means algorithm and using different values of ¯ to select the initial centroids.

Instance	p	β					
		0	0.2	0.4	0.6	0.8	1
Inst1	30	0.01667	0.00945	0.00846	0.00887	0.00856	0.00743
	40	0.01291	0.00742	0.00611	0.00718	0.00617	0.00510
Inst2	60	0.00578	0.00490	0.00465	0.00440	0.00429	0.00314
	70	0.00515	0.00417	0.00383	0.00367	0.00382	0.00277

Figure 6 shows the initial centroids (square) found by GRASP when the value of β changes and the best final configuration of centroids (circles) obtained by the p-means algorithm. In this figure we can observe that, for the instances tested, better solutions are found when most of the centroids are located in the most dense area (where there are more customers) and the initial centroids found by GRASP cannot find better solutions if centroids are widely dispersed. As a conclusion, the random selection (β = 1) got the initial configuration of centroids that shows better solutions.

For the best solution obtained for p-means (in this case using β = 1) we applied the remainder steps of the proposed procedure to evaluate the improvement in the objective function and in the computation time required. **Table 2** shows the best objective values found by the p-means algorithm using a random selection of centroids (f(GRASP)), the best objective value found after applying the others steps of the proposed heuristics (f(IGACS)) and the improvement (percentage) obtained by the latter.

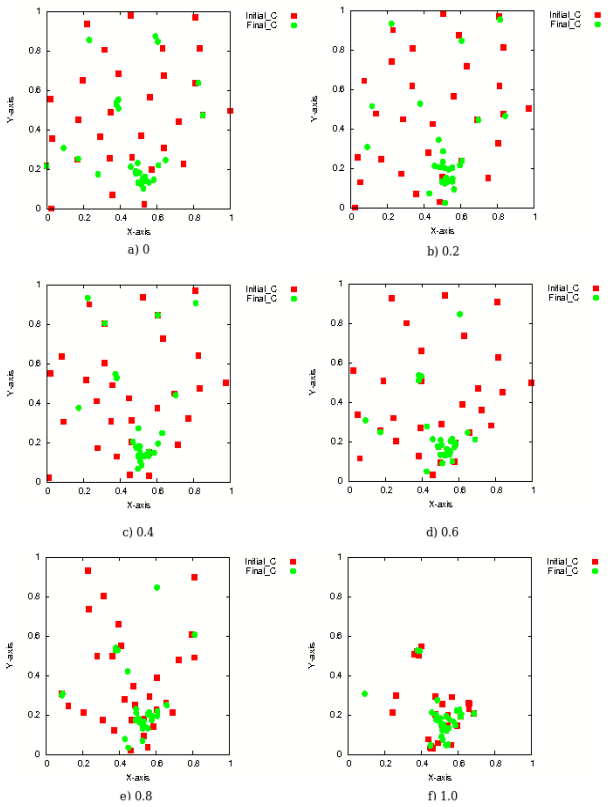

Fig. 6 Comparison between the selection of initial (squared) and final (circle) centroids using differents values of β and $\alpha 1 = \alpha 2 = \alpha 3 = \alpha 4 = 0.25$.

Table 2: Comparison between the objective function found by p-means algorithm and the objective function found by IGACS.

Instance	*p*	f(GRASP)	f(IGACS)	Improve %
Inst1	30	0.00743	0.00472	57.50
	40	0.00510	0.0033	54.54
Inst2	60	0.00314	0.00229	37.11
	70	0.00277	0.00195	42.05

In both instances the proposed approach improved the initial solution by more than 37% in less than 2968 seconds (50 minutes). **Table 3** shows the time required for each of the instances.

Table 3: Comparison between the objective function found by p-means algorithm and the objective function found by IGACS.

Instance	*p*	T_GRASP (seconds)	T_IGACS (seconds)	Total Time (seconds)
Inst1	30	172.10	754.43	926.53
	40	166.55	646.71	813.26
Inst2	٦٠	841.03	2127.27	2968.30
	70	833.95	2057.10	2891.05

Conclusions

This work addresses a real-world customer segmentation problem. We presented a mathematical model and developed an iterated greedy heuristic for this problem. We showed computacional results based on two real-world instances. Results showed that IGACS found significantly better solutions than those found by the *p*-means algorithm. An extension of this work, is the improvement of GRASP to get better results in the construction of the initial partition and reduce the computing time used by the IGACS to improve it. Another line of future work is to evaluate the solutions when the IGACS accepts certain solutions of poor quality as the work by Ruiz and Stützle [5].

Acknowledgements

This research has been supported by the Mexican National Council for Science and Technology (CONACYT) through a grant from the SEP-CONACYT 48499-Y project and a scholarship for graduate studies, and by the Universidad Autónoma de Nuevo León through the PAICYT CA1478-07 project. We also acknowledge the support of Fabián López (Grupo ARCA) for providing the information on the data sets used in this paper.

References

[1] E. A. Blackstone, A. J. Buck, S. Hakim and U. Spiegel, "Market segmentation in child adoption," *International Review of Law and Economics,* vol. 28, number 3, pp.~220–225, 2008.

[2] J. T. Bowen, "Market segmentation in hospitality research: No longer a sequential process," *International Journal of Contemporary Hospitality Management,* vol. 10, number 7, pp.~289–296, 1998.

[3] W. Xia, Z. Ping, W. Gao, and L. Jia, "Market segmentation based on customer satisfaction-loyalty links," *Frontiers of Business Research in China,* vol. 1, number 2, pp. 211–221, 2007.

[4] R. Caballero, M. Laguna, R. Martí, and J. Molina, "Multiobjective clustering with metaheuristic optmization technology". Technical Report, Department of Statistics and Operations Research, Universidad de Valencia, Valencia, Spain, 2006.

[5] R. Ruiz and T. Stützle, "A simple and effective iterated greedy algorithm for the permutation flowshop scheduling problem," *European Journal of Operational Research,* vol. 177, number 3, pp. 2033–2049, 2007.

[6] J. A. Hartigan and A. Wong, "A *k*-means clustering algorithm," J*ournal of Royal Statistical Society, Series C: Applied Statistics,* vol. 28, number 1, pp. 100–108, 1979.

[7] T. A, Feo and M. G. C. Resende, "Greedy randomized adaptive search procedures," *Journal of Global Optimization,* vol. 6, number 2, pp. 109–133, 1995.

[8] E. Erkut, "The discrete *p*-dispersion problem," *European Journal of Operationals Research,* vol. 46, number 1, pp. 48–60, 1990.

[9] E. Erkut, Y. Ürküsal, and O. Yenycerioglu, "A comparison of *p*-dispersion heuristics," C*omputers and Operation Research,* vol. 21, number 10, pp. 1103–1113, 1994.

[10] P. Hansen, N. Mladenovic, and J. A. Moreno, "Variable neighborhood search," *Revista Iberoamericana de Inteligencia Artificial,* vol. 19, pp. 77–92, 2003.

Information about (Co-)Author

Author: Diana L. Huerta-Muñoz (Graduate Student). Graduate Program in Systems Engineering, Universidad Autónoma de Nuevo León. AP111–F, Cd. Universitaria, San Nicolás de los Garza, Nuevo León, 66450, México. Tel. +52 81 1492-0383. E-mail: lucia@yalma.fime.uanl.mx

Co-author: Roger Z. Ríos-Mercado (Professor). Graduate Program in Systems Engineering. Universidad Autónoma de Nuevo León. AP111–F, Cd. Universitaria, San Nicolás de los Garza, Nuevo León, 66450, México. Tel. +52 81 1492-0383. E-mail: roger@yalma.fime.uanl.mx

Co-author: Rubén Ruiz García (Professor). Department of Applied Statistics, Operations Research and Quality. Universitat Politècnica de València. Camino de Vera S/N, 46022, Valencia, Spain, Tel. +34 96387-7007, x74946. E-mail: rruiz@eio.upv.es