Stores clustering using a data mining approach for distributing automotive spare-parts to reduce transportation costs

Mehrdad Kargari, Mohammad Mehdi Sepehri *

Department of Industrial Engineering, Tarbiat Modares University, Tehran, Iran

ABSTRACT

Clustering of retail stores in a distribution network with specific geographical limits plays an important and effective role in distribution and transportation costs reduction. In this paper, the relevant data and information for an established automotive spare-parts distribution and after-sales services company (ISACO) for a 3-year period have been analyzed. With respect to the diversity and lot size of the available information such as stores location, order, goods, transportation vehicles and road and traffic information, three effecting factors with specific weights have been defined for the similarity function: 1. Euclidean distance, 2. Lot size 3. Order concurrency. Based on these three factors, the similarity function has been examined through 5 steps using the Association Rules principles, where the clustering of the stores is performed using *k*-means algorithm and similar stores are allocated to the clusters. These steps include: 1. Similarity function based on the Euclidean distances, 2. Similarity function based on the order concurrency, 3. Similarity function based on the combination of the order concurrency and lot size, 4. Similarity function based on the combination of these three factors and 5. Improved similarity function. The above mentioned clustering operation for each 5 cases addressed in data mining have been carried out using R software and the improved combinational function has been chosen as the optimal clustering function. Then, trend of each retail store have been analyzed using the improved combinational function and along with determining the priority of the depot center establishment for every cluster, the appropriate distribution policies have been formulated for every cluster. The obtained results of this study indicate a significant cost reduction (32%) in automotive spare-parts distribution and transportation costs.

© 2011 Elsevier Ltd. All rights reserved.

1. Introduction and literature review

Customer clustering is one of the most prime and principal subjects in the Customer Relationship Management context. Actually, clustering is the process of breaking a great number of customers into several parts in a way that the stores that have been clustered in similar groups possess the same behavior. Clustering gives a holistic and high-level view of all customers databases and provides the business owners with the required authority to formulate different policies for each segment of customers. Ideally, every organization must get to know its customers but since this is not practically possible therefore, clustering makes it possible to categorize the similar customers in one segment. Similarity function differs for various industry and business types.

In this case, managing and recognition of these segments is much easier than the case of the individual customers. Clustering is applied in several ways. In some cases, considering the projected

share of profit, potential profit and definition of customer profitability a LTV model has been proposed and based on the present value, the potential value and level of customer loyalty have been segmented. Hyunseok, Taesoo, and Euiho (2004) in 2006 determined the individual value of customers considering the importance of store recognition for long-term relations, loyalty acquisition and more profitability for customer and then segmented the values according to the customer values and formulated the appropriate strategies for each segment (Kim, Jung, Suh, & Hwang, 2006).

Tsai and Chiu (2004) developed and extended a new market segmentation methodology based on the specific variables such as purchased items and the related income considering the previous transactions and then performed the market segmentation.

In some studies in the field of market segmentation methods such as *k*-means clustering model, FUZZY SOM and *k*-means have been used for the customer segmentation (Shina & Sohnb, 2004).

Also, in some other researches, customers have been homogenously segmented based on novelty, repetition and monetary value criteria and then the optimal marketing policies have been formulated (Jonkera, Piersmab, & Van den Poelc, 2004).

* Corresponding author.
  E-mail addresses: m_kargari@modares.ac.ir (M. Kargari), mehdi.sepehri@modares.ac.ir (M.M. Sepehri).

As mentioned earlier, customer segmentation is performed in several ways. At this company, level of the similarity of the customer behavior according to the data types and the available customer-related information depends on their location, customer cluster, customer city and the ordering size and time. Accordingly, in this paper, city, Euclidean distance, ordering time and order size have been used to derive the similarity function. First, two criteria of "customer city" and "order size" have been taken for customer clustering using the *k*-means algorithm. Then the Euclidean distance function has been used separately in the *k*-means algorithm for customer clustering. In order to improve the similarity function, the City, Euclidean distance, the required group of products, ordering time and order size have been combined and applied along with *k*-means algorithm for the customer clustering. When the best cluster is determined based on the clustering density criteria then, the optimum number of the clusters is determined based on the clustering quality assessment criteria.

Grouping a great number of customers with different characteristics using different customer clustering algorithms have been conducted by numerous researchers. These methods consider two major objectives of maximizing the intra-cluster similarity and maximizing the inter-cluster differences (Ajith, 2004; Jin, Zhou, & Mobasher, 2004a, 2004b; Menczer, Monge, & Street, 2002; Rich, 1999).

The customer clustering problem with the aim of minimizing the distribution costs have been studied by a great number of researchers though various methods (Gordeau, Gendreau, Laporte, Potvin, & Semet, 2002; Laporte, Gendreau, Potvin, & Semet, 2000; Salhi & Nagy, 1999).

Dondo and Gerda address the customer clustering and allocation to sales centers in their paper. They took into account the single depot and multi-depot problem. Distance between customers and distribution centers were considered as the clustering criterion (Dondo & Gerda, 2002).

Crainic et al. developed a two-step algorithm for inter-city product distribution with time windows. The first step involves the city and then customer clustering i.e. first, the products are distributed among cities and then among customers. The second step involves the routing problem within each city using the meta-heuristic methods (Feliu, Perboli, Tadei, & Vigo, 2007).

Feliu et al. developed a two-step algorithm for routing and spare parts distribution problem. In the first step, the customer clustering and distribution center determination is performed. In the second step, distribution channels are determined (Crainic, Ricciardi, & Storchi, 2007).

2. Store segmentation methodology

In this paper, first the stores are segmented and after the assessment and improvement of the distance function, the proper trend policies are formulated for each segment. Like other data mining projects, the first step deals with data preparation. Afterwards, OC (order concurrency) function has been defined based on the Association Rules principles for the assessment of the different store trends and the clustering has been performed by substituting this function with *k*-means algorithm. The improvement in the Euclidean distance function performance in *k*-means algorithm is shown using the cluster density assessment function. The number of the clusters was determined by the clustering quality assessment function. The corresponding steps are described as below.

2.1. Data preparation

In this step, the data and information included in the databases have been prepared according to the purposes and the applied

algorithms. The concepts and parameters used for the similarity function are defined as follows:

G	all of the company stores
T^C	store data center
t^c	a record from the table T^C; each record contains fields such as store code, store geographical location, order type and group, order, store city, ordering time, order size
C_i	Cluster *i*
g_{ia}	Store group "*a*" in Cluster *i*
m_{ia}	lot size of the requested group of goods "*a*" at Cluster *i*

Observation of the store trend requires the retrieving of all store information and the lot size of the requested group of goods by them for the past 3 years

Clustercode$_i$	code of Cluster *i*
Storeset$_i$	stores of Cluster *i*

Storeset$_i$ = {g_{ia}|$g_{ia} \in G$}

Volumeset$_i$ lot size of the group of goods at Cluster *i*

Volumeset$_i$ = {m_{ia}|$a = 1,2,\ldots,\|Storeset_i\|$}

GISplan$_i$ geographical location of the stores in Cluster *i*

Orderingtime$_i$ store ordering date and time at Cluster *i*

Thus, the ordered goods of stores at Cluster *i* can be defined as t_i^C and it can be saved in store's database (T^C)

$$T_i^C = \{Clustercode_i, Storeset_i, Volumeseti, Locationplan_i, Orderingtime_i\}$$

2.2. Store similarity function based on the order concurrency (OC)

In order to measure the similarity of the store trends first, the correlation of each store with other stores has been calculated based on the supportive concept in the association rules. To calculate the similarity, Eq. (1) has been applied:

$$S(g_{ia}, g_{ib}) = \frac{\|t^c \in T_i^C | t^c \quad \text{Contains } \{g_{ia}, g_{ib}\}\|}{\|T_i^C\|} \quad (1)$$

$\|T_i^C\|$ number of transactions of stores at Cluster *i*

$S(g_{ia})$ ratio of the transactions of store *a* to the total number of the stores at Cluster *i*

$S(g_{ib})$ ratio of the transactions of store *b* to the total number of the stores at Cluster *i*

$S(g_{ia}, g_{ib})$ ratio of the number of transactions including stores "*a*" and "*b*" at Cluster *i* to the total number of transactions at Cluster *i*. If these two Stores have ordered infrequently, then the value of $\gamma(g_{ia}, g_{ib})$ will be low

In OC function, the distance is determined based on the similarity of the product groups and it requires the consideration of the bilateral correlation of the stores. Therefore, the bilateral correlation of the stores at Cluster *i* (γ) is calculated using Eq. (3):

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \quad (2)$$

$$\gamma(g_{ia}, g_{ib}) = \frac{S\{g_{ia}, g_{ib}\}}{S(g_{ia}) + S(g_{ib}) - S(g_{ia}, g_{ib})} \quad (3)$$

The value of the $\gamma(g_{ia}, g_{ib})$ is between 0 and 1. If $g_{ia} = g_{ib}$ then, $\gamma(g_{ia}, g_{ib})$ = 1 i.e. stores "*a*" and "*b*" at Cluster *i* always order concurrently. If $\gamma(g_{ia}, g_{ib})$ = 0 then stores "*a*" and "*b*" never put an order concurrently. $\gamma(g_{ia}, g_{ib})$ denotes the similarity of the stores "*a*"