Uncertain Supply Chain Management 4 (2016) 55–60

Contents lists available at GrowingScience

Uncertain Supply Chain Management

homepage: www.GrowingScience.com/uscm

A new application of clustering for segmentation of banks' e-payment services based on profitability

Sorour Farokhi[a*], Babak Teimourpour[b], Fatemeh Shekarriz[c] and Maryam Masoudi[d]

[a]*Department of Industrial Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran*
[b]*Department of Industrial Engineering, Tarbiat Modaress University, Tehran, Iran*
[c]*Reactor Research School,Nuclear Science and Technology Research Institute, Tehran, Iran*
[d]*Abrar Institute of higher education, Damavand, Tehran, Iran*

C H R O N I C L E

Article history:
Received April 18, 2015
Received in revised format May 10, 2015
Accepted July 10 2015
Available online
July 15 2015

Keywords:
Clustering
K-Means
Kohonen
Point of Sales

A B S T R A C T

In today's extremely competitive environment, customers are the most important asset of any business and success of any organization depends on loyal customer retention. Therefore, it is important to learn more about different groups of customers and propose appropriate plans to take care of them. Customer segmentation is one of the most common methods to analyze customer's behavior. The market can be divided into several smaller homogeneous groups, which helps organizations deliver targeted marketing techniques and provides optimized techniques for resource allocation. This paper uses the information gathered from Point of Sales (POS) in one of Iranian private banks and using two methods of K-Means and Kohonen, customers are clustered into four segments to detect the most profitable customers.

1. Introduction

Target customer selection is one of the primary concepts in customer-based marketing and the objective of value making via the customers is to determine profitable or potentially profitable customers (Hosseini et al., 2009; Cheng & Chen, 2009; Mak et al., 2011). The capability to detect the profitable, loyal and long-term customers is the key success for customer-oriented firms. In order to access winning strategies, business owners must be looking for an appropriate strategy to determine the potential customers and to absorb them as much as possible. Some studies indicate that firms are aware of the role and importance of detecting customers who are valuable for the success of firm. In addition, the results are indicative of the fact that the strategies, which are based on locating and retaining appropriate customers will generate substantial value. Therefore, customers' segmentation is considered as an approach in marketing, which plays important role in customer relationship management (CRM). CRM is described the management skills in the organizational level achieved through a deep understanding, participating and managing the customers' requirements and it is based

* Corresponding author
E-mail address: s_farokhi@iau-tnb.ac.ir (S. Farokhi)

doi: 10.5267/j.uscm.2015.7.004

on the knowledge obtained from the customers in the direction of helping the organizational efficacy, productivity and profitability. Today, utilization of the CRM strategies plays important role as one of the primary motives behind several efforts made by organizations to build better value for their customers and have long-term revenue for them. The data extraction tool helps organizations necessary knowledge from customers' data under a CRM framework (Golmah & Mirhashemi, 2012). point of sales (POSs) plays essential role for detecting important customers. Profeta et al. (2012), for instance, determined detecting of origin of sales through POSs plays an important role in consumer decisions to purchase food. In this survey, we intend to use the information of POSs to detect important customers in banking industry in Iran. The study uses different clustering techniques to classify customers.

2. The proposed study

The proposed study of this paper uses Cross Industry Standard Process for Data Mining (CRISP-DM) (Chapman et al., 2000) for clustering customers. Fig. 1. Demonstrates the summary of the proposed study.

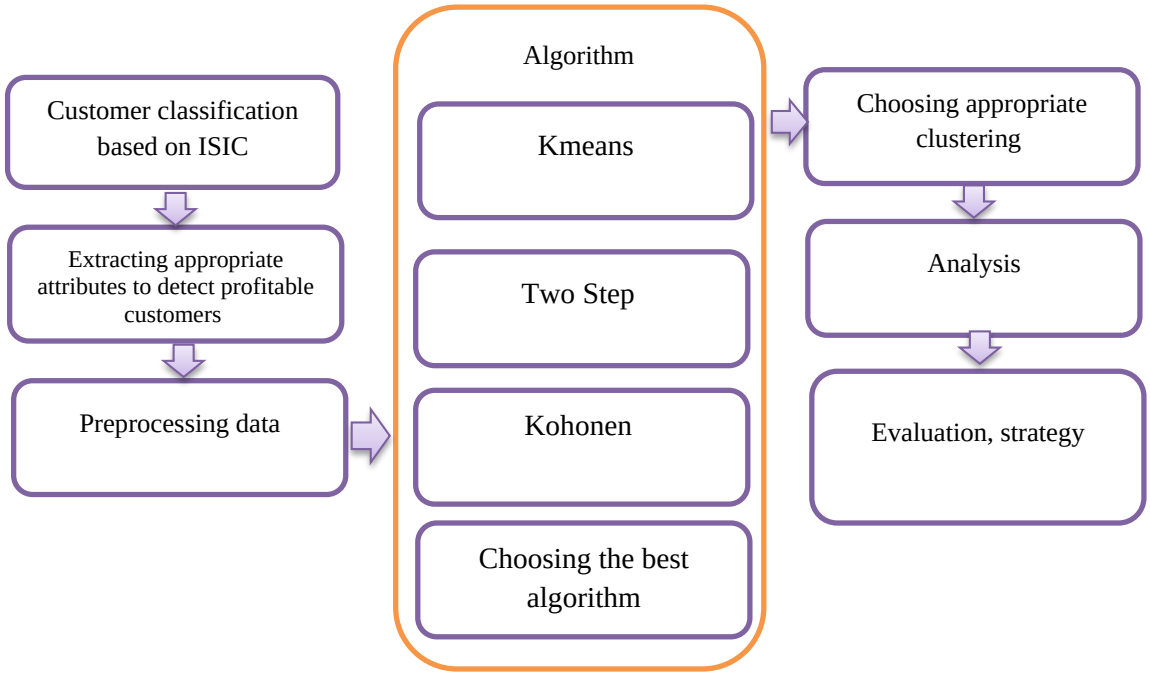

Fig. 1. The proposed clustering

The proposed study of this paper uses International Standard Industrial Classification (ISIC) to cluster different customers into four groups.

2.1 The K-means clustering algorithm

Clustering is a data-mining technique, which provides meaningful and informative clusters of objects with similar properties in an automatic mode (Garcia-Murillo & Annabi, 2002). In clustering, the main objective is to form different groups with similar characteristics. The purpose of executing K-means is to categorize samples into *k* clusters and *K* prime centers will be chosen randomly from the input data. Then the distance between each datum and each of the *K* centers are computed and they are assigned to the cluster with the least distance with its center. Finally, when the allocation of all the points to the *K* centers is accomplished, the mean of each cluster is calculated as the new center and the calculation of elements' distance and their allocation continues to new centers until there is no displacement in the clusters elements (Baradwaj & Pal, 2011).

2.2. Two step clustering

Two Step Cluster is a method for analyzing large datasets and the algorithm groups the observations in clusters, according to the approach criterion based on an agglomerative hierarchical clustering method (Şchiopu, 2010). Compared with classical techniques of cluster analysis, Two Step uses both continuous and categorical attributes. In addition, the method can automatically detect the optimal number of clusters. According to Şchiopu (2010), Two Step Cluster includes using the following steps:

- Pre-clustering;
- Solving a typical values (outliers) - optional
- Clustering.

In the pre-clustering step, it reads the data record one by one and makes a decision whether the present record can be considered to one of the previously formed clusters or it begins a new cluster, based on the distance criterion. The algorithm applies two kinds of distance measuring: Euclidian distance and log-likelihood distance. Pre-clustering procedure is constructed by forming a data structure called cluster feature (CF) tree, which includes the cluster centers. The CF tree includes of levels of nodes, each node having a number of entries where a leaf entry is considered as a final sub-cluster. For each record, beginning from the root node, the nearest child node is detected, recursively, descending along the CF tree. Once approaching a leaf node, the algorithm determines the closest leaf entry in the leaf node. If the record is around a threshold distance of the nearest leaf entry, then the record is accumulated into the leaf entry and the CF tree will be updated, otherwise, it builds a new value for the leaf node. If there is sufficient space in the leaf node to add another value, that leaf is split into two values and these values are distributed to one of the two leaves, based on the farthest pair as seeds and redistributing the other values according to the closeness criterion.

In the process of constructing the CF tree, the method applies an optional step, which leads to solve a typical outliers and considers records, which would not fit well into any other cluster. Before rebuilding the CF tree, the procedure looks for potential atypical values and separates them. After the CF tree is reconstructed, the procedure verifies whether these values may fit in the tree without increasing the tree size. Next, the values, which would not fit anywhere are considered as outliers. If the CF tree reaches the maximum permitted size, it is reconstructed according to the existing CF tree, by increasing the threshold distance, which yields a smaller CF and allows new input records. The clustering stage contains sub-clusters resulting from the pre-cluster step as input and groups them into the appropriate number of clusters. Since the number of sub-clusters is much smaller than the number of initial records, classical clustering methods can be applied accordingly. Two Step applies an agglomerative hierarchical approach, which detects the number of clusters. Hierarchical clustering method is associated with the process in which the clusters are merged, repeatedly until a single cluster may group all the records. The process begins by defining an initial cluster for each sub-cluster. Then, all clusters are compared and the one with the smallest distance is merged into one cluster. The process repeats until all clusters have are merged. Therefore, this is a simple technique to compare the solutions with various number of clusters (Şchiopu, 2010).

2.3. Kohonen Self-Organizing Feature Map

The Kohonen Self-Organizing Feature Map is a clustering and data visualization method based on a neural network viewpoint. As with other kinds of centroid-based clustering, the primary objective of SOM is to determine a set of centroids and to assign each object in the data set to the centroid, which includes the best approximation of that object. Like incremental K-means, data objects are considered one at a time and the nearest centroid is updated (Pang-Ning et al., 2006). Unlike K-means, the technique imposes a topographic ordering on the centroids and nearby centroids are also updated. The processing of points keeps updating until some predetermined limit is obtained. The final output of the

SOM method is a set of centroids that implicitly defines clusters. Each cluster consists of the points nearest to a specific centroid. SOM is a clustering technique that enforces neighborhood relationships on the resulting cluster centroids. Therefore, clusters, which are neighbors are more associated with one another than other clusters. Such relationships help the interpretation and visualization of the clustering results (Pang-Ning et al., 2006).

3. The results

We have used three methods of K-Means, Two step and Kohonen for clustering the data gathered from POSs and the preliminary results have indicated that K-means seems to perform better than other two methods. In other words, for the implementation of Two step and Kohonen, most data were located only in one cluster. Therefore, we use K-Means for further investigation using two attributes of Silhouette and Dunn. Silhouette is a method of interpretation and validation of clusters of data, which was first introduced by Rousseeuw (1987) and it provides a succinct graphical representation of how well each object is located within its cluster when the data are clustered via K-Means into *k* clusters. For each datum *i*, let *a(i)* be the average dissimilarity of *i* with other information within the same cluster. We now define the average dissimilarity of point *i* to a particular cluster *c* as the average distance from *i* to points in *c*. Moreover, let *b(i)* be the lowest average dissimilarity of *i* to any other cluster, where *i* is not a member. Now Silhouette id defined as follows,

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \tag{1}$$

where we have

$$s(i) = \begin{cases} 1 - \dfrac{a(i)}{b(i)}, & \text{if } a(i) < b(i) \\ 0, & \text{if } a(i) = b(i) \\ \dfrac{b(i)}{a(i)} - 1 & \text{if } a(i) > b(i) \end{cases} \tag{2}$$

It is clear from Eq. (2), $-1 \leq s(i) \leq 1$ and when $s(i)$ tends to 1 we need $a(i) \ll b(i)$. Since $a(i)$ is a measure of how dissimilar *i* is to its own cluster, smaller values are more desirable. Moreover, a large value of $b(i)$ represents a bad matched to its neighboring cluster. Therefore, when $s(i)$ tends to one it means that the datum is clustered, appropriately.

The implementation of Dunn's method (Dunn, 1973) is calculated as follows,

$$D = \min_{j=i+1,\ldots,n_c} \left\{ \min_{j=i+1,\ldots n_c} \left(\frac{d(n_i, c_j)}{\max\limits_{k=1,\ldots,n_c} \mathrm{diam}(c_k)} \right) \right\} \tag{3}$$

where $d(n_i, c_j)$ and $\mathrm{diam}(c_k)$ are calculated as follows,

$$\mathrm{d}(n_i, c_j) = \min_{x \in c_i, y \in c_j} d(x, y), \tag{4}$$

$$\mathrm{diam}(c_i) = \max_{x,y \in c_i} d(x, y). \tag{5}$$

Table 1 demonstrates the summary of the implementation of K-Means and Two step using Silhouette method. In addition, Fig. 1 shows the results of clustering using Dunn method.

Table 1
The summary of clustering using Silhouette attribute

Cluster	K-Means	Two step
2	0.9	0.6
3	0.9	0.5
4	0.9	0.7
5	0.8	0.7
6	0.8	0.7
7	0.8	0.5
8	0.8	0.5
9	0.8	0.5
10	0.8	0.5

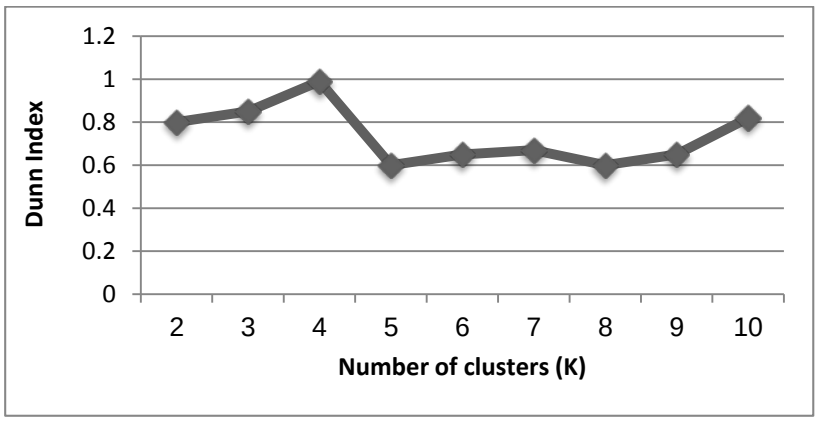

Fig. 1. The summary of the results of clusters versus Dunn

As explained earlier, the proposed study of this paper uses K-Means method to cluster different groups of customers and Table 2 demonstrates the summary of our findings. In addition, Table 3 shows the results of the distribution of the guilds in different segments after clustering.

Table 2
The summary of K-Means clustering technique

Cluster	Number	Percentage
1	40000	22.32%
2	12786	7.13%
3	15699	8.76%
4	110733	61.79%
Total	179218	%100

Table 3
The results of the distribution of the guilds in different segments after clustering

	Cluster			
Criteria	1	2	3	4
1	58808305	1819367590	86306341	11995486
2	242.3426	100	1747	47
3	13174512	1906707823	12387861	397431
4	21711	10518	153774	3818
5	130167	22369066	112599	2513
6	231925	33566002	218077	6996
7	98469	11182454	234252	-16699

As we can observe from the results of Table 2 and Table 3, cluster 2 represents high range of profitability with market share of 7.13%. The customers in this section are considered as precious asset and we need to take the necessary action for customer retention in this cluster. On the other side, customers located in cluster 4 are not profitable and they are accounted as 69% of the population of the survey. Therefore, we need to leave this cluster and increase the population of other clusters.

4. Discussion and conclusion

The implementation of K-Means clustering in this survey has provided a useful technique to detect valuable customers and categorize them into four groups. The first cluster consists of different sectors including medical equipment, laboratories and hospitals, cinemas and entertainment centers, cosmetic and health, financial services and money exchange agencies, fuel stations, automobiles, furniture and home appliances. This group is a profitable cluster and customer retention plans must be implemented. Food industry is the second cluster in our survey and customers must be taken care of appropriately. The third cluster is associated with clothing, shoes, household appliances store, flower shop, clothing stores and sporting goods, airports and air terminals, equipment and facilities. Despite the fact that these customers in this group have had relatively large numbers of POSs, they are as active as the customers in the previous clusters. Finally, tax consultant, real estate, software providers, telephone equipment sales and communication tools, store media products, car wash, auto repair shop, bookstores are considered as non-profitable cluster and bank must give up this sector.

Acknowledgement

The authors would like to thank the anonymous referees for constructive comments on earlier version of this paper.

References

Baradwaj, B. K., & Pal, S. (2011). Mining Educational Data to Analyze Students' Performance. *International Journal of Advanced Computer Science & Applications*, *2*(6), 63-69.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Sherer, C., & Wirth, R. (2000). Cross industry standard process for data mining (CRISP-DM) 1.0.

Dunn, J. C. (1973). A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*, *3*, 32-57

Garcia-Murillo, M., & Annabi, H. (2002). Customer knowledge management. *Journal of the Operational Research society*, *53*(8), 875-884.

Golmah, V., & Mirhashemi, G. (2012). Implementing a data mining solution to customer segmentation for decayable products-A case study for a textile firm. *International Journal of Database Theory & Application*, *5*(3), 73-89.

Pang-Ning, T., Steinbach, M., & Kumar, V. (2006). Introduction to data mining. In *Library of Congress* (p. 74).

Profeta, A., Balling, R., & Roosen, J. (2012). The relevance of origin information at the point of sale. *Food Quality and Preference*, *26*(1), 1-11.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53-65.

Şchiopu, D. (2010). Applying TwoStep cluster analysis for identifying bank customers' profile. *Buletinul*, *62*, 66-75.