Customer Clustering Using K-Means Clustering: Supporting Customer Relationship Management System

by Silvia Rostianingsih

Submission date: 17-Nov-2021 08:48AM (UTC+0700)
Submission ID: 1705116527
File name: ESTIC2021-TI-11.pdf (623.84K)
Word count: 2687
Character count: 13220

Customer Clustering Using K-Means Clustering: Supporting Customer Relationship Management System

Silvia Rostianingsih, Alexander Setiawan, Michael Boentarman[1, 2, 3]

[1, 2, 3]*Informatics Department, Petra Christian University*
Siwalankerto 121 – 131 Surabaya 60236

[1] *silvia@petra.ac.id*, [2] *alexander@petra.ac.id*, [3]*michaelboentarman@gmail.com*

Abstract. PT. X is a consumer good company with a variety of products such as consumption salt, packaged cooking oil, soap, sugar and so on. Currently PT. X has distributors (customer) whose role is to market products directly to consumers. With the aim of increasing sales from each distributor, PT. X provides special treatment. Special treatment is given manually for certain distributors. This is felt to be less than optimal, because the special treatment given is not right on target for each distributor who has the most sold superior product types. Therefore, we need a distributor clustering system using the K-Means method referring to sales history data. The proposed system also has features that can help distributors to place orders independently, confirm payments, submit returns, submit complaints and exchange reward points for prizes that can be settled by the company. The results of the system have been tested for system acceptance using a questionnaire that was distributed to 5 users. The results of the system acceptance test show the distributor's satisfaction level with the system. Distributors feel the convenience and benefits of the system.

INTRODUCTION

PT. X is a consumer good company with various products such as consumption salt, packaged cooking oil, soap, sugar and so on. Head office of PT. X is located at Jakarta and has two branch offices in Sidoarjo and Gresik, East Java. As a national company, PT. X has distributors spread across every city throughout Indonesia. The distributor functions as a distributor to sellers in the market, so the role of the distributor has an important role in product sales.

A system is proposed in order to perform clustering using the K-Means method [1]. The model used for clustering is Recency, Frequency, and Monetary (RFM). The application of RFM and K-Means models has been applied to microfinance institutions to cluster customers [2]. The results of the study succeeded in dividing the customer model into three clusters. The first cluster consists of 100 customers. RFM score 112-255 including the occasional customer group. The second cluster consists of 69 customers, has an RFM score 441-544 and belongs to the superstar group. The third cluster consists of 79 customers, has an RFM score 313-435 and belongs to the typical customer group.

Others research also shown that K-Means Clustering and its improvement with RFM can be used to form customer clustering in order to obtain the loyalty of customer. Sutresno discovers five clusters with different strategies according to the RFM score obtained [3]. Anitha uses Silhouette Analysis for Recency vs. Monetary using K-Means for varying number of Clusters [4]. PT. Herbal Penawar Alwahidah Indonesia Pekanbaru generates four cluster from 551 customer [5]. PT Gramedia Asri Media produce two type of clustering [6]. Customer segmentation based on the RFM produce two optimal clusters, occasional customers and dormant customers. Customer segmentation based on the number of types of books purchased produce 3 optimal cluster, namely low, medium, and high. While Syarif [7] using RFM with Fuzzy C-Means with the purpose of creating an optimal cluster and produce three customer cluster: gold, silver, bronze, and iron. Monalisa is used LRFM, which combine length of the relation between distributor and company with RFM.

In this study, RFM and K-Means models were applied to cluster distributors based on transaction data. The Frequency pattern analyzes the number of product purchases by distributors, the recency pattern analyzes the length of time for distributors to repurchase, while the monetary pattern analyzes based on the nominal amount of distributor spending on the company. The data analyzed is distributor transaction data in the 2019 period. The results of customer (distributor) clustering analysis using the K-Means method are used as a reference for giving special rewards and discounts for distributors. Rewards are calculated based on the total price of orders made by the distributor. Reward points can be exchanged for various products or vouchers that have been provided by the company. This supports Customer Relationship Management (CRM) System that they have. K-Means is chosen since it can analyze large amounts of data and it can determine the number of clusters needed.

STUDY LITERATURE

K-Means clustering is "one of the data methods" non-hierarchical clustering that attempts to partition existing data into one or more forms cluster/group [1]. This method partitions data into clusters/groups so that data that have the same characteristics are grouped into the same cluster and data that has different characteristics are grouped into other groups.

The basic K-Means algorithm is as follows [9]:

1. Determine the value of k as the number of clusters you want to form.
2. Select the initial K centroid (cluster center point) randomly. In determining the n center of the initial cluster, a random number is selected which represents the sequence of data inputs. The initial center of the cluster is obtained from the data itself, not by determining a new point, namely by randomizing the initial center of the data.
3. Calculate the distance of each data to each centroid. To measure the distance between data with the center of the cluster used Euclidian distance. Data distance calculation algorithm with the cluster center. The steps are:
   - Take the data value and the cluster center value.
   - Calculate Euclidian distance data with each cluster center. Euclidian Distance is the distance obtained from the calculation between all N data with K centroid where will obtain the level of proximity to the closest class to the data population. This Euclidian distance is useful for marking the existence of similarities between each cluster with the minimum distance and has a higher equation. Equality from the Euclidian matrix between the point and the point of the shortest distance as in equation 1.

$$d(x,y) = |x - y| = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (1)$$

4. Each data selects the nearest centroid. Distance calculation results will be carried out comparison and selected the closest distance between the data and the center of the cluster, this distance indicates that the data are in one group with a closest cluster center.
5. Determine the position of the new centroids by calculating the average value of the data located on the same centroid. The new cluster center is used for perform the next iteration, if the results obtained have not converged. Process the iteration will stop if it has met the maximum iteration entered by the user or the results achieved have converged (the new cluster center is the same as the old cluster center). If the position of the new centroids with the old centroids is not the same, then repeated (iteration).

RFM analysis is an approach to determine the nature of the customer in making a purchase of a product. This approach is very popular especially in retail and industrial companies. The RFM model is a model for determining consumer clustering based on Recency, Frequency and Monetary [10]. Recency is the time span (in units of days, months, years) from the last transaction made by consumers to date. Frequency is the total number of transactions or the average number of transactions in one period. Monetary is the average amount of costs incurred by consumers in a unit of time.

DESIGN SYSTEM

The analysis steps using the K-Means method begin with collecting distributor transaction history data [8]. Next in the process of analyzing the K-Means method is to determine the value of K for the centroid, where the

determination of K will be done randomly by the system. After the K value is obtained, a process is carried out to calculate the distance of each distributor transaction history data criterion to each centroid [7]. The process of measuring the distance between the data and the center of the cluster used Euclidian distance.

The next step is to calculate the average of each centroid to determine the new centroid. The new cluster center is used to carry out the next iteration, if the results obtained have not converged. The iteration process will stop when it meets the maximum iterations entered by the User or the results achieved have converged (the new cluster center is the same as the old cluster center). If the position of the new centroids with the old centroids is not the same, then the process is repeated until the maximum iteration. Figure 1 illustrates the flow of the distributor clustering analysis process using the K-Means method.

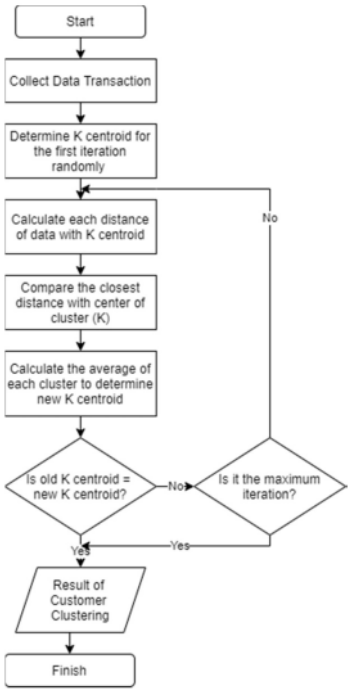

FIGURE 1. Algorithm of K-Means

IMPLEMENTATION

The calculation bellow only showing on the first 20 samples in 2019. The table of the purchase amount and the frequency can be seen in Table 1. The number will represent a distributor, the value of the purchase amount affected by recency, and frequency is the number of items purchased by the distributor.

TABLE 1. The purchase amount and the frequency

Number	Purchase Amount (Affected by Recency)	Frequency
1	99,466	9
2	284,331,164	44
3	2,163,942,618	91
4	3,192,926,795	127
5	211,188,473	123
6	205,338,416	35
7	304,840,740	46
8	297,356,106	55

Number	Purchase Amount (Affected by Recency)	Frequency
9	8,584,270,060	180
10	2,606,838,085	111
11	316,789,570	61
12	1,449,191,905	98
13	184,792,037	45
14	358,166,740	53
15	15,514,521	6
16	168,720,411	50
17	646,438	3
18	2,982,040	26
19	626,849	4
20	27,425	8

After getting the value of the number of purchases and the frequency, the next step is to calculate the distance between each distributor and each centroid (the distance calculation for each class is not shown). The number of centroids to be used is six because the desired number of classes is 6, namely: Superstar, Golden, Typical, Occasional, Everyday, and Dormant. The closest distance to the centroid will determine the class of the distributor (class determined by looking at the nearest centroid is shown). The centroid point distance table can be seen in Table 2.

TABLE 2. The centroid point distance

Number	Purchase Amount (Affected by Recency)	Frequency	Classes
1	99,466	9	5
2	284,331,164	44	2
3	2,163,942,618	91	1
4	3,192,926,795	127	1
5	211,188,473	123	2
6	205,338,416	35	2
7	304,840,740	46	2
8	297,356,106	55	2
9	8,584,270,060	180	1
10	2,606,838,085	111	1
11	316,789,570	61	2
12	1,449,191,905	98	1
13	184,792,037	45	2
14	358,166,740	53	2
15	15,514,521	6	3
16	168,720,411	50	2
17	646,438	3	4
18	2,982,040	26	3
19	626,849	4	4
20	27,425	8	6

Furthermore, by looking at Table 3, the number of purchases and the frequency in each class will be averaged so as to produce a new centroid. The new centroid point table can be seen in Table 3.

TABLE 3. The centroid point distance

Classes	Purchase Amount (Affected by Recency)	Frequency
1	3,599,433,893	121
2	259,058,184	57
3	9,248,281	16

Classes	Purchase Amount (Affected by Recency)	Frequency
4	636,644	4
5	99,466	9
6	27,425	8

The previous step will be repeated (iteration) until the result of the previous iteration class and the current iteration do not change. The following is the final result of customer clustering after six iterations. The iteration stops at the 6th iteration, because the 5th and 6th iterations produce the same class. The final result table of distributor clustering can be seen in Table 4.

TABLE 4. The final distributor clustering

Number	Purchase Amount (Affected by Recency)	Frequency	Classes
1	99,466	9	6
2	284,331,164	44	3
3	2,163,942,618	91	2
4	3,192,926,795	127	2
5	211,188,473	123	3
6	205,338,416	35	3
7	304,840,740	46	3
8	297,356,106	55	3
9	8,584,270,060	180	1
10	2,606,838,085	111	2
11	316,789,570	61	3
12	1,449,191,905	98	2
13	184,792,037	45	3
14	358,166,740	53	3
15	15,514,521	6	4
16	168,720,411	50	3
17	646,438	3	6
18	2,982,040	26	5
19	626,849	4	6
20	27,425	8	6

After the distributor is clustered, the class will be labeled with a name. The range of the number of purchases and the frequency of the class will be displayed. The table for giving levels to classes can be seen in Table 5.

TABLE 5. The label for each class

Classes	Purchase Amount (Affected by Recency)	Frequency	Label
1	8,584,270,060 - 8,584,270,060	180 - 180	Superstar
2	1,449,191,905 – 3,192,926,795	91 – 127	Golden
3	168,720,411 – 168,720,411	35 -123	Typical
4	15,514,521 – 15,514,521	6 – 6	Occasional
5	2,982,040 – 2,982,040	26 - 26	Everyday
6	27,425 - 646,438	3 - 9	Dormant

Admin from system is able to view the K-Means result. He can confirm or adjust the recommendation of distributor clustering. While on the distributor dashboard page, they can see their membership status information (Fig. 2). On the quote page, distributors can view the quote data based on distributor clustering that has been entered by the sales division. The display of the Price Offer page can be seen in Fig. 3.

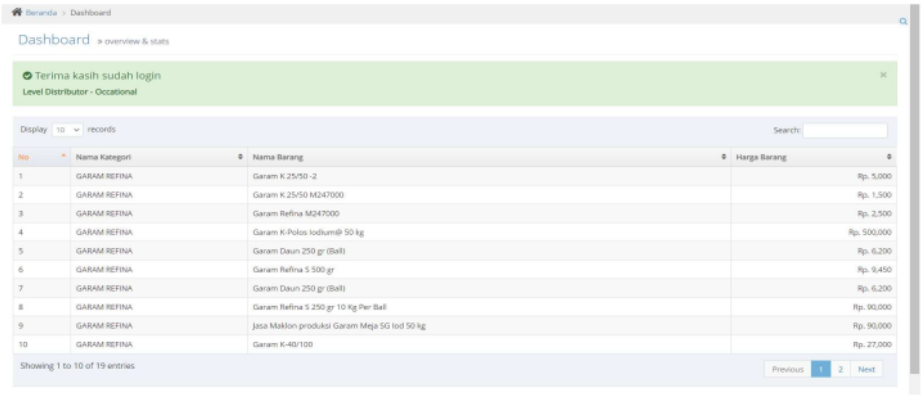

FIGURE 2. Dashboard of customer

FIGURE 3. Promotion for each clustered of distributor

CONCLUSION

Based on the results of implementation, it can be concluded that the results of this study has succeeded in clustering distributor as six cluster to determine distributor level status with K-Means. For data in year 2019, there are 509 active customers which are divided into six clusters as follows superstar cluster consists of six distributors, golden cluster consists of nine distributors, typical cluster consists of 19 distributors, occasional cluster consists of 12 distributors, everyday cluster consists of 22 distributors, and dormant cluster consists of 441 distributors. The last cluster is the default cluster for average distributor. It also has succeeded in improve the CRM system to support the company's activities. The company could give special promotion based on the cluster.

REFERENCES

1. P. Arora, Deepali, S. Varshney, "Analysis of K-Means and k-medoids algorithm for big data", Procedia Computer Science 78 (Elsevier, 2016), pp. 507-512.
2. T. Hardiani, S. Sulistyo, R. Hartanto, "Segmentasi nasabah tabungan menggunakan model RFM (Recency, Frequency,Monetary) dan K-Means pada Lembaga Keuangan Mikro", Seminar Nasional Teknologi Informasi dan Komunikasi Terapan (2015), pp. 463.
3. S. A. Sutresno, A. Iriani, and E. Sediyono, "Metode K-Means clustering dengan atribut RFM untuk mempertahankan pelanggan", Jurnal Teknik Informatika dan Sistem Informasi, 4 (2018), pp. 433–440.
4. P. Anitha and M. M. Patil, "RFM model for customer purchase behavior using K-Means algorithm", Journal of King Saud University - Computer and Information Sciences (2019).
5. F. Hadi, M. Mustakim, D. O. Rahmadia, F. H. Nugraha, N. P. Bulan, and S. Monalisa, "Penerapan K-Means clustering berdasarkan RFM Mofek sebagai pemetaan dan pendukung strategi pengelolaan pelanggan (studi

kasus: PT. Herbal Penawar Alwahidah Indonesia Pekanbaru)", Jurnal Sains dan Teknologi Industri, 15(1) (2017), pp. 69–76.
6. C.D. Rumiarti and I. Budi, "Segmentasi pelanggan pada Customer Relationship Management di perusahaan ritel PT Gramedia Asri Media", Jurnal Sistem Informasi, 13 (1), (2017).
7. S. Monalisa, "Klasterisasi customer lifetime value dengan model LRFM menggunakan algoritma K-Means", Jurnal Teknologi Informasi dan Ilmu Komputer, 5(2) (2018), pp. 247.
8. N. R. Syarif, "Implementasi algoritma Fuzzy C-Means dan Metode Recency Frequency Monetary (RFM) Pada Aplikasi Data Mining Untuk Pengelompokan Pelanggan, Sebatik 22(2) (2018), pp. 88-94.
9. P. Tan, *Introduction to Data Mining* (Pearson, 2nd Edition, 2018).
10. A. M. Hughes, *Strategic database marketing* (Probus Publishing Company, 1994).

Customer Clustering Using K-Means Clustering: Supporting Customer Relationship Management System

ORIGINALITY REPORT

18%	10%	13%	5%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Rina Harimurti, Ekohariadi Ekohariadi, Munoto Munoto, I. G. P. Asto Buditjahjanto. "Integrating k-means clustering into automatic programming assessment tool for student performance analysis", Indonesian Journal of Electrical Engineering and Computer Science, 2021 Publication	4%
2	Submitted to Universitas Amikom Student Paper	2%
3	repository.unpak.ac.id Internet Source	2%
4	Communications in Computer and Information Science, 2015. Publication	2%
5	www.neliti.com Internet Source	1%
6	Lisna Zahrotun. "Implementation of data mining technique for customer relationship	1%

management (CRM) on online shop tokodiapers.com with fuzzy c-means clustering", 2017 2nd International conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE), 2017
Publication

7	Submitted to Konsorsium Turnitin Relawan Jurnal Indonesia Student Paper	1 %
8	A A Aldino, D Darwis, A T Prastowo, C Sujana. "Implementation of K-Means Algorithm for Clustering Corn Planting Feasibility Area in South Lampung Regency", Journal of Physics: Conference Series, 2021 Publication	1 %
9	Submitted to Udayana University Student Paper	1 %
10	P. Anitha, Malini M. Patil. "RFM model for customer purchase behavior using K-Means algorithm", Journal of King Saud University - Computer and Information Sciences, 2019 Publication	1 %
11	Submitted to Universitas Dian Nuswantoro Student Paper	1 %
12	sersc.org Internet Source	1 %

13	worldwidescience.org Internet Source	<1%
14	core.ac.uk Internet Source	<1%
15	ilkogretim-online.org Internet Source	<1%
16	journal.maranatha.edu Internet Source	<1%

Exclude quotes	On	Exclude matches	< 5 words
Exclude bibliography	On		