Clustering optimization in RFM analysis based on k-means

Rendra Gustriansyah, Nazori Suhandi, Fery Antony
Faculty of Computer Science, Universitas Indo Global Mandiri, Indonesia

Article Info

Article history:

Received Jul 12, 2019
Revised Sep 15, 2019
Accepted Sep 29, 2019

Keywords:

Clustering
K-Means
Number of clusters
RFM
Validity index

ABSTRACT

RFM stands for Recency, Frequency, and Monetary. RFM is a simple but effective method that can be applied to market segmentation. RFM analysis is used to analyze customer's behavior which consists of how recently the customers have purchased (recency), how often customer's purchases (frequency), and how much money customers spend (monetary). In this study, RFM analysis has been used for product segmentation is to be arrayed in terms of recent sales (R), frequent sales (F), and the total money spent (M) using the data mining method. This study has proposed a new procedure for RFM analysis (in product segmentation) using the k-Means method and eight indexes of validity to determine the optimal number of clusters namely Elbow Method, Silhouette Index, Calinski-Harabasz Index, Davies-Bouldin Index, Ratkowski Index, Hubert Index, Ball-Hall Index, and Krzanowski-Lai Index, which can improve the objectivity and similarity of data in product segmentation so that it can improve the accuracy of the stock management process. The evaluation results showed that the optimal number of clusters for the k-Means method applied in the RFM analysis consists of three clusters (segmentation) with a variance value of 0.19113.

Corresponding Author:

Rendra Gustriansyah,
Faculty of Computer Science,
Universitas Indo Global Mandiri,
Jalan Jenderal Sudirman No. 629, Palembang, 30129, Indonesia.
Email: rendra@uigm.ac.id

1. INTRODUCTION

Clustering or data segmentation is a process of grouping (partitioning) large data sets into groups (partitions) according to their similarities. When the number of transactions becomes larger, the process of managing product databases for stock management is not an easy task. This problem can be overcome by a better approach by using the data mining method needed to segment all products into the right number of clusters according to some of their similarities. The values of various groups can then be estimated and evaluated to provide informed decisions that are useful for management in making use of resources rationally.

One simple but effective model that can be applied to product segmentation based on data attribute similarity by checking when (recency), how often (frequency), and money spent (monetary) in certain items or services is the recency (R), frequency (F), and monetary (M) model [1, 2]. This study will cluster the product data using one of the data mining methods, namely the k-Means method [3, 4] which will be proposed for RFM analysis. The use of the k-Means method for product clustering based on RFM values is expected to have better accuracy compared to manual product clustering [5-7].

Meanwhile, the determination of the optimal number of k clusters in the k-Means method will be evaluated using eight validity indices namely the Elbow Method, Silhouette Index, Calinski-Harabasz Index, Davies-Bouldin Index, Ratkowski Index, Hubert Index, Ball-Hall Index, and Krzanowski Index -Lai, which is expected to improve objectivity and accuracy in product segmentation compared to using only one method [1, 2, 8], and can simplify the stock management process [9].

2. RESEARCH METHOD

2.1. RFM Analysis

Stone and Bob (1989) first proposed the idea of the RFM method [10]. RFM is a simple but effective method that can be applied to market segmentation [10]. Hughes defines RFM analysis using information about consumer purchasing behavior in the past [11]. Recency (R) shows the period from the last purchase transaction to now. Frequency (F) is the number of purchase transactions made by customers. Monetary (M) is the total money spent by customers in a certain period of time [10].

The philosophy of RFM analysis is that products are to be arrayed in terms of recent sales, frequent sales, and the total money spent [12]. The real data set with sales dates converted to a values 1 to 5 depending on the date of sale. Therefore, the value of 5 is assigned to the top 20% of the data set in terms of the latest sales date. The value of 4 is given for the next 20% of the data set and so on, while the value of 1 refers to the oldest sales date.

For frequency, the number of transactions in a certain time period range is sorted, such as the number of transactions per month, in descending order [2, 13]. As many as 20% of the top data from the data set is given by a value of 5. The next 20% of the data set is given a value of 4 and so on, so that all the real data of the number of transactions are converted to values 1 to 5 [2].

For monetary, the average amount of money spent per month or year for all transactions is sorted in descending order [2]. As many as 20% of the top data from the data set are given a value of 5. 20% of the next data are given values 4 and so [2], so that all real data with the amount of money is transformed into values 1 to 5. Finally, all values of R, F, and M are combined to rank each product [10].

In addition, the concept of segmentation in RFM analysis will be improved to be more objective and accurate with the clustering approach using k-Means methods, so that the clusters that will be formed have the optimum data similarity. This can make the determination of the number of clusters and the interval of datasets for each cluster to be more quality and precise (not necessarily divided into 5 clusters or 20%, such as the default segmentation in RFM analysis).

2.2. K-Means Method

K-Means [14] is one of the non-hierarchical clustering data methods that partition data n into cluster k, so that the resulting intra-cluster similarity is high (minimal within-clusters sum of squares), while inter-cluster similarity is low (maximum between-clusters sum of square). K-Means is one of the most popular clustering methods, because of the simplicity of the algorithm and the speed of selecting the cluster center (centroid).

The k-Means method often applies the Euclidean distance formula to determine the similarity of data in a cluster iteratively.

Data clustering steps using the k-Means method can be done by:

a) Determine the number of clusters k;
b) Initialize k values as cluster centers (centroids) randomly;
c) Group each data into the closest cluster. The proximity of two data is calculated using Euclidean distance;
d) Recompute each centroid by computing the mean of all centroid data with current cluster members;
e) Re-clustering each data (back to step 3) using all new centroids until all centroids do not change anymore;
f) If the centroid has not changed again, the clustering process is complete.

One of the main problems of the k-Means method is how to determine the optimal number of clusters k. Research by Subbalakshmi et al. [15] have shown that the accuracy of the k-Means method can be higher, if appropriate in selecting the initial value and number of clusters [2, 13].

There are various ways that can be used to estimate the optimal number of clusters k. In this study, the optimal number of cluster k will be measured using the Elbow Method, Silhouette Index, Calinski-Harabasz Index, Davies-Bouldin Index, Ratkowski Index, Hubert Index, Ball-Hall Index, and Krzanowski-Lai Index.

2.3. Validity Index for Determining the Optimal Number of Clusters in the k-Means Method

a) *Elbow Method*: The Elbow Method (EM) [16] is a method used to determine the optimal number of clusters, by looking at the percentage of the comparison between the number of clusters that will form an angle on the curve. If the value of the first cluster with the value of the second cluster forms an angle (elbow) on the curve or value has the largest decrease, the cluster value is the best cluster value. The best number of clusters 'k' will be selected at that vertex (turning point). This method is a visual method that looks at the total intra-cluster variation or the total Within-Clusters Sum of Squares (WSS) as a function of the number of clusters. The greater the number of clusters k, the WSS value will be smaller or vice versa. The WSS formula is as follows:

$$WSS = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2 \qquad (1)$$

Where *k* = the number of clusters, *n* = the number of objects, x_i = i^{th} element in the cluster, and c_j = the centroid of j^{th} cluster.

b) *Silhouette Index*: The Silhouette Index (SI) value is used to measure how well the cluster is at a certain point [17]. Rousseeuw [18] proposed an approach that calculates the maximum index value. Silhouette refers to the method of interpretation and consistency validation in the data cluster. Silhouette functions can be calculated using Equation:

$$SI = \bar{s} = \frac{1}{n} \sum_{i=1}^{n} s(i) \qquad (2)$$

Where,

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \qquad (3)$$

$$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, i \neq j} d(i,j) \qquad (4)$$

$$b(i) = \min_{i \neq j} \frac{1}{|C_j|} \sum_{j \in C_j} d(i,j) \qquad (5)$$

With *j* is another object in one cluster $|C_i|$, *d*(*i*,*j*) is the euclidean distance between objects *i* with *j* in cluster C_i, *b*(*i*) is the distance of the average object*i* with all objects in another cluster, and the overall *SI* is the average of *s*(*i*) over all objects.

The value of the silhouette index is between -1 and 1. If one point has a silhouette index close to 1, then it is a good clustering. If the silhouette index close to -1 indicates a poor clustering (misclassification). Meanwhile, if the silhouette index close to 0 indicates an intermediate case (not good).

c) *Calinski-Harabasz Index*: Calinski-Harabasz Index (CHI) [19] evaluates cluster validity based on calculations of Between-Clusters Sum of Square (BSS) and WSS. CHI measures the separation ratio based on the maximum distance between centroids and measures compactness based on the amount of distance between each data with the centroid. Compact and well-separated configurations of clusters are expected to have high inter-cluster variance and relatively low intra-cluster variance [20, 21].

The Calinski-Harabasz Index (CHI) is calculated by the following Equations:

$$CH(k) = \frac{BSS/(k-1)}{WSS/(n-k)} \qquad (6)$$

d) *Davies-Bouldin Index*: Davies-Bouldin Index (DBI) [22] is one method used to measure cluster validity in a grouping method, cohesion is defined as the sum of the proximity of the data to the cluster center point of the cluster followed. Meanwhile, separation is based on the distance between the cluster center points to the cluster.

Measurements using DBI will maximize the inter-cluster distance between the c_i and c_j clusters and at the same time will minimize the distance between data in a cluster. If the inter-cluster distance is maximal, it means that the characteristic similarity between each cluster is small so that the differences between clusters can be seen more clearly. If the intra-cluster distance is minimal, it means that each object in the cluster has a high level of characteristic similarity. The following equation is used to calculate the DBI:

$$DBI = \frac{1}{k} \sum_{i=1}^{k} \max_{i \neq j} R_{i,j} \qquad (7)$$

Where,

$$R_{i,j} = \frac{WSS_i + WSS_j}{BSS_{i,j}} \qquad (8)$$

$$WSS_i = \frac{1}{m_i}\sum_{j=1}^{m_i} d(x_j, c_i) \quad (9)$$

$$BSS_{i,j} = d(c_i, c_j) \quad (10)$$

With $d(x,y)$ is the euclidean distance between x and y, x_i is the cluster i, c_i is the centroid of cluster x_i, and k is the number of clusters used. The smaller the *DBI* value obtained ($DBI \geq 0$), the more optimal the number of clusters is obtained.

e) *Ratkowsky-Lance Index*: The Ratkowsky-Lance index [23] is based on the mean between the sum of squares between then clusters for each data (BGSS) and the total sum of squares of each data within the cluster (TSS). The RL index is calculated by the following Equation:

$$RL = \frac{\bar{S}}{\sqrt{k}} \quad (11)$$

Where,

$$\bar{S}^2 = \frac{1}{p}\sum_{j=1}^{p} \frac{BGSS_j}{TSS_j} \quad (12)$$

$$BGSS_j = \sum_{q=1}^{k} n_q (c_{qj} - \bar{x}_j)^2 \quad (13)$$

$$TSS_j = \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2 \quad (14)$$

With k is the optimal number of clusters.

The number of clusters with the maximum the Ratkowsky and Lance index value is taken as the proposed optimal number of clusters.

f) *Hubert Index*: Hubert Index (HI) [24] is the point serial correlation coefficient between any two matrices. When the two matrices are symmetric, HI can be calculated by the following Equation:

$$\text{HI}(P, Q) = \frac{1}{N_t}\sum_{i=1, i<j}^{n-1} P_{ij} Q_{ij} \quad (15)$$

Where,
P is the proximity matrix of the data set;
Q is a matrix n x n whose element (i, j) is equal to the distance between representative data from the cluster (x_i and x_j).

Hubert index is a graphical method. A significant peak in the plot indicates the optimal number of clusters.

g) *Ball-Hall Index*: Ball-Hall Index (1965) is the mean of the distance of the items to their respective cluster centroids and is calculated using Equations:

$$BH = \frac{WSS}{k} \quad (16)$$

Where k is the number of clusters. The maximum difference in value between levels is used to show the solution for the optimal number of clusters.

h) *Krzanowski-Lai Index*: Krzanowski-Lai index (KL) [25] proposes internal indices defined by the following Equations:

$$KL(k) = \left|\frac{\text{diff(k)}}{\text{diff}(k+1)}\right| \quad (17)$$

Where,

$$diff(k) = (k\text{-}1)^{2/D}\, WSS_{k-1} - k^{2/D}\, WSS_k \text{ for } k = 2, 3, \ldots \quad (18)$$

Let *diff(k)* denote the difference in the function when the number of groups in the partition is increased from (k - 1) to k. The parameter D represents the feature dimensionality of the input object (number of attribute), WSS_k is calculated as the within-group dispersion matrix of the clustered data. The optimal number of cluster k is the value that maximizes *KL(k).*

2.4. Cluster Quality Testing

The quality of the cluster produced will be tested by evaluating the value of variance (*R*). *R* is the ratio value between the average distance of data in the same cluster (intra-cluster distance) and the average distance of data in the other clusters (inter-cluster distance) [26]. An *R* value close to 0 indicates that data in the same clusters are more similar.

$$R = \frac{{}^{1}/{}_{k}\sum_{i=1}^{k} R_k}{{}^{1}/{}_{k}\sum_{\substack{i,j=1 \\ i\neq j}}^{k} R_{ij}} \tag{19}$$

Where *R* is the variance value, *k* is the number of clusters, R_k is the average distance of data in a cluster, and R_{ij} is the average distance of data in the other clusters.

3. RESULTS AND ANALYSIS

3.1. Data Understanding

The dataset used in this study is the real-life dataset, which contained the sales data of a pharmacy in Palembang. It consist a transactional records between January and December 2015. The dataset contained 2.043 products, 399.738 sales transactions and 3.956.683 products sold [5].

3.2. Data Preparation

The input variables used for cluster analysis are recency (R), frequency (F), and monetary (M) collected from the transactions dataset. The attribute value interval for the recency of each product is 1 – 364 days. This indicates that the greater the recency value, the longer the last activity of selling the product in the period 1 January to 31 December 2015. The value interval for the frequency attribute of each product is 1 – 14.872 transactions, so the greater the frequency value indicates that the more often the product is sold in the period January 1 to December 31, 2015.

Meanwhile, the value interval for the monetary attribute of each product is Rp. 1.250 – Rp. 1.151.952.500. The greater monetary value indicates that the greater the value of sales (money paid by consumers) for a product within a period of one year. If the RFM value for each product is visualized in 3D, then Figure 1 shows the RFM value of the dataset's skewed distribution transactions.

In order for the RFM value to be normally distributed, each RFM value must be transformed into a lognormal distribution as shown in Figure 2. Generally, the use of natural logarithms such as log base 10 or log base 2 for modeling purposes does not affect the yield value [27].

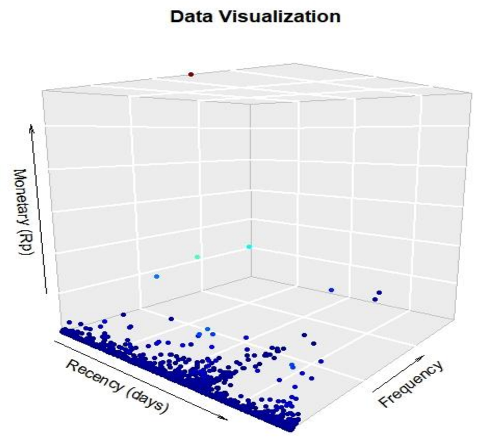

Figure 1. The data visualization based on RFM analysis

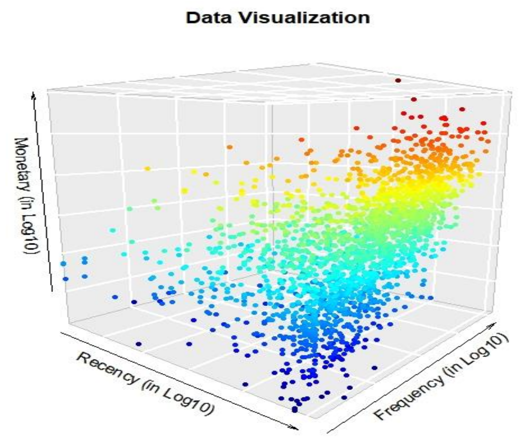

Figure 2. RFM log-transformed

Visualization of data distribution after being transformed into a lognormal distribution (Figure 2) looks better than visualizing the distribution of data before it is transformed (Figure 1). Products with high RFM values appear red and are in the upper right corner of the graph. Meanwhile, products with low RFM values appear dark blue and are in the lower left corner of the graph.

3.3. Visualize the Optimal Number of Clusters in k-Means Method

In this study, eight of indexes validity will be used to determine the optimal number of clusters (k) as shown in Figure 3. The optimal number of clusters will be measured using Elbow Method (EM), Silhouette Index (SI), Calinski-Harabasz Index (CHI), Davies-Bouldin Index (DBI), Ratkowski Index (RI), Hubert Index (HI), Ball-Hall Index (BHI), and Krzanowski-Lai Index (KLI). The number of clusters tested starts from k = 1, 2, 3, ..., 10 clusters. The evaluation results from Figure 3 show that the optimal number of clusters (k) for the k-Means method that uses eight index validity in this study is k = 3.

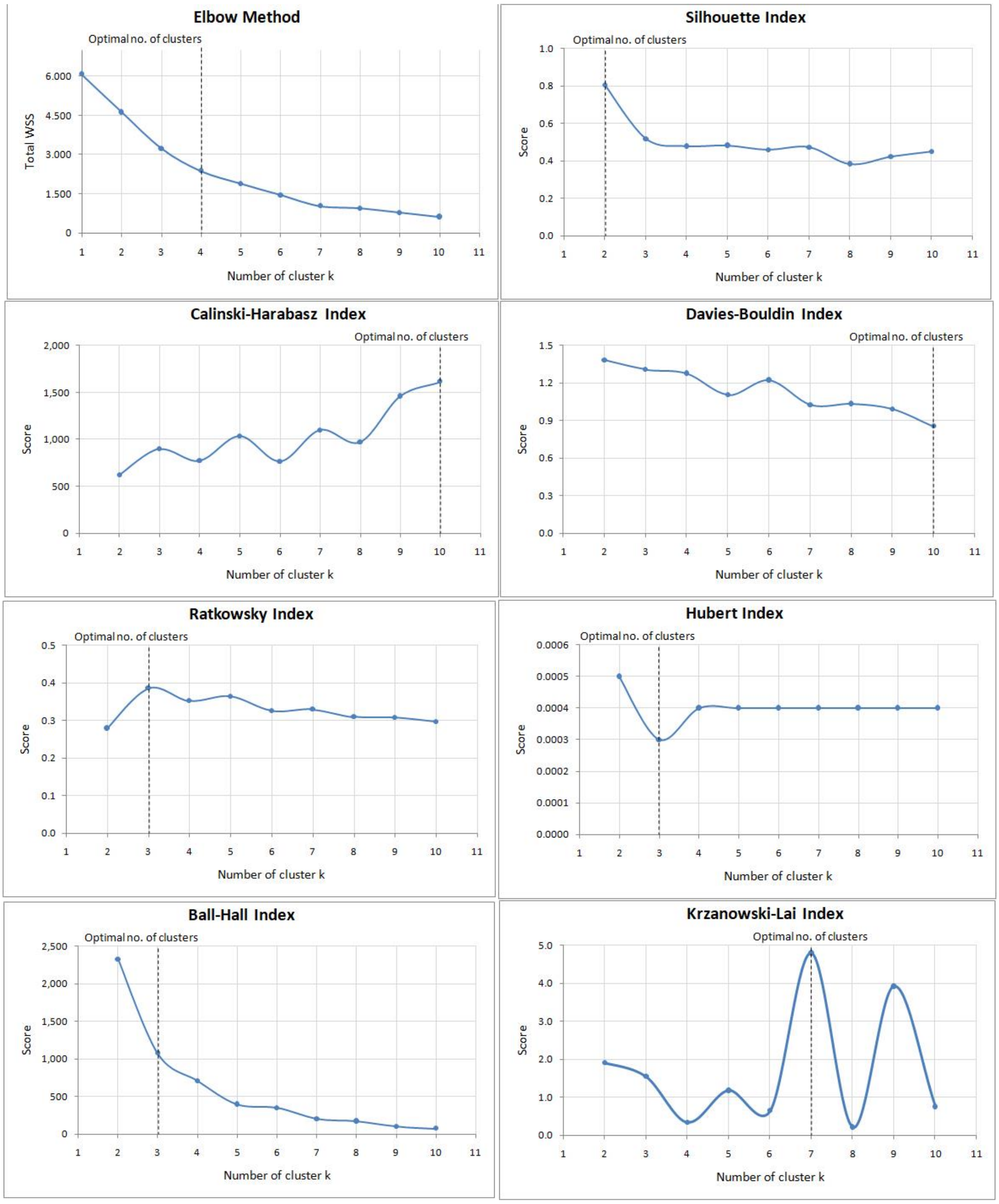

Figure 3. The optimal number of clusters (k) is calculated by EM, SI, CHI, DBI, RI, HI, BHI, and KHI

3.4. Cluster Quality Testing

In the process of forming clusters k = 3, the results of testing (evaluation) of cluster quality using (19) indicate that the variance value is 0.19113. The variance value close to 0 illustrates that members of each cluster have high similarities in data values. The test results are shown in Table 1.

Table 1. Test Results (Evaluation) for Three Clusters

Attribut	The Variance Value (R)
Recency	0.23524
Frequency	0.21875
Monetary	0.11941
Average	0.19113

3.5. Product Segmentation

From the real transactions dataset, it was obtained that the most value for recency in a year was 364 days and the least value for recency in a year was 1 day. The most value for frequency in a year was 14,872 and the least value for frequency in a year was 1. The most value for monetary in a year was Rp 1,151,952,500 and the least value for monetary in a year was Rp 1,250. The software used for clustering in this study is R Programming version 3.5.3, so the results of product clustering (segmentation) using the k-Means method (k=3) in RFM analysis can be seen in Figure 4, with interval values for each attribute RFM is listed in Table 3. The output from R Programming shows that the means of each cluster for the RFM attribute is shown in Table 2. The interval values for each cluster for the RFM attribute (Table 3) are obtained from the calculation of the lower and upper limits of each cluster in Table 2. The results of this study indicate that cluster processes become simpler and more objective than traditional approaches [5-7], so that this method can improve the research of previous RFM model [5].

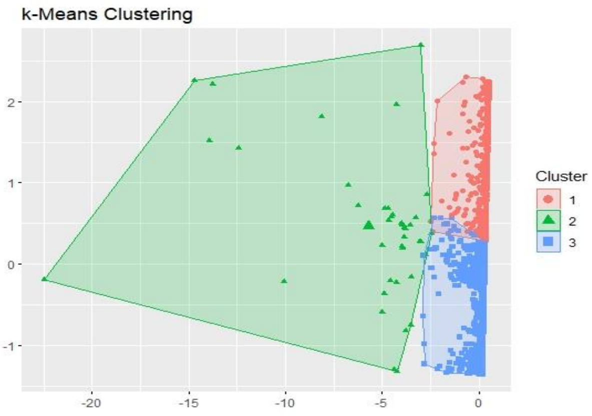

Figure 4. Visualization of clustering results using the k-Means method (k = 3)

Table 2. The Means of Each Cluster for RFM Attributes

Cluster	Recency	Frequency	Monetary (in thousands)
1	75.8167	3,436.744	3,089,608
2	224.3947	13,013.333	76,920,847
3	331.9681	107.418	286,927,000

Table 3. Cluster Intervals for Each RFM Attributes

Cluster	Recency	Frequency	Monetary (in thousands)
1	R > 299	F ≤ 213	M ≤ 6,1 7
2	149 < R ≤ 299	213 < F ≤ 6,659	6,177 < M ≤ 147,663
3	R ≤ 149	F > 6,659	M > 147,663

4. CONCLUSION

This study has produced a new procedure for RFM analysis (in product segmentation) using the k-Means method, where in the basic concept of RFM analysis, datasets are divided equally into five clusters of the same size which is 20% for each cluster. Meanwhile, the use of the k-Means method in this study (after being evaluated to obtain the optimal number of clusters with eight index validity) has resulted in a more objective product clustering with high similarity in data values, so as to increase the accuracy of the stock management process.

The evaluation results show that the optimal number of clusters for the k-Means method applied in the RFM analysis consists of three clusters (segmentation) with a variance value of 0.19113. In future work, you can use particle swarm optimization (PSO), medoid or maximizing-expectancy method as a comparison to get more optimal results, and then output compared to outcomes if using the basic RFM analysis method [28].

ACKNOWLEDGEMENTS

This study was supported by PDUPT Grant from DRPM Kemenristekdikti of the Republic of Indonesia.

REFERENCES

[1] R. Ait Daoud, A. Amine, B. Bouikhalene, and R. Lbibb, "*Combining RFM model and clustering techniques for customer value analysis of a company selling online*," in 2015 IEEE/ACS 12th International Conference of Computer Systems and Applications (AICCSA), 2015, pp. 1–6.

[2] H.-H. Wu, E.-C. Chang, and C.-F. Lo, "*Applying RFM Model and K-Means Method in Customer Value Analysis of an Outfitter*," in 16th ISPE International Conference on Concurrent Engineering, 2009, no. 2, pp. 665–672.

[3] S. Abdelaziz and S. Lu, "K-means algorithm with level set for brain tumor segmentation," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 15, no. 2, pp. 991–1000, 2019.

[4] M. Z. Hossain, M. N. Akhtar, R. B. Ahmad, and M. Rahman, "A dynamic K-means clustering for data mining," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 13, no. 2, pp. 521–526, 2019.

[5] R. Gustriansyah, D. I. Sensuse, and A. Ramadhan, "A sales prediction model adopted the recency-frequency-monetary concept," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 6, no. 3, pp. 711–720, 2017.

[6] B. H. H. Maskan, "Proposing a Model for Customer Segmentation using WRFM Analysis (Case Study: an ISP Company)," *Int. J. Econ. Manag. Soc. Sci.*, vol. 3, no. 12, pp. 77–80, 2014.

[7] S. C. Hsu, "The RFM-based Institutional Customers Clustering: Case Study of a Digital Content Provider," *Inf. Technol. J.*, vol. 11, no. 9, pp. 1193–1201, Sep. 2012.

[8] J. T. Wei, S.-Y. Lin, Y.-Z. Yang, and H.-H. Wu, "The application of data mining and RFM model in market segmentation of a veterinary hospital," *J. Stat. Manag. Syst.*, pp. 1–17, 2019.

[9] R. Gustriansyah, D. I. Sensuse, and A. Ramadhan, "*Decision support system for inventory management in pharmacy using fuzzy analytic hierarchy process and sequential pattern analysis approach*," in 2015 3rd International Conference on New Media (CONMEDIA), 2015, pp. 1–6.

[10] D. Birant, "Data Mining Using RFM Analysis," in *Knowledge-Oriented Applications in Data Mining*, no. iii, K. Funatsu, Ed. In Tech, 2011, pp. 91–108.

[11] A. M. Hughes, "Boosting response with RFM. Mark," *Tools*, vol. 5, pp. 4–10, 1994.

[12] Y.-L. Chen, M.-H. Kuo, S.-Y. Wu, and K. Tang, "Discovering recency, frequency, and monetary (RFM) sequential patterns from customers' purchasing data," *Electron. Commer. Res. Appl.*, vol. 8, no. 5, pp. 241–251, Oct. 2009.

[13] J. T. Wei, S.-Y. Lin, Y.-Z. Yang, and H.-H. Wu, "Applying Data Mining and RFM Model to Analyze Customers' Values of a Veterinary Hospital," in *2016 International Symposium on Computer, Consumer and Control (IS3C)*, 2016, pp. 481–484.

[14] M. J. Garbade, "Understanding K-means Clustering in Machine Learning," *Towards Data Science*, 2018. [Online]. Available: https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1.

[15] C. Subbalakshmi, G. Rama Krishna, S. Krishna Mohan Rao, and P. Venketeswa Rao, "A method to find optimum number of clusters based on fuzzy silhouette on dynamic data set," *Procedia Comput. Sci.*, vol. 46, no. Icict 2014, pp. 346–353, 2015.

[16] M. A. Syakur, B. K. Khotimah, E. M. S. Rochman, and B. D. Satoto, "*Integration K-Means Clustering Method and Elbow Method for Identification of the Best Customer Profile Cluster*," IOP Conf. Ser. Mater. Sci. Eng., vol. 336, no. 1, 2018.

[17] A. Starczewski and A. Krzyżak, "Performance Evaluation of the Silhouette Index," 2015, pp. 49–58.

[18] P. J. Rousseeuw, "Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis," *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, Nov. 1987.

[19] T. Calinski and J. Harabasz, "A dendrite method for cluster analysis," *Commun. Stat.*, vol. 3, no. 1, p. 1974, 1974.

[20] N. Tomašev and M. Radovanović, "Clustering evaluation in high-dimensional data," *Unsupervised Learn. Algorithms*, pp. 71–107, 2016.

[21] M. E. Celebi and K. Aydin, *Unsupervised learning algorithms*. 2016.

[22] D. L. Davies and D. W. Bouldin, "A Cluster Separation Measure," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-1, no. 2, pp. 224–227, Apr. 1979.

[23] R. S. Hill, "A Stopping Rule for Partitioning Dendrograms," *Bot. Gaz.*, vol. 141, no. 3, pp. 321–324, Sep. 1980.

[24] R. C. Dubes, "How many clusters are best? - An experiment," *Pattern Recognit.*, vol. 20, no. 6, pp. 645–663, 1987.

[25] W. J. Krzanowski and Y. T. Lai, "A Criterion for Determining the Number of Groups in a Data Set Using Sum-of-Squares Clustering," *Biometrics*, vol. 44, no. 1, p. 23, 2006.

[26] T. Handhayani, I. Wasito, M. Sadikin, and Ranny, "*Kernel based integration of Gene expression and DNA copy number*," Int. Conf. Adv. Comput. Sci. Inf. Syst., pp. 303–308, 2013.

[27] N. Zumel and J. Mount, *Practical Data Science with R*, Second. Shelter Island, New York: Manning Publications Co., 2014.

[28] R. Gustriansyah, N. Suhandi, and F. Antony, "The Design of UML-Based Sales Forecasting Application," *Int. J. Recent Technol. Eng.*, vol. 7, no. 6, pp. 1507–1511, 2019.