Journal of King Saud University – Computer and Information Sciences 33 (2021) 1251–1257

Contents lists available at ScienceDirect

Journal of King Saud University – Computer and Information Sciences

journal homepage: www.sciencedirect.com

RFM ranking – An effective approach to customer segmentation

A. Joy Christy [a,*], A. Umamakeswari [a], L. Priyatharsini [b], A. Neyaa [b]

[a] Department of CSE, School of Computing, SASTRA Deemed to be University, Thanjavur, India
[b] School of Computing, SASTRA Deemed to be University, Thanjavur, India

ARTICLE INFO

Article history:
Received 1 June 2018
Revised 26 August 2018
Accepted 4 September 2018
Available online 5 September 2018

Keywords:
Customer segmentation
RFM analysis
K-Means
Fuzzy C-Means
Initial centroids

ABSTRACT

The efficient segmentation of customers of an enterprise is categorized into groups of similar behavior based on the RFM (Recency, Frequency and Monetary) values of the customers. The transactional data of a company over is analyzed over a specific period. Segmentation gives a good understanding of the need of the customers and helps in identifying the potential customers of the company. Dividing the customers into segments also increases the revenue of the company. It is believed that retaining the customers is more important than finding new customers. For instance, the company can deploy marketing strategies that are specific to an individual segment to retain the customers. This study initially performs an RFM analysis on the transactional data and then extends to cluster the same using traditional K-means and Fuzzy C- Means algorithms. In this paper, a novel idea for choosing the initial centroids in K- Means is proposed. The results obtained from the methodologies are compared with one another by their iterations, cluster compactness and execution time.

1. Introduction

In recent years, there has been a massive increase in the competition among firms in sustaining in the field. The profits of the company can be improved by a customer segmentation model. Customer retention is more important than the acquisition of new customers. According to the Pareto principle (Srivastava, 2016), 20% of the customers contribute more to the revenue of the company than the rest. Customer segmentation can be performed using a variety of unique customer characteristics to help business people to customize marketing plans, identify trends, plan product development, advertising campaigns and deliver relevant products. Customer segmentation personalizes the messages of individuals to better communicate with the intended groups. The most common attributes used in customer segmentation are location, age, sex, income, lifestyle and previous purchase behavior.

Here, segmentation is done using behavioral data since it is commonly available and continuously evolving with time and purchase history. RFM (Recency, Frequency, and Monetary) analysis is a renowned technique used for evaluating the customers based on their buying behavior. A scoring method is developed to evaluate scores of Recency, Frequency, and Monetary. Finally, the scores of all three variables are consolidated as RFM score ranging from 555 to 111 (Haiying and Yu, 2010) which is used to predict the future patterns by analyzing the present and past histories of the customer. In this context, it has been observed that the scores of three factors Recency, Frequency and Monetary directly proportional to customer's lifetime and retention.

Once the values of recency, frequency and monetary are calculated, the K-Means algorithm is applied to the variables to clusters of the customer base. The behavior of each cluster is analyzed to find the group of customers who give more profits to the company. Similarly, clustering is performed using two other algorithms namely, Fuzzy C – Means clustering and the proposed method with chosen initial centroids in the existing K– Means algorithm. The motivation of the paper is to propose a method for choosing initial centroids for K-means algorithm and to impose the method to segment the customer with reduced iteration and time. Now that clusters of customers are found, it is necessary to understand the differences between these groups of customers. A thorough analysis is performed on the clusters to aid in finding the targeted customers and bestows them with appropriate promotions and offers. Also, a novel Repetitive Median based K-Means algorithm is

* Corresponding author.
E-mail address: joychristy@cse.sastra.edu (A.J. Christy).
Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

https://doi.org/10.1016/j.jksuci.2018.09.004
1319-1578/

proposed with an intension to reduce the number of iterations than the traditional clustering algorithms. The outcome of the proposed work is a meaningful customer segmentation which will be useful for marketing people. The rest of the study focuses on analyzing all the three clustering approaches regarding iterations, cluster compactness, execution time and various other factors.

2. Literature review

Jiang and Tuzhilin (2009) identified that both customer segmentation and buyer targeting are necessary to improve the marketing performances. These two tasks are integrated into a step by step approach, but the problem faced is unified optimization. To solve the problem, the author proposed the K-Classifiers Segmentation algorithm. This approach focuses on distributing more resources to those customers who give more returns to the company. A sizable amount of authors had written about different methods for segmenting the customers.

He and Li (2016) suggested a three-dimensional approach to improving the customer lifetime (CLV), the satisfaction of the customer and customer behavior. The authors have concluded that the consumers are different from one another and so are their needs. Segmentation assists in finding their demand and expectations and proving a good service.

Cho and Moon (2013) proposed a customized recommendation system using weighted frequent pattern mining. Customer profiling is performed to find the potential customers using the RFM model. The author has defined varied weights for each transaction to generate weighted association rules through mining. Using the RFM model will provide a more accurate recommendation to the customer which in turn increases the profit of the firm.

Zahrotun (2017) used the customer data from online to identify the finest customer using Customer Relationship Management (CRM). By applying the CRM concept for online shopping, the author identifies the potential customers by segmenting them which helps us in increasing the profits for the company. So to perform customer segmentation and marketing to customers in an accurate way the Fuzzy C-Means Clustering Method is used. Thus, this helps the customers to get special facilities in more than one category in the appropriate marking strategies according to their needs.

Shah and Singh (2012) proposed a new clustering algorithm which executes similar to the K-means algorithm and K-medoids algorithms. Both the methods are partitional approaches. The proposed algorithm does not provide an optimal solution in all case, but it reduces the cluster error criterion. Saurabh observes that as the number of clusters increases the new method takes lesser time to execute than the traditional methods.

Sheshasaayee and Logeshwari (2017) designed a new integrated approach by segmentation with the RFM and LTV (Life Time Value) methods. They used a two-phase approach with the first phase being the statistical approach and the second phase is to perform clustering. They aim to perform K-means clustering after the two-phase model and then use a neural network to enhance their segmentation.

Lu et al. (2014) analyzed the customer churn prediction. The authors used logistic regression and isolated the transactional data for creating a new distinct prediction model. With his experimental implementation, it is observed that customers with the utmost churn value can be identified and can be retained using individual marketing strategies. Zhang believes in deducing the cause for the churn behavior of a customer and fulfilling the individual needs is necessary for the company's long existence.

Jiang and Tuzhilin (2009) presents a direct clustering approach that clusters the customers not based on computed statistics, but by combining transactional data of several customers. The authors also showed that it is NP-hard to find an optimal segmentation solution. So, Tuzhilin came up with different sub-optimal clustering methods. The authors then experimentally examined the customer segments obtained by direct grouping, and it is observed to be better than the statistical approach.

3. Algorithm description

The transactional dataset of the customers of a company is used to perform the segmentation process. In this research, three different algorithms have been used to cluster the customers based on RFM analysis. The data is initially pre-processed to remove outliers and to filter meaningful instances. The outliers are detected using the z-core to identify the relationship of data with its mean and standard deviation. The relationship between mean and standard deviation are mapped to 0 and 1respectively. The data that is too far from the mean (zero) are considered as outliers. The pre-processed information is then fed into the RFM model to calculate the recency, frequency, and monetary values. The three attributes are then passed to three clustering algorithms namely K-Means, Fuzzy C-Means and Repetitive Median based K-Means (RM K-Means) clustering algorithm. These algorithms cluster the customers into segments. The workability of the clustering algorithms is then analyzed regarding the number of iterations, cluster compactness and the time taken for execution. The Fig. 1 gives a brief view of the proposed customer segmentation system.

3.1. RFM analysis

Recency, frequency and monetary (RFM) analysis is a powerful and recognized technique in database marketing. It is widely used to rank the customers based on their prior purchasing history. RFM analysis finds use in a wide range of applications involving a large number of customers such as online purchase, retailing, etc. This method groups the customers based on three dimensions, recency(R), frequency (F) and monetary (M).

3.1.1. Recency – When was the last time the customer made a purchase?

Recency value is the number of days a customer takes between two purchases. A smaller value of recency implies that the customer visits the company repeatedly in a short period. Similarly, a greater value implies that the customer is less likely to visit the company shortly.

3.1.2. Frequency – How many times did the customer purchase?

Frequency is defined as the number of purchases a customer makes in a specific period. The higher the value of frequency the more loyal are the customers of the company.

3.1.3. Monetary – How much money did the customer spend?

Monetary is defined as the amount of money spent by the customer during a certain period. The higher the amount of money spent the more revenue they give to the company.

Each customer is assigned with three different scores for recency, frequency, and monetary variables. Scoring is done in the scale from 5 to 1. The top quintile is given a score of 5, and the others are given 4, 3, 2 and 1. The scores can be assumed to have unique characteristics as given in Table 1.

Finally, all the customers are provided with scores 555,554... 111. The customers with the score 555 can be called as the potential customers of the company since they are likely to give more profit to the company and vice versa goes with the customers

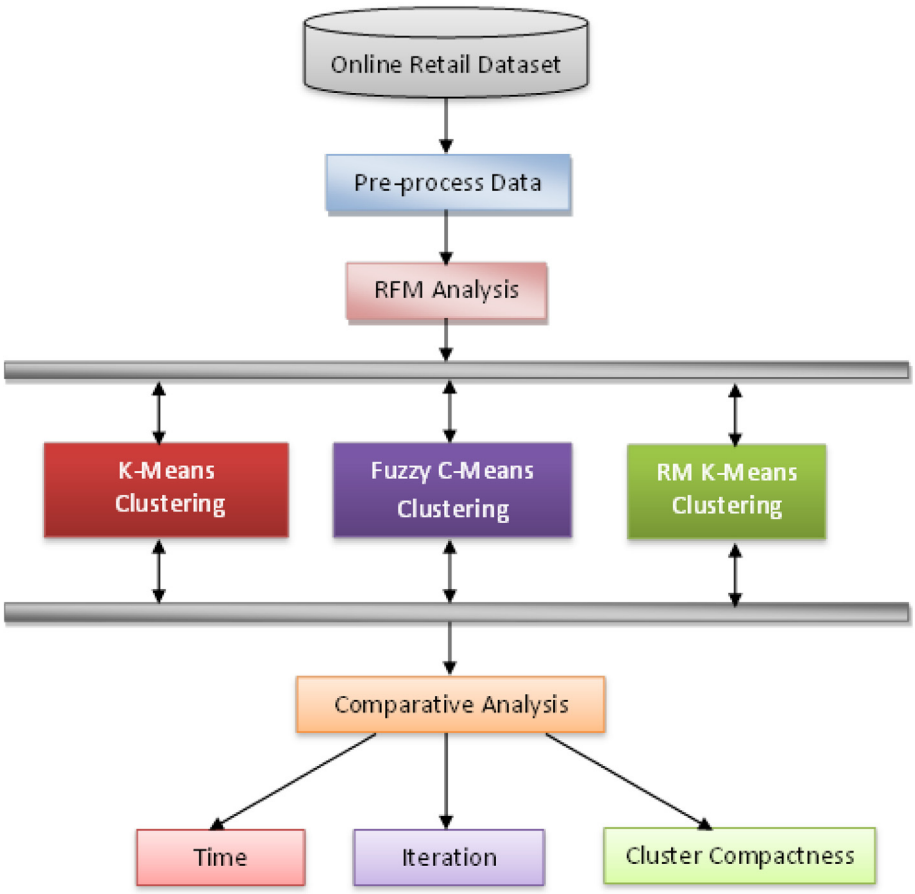

Fig. 1. The framework of RM K-Means Analysis.

Table 1
RFM Score Description.

Score	Characteristics
5	Potential
4	Promising
3	Can't-lose them
2	At risk
1	Lost

having a score of 111. Depending on this RFM score, each customer can be put into a different segment.

3.2. K-Means clustering

K-Means is a standard algorithm which takes the parameters and the number of clusters as inputs and partitions the data into the defined number of clusters such that the intra-cluster similarity is high. K-Means is an iterative approach which computes the value of centroids before each iteration. The data points are moved among different clusters depending on the centroids calculated at each iteration. The process is repeated until the sum cannot be decreased any more. K- Means algorithm is shown in Algorithm 1.

The values of the variables recency, frequency and monetary are normalized using min-max normalization. This is performed since skewed values may be problematic. Now the clustering algorithm is applied to the scaled data. The number of clusters is limited to 10. The amount of money earned by each segment of customers is calculated to find the segment of customers which give more revenue to the company. The complexity of K-means is O (n + k + i). Where 'n' refers to the number of instances, k refers to the number of clusters, and i refers to the number of iterations.

Algorithm 1 (*K-Means*).

Input:
⇒ Customer Dataset containing 'n' instances
⇒ k: the number of clusters
Output:
⇒ Customer data Partitioned to k clusters
Algorithm:
1. Initially, depending on the value of k, k random points are chosen as initial centroids.
2. The distances of each data point from the centroids chosen earlier are evaluated using the Euclidian distance.
3. The distance values are compared and the data point is assigned to the centroid which has shortest Euclidian distance value.
4. The previous steps are repeated. The process is stopped if the clusters obtained are same as that of the previous step.

3.3. Fuzzy C-Means

Fuzzy C-Means is a clustering approach (Memon and Lee, 2017) which permits a specific data to be present in more than one cluster. It does not decide the membership history of a data point to a given cluster. Instead, the likelihood that a specific data point will belong to that cluster is calculated. The advantage that Fuzzy C-Means has over K-Means is that the result obtained for the large

and similar dataset is better than K-means algorithm because in K-means a data point must entirely be present in only one cluster. In this study, a customer may belong to more than one cluster which increases the chance of retaining the customers by treating them with different offers for each segment. The time complexity of Fuzzy C-Means is O ($n + k + d^2 + i$), where d is the number of iterations.

Similar to the previous algorithm, the variables are scaled using min-max normalization. Now the customers are clustered based on Fuzzy C-Means clustering (Zahrotun, 2017) based on the recency, frequency, and monetary values.

Algorithm 2 (*Fuzzy C-Means*).

Input => Customer Dataset containing 'n' instances
=> k: the number of clusters
Output:
=> Customer data Partitioned to k clusters
Algorithm:
1. Randomly selects k initial centers.
2. Calculate the fuzzy membership matrix μ_{ij}.
$\mu_{ij} = 1/\sum_{c=1}^{k}\left(\frac{d_{ij}}{d_{ic}}\right)\left(\frac{2}{m}-1\right)$
3. Compute the cluster centers v_j.
$v_j = \left(\sum_{i=1}^{n}\left(\mu_{ij}\right)^{m} x_i\right)/\left(\sum_{i=1}^{n}(\mu_{ij})^{m}\right)$
4. Repeat steps 2 and 3 until lowest value of j is achieved, where j is the objective function.

3.4. Repetitive median K-Means

Although K- Means algorithm is traditionally used for grouping, it has few disadvantages. K-Means chooses the initial centroids in a random fashion. Then the distance of each data point from the centroid is calculated by Euclidian distance, and each point is allocated to the closest centroid which forms a cluster. The problem with choosing initial centroids randomly is that the centroid may bundle closer to each other causing the clusters to be less meaningful. Initial centroids determine the goodness of cluster such as reducing number of iterations, global optimum solutions, and cluster compactness. The performance of K-Means is degraded by random initial centroids (Liu et al., 2014).

Table 2
Online Retail Dataset Description.

No.	Attribute Name	Description	Data type
1	InvoiceNumber	6-digit unique number for each transaction	Nominal
2	StockCode	5-digit unique number for each product	Nominal
3	Description	Product Name	Nominal
4	Quantity	Quantity of product per transactions	Numeric
5	InvoiceDate	Invoice Date and Time	Numeric
6	UnitPrice	Product price per unit	Numeric
7	Customer Id	5-digit unique number for each customer	Nominal
8	Country	Country Name	Nominal

This paper proposes a new way for choosing the initial centroids for the K- Means algorithm. The three variables Recency (R), Frequency(F) and Monetary(M) that are to be clustered are sorted and stored in ascending order in three vectors as R', F' and M'. The median value of each vector is found and assigned as the initial centroids for the K-Means algorithm. Iteratively the median values are calculated from the R', F' and M' values k number of times depending the value of k (number of segments). Choosing initial centroids with its mean distribution reduces the number of iterations and computational time of traditional K –Means algorithm. It is observed the clusters that are obtained through the modified approach is more meaningful and appropriate compared to the traditional method by choosing centroids randomly. The complexity of RM K-Means is as same as K-Means, which is O ($n + k + i$). Since the initial random centroids are computed using a median based method, the proposed RM K-means algorithm reduces the number of iterations with K-means.

Algorithm 2 (*RM K-Means*).

Input:
=> Customer dataset containing n instances
=> K: the number of clusters
Output:
Customer data partitioned to k clusters
Algorithm steps:
1. Upload the Customer transactional dataset
2. Pre-process the dataset by removing outliers and null valued instances
3. Compute R, F and M Score for each instance
4. Order RFM score in sequence as R', F' and M' and store it in a vector
5. Let S = total number of instances/ k
6. Split R'F'M' vector with k segments where each segment consists of S instances
7. For i = 1 to k do
   6.1. Compute median for each segment i
   6.2. Store the median in the vector m[i].
8. Let the values of m vector be the initial centroids for K-means
9. Compute the distance of each object's RFM with centroids
10. Group the object based on the minimum distance
11. Recompute cluster centroids
12. Repeat steps 8 to 10 until there is no change in the cluster members or centroids

4. Experimentation and result discussion

The performance of the proposed methodology is evaluated by working on the transactional data set of the customers of an online retail store for one year is obtained from the University of California Irwin (UCI) repository. The step-by-step process of customer segmentation is presented in this section. The dataset consists of eight attributes including the customer ID, product code, product

Table 3
RFM Calculator.

Parameter	RFM Score				
	5	4	3	2	1
Recency (days)	7	30	90	180	365
Frequency (number of purchases)	15	12	9	6	3 and lesser
Monetary (in dollar)	Above 12,000	9000 – 12,0000	6000–9000	3000–6,0000	Below 3000

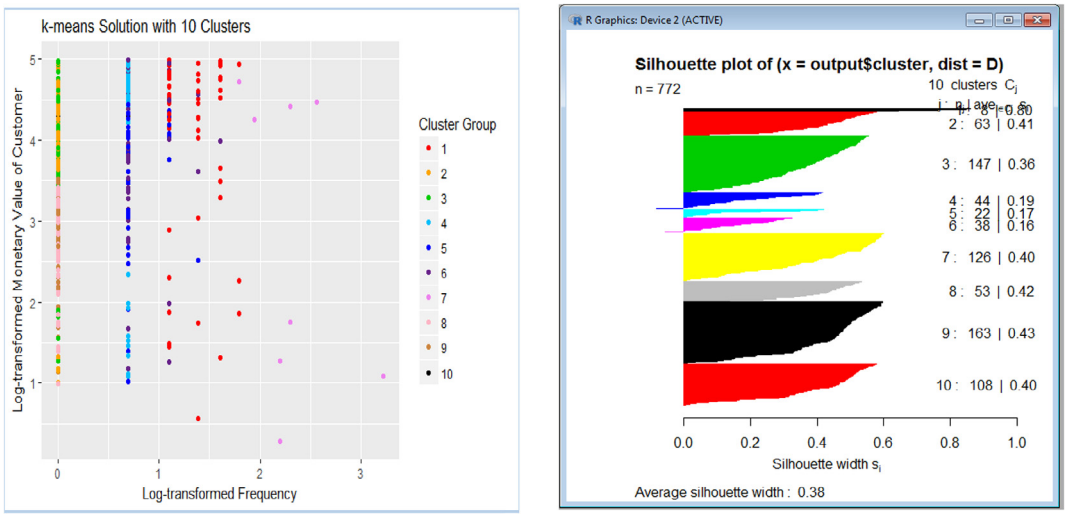

Fig. 2a. K-Means Clustering.

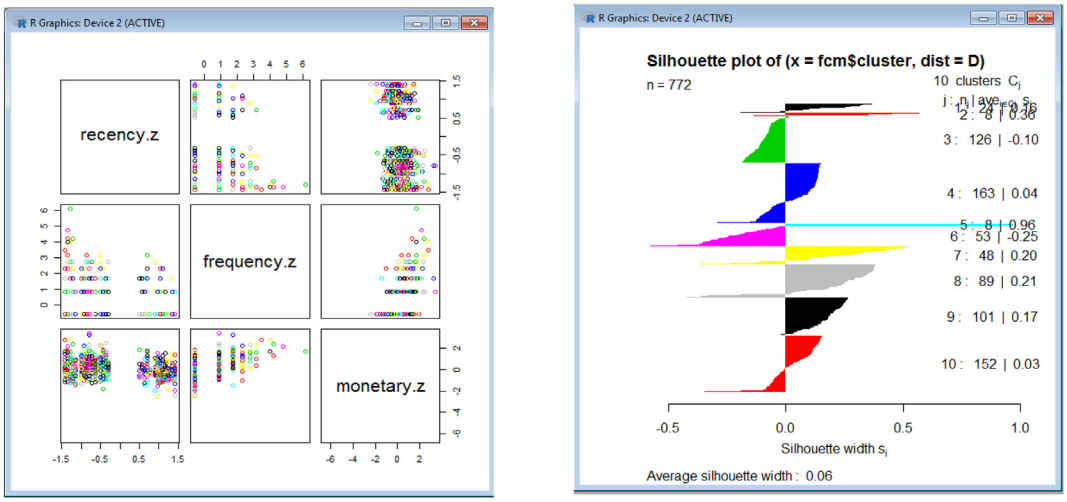

Fig. 2b. Fuzzy C- Means Clustering.

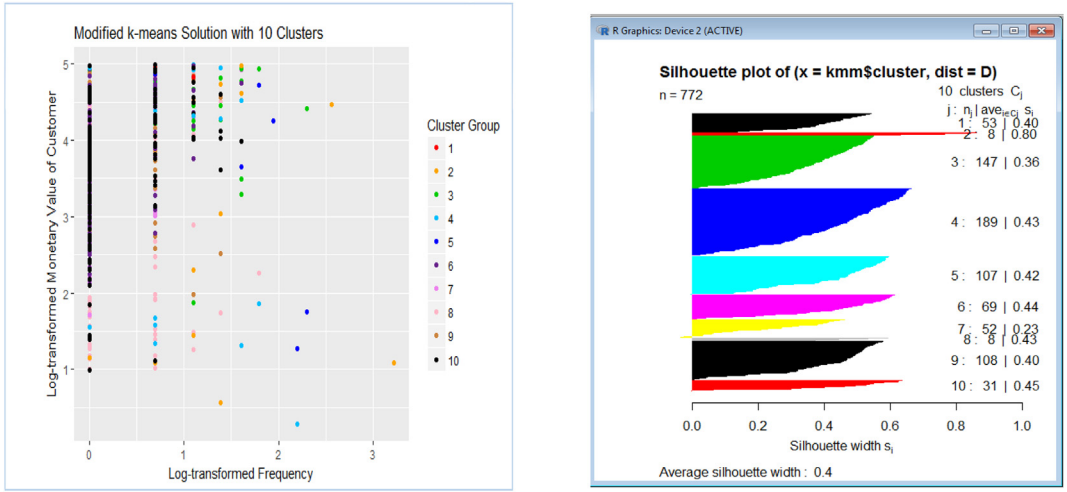

Fig. 2c. RM K- Means Clustering.

Table 4
Comparative Analysis of RM K-Means.

	K-Means	Fuzzy C-Means	RM K-Means
Iterations	4	193	2
Time Taken (in seconds)	2.0035	24.7988	1.4917
Average silhouette width	0.33	0.43	0.49

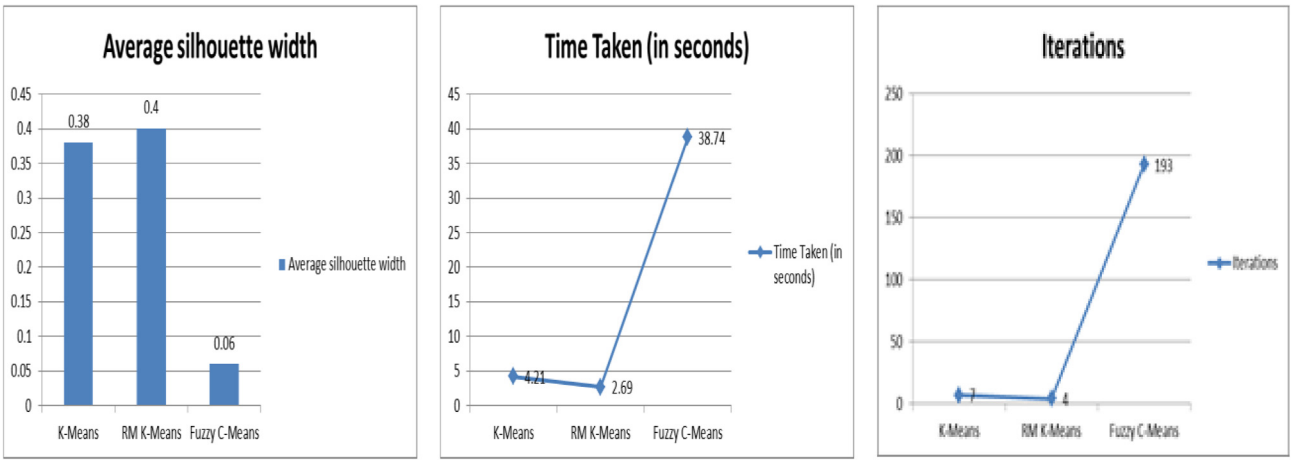

Fig. 3. Result Analysis of RM K-Means.

name, the price of the product, date and time of purchase, etc. The original data set consists of 18,267 instances with eight attributes. The dataset contains the purchase of information of customers from 1-12-2010 to 09-12-2011. The instances with missing values in important attributes, unit price and quantity less than 0 and the date exceeding the current date are all removed during data pre-processing. To identify the outliers, the Z-Score analysis is also performed as an additional step in data pre-processing. The meaningful instances such as invoice data and time, the quantity of product per transaction, product price per unit concerning recency, monetary and frequency are filtered, and only those records have been inputted into the benchmark algorithms. The modified dataset contains 772 instances with three additional attributes recency, frequency and monetary derived from RFM calculation. The description of the original dataset is shown in Table 2.

4.1. RFM calculator

Table 3 denotes the precise calculation for computing the RFM score for each instance, where the score 5 in each parameter is the highest.

The output plots obtained from K- Means, Fuzzy C-Means and RM K- Means are shown in Fig. 2.

The execution time for each algorithm is calculated from the system time. It is observed that the proposed RM K- Means consumes lesser time than the other two techniques because of the lesser number of iterations. The number of iterations is reduced in the RM K-Means because the initial centroids are calculated based on median values. The silhouette width is used for studying the average distance between the resulting clusters. Silhouette plot visually analyses the clustering outcome and displays the number of customers in each cluster and also the minimum distance from the point in the cluster to that of another cluster. A higher value of average silhouette width indicates that the data points within a cluster are closer to each other but not to the points in other clusters. The average silhouette width is calculated for the resulting clusters obtained by both K-means clustering technique and by the RM K-Means and K-Means technique. It is observed that the average silhouette width of RM K-Means is greater than that of Fuzzy C-Means clustering and the K – Means clustering. The results given in Table 4 are plotted in Fig. 3.

5. Conclusion

Segmenting the customers will deepen the relationships with customers. Finding new customers for the enterprise is vital, meanwhile retaining the existing clients (Tong et al., 2017) is even more important. In this paper, segmentation is done using RFM analysis and then is extended to other algorithms like K –Means clustering, Fuzzy C – Means and a new algorithm RM K-Means by making a minor modification in the existing K – Means clustering. The working of these approaches is analyzed. The time taken by each algorithm to execute is analyzed, and it is observed that the proposed K –Means approach consumes lesser time and also reduces the number of iterations. The proposed algorithm is more effective because the centroids are more meaningful and are calculated at the beginning based on the effective medians of data distribution. Since segmentation is done based on the values of recency, frequency, and monetary values, the company can customize their marketing strategies to the customers based on their buying behavior. Future work includes studying the performance of the customers in each segment such as the products which are bought frequently by the members of each segment. This would help better in providing better promotional offers to specific products.

References

He X., Li, C., 2016. The research and application of customer segmentation on e-commerce websites. In: 2016 6th International Conference on Digital Home (ICDH), Guangzhou, pp. 203–208. doi: 10.1109/ICDH.2016.050.

Haiying, M., Yu, G., 2010. Customer Segmentation Study of College Students Based on the RFM. In: 2010 International Conference on E-Business and E-Government, Guangzhou, pp. 3860-3863. doi: 10.1109/ICEE.2010.968.

Sheshasaayee, A., Logeshwari, L., 2017. An efficiency analysis on the TPA clustering methods for intelligent customer segmentation. In: 2017 International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), Bangalore, pp. 784–788.

Srivastava, R., 2016. Identification of customer clusters using RFM model: a case of diverse purchaser classification. Int. J. Bus. Anal. Intell. 4 (2), 45–50.
Memon, K.H., Lee, D.H., 2017. Generalised fuzzy c-means clustering algorithm with local information. In: IET Image Processing, vol. 11, no. 1, pp. 1-12, 1.
Zahrotun, L., 2017. Implementation of data mining technique for customer relationship management (CRM) on online shop tokodiapers.com with fuzzy c-means clustering. In: 2017 2nd International conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE), Yogyakarta, pp. 299–303.
Tong, L., Wang, Y., Wen, F., Li, X., Nov. 2017. The research of customer loyalty improvement in telecom industry based on NPS data mining. China Commun. 14 (11), 260–268. https://doi.org/10.1109/CC.2017.8233665.
Shah, S., Singh, M., 2012. Comparison of a Time Efficient Modified K-mean Algorithm with K-Mean and K-Medoid Algorithm. In: 2012 International Conference on Communication Systems and Network Technologies, Rajkot, pp. 435–437.
Liu, C.C., Chu, S.W., Chan, Y.K., Yu, S.S., 2014. A Modified K-Means Algorithm – Two-Layer K-Means Algorithm. In: 2014 Tenth International Conference on Intelligent Information Hiding and Multimedia Signal Processing, Kitakyushu, pp. 447–450. doi: 10.1109/IIH-MSP.2014.118.
Cho, Young, Moon, S.C., 2013. Weighted mining frequent pattern-based customer's RFM score for personalized u-commerce recommendation system. J. Converg. 4, 36–40.
Jiang, T., Tuzhilin, A., March 2009. Improving personalization solutions through optimal segmentation of customer bases. IEEE Trans. Knowledge Data Eng. 21 (3), 305–320. https://doi.org/10.1109/TKDE.2008.163N.
Lu, H., Lin, J.Lu., Zhang, G., May 2014. A customer churn prediction model in telecom industry using boosting. IEEE Trans. Ind. Inf. 10 (2), 1659–1665. https://doi.org/10.1109/TII.2012.2224355.