Data Mining Approach for Customer Segmentation in B2B Settings using Centroid-Based Clustering

Nadhira Riska Maulina
Department of Industrial Engineering
Faculty of Engineering, Universitas Indonesia
Depok, Indonesia
nadhira.riska@ui.ac.id

Isti Surjandari
Department of Industrial Engineering
Faculty of Engineering, Universitas Indonesia
Depok, Indonesia
isti@ie.ui.ac.id

Annisa Marlin Masbar Rus
Department of Industrial Engineering
Faculty of Engineering, Universitas Indonesia
Depok, Indonesia
annisamarlin@ui.ac.id

***Abstract*— Big data and advanced analytics in organizations are dominant in customer-centric departments such as marketing, sales, and customer service. For company, designing marketing strategies using customer segmentation is useful to improve business revenue. Clustering algorithms able to deal with large data set to recognize patterns and identify customer segments. In this paper, different clustering algorithms will be compared, specifically centroid-based clustering K-Means, CLARA, and PAM with Fuzzy C-Means clustering. The purpose of this research is to find optimum number of clusters using clustering algorithm with the best validation measure score. Dataset is acquired from Tech Company in Indonesia that provide machine with Point of Sale system for food and beverages merchants, since the company in B2B settings. Among three clustering methods, K-Means have the best validation measure score. After compared to Fuzzy C-Means, K-Means outperforms FCM based on time complexity and quality of clustering. Cluster analysis is done to identify customer information. Therefore, this research able to deliver an insightful understanding about customer characteristics using big data analytics and provide an effective Customer Relationship Management Systems.**

Keywords—Data Mining, Customer Relationship Management, Customer Segmentation, Centroid-Based Clustering

I. INTRODUCTION

Over the past decade, the rapid growth of technology has transformed the way enterprises manage their marketing strategies. In a mature market, it is difficult for company to acquire new customers. Each company will compete to gain new customers or retain the existing one by incorporating aggressive strategies, massive advertisements, and discount vouchers [1]. In such competitive and saturated markets, company should be able to understand the characteristics of their customers. Theoretically, one-to-one marketing approach is an effective way to treat customers in order to keep their loyalty [2]. However, not all company possess enough resources to carry out a very specific targeted marketing. One of the solution is to segment the customers so that company can develop strategies that can be intended to specific group of customers.

Company utilize information extracted from big data to identify customers' purchase pattern and characteristics [3]. Big Data consisted of large amount of data that are difficult to process with traditional processing tools to gain insights from it [4]. The characteristic of big data is known as 5 Vs model (Volume, Velocity, Variety, Value, and Veracity) [5]. Many company invested on CRM with big data enabled to log customer's data such as complaints, active status, behavior, and payments. These data can give valuable insight for company to run their business [6].

Since the scope of CRM is a 'personalized' approach, identification of customers' needs, preferences, and behaviors must be done first. Customer segmentation is the identification process of customers that have common characteristics and put them in the same group. These characteristics can be obtained from customer's behavior such as transaction frequency and monetary value [7]. In order to group set of data objects with common characteristics into several groups or clusters, clustering is the most popular technique in customer segmentation literatures [8]. Most of clustering techniques are developed based on the similarity and distance so that data with similar characteristics are grouped near to each other and dissimilar data are placed far away from the objects. [9].

According to [10], there are issues encountered when using clustering algorithms such as the analysis of similarity measurement, the discovery of the optimum cluster number for specific dataset, the efficient way to cluster a large data set, and the evaluation of clustering results. Centroid-based clustering is used to identify customer segmentation in this study. In order to answer the issues of clustering algorithms, partitioning clustering algorithm such as K-Means and K-Medoids algorithms which possess well-defined boundary [12] are compared to soft clusters algorithm which devoid distinct boundaries, Fuzzy C-Means Algorithm.

TABLE I. RFM ATTRIBUTES DEFINITIONS

Attributes	Definition
Recency	Period between the last transaction and the specified time period for this research (March 2019)
Frequency	Average number of transactions in a certain period of time
Monetary	Average money paid to company for service fee per day in a certain period of time

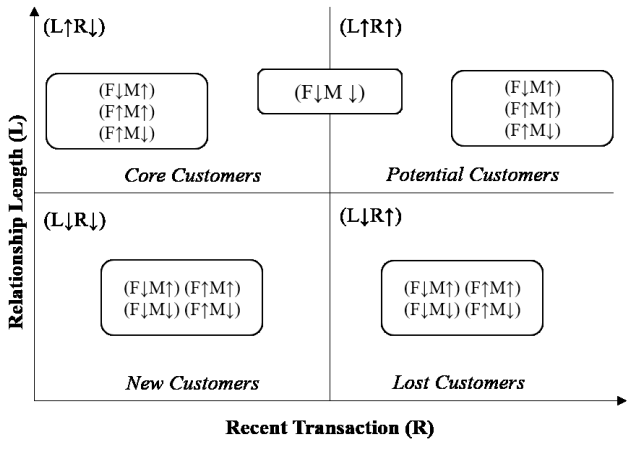

Fig. 1. Customer Loyalty Matrix

978-1-7281-1941-0/19/$31.00 ©2019 IEEE

TABLE II. RFM QUANTILE SCORES

Score	Score Name	*Recency* (days)	*Frequency*	*Monetary* (IDR)
5	Very High	0 – 32	14217 – 17775	720361 – 806200
4	High	33 – 64	10663 – 14216	634521 - 720360
3	Medium	65 – 95	7109 – 10662	548681 - 634520
2	Low	96 – 127	3555 – 7108	462841 - 548680
1	Very Low	128 - 160	0 – 3554	377000 – 462840

The purpose of this research is to compare comprehensive centroid-based clustering algorithms with Fuzzy C-Means Algorithm to search for the best clustering results. Therefore, this research able to deliver an insightful understanding about customer characteristics and examine the meaning of big data through CRM. Customer segmentation in this research is done through behavioral attributes such as Recency, Frequency, and Monetary [13].

Number of clusters *k* is defined through internal cluster validity indices. After that number of cluster has been defined, customers' data are then clustered with partitioning clustering algorithms and Fuzzy C-Means Algorithms. Validation measures are taken into account to find out which algorithm performs the best. Algorithm with optimum validation measures, will be compared to Fuzzy C-Means' time complexity. Once the clusters are formed, customer segments are analyzed and evaluated to gain insights about the behavioral propensity of each customer segment. Further details about customer information from each cluster enable company to develop differentiated strategy to specific group of customers.

The rest of this paper is structured as follows: Section 2 gives an explanation on how the data is processed. In Section 3, experiments and results are presented. The results are discussed in section 4, followed by section 5 as a conclusion for this research.

II. RESEARCH METHODOLOGY

Research methodology in this paper is carried out from data collection to data processing stage in order to generate the expected results.

A. Customer Segmentation

First Company needs to identify the characteristics of their customers' needs, preferences, views, and opinions about the service or products given. Therefore, transactional data can be used to develop effective segmentation solutions. Commonly, transactional data is used to employ customer behavior such as: frequency, recent purchases, and total spending amount.

In retail or restaurant industry, these kinds of transactional data are stored at the Point of Sale and logged every details about customer transaction history which allow company to make further analysis about their customer's trends. The object of this research is categorized as Business-to-Business setting (B2B). Products from B2B company settings are purchased by merchants as an intermediate tools to serve their customers. While Business-to-Customer (B2C) directly sell to end customers [16].

Common approach to understand customer behavior is RFM Analysis which consists of three predefined variables obtained from customers' transactional data. Table 1 explains the description of each attributes. RFM analysis usually includes grouping customers into 5 equal size (quantiles) in which the partition shown in Table 2. For instance, the breakdown into five groups results in quantiles of 20%. For the first 20% part of the data will belong to Group 1, the next 20% will belong to Group 2, and so on. These numbers will become RFM score or RFM status for each cluster so that customers can be categorized using customer loyalty matrix as shown in Fig. 1.

B. Centroid-based Clustering Techniques

In this research, Fuzzy C-Means Clustering are taken into account to be compared with common clustering techniques such as K-Means, and K-Medoids (PAM and CLARA) for customer segmentation. Centroid-Based Clustering is widely known as a popular algorithm to minimize distance between data points through several iterations until convergent stage has been reached [16].

1) K-Means

As the simplest unsupervised learning method, K-means clustering is a common algorithm used on data that remains unlabeled to be clustered into similar groups [17]. The purpose of k-means itself divides a data in several clusters (groups) as many as k, which is the number of k determined by you and represented by the Mean (Average). Means of each cluster is assumed to be a good parameter of each observation of the cluster.

2) K-Medoids

K-Medoids algorithm is almost similar to K-Means. However, K-Means algorithm uses mean and the measurement of distance metric is calculated from each data average [16]. Whereas K-Medoids uses Medoid (median) to become the center point of a cluster. K-Medoids is also called PAM (Partition around Medoids). The K-Medoids algorithm has ability to overcome noise and outliers, where objects with large values have possibility to deviate from data distribution.

Another advantage is that the results of the Clustering process do not depend on the entry sequence of the dataset [18]. Compared to K-Means, K-Medoid is stronger when dealing with noise and outliers because it minimizes a number of paired dissimilarities, not the sum of squares of Euclidean distances. For large database, K-Medoids algorithm such CLARA (Clustering LARge Applications) can be utilized. The effectiveness of this algorithm depends on the sampling size. CLARA uses dataset samples randomly and looking for the best k-Medoids between the chosen sample datasets.

For validating cluster results, internal properties can be used [19]. Measurement criteria for internal validation reflect on compactness, connectivity, and cluster partition separations [20]. Compactness assess the cluster homogeneity and estimating the intra-cluster distance variance. Whereas separation measures the distance between

cluster centroids. These two indices measure two opposite trend which is connectivity increases as the number of cluster add up, while separation is contrary.

Therefore, popular method to combine these two indices are Dunn Index and Silhouette Width. These methods combine compactness and separation non-linearly. Silhouette Value measures the degree of confidence in the clustering assignment [21]. While Dunn Index is the ratio of the smallest distance between observations in different cluster so that maximized intra-cluster distance can be obtained. Silhouette Value and Dunn index should be maximized.

3) Fuzzy C-Means

As an extension of K-Means algorithm, Fuzzy C-Means clustering (FCM) was introduced by Dunn in 1974 and extended by Bezdek in 1981 [22]. It is can also be used for pattern recognition for large data set. According to [23], Euclidean distance is a good distance metric for this technique. FCM also depends on the feature-weights values in the interval [0, 1]. Objects that are located on the boundaries between more than one groups are not forced to belong only in one group. Data points must not exclusively belong to one cluster center and can be assigned membership to other cluster center. So that each data point may belong to two or more clusters with different degrees of membership [24].

This algorithm works by finding a group of fuzzy cluster iteratively. The chosen corresponding cluster center represents the best structure of data. From the research done by [25], FCM able to store more data compared to K-Means algorithm. Fuzzy C-Means algorithm are as follow [26]:

1. Initialize the membership matrix randomly

$$\sum_{j=1}^{c} = \mu_j(x_i) = 1 \quad (1)$$

2. Calculate the Centroid

$$C_j = \frac{\sum_i [\mu_j(x_i)^m x_i]}{\sum_i [\mu_j(x_i)]^m} \quad (2)$$

3. Calculate dissimilarity between the data points and centroid using Euclidean Distance

$$D_i = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \quad (3)$$

4. Update the new membership matrix

$$\mu_j(x_i) = \frac{[\frac{1}{d_{ji}}]^{\frac{1}{m-1}}}{\sum_{k=1}^{c} [\frac{1}{d_{ki}}]^{\frac{1}{m-1}}} \quad (4)$$

5. Repeat step number 2, unless the centroids are not changing.

with,

m = Any real number greater than 1,

μ_{ij}= The degree of membership of x_i in the cluster j,

xi = The ith of d-dimensional measured data,

cj = The d-dimension center of the cluster.

C. Centroid-based Clustering Techniques

After completing clustering process, performance evaluation is needed to compare clustering algorithms. Validation is an evaluation process to find out the goodness of clustering result [27]. As one of the unsupervised learning algorithms, cluster analysis usually use the internal index validation to evaluate the clustering results [27].

There have been developed many internal indices to validate clustering results for working with common centroid-based algorithms. However these internal indices cannot be used for fuzzy clustering results [28]. Several validity indices for fuzzy clustering according to Bezdek (1974) are the Partition Entropy (PE) [27] and Partition Coefficient (PC)[29].

PC and PE are defined as (5) and (6) [27]. By maximizing the number of V_{PC} and minimizing the number of V_{PE} with respect to c = 2,3,…,C_{max}, an optimal partition is reached.

$$V_{PC} = \frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^2 \quad (5)$$

$$V_{PE} = -\frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^2 \, log_a \, u_{ij} \quad (6)$$

For the rest of the other centroid-based clustering techniques, the validity index can be determined by calculating the intra-cluster distance and inter-cluster distance. Intra-cluster distances are used to measure the distance between data vectors to centroid in a cluster. The aim is to have minimum distance. While Inter-cluster distance is the distance between the centroids of the clusters and aim to have maximum distance. These distance are used to evaluate the clustering performance to ensure compactness intra-cluster distance and separation of inter-cluster distance.

III. Experimental Results

Dataset used in this research was obtained from Tech Company in Indonesia which provide Point of Sale system especially for food and beverages industry. The merchants consist of several category such as, coffee shop or café, casual dining restaurant, food stall, etc. This Tech Company runs in B2B settings. Their products are Merchant-Facing and not directly to the end customers. Customers' data from each merchant are collected at the POS system and all the details of transactions will be logged. The data contains product bought from the merchants, exact time and time of the transaction, and the money paid with total 4.572 customer records. All data processing steps are implemented in R*Studio* version 1.1.442.

A. Data Pre-Processing

Data for this research is obtained from data warehouse which has logged all transactional data from merchants in the period of October 2018 – February 2019. The steps taken for data pre-processing are removing data with missing values and outliers. Next step is to define attributes RFM from the datasets. The specified period ends in the beginning of March 2019 following the start of this research. Data normalization using Min-Max normalization is taken to uniform the data range.

B. Selection Number of Clusters

Before the clustering technique is done, the number of clusters must be defined first. Optimal number of clusters is a fundamental part in clustering.

TABLE III. VALIDATION MEASURES FOR PARTITIONING CLUSTERING ALGORITHMS

Algorithms	Validation Measures	Cluster Number				
		2	3	4	5	6
K-Means	Connectivity	13.661	17.395	37.1242	38.812	43.603
	Dunn	0.0438	0.0821	0.0703	0.0664	0.0612
	Silhouette	0.416	0.561	0.428	0.3619	0.332
CLARA	Connectivity	27.619	31.074	36.233	41.128	44.577
	Dunn	0.0291	0.0569	0.0671	0.0753	0.0768
	Silhouette	0.519	0.534	0.483	0.423	0.505
PAM	Connectivity	17.819	26.973	34.971	44.102	47.568
	Dunn	0.0476	0.0439	0.0261	0.02591	0.02113
	Silhouette	0.401	0.506	0.418	0.543	0.493

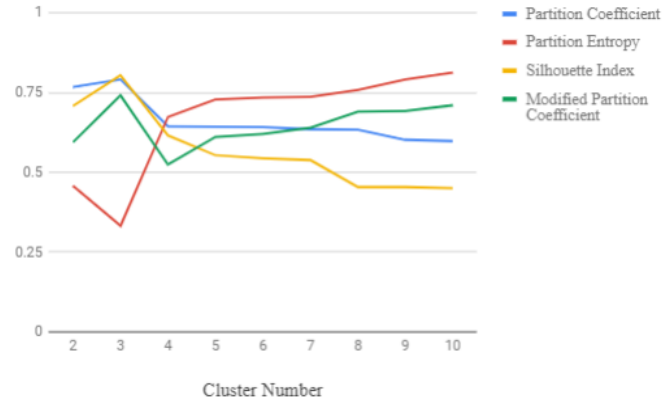

Fig. 2. Validation Indices Fuzzy C-Means

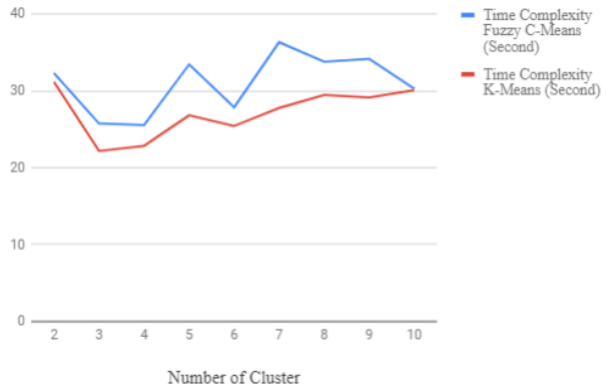

Fig. 3. Time Complexity Comparison

Calculations for k=2 until k=6 were done using K-Means, CLARA, and PAM algorithms. The results are shown in Table 3. The optimal number of clusters seems to be three from the validation measures indices since the highest Dunn score is obtained by K-Means clustering when k = 3 and also the highest Average Silhouette Width when k = 3.

C. *Comparison Between K-Means Clustering and Fuzzy C-Means Clustering*

Among three clustering algorithms, the result shows that the optimum validation measures scores are obtained mostly by K-Means. Therefore K-Means algorithm will be compared to Fuzzy C-Means algorithm. The experimental results of Fuzzy C-Means clustering are shown in Fig. 3. It shows the plots of validation indices for Fuzzy C-Means. The highest partition Coefficient Index, silhouette index, and modified partition coefficient are obtained when k = 3. Also, the lowest Partition Entropy Index is obtained when k is equal to three.

Since the optimum number of clusters obtained by K-Means and FCM algorithm are the same, both algorithms are compared using time complexity. According to [30] time complexity states how long an algorithm runs at runtime based on the input given. We can conclude from the results that K-Means algorithm takes less time to compute the data compared to FCM algorithm. FCM has closer results to K-Means clustering but still needs more time to do computations since Fuzzy measures calculations involvement in the algorithm. Based on the obtained result, it can be safely stated that K-Means clustering performs better in terms of time complexity compared to FCM. Therefore, further analysis about customers' data in each cluster is done using K-Means algorithm.

IV. ANALYSIS AND DISCUSSION

The total of customers in cluster 1 is 1626 merchants, total customers in cluster 2 is 2503 merchants, and the total customers in cluster 3 is 443 merchants. Means for each attribute can be used to give RFM score to each cluster. The RFM score given to each cluster is based on Table 2. The customer segmentation results and the descriptions are shown in Table 4 and Table 5.

RFM score obtained by each cluster are then categorized into two types. If the attribute score is equal to or greater than 3, then the attribute will be denoted as (↑) or the value is greater than average. While sign (↓) denotes that the value of an attribute is less than 3 or smaller than the group's average. For the group description, referring to customer loyalty matrix (Fig.1) from [31] and [14], they defined 4 types of customers.

From the three clusters, there are 3 different RFM status. All clusters have average Recency score above the average. The most valuable customers or core customers group is obtained by *cluster* three with high Recency, high frequency, and high monetary. Customers belong to this group often use the POS machine for transaction and give the biggest value to company called as core customers [32].

TABLE IV. AVERAGE RFM ATTRIBUTES AND RFM STATUS

Cluster	Customer Number	Recency (Avg.)	Score	Frequency (Avg.)	Score	Monetary (Avg.)	Score	RFM Status
1	1626	13.2	5	682.2	1	586,777.2	3	↑↓↑
2	2503	9.2	5	686.9	1	493,768.6	2	↑↓↓
3	443	7.6	5	7569.5	3	689,714.8	4	↑↑↑
Total Average		10.0		1646.2		590,086.9		

TABLE V. CUSTOMER GROUP DESCRIPTIONS

Customer Cluster	RFM Status	Group Description
1	R↑F↓M↑	Recent purchases with low frequency of transactions but big spender with high monetary values; Belong to potential customers
2	R↑F↓M↓	Loyal customers with low frequency of transactions and low monetary value; Belong to lost customers
3	R↑F↑M↑	High value customers; Frequent transactions and high monetary value; Belong to Core customer group

Customers in *cluster* 2 have Recency above the average, Frequency below the average, and monetary below the average. This group is labeled as New Customers since they not done many transactions and the money paid to the company is relatively low. However, customers in this group can be converted into potential customers, since their Recency score is high.. Meaning that the usage of POS machine is used constantly every day.

Lastly, *cluster* 1 consists of 1626 customers with high recency above average and high monetary value. However, the frequency is below the average. This group is labeled as potential customers. Although they have low frequency score, this group is considered as profitable group. Based on [33] potential customer group can be converted in golden or core customers with attractive product offers to increase the number of transactions.

V. CONCLUSION

Company needs to identify the characteristics of their customers' needs, preferences, views, and opinions about the service or products given. Common approach to understand customer behavior is RFM Analysis which consisted of three predefined variables obtained from customers' transactional data. Dataset is acquired from Tech Company in Indonesia that provide machine with Point of Sale system for food and beverages merchants, since the company in B2B settings. In previous data mining research, the domain has not yet been investigated well. In this paper, a novel approach is done to compare various clustering techniques to discover which algorithm performs best.

The purpose of this research is to compare comprehensive centroid-based clustering algorithms with Fuzzy C Means Algorithm to search for the best clustering results. Number of clusters k is defined through internal cluster validity indices using partitioning clustering algorithm (K-Means, CLARA, and PAM). From the results obtained, the optimum number of cluster is equal to three based on connectivity, Dunn Index, and Silhouette Coefficient. Validation measures are taken into account to find out which algorithm performs the best. Since the best validation measure score is obtained by K-Means, for the next step of experiment, it is used to be compared to Fuzzy C-Means algorithm. Experimental results of clustering using Fuzzy C-Means from $k=2$ to $k=10$ shows that the best validation indices score is obtained when $k=3$.
Algorithm with optimum validation measures, will be compared to Fuzzy C- Means based on time complexity.

The comparison of time complexity resulted in K-Means algorithm outperforms Fuzzy C-Means. Therefore, cluster analysis was done using the best clustering algorithm, K-Means. Once the clusters are formed, customer segments are analyzed and evaluated to gain insights about the behavioral propensity of each customer segment. Further details about customer information from each cluster enable company to develop differentiated strategy to specific group of customers.

For future work, further analysis can be done. In B2B settings, company have different characteristics of customers compared to B2B settings, therefore work need to be improved. Other variables such as residence and business category can be included for customer segmentation and might be one of the way for company to understand their customers.

ACKNOWLEDGMENT

Authors would like to express gratitude and appreciation to Universitas Indonesia for funding this research through PIT-9 Research Grants Universitas Indonesia No: NKB-0061/UN2.R3.1/HKP.05.00/2019

REFERENCES

[1] C. W. Lamb, J. F. Hair, and C. McDaniel, "Essentials of Marketing", Lachina Publishing Services, 2011

[2] S. Bhayani, "Internet Marketing Vs. Traditional Marketing: A Comparative Analysis", FIIB Business Review. Volume 3 (3), pp. 53-63, 2018

[3] H. J. Watson, "Tutorial: big data analytics: concepts, technologies, and applications, Community Association Information System". 34 (65), pp. 1247–1268, 2014

[4] F. Ohlhorst, "Big data analytics: turning big data into money", 2013

[5] Ishwarappa and J. Anuradha, "A Brief Introduction on Big Data 5Vs Characteristics and Hadoop technology", *Procedia Computer Science*, 48, pp. 319-324, 2015.

[6] K. Tsiptsis and A. Chorianopoulos,"Data mining techniques in CRM: Inside customer segmentation", Hoboken, NJ: Wiley, 2010

[7] X. Zhou, Z. Zhang, and Y. Lu, "Review of customer segmentation method in crm," in *Computer Science and Service System* (CSSS), pp. 4033–4035, 2011.

[8] A. Hiziroglu, "Soft computing applications in customer segmentation: State-of-art review and critique," *Expert Systems with Applications*, vol. 40, no. 16, pp. 6491–6507, 2013.

[9] A. Joshi, and R. Kaur,"A Review: Comparative Study of Various Clustering Techniques in Data Mining", *International Journal of Advanced Research in Computer Science and Software Engineering*, 3(3), pp. 55-57, 2013.

[10] G. Gupta, "Data Mining: An Introduction – Case Study. Introduction to Data Mining and Its Applications Studies in Computational Intelligence", pp. 217-229, 2006.

[11] S. Lloyd, "Least squares quantization in PCM," in IEEE Transactions on Information Theory, vol. 28, no. 2, pp. 129-137, 1982.

[12] J.B. MacQueen, "Some Methods for classification and Analysis of Multivariate Observations", *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281–297, 1967.

[13] A.J. Christy, A. Umamakeswari, L. Priyatharsini, A. Neyaa, "RFM Ranking – An Effective Approach to Customer Segmentation", *Journal of King Saud University - Computer and Information* Sciences, 2018

[14] D. A. Kandeil, A.A. Saad, and S. M. Youssef, "A Two-Phase Clustering Analysis for B2B Customer Segmentation", *International Conference on Intelligent Networking and Collaborative Systems*, 2014

[15] A. Parvaneh, H. Abbasimehr, and M. Tarokh, "Integrating AHP and data mining for effective retailer segmentation based on retailer lifetime value," Journal of Optimization in Industrial Engineering, vol. 5, no. 11, 2012, pp. 25–31.

[16] A. K. Mann and N. Kaur, "Survey Paper on Clustering Techniques", *International Journal of Science, Engineering and Technology Research* (IJSETR) Volume 2, Issue 4, 2013.

[17] K. Singh, D. Malik, N. Sharma, "Evolving limitations in K-means algorithm in data mining and their removal", IJCEM International *Journal of Computational Engineering & Management*, 2011.

[18] M.T Furqon, L. Muflikhah., "Clustering the potential risk of tsunami using Density-Based Spatial clustering of application with noise (DBSCAN)", *Journal of Environmental Engineering & Sustainable Technology (JEEST)* Vol. 03 No. 01, pp. 1-8, 2016.

[19] S. Datta and S. Datta, "Methods for evaluating clustering algorithms for gene expression data using a reference set of functional classes", *BMC Bioinformatics,7* (1), 2006.

[20] J. Handl, J. D. Knowles, D. B. Kell. "Computational Cluster Validation in Post-Genomic Data Analysis". *Bioinformatics*, 21(15), pp. 3201–12, 2005.

[21] G. Brock, V. Pihur, S. Datta, S. Datta, "ClValid: AnRPackage for Cluster Validation." *Journal of Statistical Software,25*(4), 2008.

[22] X. Wang, Y. Wang, and L. Wang, "Improving fuzzy c-means clustering based on feature-weight learning." *Pattern Recognition Letters,* 25(10), pp. 1123-1132, 2004.

[23] A. Dik, A. El Moujahid, A. Bouroumi, A. Ettouhami, "Weighted distances for fuzzy clustering." *Applied Mathematical Sciences*. Vol.8, pp. 147-156, 2014.

[24] T. Singh and M. Mahajan. "Performance Comparison of Fuzzy C Means with Respect to Other Clustering Algorithm." *International Journal of Advanced Resarch in Computer Science and Software Engineering.* Vol 4(5), 2014.

[25] Y. Zhang, D. Huang, M. Ji and F. Xie, "Image segmentation using PSO and PCM with Mahalanobis distance." *Expert Systems with Applications*, 38 (7), pp. 9036-9040, 2011.

[26] R. Shanthi and R. Suganya, "Enhancement of Fuzzy Possibilistic C-Means Algorithm using EM Algorithm (EMFPCM)." *International Journal of Computer Applications, 61*(12), pp. 10-15, 2013

[27] J. C. Bezdek†, "Cluster Validity with Fuzzy Sets." *Journal of Cybernetics* 3(3), pp. 58-73, 1973.

[28] M. B. Ferraro and P. Giordani, "A toolbox for fuzzy clustering using the R programming language." *Fuzzy Sets and Systems,* 279, pp. 1-16, 2015.

[29] J. C. Bezdek. "Numerical taxonomy with fuzzy sets". *Journal of Mathematical Biology*, vol. 1, pp. 57–71, 1974.

[30] P. I. Dalatu, "Time Complexity of K-Means and K-Medians Clustering Algorithms in Outliers Detection". *Global Journal of Pure and Applied Mathematics*. ISSN 0973-1768 Volume 12, Number 5, pp. 4405–4418, 2016.

[31] H. H. Chang and S. F. Tsay. "Integrating of som and k-means in data mining clustering: An emperical study of crm and profitability evaluation," *Journal of Information Management*, vol. 11, no. 4, pp. 161–203, 2004.

[32] D. R. Liu and Y. Y. Shih. "Hybrid Approaches to Product Recommendation based on Customer Lifetime Value and Purchase Preferences". *Journal of Systems and Software*, 77 (2), pp. 181-191, 2005.

[33] S. Babak and K. Amir, "Customer lifetime value (CLV) measurement based on RFM model". Iranian Accounting & Auditing Review. 14, 2007.