JURNAL RISET INFORMATIKA **P-ISSN: 2656-1743 |E-ISSN: 2656-1735**
Vol. 2, No. 2 March 2020 DOI: https://doi.org/10.34288/jri.v2i2.38

A STUDY OF COMPARING CONCEPTUAL AND PERFORMANCE OF K-MEANS AND FUZZY C-MEANS ALGORITHMS (CLUSTERING METHOD OF DATA MINING) OF CONSUMER SEGMENTATION

Yunita[1]; Sukrina Herman[2]; Ahsani Takwim[3]; Septian Rheno Widianto[4]

Magister of Information System
STMIK LIKMI, Bandung, Indonesia
www.stmiklikmi.ac.id
yunita.nambela@gmail.com[1*], sukrinahermanheral01@gmail.com[2],
ahsanitakwim10@gmail.com[3], septian.rheno@likmi.ac.id[4]
*Corresponding Author

Abstract

Consumers, especially potential customers, are an important asset in a company that should be maintained properly. The tight competition requires companies to focus on the customer's needs. Consumer segmentation is one of the processes carried out in the marketing strategy. Consumer or consumer segmentation data mining supports the grouping process results. Based on mapping studies on data mining in support of consumer segmentation, two algorithms are often used: K-means clustering and Fuzzy C-means clustering. The attributes used for mining in customer segmentation processes are customer data, products, demographics, consumer behavior, transactions, RFMDC, RFM (Recency, Frequency Monetary), and LTV (Life lifetime value). It is important to combine the clustering algorithm to algorithm Classification, Association, and CPV to get the potential value of each cluster.

Keywords: Data Mining, Consumer Segmentation, Algorithm, Clustering.

Abstrak

Konsumen merupakan asset penting dalam perusahaan yang harus dijaga dengan baik terutama pelanggan potensial. Persaingan usaha yang ketat mengharuskan perusahaan untuk berfokus kepada kebutuhan yang diinginkan oleh pelanggan. Segmentasi konsumen merupakan salah satu proses yang dilakukan dalam strategi pemasaran. Untuk mendukung hasil yang proses pengelompokan konsumen atau segmentasi konsumen ini maka dukungan data mining sangat berperan penting. Berdasarkan pemetaan penelitian mengenai dukungan data mining pada segmentasi konsumen didapat ada dua algoritma yang sering digunakan untuk segmentasi konsumen antara lain K-Means Clustering dan Fuzzy C-Means clustering. Adapun atribut-atribut yang digunakan untuk proses mining pada segmentasi konsumen adalah data konsumen, produk, demografi, perilaku konsumen, transaksi, RFMDC, RFM (Rerency, Frequency Monetery), dan LTV (Life Time Value). Dan penting untuk menggabungkan algoritma clustering dengan algoritma Classification, Association, dan CPV untuk mendapatkan nilai potensial dari tiap cluster.

Kata kunci: Data Mining, Segmentasi Konsumen, Algoritma, Pengelompokan

INTRODUCTION

Consumers, especially potential customers, are an important asset in a company that should be maintained properly. The tight competition requires companies to focus on the customer's needs. With many customers, a company needs a strategy to define the Company's potential customers by way of grouping customers. Grouping customers based on Their characteristics will affect a company's marketing management. Therefore, to classify customers based on their respective characters needed methods, one of which is Data Mining. Data mining is one of the sciences in the informatics study of data mining, and text documents are a mined science. Clustering is a technique of one of the functionalities of data mining, and a clustering algorithm is an algorithm of grouping several data into groups - a group of specific data.

Data mining is a theory that uses historical data to find a pattern to assist decision-making. The results of the mining process are patterns or patterns. These patterns are analyzed to find new

knowledge useful for managers to make decisions. Support for the decision of the mining process is normally used to help solve strategic problems for ordinary results found in the form of a prediction for the future. Business at present is very tight. Competition is everywhere. Especially with the open market, a company must conduct a process to improve its competitive advantage. One of the ways that we engaged one is to build a system called Business Intelligence (BI). Business Intelligence (BI) is an integrated system that the Company coordinates. BI is built to support decision-making in all areas of the Company, including the top manager. BI combines the data storage and management knowledge to analyze it. The technique is done in between other BI ETL (Extract Transformation Loading), data warehouse, data mining and data visualization such as OLAP (Online Analytical Processing). This paper focuses on the area of data mining techniques in particular marketing strategy consumer segmentation. (Ranjan, 2007)

Clustering is a technique of one of the functionalities of data mining, clustering algorithm is an algorithm of grouping a number of data into groups - specific data group (cluster). There are many algorithms used in clustering, including the algorithm K-Means and Fuzzy C-Means (FCM). FCM is the algorithm clustering where one object can be members cluster and restrictions cluster FCM is sketchy. The basic concept of the FCM first is to determine the center of cluster. And each data point has a degree of membership for each cluster. The degree of membership in FCM algorithm is between 0 and 1. (Bora & Gupta, 2014) Algorithm K-Means is one algorithm clustering which is used to partition the data into some cluster, where the data has a high degree of similarity are grouped in one cluster whereas data that have different characteristics are grouped into cluster different (Elhabbash, 2010)

MATERIALS AND METHODS

On this research conducted a comparison between the clustering of the K-means and fuzzy clink C-means towards consumer segmentation. By comparing some previous research on the concept and performance of both methods of the clustering algorithm. Then after the comparison of the authors determine which method is better in the consumer segmentation process and whether there is a need to be done in improving the performance of the clustering method, or combine both the algorithm and Other data mining methods to get the best consumer segmentation. The following will describe each algorithm's concept and completion size in the clustering method.

The algorithms used in clustering are as follows.

A. K-Means

In the method of clustering, the main concept is emphasized is the central search cluster iteratively, where the center determined based on the minimum distance of each data at the center cluster. Measures completion algorithm K-Means can be described as follows:

1. Identification data will cluster and specify the number of cluster Data, Xij (i = 1,2, n; j = 1,2, m) where n is the number of data to be clustered and m is the number of data variables.
2. At the beginning of each iteration center cluster determined
   Random (Random), C_{kj} (k = 1k; j = 1 m).
3. Determine the distance of each data against center cluster by using Euclidean formula following:

$$\boldsymbol{d_{ik}} = \sqrt{\sum_{j=1}^{m}(\boldsymbol{X_{ij}} - \boldsymbol{C_{kj}})^2} \quad \text{.....(1)}$$

4. Group data based on the minimum distance of data toward the center cluster.
5. Check the condition if no data is still moving in the other group, if he did iteration next to the counting center cluster based on the average value of the data which are members of cluster which formed the previous iteration results using the formula:

$$\boldsymbol{C_{kj}} = \frac{\sum_{i=1}^{p} \boldsymbol{x_{ij}}}{\boldsymbol{p}} \quad \text{..... (2)}$$

   Where X_{ij} is cluster to-k and number of cluster.

6. If no more data is moved on cluster the other, then the iteration process stops. In algorithm K-Means the data are regrouped on a cluster if the data has a minimum distance to the center cluster which can be calculated using the formula (2).

B. Fuzzy C-Means

Fuzzy C-Means is one of the methods fuzzy clustering who first developed by (Dewangan & Ambhaikar, 2013) and later amended by (Dulyakarn & Rangsanseri, 2001) as a method often used in pattern recognition (pattern recognition). FCM clustering in the data was based

on the degree of membership value between 0 and 1.

In the FCM algorithm the first step is to determine the center of the cluster that will mark the average location for each cluster. In the initial condition, the cluster center is still inaccurate. Each data has a degree of membership for each cluster. By improving the cluster center and the membership value of each data repeatedly, it can be seen that the cluster center will go to the right location.

This loop is based on the minimization of the objective function that describes the distance from the data supplied to centers cluster which is weighted by the degree of membership of the data points. The output of the FCM is not a fuzzy inference system, but it is a center row cluster and some degree of membership for each data point. Steps to resolve the algorithm FCM algorithm in the first step Fuzzy C-Means can be described as follows:

1. Identification data will cluster in the form of a matrix size nxm (Xij = sample data to i (i = 1,2, ... n) and attribute j (j = 1,2, ... m).
2. Determine the number cluster (c), rank (w), maximum iteration (maxiter) error the smallest expected (•), the initial objective function (P0 = 0 and the initial iteration (t = 1).
3. Generating numbers random μik (i = 1,2, ..n; k =1,2, .., c) as elements of the initial partition matrix μ. Count the number of each column with the formula:

$$\boldsymbol{Q_j} = \sum_{k=1}^{c} \boldsymbol{\mu_{ik}} \quad \text{.....(3)}$$

Calculate the value of the matrix element partitioning member of the set U by the formula:

$$\boldsymbol{\mu_{ik}} = \frac{\boldsymbol{\mu_{ik}}}{\boldsymbol{Qi}} \quad \text{.....(4)}$$

4. Calculating the value of the center cluster all k (V_{kj} with k = 1,2, .., c; j = 1,2, ..., m)
5. Calculate the objective function at the iteration to-t (Pt):

$$\boldsymbol{P_t} = \sum_{i=1}^{n} \sum_{k=1}^{c} ([\sum_{j=1}^{m} (\boldsymbol{X_{ij}^{-(2.6)}} - \boldsymbol{V_{kj}})^2](\boldsymbol{\mu_{ik}})^{\boldsymbol{w}}) \quad \text{..... (5)}$$

6. Counting of change the partition matrix μ.:

$$\boldsymbol{\mu_{ik}} = \frac{[\sum_{j=1}^{m} (x_{ij} - \boldsymbol{V_{kj}})^2]^{\frac{-1}{w-1}}}{\sum_{k=1}^{c} [\sum_{j=1}^{m} (\boldsymbol{X_{ij}} - \boldsymbol{V_{kj}})^2]^{\frac{-1}{w-1}}} \quad \text{.....(6)}$$

7. Check the condition of stopping:
   a. If (Pt - Pt-1) <• or (t <Max Iter) then the iteration stops.
   b. If not: t = t + 1, repeat steps 4 to 7.

In the algorithm Fuzzy C-Means the data is grouped in a cluster, if the data has a maximum distance value partition matrix U (μik) towards the center cluster.

RESULT AND DISCUSSION

Several previous studies on the utilization of the K-Means algorithm and Fuzzy C-Means on consumer segmentation process, where the results of the study stated that the process of customer segmentation using K-Means algorithm and Fuzzy C-Mean The successful and unsuccessful. Following a discussion of some previous research.

A. Utilization of The Algorithm K-Means Clustering on Customer Segmentation

In (Amborowati & Winarko, 2014) Most consumer segmentation process uses an algorithm K Means Clustering, and the results declared that successful and effective segmentation. K-Means Clustering Algorithm has several advantage, among others, provide a good solution to the problem of clustering for data objects that have a numeric attribute, relative scalable and efficient in processing a large data set, the algorithm is not sensitive to input the data, including fast algorithms in modeling and easy to understand. Perform customer segmentation algorithm using K Means Clustering with RFMDC attributes. RFMDC is the development of RFM models (recency, frequency, monetary) used to do segmentation consumer Where development is diversity and continuousness.

In (Shashidhar & Varadarajan, 2011) to segment consumers by uses two data mining algorithms, the first to do is to segment with Neural network to have equal type at the time of sample data used in the different categories. Results from the first step is used as input in the K-Means algorithm Clustering.

The use of K-Means (Khajvand & Tarokh, 2011), where for customer segmentation using K Means Clustering and a Decision Tree for the classification. The first step was to determine RFM parameters by making scale of based transactions, demographics, and products. The second step of the process using K means clustering segmentation based on the results of FRM. As with the (Amborowati & Winarko, 2014) in his paper did consumer segmentation with using the two-step.

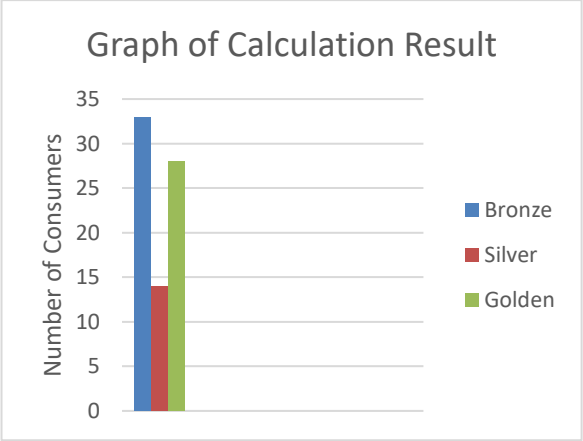
The first step uses K-Means Clustering to produce cluster based on the demographic and step both using Neural Network to calculate the potential value of each cluster.

While Li, 2010, do customer segmentation K-Means Clustering for get four consumer groups, and predict using C5.0, Neural Network, Chi squared automation interaction detector (CHAID). The result is a new consumer and potential value.

B. Utilization Of Fuzzy C-Means Algorithm Clustering On Consumer Segmentation

Research by Megawati, Mukid, and Rahmawati about the use fuzzy c-means algorithm for market segmentation. Here, researchers used algorithms fuzzy c means for classifying consumers into two clusters based on 10 variables psikografik. Data were obtained through questionnaires Pasaraya RITA Cilacap. Results of research on this study show that consumer segmentation is divided into 2 cluster. respondents in cluster 1 pay attention to the level of low prices, the completeness of goods - goods, large discounts, service when shopping satisfying, convenient location, spacious parking, convenience when shopping, adequate public facilities, facilities complete payment, and the cleanliness of the room compared to respondents in cluster 2. With this research, the right target market can be applied to a self-service by case studies taken by the researcher.

Dewi Astria and Suprayogi researched the application of the fuzzy c-means algorithm Clustering in CV Mataram Jaya Bawen. This study uses Fuzzy C-Means (FCM) in customer data processing transactions. The calculation algorithm consists of initialization data such as membership data, calculating a centroid, calculating the distance to the centroid of data, calculating the value of membership, and calculating the value of the objective function. The objective function value affects the iteration because if the change has not reached the objective function value, the smallest positive value iteration will continue to do so. Once the objective function value reaches a small positive value, cluster data can be determined. End-formed clusters show that the number of consumers is 28 golden, 14 silver, and 33 bronze. Here are the results of the graphical display that shows the number of customers in a cluster and makes CV Mataram Jaya Bawen to monitor customers.



Source: (DewiAstria, 2017)

Figure 1. Cluster Number of customers CV Mataram Jaya

To determine the validity of cluster in formation cluster customers on the system, Dewi Astria and Suprayogi calculate partition coefficients or partition coefficient (PC) to evaluate the value of membership data at each cluster. Values greater (closer to 1) means that the quality cluster obtained the better. results to calculate the accuracy of the PCI Index cluster is 0.596277. In the study, the Fuzzy C-Means algorithm works well in clustering customers. The algorithm classifies customers into three clusters (golden, silver and bronze), with the accuracy of the cluster being 0.596277, which means a pretty good degree of accuracy.

C. The attributes used for the segmentation of consumer

At the time of clustering, it will be used attributes for the mining process. The attributes that are often used to segment customers include consumer data, demographic factors, type of product, consumer behaviour, consumer transactions to the product, RFM, (RFMDC (recency, frequency, monetary, diversity and continuousness) and LTV (lifetime value), value information, and behaviour information.

Table 1. Attributes are used to perform customer segmentation

No	Author Name	Year	Atribut
1	Lin, Jian Bang., Liang Tehsin, Lee, Yong-goo, 2012	2012	Demografi, Product
2	Ye, Luo., Qiu-ru, Cai., Etl, 2012	2012	Customer: identifikasi konsumen, metode kontak,

No	Author Name	Year	Atribut
			waktu akses
			Value Information: Biaya bulanan individu atau perusahaan
			Behavior Information: durasi telepon, waktu telepon
3	lu, ke., furukawa, tetsuya, 2012	2012	Customer, Transaksi
4	hajiha, Ali, Radjar, reza, malayeri., samira s., 2011	2011	RFMDC
5	ren, shuxia, sun qiming, shi, yuguang., 2010	2011	Customer, Customer Account, Pinjaman
6	chang-shun, yan, yu-liang, shi, yuan-yuan, sun, 2011	2011	Customer, Penyakit
7	zhou, shuwen, Lei, guanghong, 2011	2011	Customer
			Data CRM: umur, jenis kelamin, pendidikan, pekerjaan, status pernikahan, pendapatan
8	bonsnjak, z., grljevic, o., 2011	2011	Customer, Pinjaman
9	yao, Zhiyuan, eklund, thomas., back, barbro, 2010	2010	Demografi, Perilaku konsumen
10	namvar, morteza, glolamian, M.R, 2010	2010	RFM (Ferency, frequency monetary) Demografi, LTV (lifetime value)
11	yu, gu, jiahui, li, 2010	2010	customer value
12	Bi, jianxin, 2010	2010	customer value, product, pangsa pasar
13	wang, wei, fan, shidong, 2010	2010	customer, transaksi
14	Yihua, Zhang, 2010	2010	customer, product, transaksi
15	Li, We, Wu, Xuemei, 2010	2010	Application Source, Number Credit Card
			frekuensi penggunaan, demografi konsumen

Source: (Amborowati & Winarko, 2014)

CONCLUSIONS AND SUGGESTION

Consumer segmentation is fundamental to the marketing strategy. To support the process of grouping the results of the consumer or This consumer segmentation then support data mining is very important. The data mining algorithm that is most appropriate and often used for segmentation is a K-Means Clustering and Fuzzy C Means. Advised on the comparative performance of the two algorithms are the grouping to do the merge customer data clustering algorithm (K means Clustering and Fuzzy C-Means) with some data mining algorithms such as Classification, Association, and CPV matrix to obtain the potential value of each cluster. The attributes used to process mining on consumer segmentation are customer data, product, demographics, consumer behaviour, transactions, RFMDC, RFM (Recency, Frequency Monetary) and LTV (Life Time Value).

REFERENCES

Amborowati, A., & Winarko, E. (2014). Review Pemanfaatan Teknik Data Mining Dalam Segmentasi Konsumen. *Prosiding Seminar Ilmiah Nasional Komputer Dan Sistem Intelijen (KOMMIT 2014)*, *8*(Kommit), 66–73.

Bora, D. J., & Gupta, D. A. K. (2014). A Comparative Study Between Fuzzy Clustering Algorithm and Hard Clustering Algorithm. *International Journal of Computer Trends and Technology*, *10*(2), 108–113. https://doi.org/10.14445/22312803/ijctt-v10p119

Dewangan, O., & Ambhaikar, P. A. (2013). Extended Fuzzy C-Means Clustering Algorithm in Segmentation of Noisy Images. *International Journal of Scientific Engineering and Research (IJSER)*, *1*(1–3), 16–19.

DewiAstria, S. (2017). Penerapan Algoritma Fuzzy C-Means Untuk Clustering. *Universitas Dian Nuswantoro Semarang*, *6*(Maret 2017), 169–178.

Dulyakarn, P., & Rangsanseri, Y. (2001). Fuzzy c-means clustering using spatial information with application to remote sensing. *22nd Asian Conference on Remote Sensing*, *2*(November), 2–5.

Elhabbash, A. H. (2010). *Enhanced K-means Clustering Algorithm*. The Islamic University of Gaza. Retrieved from

https://iugspace.iugaza.edu.ps/handle/20.500.12358/18772

Khajvand, M., & Tarokh, M. J. (2011). Analyzing Customer Segmentation Based on Customer Value Components (Case Study: A Private Bank) (Technical note). *Journal of Industrial Engineering*, 79–93. Retrieved from https://www.sid.ir/en/journal/ViewPaper.aspx?id=351714

Ranjan, J. (2007). Applications of Data Mining Techniques in the Pharmaceutical Industry. *Journal of Theoretical & Applied Information Technology*, *3*(4), 61–67.

Shashidhar, H., & Varadarajan, S. (2011). Customer Segmentation of Bank based on Data Mining – Security Value-based Heuristic Approach as a Replacement to K-means Segmentation. *International Journal of Computer Applications*, *19*(8), 13–18.