Integration K-Means Clustering Method and Elbow Method For Identification of The Best Customer Profile Cluster

IOP Conf. Series: Materials Science and Engineering **336** (2018) 012017 doi:10.1088/1757-899X/336/1/012017

Integration *K-Means Clustering Method and Elbow Method* For Identification of The Best Customer Profile Cluster

[1]M A Syakur, [2]B K Khotimah, [3]E M S Rochman, [4]B D Satoto
[1,2,3,4]Faculty of Engineering, University of Trunojoyo Madura

Abstract. Clustering is a data mining technique used to analyse data that has variations and the number of lots. Clustering was process of grouping data into a cluster, so they contained data that is as similar as possible and different from other cluster objects. SMEs Indonesia has a variety of customers, but SMEs do not have the mapping of these customers so they did not know which customers are loyal or otherwise. Customer mapping is a grouping of customer profiling to facilitate analysis and policy of SMEs in the production of goods, especially batik sales. Researchers will use a combination of K-Means method with elbow to improve efficient and effective k-means performance in processing large amounts of data. K-Means Clustering is a localized optimization method that is sensitive to the selection of the starting position from the midpoint of the cluster. So choosing the starting position from the midpoint of a bad cluster will result in K-Means Clustering algorithm resulting in high errors and poor cluster results. The K-means algorithm has problems in determining the best number of clusters. So Elbow looks for the best number of clusters on the K-means method. Based on the results obtained from the process in determining the best number of clusters with elbow method can produce the same number of clusters K on the amount of different data. The result of determining the best number of clusters with elbow method will be the default for characteristic process based on case study. Measurement of k-means value of k-means has resulted in the best clusters based on SSE values on 500 clusters of batik visitors. The result shows the cluster has a sharp decrease is at K = 3, so K as the cut-off point as the best cluster.

1. Introduction

Clustering is in daily life, because it could not be separated with a number of data that produce information to meet the needs of life. One of the most important tools in relation to data is to classify or classify the data into a set of categories or clusters [1]. One clustering technique is the method of K-means algorithm using the process repeatedly. The K-Means method is the simplest and most common clustering method. K-means has the ability to group large amounts of data with relatively fast and efficient computation time [2]. However, K-Means has a disadvantage depending on the initial cluster center determination. K-Means cluster test results in the form of solutions that are local optimal. From the trial process is expected to have similarities or closeness between data so that can be grouped into several clusters, where among cluster members have a high level of similarity [3]. According to (Celebi et al., 2013) the K-Means algorithm is also versatile, which is easy to modify at every stage of the process, simple in the distance calculation function, and depends on iteration termination criteria. K-Means Clustering is a localized optimization method that is sensitive to the selection of the starting position from the midpoint of the cluster. So choosing the starting position from the midpoint of a bad

cluster will result in K-Means Clustering algorithm being trapped in the optimal locale [4]. The K-means method will choose the pattern up to k as the starting point of the centroid randomly or randomly. The number of iterations with the centroid cluster will be affected by the initial centroid cluster at random. So that can be fixed by determining the centroid cluster in the high initial data to get higher performance [5]. Consumer segmentation based on consumer behaviour measured by consumer profiling. Customer profiling is built through information with criteria: age, gender, and residential area information. To process customer profiling using data mining by segmenting consumers can be estimated through the consumer profile, which conducted using clustering algorithm for bank customer segmentation [6]. Kaur et al, 2013 proposed improvements to the classic K-Means algorithm to produce more accurate clusters. The proposed algorithm is based on data separation, to find the initial centroid according to the data distribution. The results of this study have resulted in better clusters in a short calculation time. The main objective of this research is the segmentation of bank customers to find the transactional relationship between customer and company to provide mutual solution [14]. In other words, improving customer relationships and evaluating customer segments by predicting credentials for each group of customers and will provide a more appropriate type of transaction model with the customer [6-7]. Clustering results vary depending on the number of clustering parameter changes. The k-means method is a simple clustering technique and quickly takes care of the problem to determine the exact number of clusters in the data set. Customer segmentation research is proposed for the use of various regulations for various customers with high risk to the bank [8]. Elbow proposes several ways to determine k as the number of dynamically formed clusters, one of which is the elbow method [9].

This research will combine K-Means with elbow method to determine the actual number of clusters Customer profiling segmentation in SME. The value on k will continue to increase in each process and decrease with great value. The graph shows the elbow n of all the k values obtained. This research will find the best value of k by using elbow method. The elbow method is easy to implement by looking at the ideal k value graph with the position on the elbow along with the SSE (Sum of Square Error) which is less than 1. The best cluster k result will be the basis for clustering. The smaller the value of SSE and the elbow graph decreases the better the cluster results.

2. Literature Review

2.1. Customer Segmentation

One of the ways used to know customer characteristics information by doing customer segmentation. One way that can be used to manage the relationship between customers and companies is to provide different treatment according to customer characteristics of each segment. Customer segmentation is done with data mining to know the customer characteristics information hidden inside. The way to find out the customer segments of a company is clustering analysis. Clustering is the process of forming segments of a set of data by measuring similarities between data with other data [10]. The purpose of segmentation is to customize the products, services, and marketing messages for each segment. The segmentation process puts customers in line with the characteristics of similar customer groups. Customer segmentation is a preparatory step to classify each customer according to a defined customer group. Customer segmentation based on market research and demography requires understanding the characteristics of all customers to be more effective. Customer segmentation also identifies segmentation in customer behaviour. In addition, customer payment transaction data is used to gain insight into customer behaviour. Customer segmentation to form groups based on their income and expenses. It can be used to identify high-value customers and prioritize service [11].

2.2 K-Means Clustering

The K-means algorithm is one of the algorithms with partition, since K-Means is based on determining the initial number of groups by defining the initial centroid value [12]. The K-Means algorithm requires precise numbers in determining the number of clusters k, since the initial cluster centre may change so that this event may result in unstable grouping of data [13]. The output of K-Means depends on the selected centre values on clustering. This algorithm the initial value of the cluster's centre point becomes the basis for the cluster determination. The initial cluster centroid cluster randomly assigns an impact to the performance of the cluster (14-16). K-Means Clustering algorithm is one of the clustering methods by partitioning from set data into cluster K. It is a distance-based clustering algorithm that divides data into a number of clusters in numerical attributes.

1. Determine the number of clusters K and the number of maximum iterations.
2. Perform the initialization process K midpoint cluster, then the equation of centroid count feature:

$$C_i = \frac{1}{M}\sum_{j=1}^{M} x_j$$

   Equation 1 is done as much as p dimensions from i = 1 to i = p
3. Connect any observation data to the nearest cluster. Euclidean distance spacing measurements can be found using equation 2.

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

4. Reallocation of data to each group based on comparison of distance between data with each group's centroid [9].

$$a_{ij}\begin{cases}1 & d = \min\{D(x_i, c_i)\\ 0 & otherwise\end{cases}$$

5. Recalculate the cluster midpoint position.
   a_{ij} is the value of the membership of point x_i to the centres of the group c_1, d is the shortest distance from the data x_i to the group K after being compared, and c_1 is the centre of the group to 1. The objective function used by this method is based on the distance and the value of the data membership in the group. The objective function according to MacQueen (1967) can be determined using equation.

$$J = \sum_{i=1}^{n}\sum_{l=1}^{k} a_{ic} D(x_i, c_1)^2$$

   n is the amount of data, *k* is the number of groups, a_{i1} is the membership value of the data point x_i to the c_1 group followed *a* has a value of 0 or 1. If the data is an anngota of a group, the value $a_{i1} = 1$. If not, the value $a_{i1} = 0$.
6. If there is a change in the cluster midpoint position or number of iterations <the maximum number of iterations, return to step 3. If not, then return the clustering result.

2.3 Elbow Criterion

Illustration of K value on Elbow combination with K-Means was graph of cluster relationship with error decreasing, increasing value of K then graph will decrease slowly until result of value of K is stable. For example, the value of the cluster K = 2 to K = 3, then from K = 3 to K = 4, it shows a drastic decrease to form the elbow at point K = 3 then the ideal cluster k is K = 3 [11]. The combined Elbow and K-Means Methods can determine the value of K at the best cluster.

1. Find k as the number of clusters formed. This study will use the elbow criterion method to select the number of k clusters to be used for grouping data on the K-Means algorithm. The elbow method is expressed by Sum of Squared Error (Irwanto, et al, 2012):

$$SSE = \sum_{K=1}^{K} \sum_{x_i \in S_K} \|X_i - C_k\|_2^2$$

With k = many clusters formed Ci = the i-th cluster, x = the data present in each cluster.

2. Determine the cluster's center point at the beginning at random. Early centroid determination is done randomly from the available objects as much as cluster k, then to calculate the next i-cluster centroid, by the following formula:
   $v = \frac{\sum_{i=1}^{n} x_i}{n}$; $i = 1,2,3, \ldots, n$
3. Calculate the distance of each object to each centroid using the Euclidian Distance.
   $d(x, y) = \|x - y\| = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$; $i = 1,2,3, \ldots, n$
   with x_i : Variable on Object x to-i and y_i : Variable output y
   n : The number of objects
4. Allocate each object into the nearest centroid.
5. Allocation of objects into each cluster at iteration with k-means. Where each cluster member object has been measured the proximity distance to the cluster's center point.
6. Perform iteration, then process determine the position of new centroid by using equation (2.2).
7. Repeat step 3 if the new centroid position with the old centroid is not the same.

3. **Results and Discussion**

K-means clustering research is done by taking data during the last 1 month for profiling customer with parameters as attribute in the form of customer criteria. Where each criterion has a predetermined category range, for gender criteria and activity has sub criteria, following sub criteria for gender and activity. Customers have been filled out the questionnaire using 6 criteria, according to table 2.

Table 1. Sampel data profiling of customer

No	User Name	Age	Profession	Income (million)	Education	Quality	Gender
1	Moawi	25	Employee	1.500.000	high school	3	Female
2	Windaryanti	47	Teacher	7.000.000	S1	1	Male
3	Sulistiara	60	Pension	3.000.000	S1	5	Female
4	Mailinda	70	Entrepreneur	12.000.000	Junior high school	6	Female
5	Fania Dara	56	Farmer	1.000.000	primary school	4	Male
6	Fleni Dwi	63	Government employees	6.000.000	S1	2	Female
7	Imroatun	48	Trader	11.000.000	primary school	6	Female
8	Marjatul	20	Student	500.000	high school	1	Male
9	Yayuk	29	Employee	6.000.000	D3	3	Female
10	Siti Zainab	42	Teacher	5.000.000	S1	1	Female

Table 2. Process of data transformation with age group

No.	Age	Frequency/Initial
1	15-25	14
2	25-35	13
3	35-45	2
4	45-55	31
5	55-65	25
6	x≥65	16

The data contained in table 1 would not be directly processed, because there is a large number of numbers between the variables. The difference in distance or the magnitude of this number can be quite difficult in the process of grouping. One of the solutions used to minimize the number of variables between the variables with the equation 5 [11].

$$\text{Normalization}\,\frac{(x_i - x_{min})}{(x_{max} - x_{min})}$$

The value of the variables is normalized to the range 0 - 1. Normalization of numbers on each variable before the calculation process is done so that the centroid value on the parameter does not exist that predominates in the calculation of the distance between data [11]. The data used in the form of data on the number of visits batik sales with criteria : Profession, Income, Quality of Batik, and Gender Education. The result combined system test of K-Means method with Elbow according to Table 3.

Table 3. Result of anaisa cluster of K-Means data method with Elbow

No	Age	Profession	Income (million)	Education	Quality	Gender	Cluster	Number
1	25≤x≤40	Employee, Teacher	x≤2	High School	3	Female	3	132
2	41≤x≤55	Teacher, Entrepreneur	5≤x≤10	S1	1	Male	2	223
3	15≤x≤24	Student	2≤x≤5	S1	5	Female	3	112
4	x≤65	Pension, Entrepreneur	x≥10	Junior High School	6	Female	1	25

Furthermore, this research performs the best cluster optimization on K-means with Elbow method. The determination of k value with Elbow Method by choosing the number of clusters or k values by seeing the decrease significantly.

Table 4. Sum of Square Error results from each cluster for 100 and 300 Customer Profiling

Cluster	Result of Sum Square Error for 100 data	The difference between SSE	Result of Sum Square Error for 300 data	The difference between SSE
C2	845,432	845,432	786,563	786,563
C3	532,133	313,299	665,434	121,129
C4	431,429	100,704	301,341	364,093
C5	367,232	64,197	234,453	66,888
C6	334,451	32,781	278,842	-44,389
C7	243,341	91,110	289,678	-10,836
C8	231,211	12,130	234,563	55,115

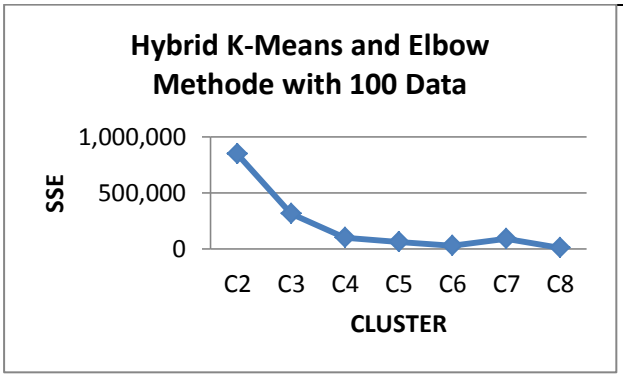

Figure 1. Graph of Sum of Square Error 100 Data

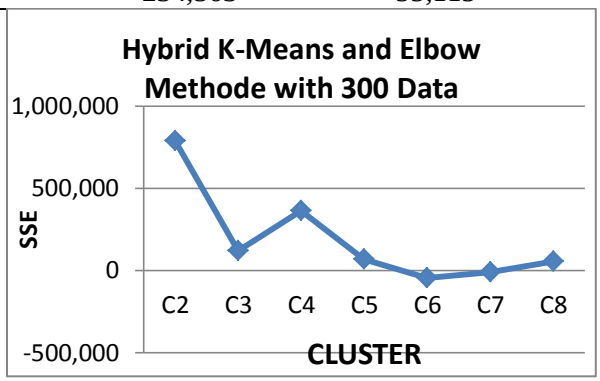

Figure 2. Graph of Sum of Square Error 300 Data

Test Data used k-means clustering and uses the Elbow method at 100 and 300 Customer Profiling. The K-Means Clustering process uses the Elbow method to determine the value of k. The result of the cluster formed will be labelled or named to facilitate the company in considering the characteristics of its customers. Performance tests have used 100 and 300 customer data purchases of goods. The results of Sum of Square Error calculations of each cluster have experienced the greatest decrease in K = 3 can be seen in Table 4, Figure 1 and Figure 2. In this test will find the performance of each number of clusters that are adjusted to the range of values on the Elbow method. The graph contained the SSE value in the experimental number of clusters between 2 and 8. The number of segments of 3 SSE values is 313,29, the value is not the lowest SSE value but the lower value.

4. **Conclusion**

The results obtained from the process in determining the best number of clusters with elbow and K-Means methods that the determination of the best number of clusters can produce the same number of clusters K on the amount of different data. The result of determining the best number of clusters with elbow method will be the default for characteristic process based on case study.

References

[1] Bataineh K M, Naji M and Saqer M 2011 A Comparison Study Between Various Fuzzy Clustering Algorithms. Jordan Journal of Mechanical and Industrial Engineering (JJMIE) **5** p335
[2] Bain K K, Firli I, And Tri S 2016 Genetic Algorithm For Optimized Initial Centers K-Means Clustering In SMEs, Journal of Theoretical and Applied Information Technology (JATIT) **90** p 23
[3] Bain K K 2015 Customer Segmentation of SMEs Using K-Means Clustering Method and modeling LRFM, International Conference on Vocational Education and Electrical Engineering, Universitas Negeri Surabaya
[4] Celebi M E, Kingravi H A, and Vela P A 2013 A comparative study of efficient initialization methods for the k-means clustering algorithm, Expert Systems with Applications **40** p 200
[5] Neda S and Farzad M S 2016 Customer Segmentation of Bank Based on Discovering of Their Transactional Relation by Using Data Mining Algorithms, Modern Applied Science **10** p 283
[6] Gursharan S, Harpreet K and Gursharan S 2014 A Novel Approach Towards K-Mean Clustering Algorithm With PSO, International Journal of Computer Science and Information Technologies (IJCSIT) **5** p 5978
[7] Antreas D A 2000 Customer Satisfaction Cues To Support Market Segmentation and Explain Switching Behavior, Journal of Business Research **47** p 191
[8] Madhulatha T S 2012 An Overview On Clustering Methods, IOSR Journal of Engineering **4** p 719
[9] Joshi K D and Nalwade P S 2013 Modified K-Means for Better Initial Cluster Centres. International Journal of Computer Science and Mobile Computing II **7** p 2
[10] Singh H and Kaur K 2013 New Method for Finding Initial Cluster Centroids in K-means Algorithm. International Journal of Computer Applications, LXXIV **6** p 27
[11] Sujatha S and Sona A S 2013 New Fast K-Means Clustering Algorithm using Modified Centroid Selection Method. International Journal of Engineering Research & Technology II **2** p1
[12] Eltibi M F and Ashour W M 2011 Initializing K-Means Clustering Algorithm using Statistical Information. K-means clustering algorithm is one of the best known, XXIX **7** p 51
[13] Aristidis L, Nikos V, Jacob J V 2011 The global k-means Clustering algorithm IAS technical report series, IAS-UVA-01-02
[14] Cosmin M P, Marian C M, Mihai M An Optimized Version of the K-Means Clustering Algorithm, Proceedings of the 2014 Federated Conference on Computer Science and Information Systems (ACSIS) **2** p 695