# Analysis of Centroid Value Variations Against the Number of Iterations Using the Clustering K-Means Algorithm

H.Simanjuntak, M Zarlis, P H Putra

***Abstract*:. In this study, the researchers used the K-means Algorithm to see and examine the effect of variations in centroid values on the number of iterations. The results of this study were: Different number of centroid results in different number of iterations. The large number of centroid values did not always cause the number of iterations to increase. Testing with the number of centroid values 3, 8, 10 has a smaller number of iterations, namely the 3 iteration with the level of similarity of the previous data, compared with the number of centroid values 2, 3, 4, 5, 6, 7, 8, 9, 10. Testing with the number of centroid 2 values had a greater number of iterations, reaching the 9th iteration to reach the previous level of data similarity.***

## I. INTRODUCTION

One technique known in data mining is clustering. Understanding of scientific clustering in data mining grouping a number of data or objects into the cluster contains data as closely as possible and different from objects in other clusters. Until now, scientists are still making various efforts to improve the cluster model and calculate the optimal number of clusters so that the best cluster can be produced.

In this study, researchers used the K-means Algorithm to see and examine the effect of variations in centroid values on the number of iterations. Case study taken by researchers was at PT. Auto 2000 Medan which had a number of delinquent car loan customers. The problem arised when the number of customer data whose loans were jammed was still random, made it difficult for the company to sort customer data with the same arrears. Therefore, it was necessary to group the customer data with the same amount of credit arrears, so that the company would be easier to make effective actions on the credit congestion that occurs. To overcome this problem, this problem becomes important to be analyzed by using the centroid variation approach method to find out at which iteration point will be found the same cluster.

## II.RELATED RESEARCH

Ting Li, et al, on research User interest domains were based on log data operations using groupings such as k-means to analyze users, then for each user group, analyze and tag them. After analysis, users can be grouped based on their behavior characteristics. First, the concept of silhouette coefficients was introduced and the optimal number of groupings was included in the set [1].

Unnati R Raval, Chaita Jani, This study described techniques that improve techniques to determine the initial centroid and determine data pointing to the nearest group more accurately with O (n) time complexity faster than traditional k-means [2].

Pooja Pandey, Ishpreet Singh, In this study described the k-means algorithm, analyzed its deficiencies and also the alternative goals for this deficiency. This algorithm created centroids at random which was not practical and also required in-depth knowledge of groupings. Second, re-assigned datapoin several times to make this algorithm low efficiency [3].

## III. PROPOSED METHOD

### a)Algorithm Clustering

Clustering is a method of analyzing data. The aim is to group data with similar criteria into 'the same area' and data with different criteria into 'different regions' [4].

Clustering is an important part of data analysis and data mining applications. Data is divided into groups based on the same features. Eachdata group with similar objects are clusters. It means clustersare the ordered set of data which have the familiarcharacteristics. Clustering is a process of unsupervisedlearning. Highly superior clusters have high intra-classsimilarity and low inter-class similarity [5].

### b)K-Means

K-Means is a commonly used grouping technique. This is really based on the division methodology.

This partition n data items into group k where k shows the number of clusters specified by the user. Clusters are formed in such a way that each item in the cluster has a minimum distance from the centroid..For calculate the distance between items and centroids, k-algorithm algorithm uses Euclidean distance measurements. With the aim of minimizing the sum of squared distances between all cluster points and centres [3].

K-means is a grouping technique commonly used. This algorithm is the most popular grouping tool used in scientific and industrial applications. This is a cluster analysis method that aims to partition 'observations into groups where each observation belongs to a group with the closest mean.[5].

K-means is the simplest used partitioning technique. This is one method of grouping based on partitions. Given a set of numerical objects A and integer number k, the k-means algorithm looks for partitions of the Ainto k cluster

Revised Manuscript Received on December 22, 2018

**H.Simanjuntak,**Student, Department of Information Technology, Faculty of Computer Science and Information Technology, Universitas Sumatera Utara, Indonesia

**P H Putra,** Student, Department of Information Technology, Faculty of Computer Science and Information Technology, Universitas Sumatera Utara, Indonesia

**M Zarlis ,**Department of Information Technology, Faculty of Computer Science and InformationTechnology, Universitas Sumatera Utara, Indonesia

which minimizes the number of errors in a group. The k-means algorithm starts with the cluster cluster initialization [1].

## IV. RESEARCH METHODOLOGY

This research method included the stages of analysis carried out when the planning stage has beencompleted. At this stage carried out further research to obtain more detailed data, which aimed for technical system development needs.

Furthermore, this research was to analyze the influence of centroid on the number of iterations by clustering method. The author wanted to find out whether there was a difference between the addition of the centroid value and the number of iterations.

## V. RESULTS AND DISCUSSIONS

In this study will be analyzed the effect of the variation of centroid values on the number of iterations by using method clustering k-means. To find out the effect of the centroid value, the number of centroids was varied from the centroid 2 to the centroid 20 value with the total data 100 and the maximum iteration number 15.

The following were the stages of the clustering process with centroid 2 values, 100 data numbers and a maximum iteration number of 15.

**Table 1 (a).**The initial cluster center with the 1st iteration

| Early Cluster Center | | |
|---|---|---|
| c11=559 | c12=33 | c13=8 |
| c21=1314 | c12=33 | c23=5 |

In table 1 (a) the number of centroid 1 = 10, the number of centroid 2 = 90 with the iteration to 1.

**Table 1 (b).** New cluster center with 2nd iteration

| Calculate the New Cluster Center | | |
|---|---|---|
| c11=9469920 | c12=24 | c13=2.6 |
| c21=4570802.89 | c22=38.53 | c23=3.09 |

In table 1 (b) the number of centroid 1 = 15, the number of centroid 2 = 85, with the second iteration where the level of similarity with the previous data is 85 data.

**Table 1 (c).**New cluster center with 3rd iteration

| Calculate the New Cluster Center | | |
|---|---|---|
| c11=8721673.33 | c12=28 | c13=2.93 |
| c21=4414663.06 | c22=38.68 | c23=3.06 |

In table 1 (c) the number of centroid 1 = 23, the number of centroid 2 = 77, with the 3rd iteration where the level of similarity with the previous data is 92 data.

**Table 1 (d).**New cluster center with 4th iteration

| Calculate the New Cluster Center | | |
|---|---|---|
| c11=8089543.48 | c12=29.22 | c13=2.91 |
| c21=4155999.48 | c22=39.43 | c23=3.08 |

In table 1 (d) the number of centroid 1 = 27, the number of centroid 2 = 73, with the 4th iteration where the level of similarity with the previous data is 96 data.

**Table 1 (e).** New cluster center with 5th iteration

| Calculate the New Cluster Center | | |
|---|---|---|
| c11=7857522.22 | c12=30.22 | c13=2.89 |
| c21=4026278.91 | c22=39.62 | c23=3.1 |

In table 1 (e) the number of centroid 1 = 38, the number of centroid 2 = 62, with the 5th iteration where the level of similarity with the previous data is 89 data.

**Table 1 (f).**New cluster center with 6th iteration

| Calculate the New Cluster Center | | |
|---|---|---|
| c11=7335652.63 | c12=32.21 | c13=2.89 |
| c21=3666397.74 | c22=40.06 | c23=3.13 |

In table 1 (f) the number of centroid 1 = 47, the number of centroid 2 = 53, with the 6th iteration where the level of similarity with the previous data is 91 data.

**Table 1 (g).**The center of the new cluster with the 7th iteration

| Calculate the New Cluster Center | | |
|---|---|---|
| c11=7019893.62 | c12=36.26 | c13=3.13 |
| c11=3323329.44 | c22=37.81 | c23=2.96 |

In table 1 (g) the number of centroid 1 = 53, the number of centroid 2 = 47, with the 7th iteration where the level of similarity with the previous data is 94 data.

**Table 1 (h).** New cluster center with 8th iteration

| Calculate the New Cluster Center | | |
|---|---|---|
| c11=6822243.4 | c12=38.72 | c13=3.21 |
| c21=3074309.79 | c22=35.23 | c23=2.85 |

In table 1 (h) the number of centroid 1 = 56, the number of centroid 2 = 44, with the 8th iteration where the level of similarity with the previous data is 97 data.

**Table 1 (i).**The center of the new cluster with the 9th iteration

| Calculate the New Cluster Center | | |
|---|---|---|
| c11=6725608.93 | c12=38.57 | c13=3.2 |
| c21=2941758.18 | c22=35.18 | c23=2.84 |

In Table 1 (i) the number of centroid 1 = 56, the number of centroid 2 = 44, with the 9th iteration where the level of similarity with the previous data is 97 data. After the 9th iteration is still not found in common with the previous data, then calculating the center of the new cluster with the 10th iteration so that the level of similarity with the previous data reaches 100.

Furthermore, the authors conducted several tests of the effect of centroid values on the number of iterations by clustering method where the number of centroid 2 values to the centroid 10 with the amount of data 100 and the number of iterations maximum 15. Test results can be seen in table 2.

**Table 2.**Effect of centroid values 2 to 10 on the number of iterations

| Number of Centroid Value | Number of Iterations |
|---|---|
| 2 | 9 |
| 3 | 3 |
| 4 | 8 |
| 5 | 5 |
| 6 | 6 |
| 7 | 5 |
| 8 | 3 |
| 9 | 4 |
| 10 | 3 |

In table 2, it can be seen that testing with the number of centroid values 3, 8, 10 has a smaller number of iterations, ie reaching the 3rd iteration with the previous data similarity level, rather than testing using the number of centroid values 2, 4, 5, 6, 7 , 8, 9, 10 In table 2, it can be seen that the test with the number of centroid 2 values has a greater number of iterations, namely reaching the 9th iteration to reach the previous level of data similarity.

In table 2, it can be seen that the number of different centroid values will produce a different number of iterations. The large number of centroid values does not always cause the number of iterations to increase. A comparison chart of the effect of the number of centroid values with the number of iterations can be seen in Figure 1.

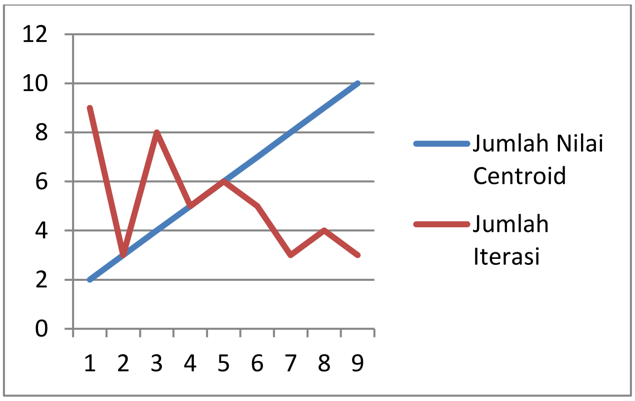

**Figure 1:** Comparison graph of the effect of the number of centroid values with the number of iterations

From Figure 1 it can be seen that the number of different centroid values will produce a different number of iterations. Testing with a large number of centroid values does not always cause the number of iterations to increase.

## VI. CONCLUSION

From the results of the study some conclusions can be drawn, among others: The number of different centroid values will produce a different number of iterations. The large number of centroid values does not always cause the number of iterations to increase. Testing with the number of centroid values 3, 8, 10 has a smaller number of iterations, namely the 3 iteration with the level of similarity of the previous data, compared with the number of centroid values 2, 3, 4, 5, 6, 7, 8, 9, 10. Testing with the number of centroid 2 values having a greater number of iterations, reaching the 9th iteration to reach the previous level of data similarity.

## REFERENCES

1. Jiaoling Ting Li, et al, 2018, *State Grid Office System User Clustering Analysis Based On K-Means Algorithm*, International Conference On Big Data Analysis.
2. Unnati R Raval, Chaita Jani, 2016, *Implementing & Improvisation Of K-Means Clustering Algorithm*, International Journal of Computer Science and Mobile Computing .

3 Pooja Pandey, Ishpreet Singh, 2016, *Comparison Between Stardard K-Mean Clustering And Improved K-Mean Clustering*, International Journal of Computer Applications.

4 Purnawansyah, Haviluddin, 2016, K-Means Clustering Implementation in Network Traffic Activities, IEEE.

5 Anshul Yadav, Sakshi Dhingra, 2016, *A Review On K-Means Clustering Technique*, International Journal of Latest Research in Science and Technology.