Accuracy of Data Cluster Using Modify K-Mean Algorithm by Local Deviation Method

S. Sriadhi[1*], Syawal Gultom[1], M. Martiano[2]

[1]Faculty of Engineering, Universitas Negeri Medan, Indonesia
[2]Universitas Muhammadiyah Sumatera Utara, Indonesia, Indonesia

*Corresponding Author: sriadhi@unimed.ac.id

Abstract. *Data clustering requires accuracy and consistency to provide unbiased results. One of the most used methods is K-Means algorithm although it still has a fairly high error rate. The purpose of this research is to produce an accurate and consistent formulation in data cluster through K-Means modification named K-Means algorithm with Local Deviation Method (K-Means LDM). This study used credit of study load and study period (semester) variables from the data of two batch students totalling 1089 data. The data analysis includes a mean deviation of two tests for credit and semester variables as well as comparative test results of the two methods, namely the K-Means algorithm and K-Means LDM algorithm. The test result shows that the K-Means LDM algorithm may reduce the error with MSE 290.95 in the first and second tests, while the MSE value of the K-Means Algorithm is 508.54 in the first test and 881.13 in the second test. The result of the study suggests the use of K-Means LDM algorithm because it may reduce error index by 58.13% and is more accurate and consistent compared to the K-Means algorithm in the big data clustering process.*

Keyword: K-Means, K-Means LDM, accuracy, cluster

1. Introduction

Data clustering can be done in several methods, one of which is the K-means algorithm. K-means is an input data clustering algorithm that is divided into several groups without pre-checking the target class [1,2]. This method includes *unsupervised learning*, which will group data or object into *k* groups. In each cluster, there is a central point (*centroid*) that represents the cluster [1,3,4].

The process of grouping data into a *cluster* can be done by calculating the closest distance of a data to a *centroid* point [7,8]. The Minkowski distance calculation can be used to calculate the distance between two data with the following formula:

$$d(x_i, x_j) = \left(|x_{i1} - x_{j1}|^g + |x_{i2} - x_{j2}|^g + \ldots + |x_{ip} - x_{jp}|^g\right)^{1/g} \quad (1)$$

Incorrect data grouping will produce inaccurate output. The K-Means algorithm receives input data without class labels. Previous research showed the accuracy of K-Means in the range value of 0.3 and 0.6 [8]. The randomly selected data will affect various results and affect the Means Square Error (MSE value). Therefore, a method is needed to improve the accuracy of the data clustering process. As research by Hassan Ismkhan [9]. Sriadhi [10] and Gultom [11] which show that a modified algorithm is needed to improve accuracy in big data clustering process.

Data clustering using K-Means method is basically done by algorithmic procedures (1) Determine the number of clusters first, (2) Allocate random data into clusters, (3) Perform centroid or average calculation of data in each cluster, (4) Allocate each data into the nearest centroid, (5) Repeat step (3) if there is still data that moves the cluster or changes the centroid value or changes in the value of the objective function above the specified threshold value [12-14]. Referring to the basic function of the K-Means Clustering method that is to minimize the objective function specified in the clustering process by minimizing variations between data in a cluster and maximizing variation with data in other clusters.

ISSN: 2005-4238 IJAST

In the clustering process, the K-Means looks for the value of proximity between the objects and the center of the selected cluster. The proximity value is the center point of the cluster on the object. A deviation method is needed in determining the center point that will be used as the center of the cluster to minimize the value of MSE in the clustering process [12,15,16]. This research will produce the exact formulation to improve the accuracy of data grouping which focused on variability (mean deviation), clustering with K-Means Deviation, and clustering performance by measuring the MSE.

2. Research Method

The process of data clustering using K-Means algorithm begins by determining the input parameter (k) i.e. the number of desired clusters. The amount of data or objects (n) will be grouped into (k) clusters to produce high intra-cluster similarities but the similarity between clusters is low. Similarity can be determined by cosine, covariance, and correlation, while to measure dissimilarity you can use distance. The closer the distance means the higher the similarity and conversely the farther the distance the lower the similarity [14.17,18]. Furthermore, the accuracy improvement is done by the Local Deviation Method.

This research was carried out with two main variables, namely study load (semester credit unit) and study period (semester) for two batch students totalling 1,089 data. The data analysis was carried out starting from the mean deviation of two observations individually with the data center for the credit and semester variables are $x_1 - x_n$ and $y_1 - y_n$. The use of Local Deviation Method was done by referring to the following analysis

2.1. Mean Deviation

The credit and semester variables have sequences on the same variable, namely $x_1 - x_n$ and $y_1 - y_n$. The average deviation involved all data observation in the calculation and the variability was measured by comparing observation data individually to the canter of the data with the following formula:

$$mean.dev = \sum \vee X_i - \frac{\acute{X}\vee}{n} \quad (2)$$

The pseudo code of the deviation formula is as follows:

```
Begin {
load data
 mahasiswa where thnmsk='2011' limit 50;
row->NIMMHS;
mahasiswa where thnmsk='2010' limit 50;
row->NIMMHS;
insert kmeans(NIMMHS,ileterasi)values (row->NIMMHS;);

load data
trs_ipk where kdmhs='row->NIMMHS;'
$Tahun =row->Tahun;
if ($tahun==''){}else{
$tahun1=$tahun1+5;
$tahuna=$tahuna+1;
}
$Tahun2=$Tahun1*$Tahun1;

$Tahun5=abs(($Tahuna*$Tahun1)-$Tahun2);
$tahun4=$Tahuna-1;
$Tahun3=$Tahun5/$Tahuna*$tahun4;

$Tahun7=sqrt($Tahun3)^2; }
```

2.2. K-Means with Deviasi

The value of x and y on each formed object were then grouped using K-Means as the *clustering* method. The K data element was chosen as the starting point and the distance of all data element was calculated using the Euclidean distance formula. The data element

which smaller than the centroid distance was transferred to a suitable cluster until there was no more alteration in the [k-1] group which is the basic technique [8,19].

The K-Means algorithm procedure began by sorting and summing objects and then dividing it by the number of clusters to form groups based on the number of the cluster which was calculated using the mean deviation formula as the center of the cluster. The next step was to calculate the distance of each input data to the centroid using Euclidean Distance until the closest distance was found using the following Euclidean Distance formula:

$$d(x_i, \mu_i) = \sqrt{(x_i - \mu_i)^2} \quad (3)$$

From the above formula, (xi, μi) is the distance between *cluster* x to the center of *cluster* μ at the i[th] word; xi is the weight of the i[th] word in the cluster of which the distance is being measured; and μi is the weight of the i[th] data at the center of the cluster. The next step was to classify each data based on the proximity to the centroid and updated it based on the number of data in the cluster (n_k) and the total distance value data of each cluster using the following formula

$$C_k = \frac{1}{n_k} \sum d_i \quad (4)$$

The above process was repeated until each member of the cluster did not change and the average value of the center of the cluster (μj) in the last iteration would be used as a parameter to determine the data classification.

2.3. Performance Measurement

The program performance measurements in the K-Means method used Local Deviation method (K-Means LDM) compared with K-Means method [12] which measured with Mean Square Error method using the following formula

$$MSE: \frac{1}{n} \sum_j \left(y_{ij} - y_j\right)^2 \quad (5)$$

From the above formula, y_{ij} is the actual value, y_j is the achieved value, and n is the total number of the data.

3. Result and Discussions

The research results show the two tests produced three clusters, namely two clusters in the test with the K-Means algorithm and one cluster from the test results using the LDM K-Means algorithm. The results of this test are to determine the level of consistency and clustering accuracy of two methods, namely K-Means and K-Means LDM based on Means Square Error (MSE) of the test results. Comparison of the two methods can be seen in Table 1.

Table 1. Comparison of Means Square Error (MSE) for Two Test Method

Iteration	Means Square Error			
	K-Means		K-Means Modified	
	1st Test 1	2nd Test 2	1st Test	2nd Test
1	321.30	2,088.00	231.20	231.20
2	756.00	543.80	643.34	643.34
3	632.00	763.40	212.00	212.00
4	611.10	578.98	190.14	190.14
5	419.10	431.50	178.10	178.10
6	410.10	-	-	-
MSE	508.54	881.13	290.95	290.95

It can be seen from Table 1 that there are significant and large differences between K-Means calculations in test 1 and test 2 of cluster values generated in each iteration. The second test produces a cluster value that is greater than the results of the first test, except for the second and the fourth iterations. The difference in cluster values in the test using the K-Means algorithm is shown in the graph as stated in Figure 1.

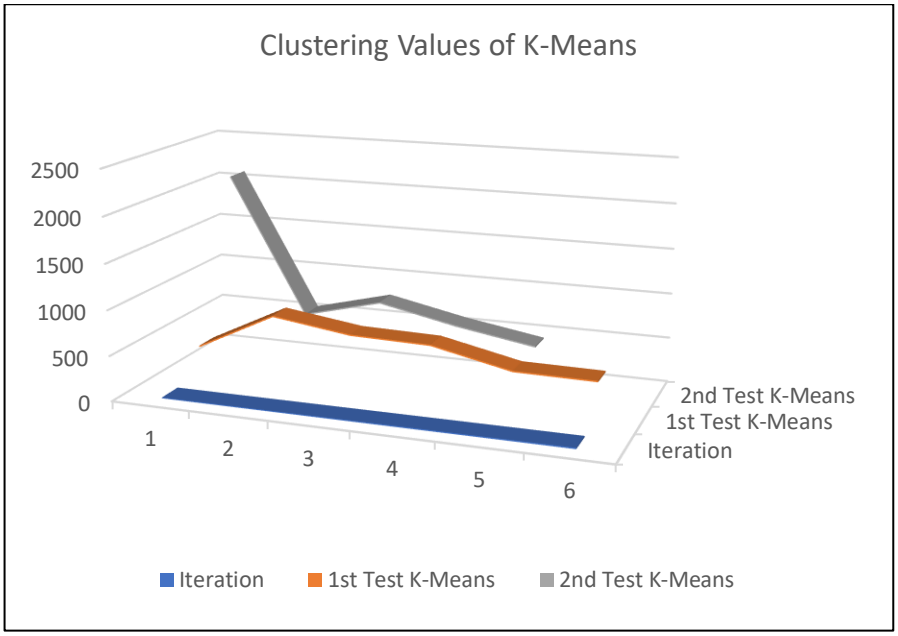

Figure 1. Data Clustering Value using K-Means Algorithm

The decrease in cluster value in the second and the fourth iterations of the second test is much smaller than the increase in cluster value in the first and third iterations, while the sixth iteration is not measurable. This is different from the test results using the K-Means LDM algorithm which shows constant cluster values in the first test and the second test. Figure 2 shows the accuracy and consistency of cluster values through two tests using the LDM K-Means algorithm.

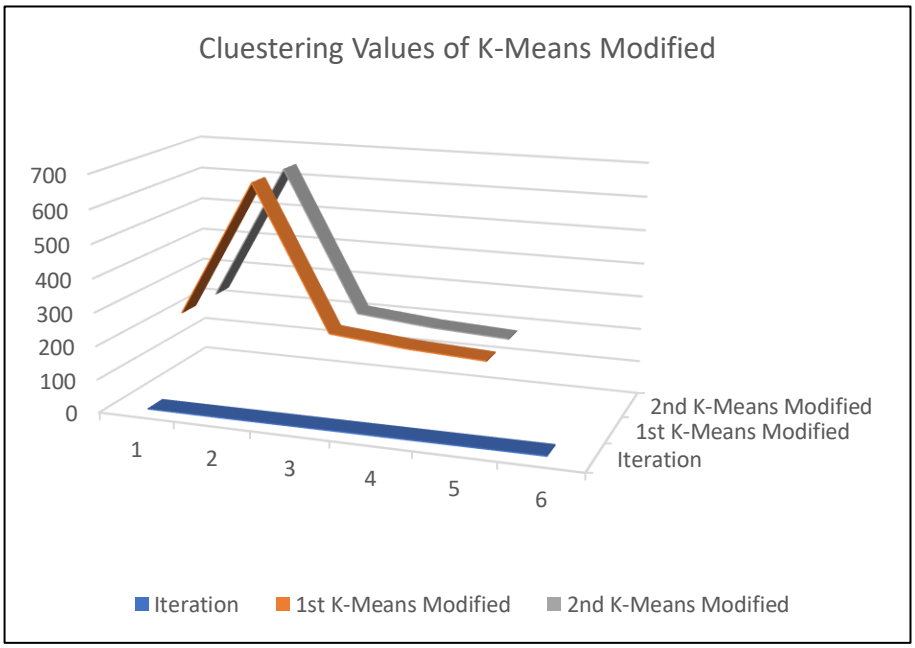

Figure 2. Data Clustering Method using K-Means LDM Algorithm

ISSN: 2005-4238 IJAST

The test results using two algorithms, namely K-Means and K-Means LDM, show significantly different results. The magnitude of the difference in cluster values using the K-Means algorithm in the first test and the second test proves that in the clustering process the system still has weaknesses in the accuracy of data grouping. The high error value in these two tests is caused by the distance of the predicted cluster canter point ($\acute{y}$) with the actual value (y) on the object. That is why it is necessary to modify the K-Means algorithm, one of which by using K-Means LDM algorithm. In this study, the K-Means LDM is able to prove the accuracy and consistency of clustering results compared to the K-Means algorithm. The accuracy and consistency are shown by the same test results in the first and second tests. This is because the cluster point selected in the K-Means LDM algorithm places the distance between the predicted cluster points (y) with the actual object value (y) closer together so that the error value can be reduced. This is in line with the results of studies that have succeeded in increasing accuracy in large data clustering using the development of the K-Means and K-Medoids algorithms [10,11,19,20].

As with the test results in each iteration, the cumulative results obtained through two tests on the K-Means algorithm and the K-Means LDM algorithm produced MSE values that are in line with the results of the iteration stage testing. As shown in Table 1, the overall MSE value for the K-Means algorithm has a significant difference between the first test and the second test, with MSE 508.54 in the first test and increasing to 881.13 in the second test. This is not the case with the K-Means LDM algorithm, the modified K-Means, which results in a consistent value of MSE 290.95 both in the first and the second test. Figure 3 shows the accumulative MSE of two algorithms on two tests.

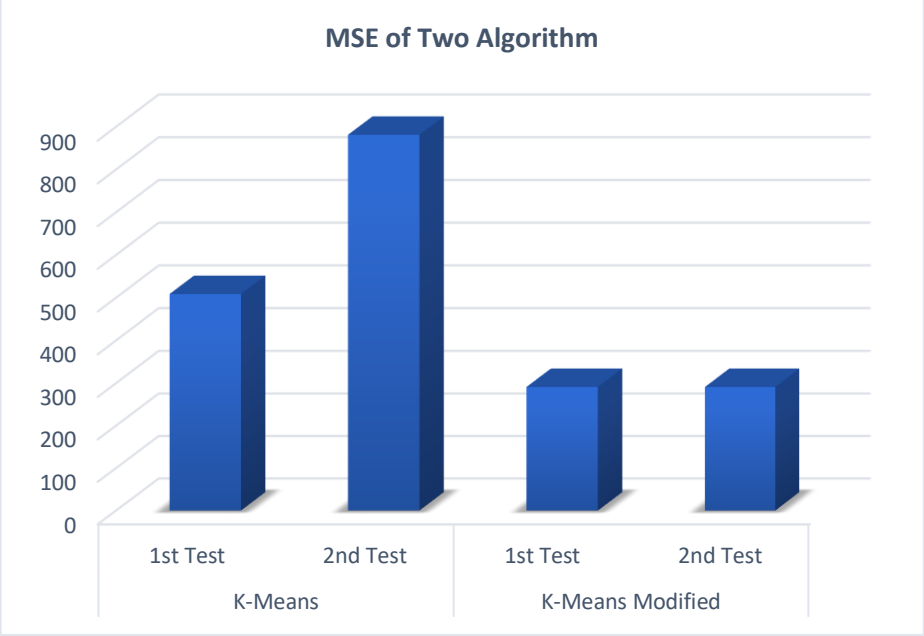

Figure 3. Comparison of MSE Value for Two Test and Two Clustering Method

The results of testing the data on two variables (credits and semester) for two years show that the Means Square Error value is different in the K-Means algorithm and the K-Means LDM algorithm. Data clustering results using the K-Means algorithm have a larger and inconsistent MSE error value (508.54 and 881.13) compared to the K-Means LDM algorithm which is able to reduce errors with MSE much smaller and consistent (290.95) for the first and second test. It can be concluded that, overall, the developed K-Means LDM algorithm may reduce the error value by 58.13% compared to the K-Means algorithm in the big data clustering process. The results of this study support the results of previous studies, namely Laurence Morissette and Sylvain Chartier [2], Preeti Arora. Deepali. and S. Varshney [3], Hassan Ismkhan [9], Sriadhi [10], Melnykov. V, Zhu. X

[21] who developed algorithms and compared them with the K-Means algorithm for clustering large amount of data.

The results of the study show that the K-Means algorithm that has been used for a large data clustering still has many weaknesses, especially accuracy and consistency. Moreover, the development of the K-Means LDM provides proves of the increase in accuracy and consistency of cluster data for large amount of data. This is in line with other methods that have been developed, such as K-Medoid with Ecluidience Distance Algorithm and K-Means Method with Linear Search Algorithm [17,22]. The concept that underlies accuracy and consistency in the data clustering process is the reduction of the Means Square Error (MSE) value by decreasing the distance between the predicted cluster points (y) and the actual object value (y).

4. Conclusion

The clustering process in a large amount of data needs to pay attention to the accuracy and consistency in order to obtain valid measurement results. This study produces the K-Means LDM algorithm that may improve the accuracy and consistency in clustering big data. The K-Means LDM method is able to reduce the error rate (MSE) to 58.13% of the original method, namely K-Means. In addition, the K-Means LDM are also able to improve consistently compared to the K-Means method which always produces variable data values (inconsistencies). This study succeeded in developing the K-Means LDM algorithm which is able to answer the problems of data accuracy and consistency in clustering a large amount of data.

5. References

[1] Kahkashan Kouser and Sunita, "A comparative study of K Means Algorithm by Different Distance Measures", International Journal of Innovative Research in Computer and Communication Engineering, vol.1, no.9, **(2013)**, pp.2443-2447.

[2] Laurence Morissette and Sylvain Chartier, "The k-means clustering technique: General considerations and implementation in Mathematica", Tutorials in Quantitative Methods for Psychology, vol. 9, no.1, **(2013)**, pp. 15-24.

[3] Preeti Arora. Deepali. and S. Varshney, "Analysis of K-Means and K-Medoids Algorithm for Big Data," Phys. Procedia, vol.78, no.1, **(2016)**, pp.507–512.

[4] Shyr-ShenYu, Shao-WeiChu, Chuin-MuWang, Yung-KuanChan, Ting-ChengChang, Two improved k-means algorithms", Applied Soft Computing, Elsevier", vol.68, (2018), pp.747-755.

[5] Wang Meng, Dui Hongyan, Zhou Shiyuan, Dong Zhankui, Wu Zige, "The Kernel Rough K-Means Algorithm", Journal of Recent Advances in Computer Science and Communications, vol.13, no.2, **(2019)**, pp.253 – 259.

[6] Gopal Behera, Ashok Kumar Bhoi, "General Applicability of K-means Algorithm with Enhanced Centroids", International Journal of Latest Technology in Engineering, Management & Applied Science (IJLTEMAS), vol.vii, issue.iii, **(2018)**, pp.201-205.

[7] X. Wu and V. Kumar, eds, "The Top Ten Algorithms in Data Mining*",* Chapman and Hall, **(2009).**

[8] Han. J and Kamber. M, "Data Mining Concepts and Techniques". Elsevier, Amsterdam **(2012).**

[9] Hassan Ismkhan, 2018. "I-k-means−+: An iterative clustering algorithm based on an enhanced version of the k-means", Pattern Recognition, Elsevier, vol.79, **(2018),** pp.402-413.

[10] S. Sriadhi, Syawal Gultom, M. Martiano, Robbi Rahim and Dahlan Abdullah, "K-Means Method with Linear Search Algorithm to Reduce Means Square Error (MSE) within Data Clustering", Journal of Materials Science and Engineering (Conf. Series), vol.434, no.1, **(2018)**, p. 2032

[11] Syawal Gultom, S. Sriadhi, M. Martiano, Janner Simarmata, "Comparison analysis of K-Means and K-Medoid with Ecluidience Distance Algorithm, Chanberra Distance, and Chebyshev Distance for Big Data Clustering", Journal of Materials Science and Engineering (Conf. Series), vol.420, no.1, **(2018)**, p.012092

[12] Rougher. J, "Ensemble Averaging and Mean Squared Error", Jurnal of Climate, vol.29, no.4, **(2016)**, pp.1-6.

[13] G. Karthikeyan and Rizwana. A, "Modify EncriptionAlgorithm for VANET using Network Simulator Tool", International Journal of Advanced Science and Technology, vol.28, no.1, **(2019)**, pp.416-424.

[14] Ali Abdul-hussian Hassan, Wahidah Md Shah, Ali Mohamed Husein, Hyder Abdul Hussein Hassa, Mohd Fairuz Iskandar Othman, "An Improved LEACH Algorithm based on Fuzzy C-Means Algorithm and Distributed Cluster Head Selection Mechanism", International Journal of Advanced Science and Technology, vol.28, no.1, **(2019),** pp.406-415.

[15] Chen Chung Liu, Shao Wei Chu, Yung Kuan Chan, Shyr Shen Yu, "A Modified K-Means Algorithm - Two-Layer K-Means Algorithm", Tenth International Conference on Intelligent Information Hiding and Multimedia Signal Processing, IEEE Xplore, (2014).

[16] Cuong Duc Nguyen, Trong Hai Duong, "K-means** – a fast and efficient K-means algorithms", International Journal of Intelligent Information and Database Systems, vol. 11, n 1, **(2018),** pp.27-45.

[17] X.-J. Liu, J.-B. Yuan, J. Xu, B.-J. Duan, "Quantum k-means algorithm", Journal of Jilin University (Engineering and Technology Edition), vol.48, no.2, **(2018),** pp.539-544.

[18] M. Krishnamoorthy, A. Noble Mary Juliet, C. Keerthana, R. Usha Nandhini, "Diseases Identification in Plants Using K-Means Algorithm", International Journal of Computer *Sciences and Engineering*, vol.7, no.2, **(2019),** pp.458-462

[19] Brian Morris, Zachary H Levine, "The Poisson-Influenced K-Means Algorithm, The Mathematics Journal", vol.18, **(2016),** pp. 1-28.

[20] Sharfuddin Mahmood, Mohammad Saiedur Rahaman, Nandi Mashiour Rahman, "A Proposed Modification of K-Means Algorithm", International Journal of Modern Education and Computer Science, vol.6, **(2015)**, pp. 37-42. \

[21] Melnykov. V, Zhu. X, "An extension of the K-means Algorithm to Clustering Skewed Data", Comput Stat, vol.34, **(2019),** pp.373–394

[22] SK.Ahammad Fahad, Md. Mahbub Alam, "A Modified K-Means Algorithm for Big Data Clustering", International Journal of Computer Science Engineering and Technology (*IJCSET), v*ol 6, issue 4, (2016), pp.129-132.