K-means method with linear search algorithm to reduce Means Square Error (MSE) within data clustering

S Sriadhi*, S Gultom², M Martiano¹, R Rahim³ and D Abdullah⁴

¹ Department of ITC Education, Universitas Negeri Medan, Medan, Indonesia
² Department of Mathematics Education, Universitas Negeri Medan, Medan, Indonesia
³ Department of Informatics Technology, Institut Tekbologi Medan, Medan, Indonesia
⁴ Department of Informatics, Universitas Malikussaleh, Medan, Indonesia

*riadhi@unimed.ac.id

Abstract. K-means method is limited in identifying and grouping the data by characteristics similarity in clustering. This study develops K-Means method with LSA to fix the issue of objectivity in data clustering as compared to K-Means method used lately. Data variables used are study load (credits) and study period (semester) of students in two academic years, which is amounted to 1,089 records. Data is analysed by using comparative statistic between the results of clustering test using K-Means method and K-Means method with LSA. The test findings show that the clustering using K-Means only groups the data into 3 clusters while the use of K-Means method with LSA produces 5 clusters. There are 327 different characteristics data identified by K-Means method with LSA which are grouped in two new clauses so it results in five clusters, for what is rated similar by K-Means method which only produces three clusters. This study concludes that K-Means method with LSA is more objective in clustering the data clustering and reducing MSE level error due to the sensitivity of data similarity within the cluster as always happened with K-Means method. Therefore, it is recommended that K-Means method with LSA be used in clustering to objectively identify the data and avoid any errors in the clustering process for more optimal data utilization.

1. Introduction

K-Means is one of the methods in the process of clustering [1], known as unsupervised learning for the groups label isn’t known [2]. The purpose of the clustering process is identifying group or cluster similarity to similar object from each of the objects on other clusters [3].

Although K-means algorithm method is widely used, it still has some weaknesses. Among them is being sensitive to noisy data and outliers. For example, in the process of clustering, there are many objects which have the similar value to the specified cluster value, causing the Mean Square Error (MSE) value of the K-means algorithm to increase. Data with varied range of values but having many similarities are often trapped to sensitivity, so the object goes into some clusters which do not belong to their group. Figure 1 represents one example of error in clustering.

K- Mean’s method is easily trapped in local minimum data related to measurement, especially for the number of quadratic errors [4]. To deal with this problem, filtering effort is needed to avoid the sensitivity error of similar object value in some clusters.
This study aims to develop a method to fix the error due to object value sensitivity by combining K-Means method with Linear Search Algorithm (K-Means with LSA) to identify the similar object value. The encryption of this algorithm is used to find a particular element in an array, but not to change the data arrangement, be it ascending or descending [5]. It is done by counting the number of cluster center point within the built equation. The use of linear search algorithm with K-Means method will result in more accurate findings and reduced error.

2. Research methods
This study is conducted at Universitas Negeri Medan in 2018. The object variables chosen are study load (credits) and study period (semester). The data is gained from the university academic information system by taking the samples of students in the academic year of 2010 and 2011 who is amounted to 1,089 records.

Data analysis is done by using K-Means method combined with Linear Search Algorithm (K-Means with LSA), which is grouping the objects far from the cluster point through the following procedures:

a. Determining the number of k-clusters to be formed.
b. Generating random values for the center of the initial cluster (centroid) as much as k-clusters.
c. Calculating the distance of each input data on each centroid using the Euclidean Distance formula to it finds the closest distance of each data with the centroid. Here is the Euclidean Distance equation

\[ d(x_i, \mu_i) = \sqrt{(x_i - \mu_i)^2} \]  

(1)
d. With \( d (x_i, \mu) \) is the distance between the cluster \( x \) and the cluster center \( \mu \) in the word \( x_i \) is the i-th weight of the cluster to be searched for distance, \( \mu i \) the weight of the word to i at the center of the cluster.
e. Classifying each data based on its proximity to the centroid (smallest distance).
f. Updating the centroid value. The new centroid value is obtained from the average cluster in question by using the formula:

\[ C_k = \frac{1}{n_k} \sum d_i \]  

(2)

Where:
\( n_k \) : the amount of data in the cluster
\( d_i \) : the sum of the value of the incoming distance in each cluster

After the above process, the next two final steps are written as follows:

- Checking the result of the cluster within the criteria of distance K value equal to the other K value by using the LSA method.
- If the same value is identified, the data is then clustered and calculated of its cluster center point by using the equation 4 so that the addition of the cluster center point takes place.
The results of measurement by using the K-Means method with deviation are compared to the results of measurements which include the Mean Square Error (MSE) Method [6]. The formula used for calculating MSE is:

\[
MSE = \frac{1}{n}\sum_{j}(y_{ij} - y_{j})^2
\]  

(2)

Where:
- \( y_{ij} \): Actual value
- \( y_{j} \): Value reached
- \( n \): Amount of data

The result of clustering will stop if testing results in a constant value. The comparison of clustering level error which is the identification of objects in the same cluster yet different based on the testing with K-Means method and K-Means method with LSA is deemed as the conclusion of the testing result in this study.

3. Results and discussion

The results of data analysis use the K-Means method which is combined with Linier Search Algorithm (K-Means with LSA). K-Mean is a method that has the object grouping process within a dataset into some groups, and every group member has similar data characteristics among one another but different characteristics with other group data. K object data is selected as cluster initial point, then distance on every object is calculated by using Euclidean Distance formula. Characteristics data which is far from K object will move to another K object. Grouping process is continued to K object which is not changing.

Linear Search Algorithm (LSA) method is a linear or sequential searching which checks the data in an array or link-list on each element and compares it to the searching element. Vector V is given in a search algorithm (V, N and X) containing N element. This algorithm looks for V and is given to the value of X within pseudo code linear search algorithm [7] as follows.

[Search for the location of value X in vector V]
FOUND ← false
I ← 1
Repeat while I ≤ N and not FOUND
If V [I] = X
then FOUND ← true
LOCATION ← I
Exit
else
I ← I + 1
Write ('Value of', X, 'NOT FOUND')
2. [Finished]

K-Means method with LSA is a combination of K-Means method and Linear Search Algorithm developed this study to reduce the error due to the sensitivity of data similarity in the cluster or groups. Performance measurement is done by comparing K-Means method deviation with the measurement result of Mean Square Error method within this pattern \( MSE = \frac{1}{n}\sum_{j}(y_{ij} - y_{j})^2 \). The experiment result using K-Mean methods with the random selection of centroid centre experience some obstacles in determining the distance because of the objects outside the criteria. The testing results stall on iteration 5 resulting in three clusters as shown on table 1.

| No | object | Variables | Cluster | Central Point Distance | Distance | Cluster |
|----|--------|-----------|---------|------------------------|----------|---------|
| 1  | a8047e | x 6 y 163| C1 35   | C2 18                  | C3 24    | C2      |
| 2  | a9321d | x 4 y 119| C1 2147483 | C2 2147483           | C3 2147483 | C1      |
Table 1. Cont.

| No | Object | Variables | Cluster | Central Point | Distance | Error |
|----|---------|-----------|---------|--------------|----------|-------|
| 3  | a9b304  | 4         | 137     | 2147483      | 2147483  | C1    |
| 4  | aa353f  | 5         | 147     | 2147483      | 2147483  | C1    |
| 5  | aab787  | 5         | 146     | 2147483      | 2147483  | C1    |
| 6  | ab3a48  | 5         | 148     | 2147483      | 2147483  | C1    |
| 7  | abba0   | 5         | 154     | 2147483      | 2147483  | C1    |
| 8  | ac3e05  | 5         | 146     | 2147483      | 2147483  | C1    |
| 9  | acfb7   | 3         | 124     | 2147483      | 2147483  | C1    |
| 10 | ad9567  | 4         | 120     | 2147483      | 2147483  | C1    |
| 11 | ae16a9  | 5         | 141     | 2147483      | 2147483  | C1    |

On table 1, there are objects incorporated in clusters 1, 2 and 3. Although in the K-Means method some attempts to minimize the error rate for each iteration are performed, the objects incorporated in clusters 1, 2 and 3 have as many as 327 similar object members in the iteration 5. While in the K-Means method with LSA, the grouping process is done to find and group the objects at each iteration and form a new cluster in an effort to minimize the object which has the same value, resulting in five clusters and stopping on iteration 6, as shown on table 2.

### Table 2. Results K-Means method with LSA grouping (Sequential Cluster) six iteration.

| No | Object   | Variables | Cluster | Central Point | Distance | Error |
|----|----------|-----------|---------|--------------|----------|-------|
| 1  | a8047c   | 6         | 163     | 35           | 24       | 2147483 2147483 C2 201.6 |
| 2  | a9321d   | 4         | 119     | 24           | 100      | 2147483 2147483 C1 184.9 |
| 3  | a9b304   | 4         | 137     | 32           | 24       | 2147483 2147483 C1 19.36 |
| 4  | aa353f   | 5         | 147     | 18           | 2147483 2147483 C1 207.3 |
| 5  | aab787   | 5         | 146     | 13           | 2147483 2147483 C1 19.36 |
| 6  | ab3a48   | 5         | 148     | 19           | 2147483 2147483 C1 237.1 |
| 7  | abba0   | 5         | 154     | 21           | 2147483 2147483 C1 457.9 |
| 8  | ac3e05   | 5         | 146     | 13           | 2147483 2147483 C1 179.5 |
| 9  | acfb7   | 3         | 124     | 2147483      | 2147483 2147483 C1 73.96 |
| 10 | ad9567  | 4         | 120     | 2147483      | 2147483 2147483 C1 158.7 |
| 11 | ae16a9  | 5         | 141     | 8            | 2147483 2147483 C1 70.56 |

The clustering process which is done by using K-Means method and K-Means method with LSA have 3 clusters (red, green, blue) and 5 clusters (red, yellow, light blue, green, blue) which its nearness values is identified by clicking the formula Euclidean Distance

\[ d(x_i, \mu_i) = \sqrt{(x_i - \mu_i)^2} \]

thus forming such a group as seen on table 2.

The results of object clustering with K-Means method only produce three clusters while the K-Means method with LSA produce five clusters. K-Means method with LSA is proven to be more thorough in clustering. The visualization of the data object clustering comparison is shown on figure 2.
There are 327 data objects of different characteristics are clustered in the same group as other data within one cluster by K-Means method, meanwhile K-Means method with LSA distinguish those data in new
clusters because they have different characteristics. Further, the MSE value is obtained through the formula $MSE = \frac{1}{n} \sum_j (y_{ij} - \bar{y}_j)^2$ with the results as shown on table 3.

**Table 3.** Number of means square error comparison.

| Iteration | K-Means Number Cluster 3 | K-Means with LSA Number 5 Cluster |
|-----------|---------------------------|----------------------------------|
| 1         | 8 03.11                   | 300.67                           |
| 2         | 403.21                    | 305.45                           |
| 3         | 302.11                    | 202.11                           |
| 4         | 2 02.22                   | 110.45                           |
| 5         | 2 03.23                   | 68 .. 23                         |
| 6         | -                         | 6 0.11                           |
| AVG MSE   | 382.776                   | 174.503                          |

It can be seen on table 3 that error rate on iteration using K-Mean methods is bigger than K-Means method with LSA. Thus, it can affect the average level of Means Square Error, relevant eighth the study by Singh and Rougher [4,6]. This is because the object calculated by using the Euclidian Distance formula is too far from the cluster to be searched leading to a larger degree of error. Error level rate in both methods can be seen in Figure 2 below.

![Figure 3. Comparison of means square error rate.](image)

On figure 2, it can be seen that there is a different performance between K-Means method with K-Means method with LSA. The average number of errors on each iteration is calculated, so it generates the data as seen on table 4.

**Table 4.** Comparison of means square error number.

| No | Methods                  | Avg MSE |
|----|--------------------------|---------|
| 1  | K-Means                  | 382,776 |
| 2  | K-Means Linear Search    | 174,503 |

The use of K-Means method with LSA is proven to enable the reduction of error level due to the sensitivity of objects value characteristics similarity which actually have different characteristics. The length of iteration range will determine the error level of cluster identification. On certain point, where the clustering results shows the same data, testing will stop because of the data obtained will not change anymore. Besides, the constant saturation value from the testing results is obtained due to the threshold value similarity before and after the testing. The results of this study is a development of previous research [4,5,7].

4. **Conclusion**

K-Means method is likely to produce the errors due to the sensitivity of data characteristics similarity, particularly the data outside the specified cluster criteria. By combining K-Means method with LSA, the identification process of data characteristics similarity, the undefined outlier data are grouped by
using sequential linear method. The results obtained is greatly significant that the use of K-Means method produces 327 data objects in a new cluster that is not detected as different by K-Means method. This accuracy will minimize the occurrence the value of Means Square Error (MSE). As such, this study provide the solutions for programmers to minimize error in the data clustering system that has similar characteristics in its group, by combining the K-Means method with LSA.

Acknowledgments
The authors would like to address the gratitude to the leader of Universitas Negeri Medan who has supported the implementation of this research and the academic and ICT Siakad staff who have provided the data in this research process.

References
[1] Windha 2015 Clustering menggunakan metode K-Means untuk menentukan status gizi balita J. Informatika 15(7)
[2] H J Kamber M 2012 Data Mining Concepts And Techniques Elsevier
[3] Aggarwal C C and Reddy C K Eds 2013 Data Clustering: Algorithms And Applications (Boca Raton: CRC Press FL)
[4] Singh and Yadav 2013 Study of K-Means and enhanced K-Means clustering algorithm Int. J. Advanced Research In Computer Science 10(5)
[5] Aurora B and Sharma 2014 Two way linier search algorithm Int. J. Computer Applications 107(21)
[6] Rougher J 2016 Ensemble averaging and mean squared error J. Climate. 29(4)
[7] Parmar and Kumbharna 2015 Comparing Linear Search and Binary Search Algorithms to Search an Element from a Linear List Implemented through Static Array Dynamic Array and Linked List Int. J. Computer Applications 3