Modification K-Means Model with Local Deviation Method to Improve the Accuracy in Forming Clusters

Martiano[1], Muhammad Zarlis[2], Sutarman Wage[3]
[1,2]Computer Science Department, Universitas Sumatera Utara, Jl.Dr.Mansyur, Medan, Indonesia
[3]Mathematic Department, Universitas Sumatera Utara, Jl.Dr.Mansyur, Medan, Indonesia

martino84@gmail.com

Abstract. K-Means is one method in data mining that can be used to clustering data (Winda, 2015). But in the process of clustering K-Means Trying to minimize the number of Euclidean distance from the mean (Ismkhan, 2018) and it depends on the selection of cluster starting point (Han & Kamber, 2012). In this paper the authors try to use local deviation in calculating the cluster center based on distance variables are calculated and determined the mean so that will form the result between Σx and Σy. The results obtained from the modification algorithm is able to reduce the level of MSE (Means Square Error) performed on tests 1 and 2 that have a value of 290.95, while at K-Means MSE levels change in test 1 reaches 508.54 and test 2 reached 881.13 which gave MSE results higher than K-Means.

Keyword: K-Means, Clustering, Deviation, and MSE

1 Introduction

K-Means is one method in data mining that can be used to perform grouping/clustering of data (Winda, 2015). Clustering K-means is known as unsupervide learning because the class information on unknown object (Han and Kamber, 2012) But based on the results of research has been done K-Means method gives the result accuracy 30.00% - 60.00%. This is caused by randomly selected so it affects the different results (Han and Kamber, 2012) that affect the value of MSE (Means Square Error). In the clustering process K-Means looks for the value of proximity between objects with the selected cluster center. The value of the proximity is the central point of the cluster that mediates the object. So it takes the method deviation in determining the middle value which will serve as the center of cluster that can reduce the Means Square Error value in the clustering process. The purpose of this research is to improve the performance between K-Means method and K-Means method with Deviation in clustering and minimize means square error. So as to improve the accuracy of the K-Means method.

ICASI 2019, July 18-19, Banda Aceh, Indonesia

DOI 10.4108/eai.18-7-2019.2288601

2 Research Methods

2.1 Data Collection Methods

The data used in the study is data sourced from the academic data of the state university field that is student data force 2010 and 2011 which has a total of 1089 data. Variables taken in this research are Semester Credit Unit (SKS), and semester.

2.2 Data Analysis Method

2.2.1 Average Deviation

The size of variability is crucial for drawing a series of data, especially if one wants to compare two or more data sets. As in this case the variables of sks and semesters have the same variables as $x_1 - x_n \; and \; y_1 - y_n$. The mean deviation involves all observational data in its calculations and variability measured by comparing individual observation data with its data center (Bambang, 1994) The mean deviation is formulated in:

$$\text{Average Deviation} = \frac{\sum |X_i - \overline{X}|}{n} \quad (1)$$

X_i : the i data of random variable X
X : Average Sample
n: Sample Size

pseudo code deviation calculation is

```
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;
End

2.2.2 K-Means

K-means is one of the clustering methods. Clusters depend entirely on the selection of early centric groups. The data element K is selected as the starting center; Then the distance of all Element data is calculated by Euclidean distance formula. Data elements less than centroids distance are moved to the appropriate cluster. The process continues until no more changes occur in the [k-1] group. Grouping this partition is the most popular and fundamental technique (Han & Kamber, 2012). Here are the steps of the K-Means algorithm (Rahmawati, 2016):

a) Determine the many k-clusters to be formed.
b) Generating random values for the center of the initial cluster (centroid) as much as k-clusters.
c) Calculates the distance of each input data on each centroid using the Eucledian Distance formula until it finds the closest distance of each data with the centroid. Here is the Eucledian Distance equation:

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

With d (xi, μi) is the distance between the cluster x and the cluster center μ in the word ke i x_i is the i-th weight of the cluster to be searched for distance, μi the weight of the word to i at the center of the cluster.

d) Classify each data based on its proximity to the centroid (smallest distance).
e) Updating the centroid value. The new centroid value is obtained from the average cluster in question by using the formula:

$$C_k = \frac{1}{n_k} \sum d_i \quad (3)$$

Where:
nk : the amount of data in the cluster.
di : the sum of the value of the incoming distance in each cluster.
f) Doing looping from steps 2 to 5 until the members of each cluster nothing has changed.
g) If step 6 has been fulfilled, the mean value of the cluster center (μj) in the last iteration will be used as a parameter to determine the classification of the data.

2.2.3 K-Means with Deviation

After the value of semester (x) and SKS (y) is obtained, then data processing by using deviation method in finding the center point of the cluster. These are the steps of the K-Means Deviation:

a) Sort and submit an object which is then divided by the number of clusters specified so as to form a group based on the number of clusters formed.
b) Then the result of l is calculated by using the formula:

$$\text{Average Deviation} = \frac{\sum|X_i - \overline{X}|}{n} \quad (4)$$

X_i : the i data of random variable X
X : Average Sample
n : Sample Size
The results of these calculations serve as the center of the cluster.

c) Calculates the distance of each input data on each centroid using the Eucledian Distance formula until it finds the closest distance of each data with the centroid. Here is the Eucledian Distance equation:
$$d(x_i, \mu_i) = \sqrt{(x_i - \mu_i)^2} \quad (5)$$

With d $(xi, \mu i)$ is the distance between the cluster x and the cluster center μ in the word ke i x_i is the i-th weight of the cluster to be searched for distance, μi the weight of the word to i at the center of the cluster.

d) Classify each data based on its proximity to the centroid (smallest distance).
e) Updating the centroid value. The new centroid value is obtained from the average cluster in question by using the formula:
$$C_k = \frac{1}{n_k}\sum d_i \quad (6)$$

Where:
nk : the amount of data in the cluster
di : the sum of the value of the incoming distance in each cluster
f) Doing looping from steps 2 to 5 until the members of each cluster nothing has changed.
g) If step 6 has been fulfilled, the mean value of the cluster center (μj) in the last iteration will be used as a parameter to determine the classification of the data.

2.2.3 Performance Measurement

Measurement of performance on K-Means method by using deviation compared with K-Means method, measured using using Mean Square Error (MSE) method. The formula used to calculate Mean Square Error (MSE) can be seen below (Rouhgier, 2016):
$$MSE = \frac{1}{n}\sum_j (y_{ij} - y_j)^2 \quad (7)$$

Where :
y_{ij} : Actual value
y_j : Value reached
n : Amount of data

3 Results and Discussion

In this study the number of clusters formed are 3 clusters consisting of 2 tests. The test was conducted to determine the level of consistency of accuracy between K-Means method with K-Means Modification method as measured by the number of Means Square error generated. Comparison of both methods can be seen from the table below:

Table 1. Number of Means Square Error Comparison

Iteration	K-Means		K-Means Deviation	
	test 1	test 2	test 1	test 2
1	321,3	2088	231,20	231,20
2	756	543,8	643,34	643,34
3	632	763,4	212,0	212,0
4	611,10	578,98	190,14	190,14
5	419,10	431,5	178,10	178,10
6	410,10	-	-	-
MSE	508,54	881,13	290,95	290,95

From Table 1 it can be seen that the very large difference between the number of K-Means calculations in tests 1 and 2 provides very different MSE levels. And the MSE level yields greater value in the second test. The high error rate in K-Means in tests 1 and 2 is due to the distance of predicted cluster center point ($\bar{y}$) with the actual value (y) on the object. However, the K-Means Modification method in tests 1 and 2 gives the same result, this is because the selected cluster point using this method gives the same value in tests 1 and 2 and the selected cluster values provide the distance between predicted cluster points (y) with the actual object value (y) closer, resulting in a lower MSE value than the K-Means method. The following is the rate of development of MSE:

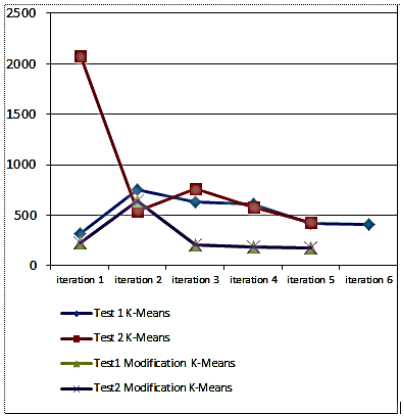

Fig.1. Means Square Rate Comparison Error Using K-Means and K-Means Deviation

From Table 1 it can be seen that the very large difference of MSE development rate so that we can average the rate of MSE development on test 1 and 2. So the bar graph can be seen below:

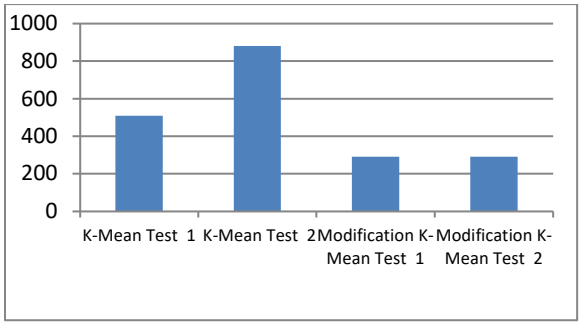

Fig.2. Number of Means Square Error Comparison Using K-Means and K-Means Deviation

Table 2. Number of Means Square Error Comparison

No	Methods	test 1	test 2
1	K-Means	508,54	881,13
2	K-Means Deviation	290,95	290,95

4 Conclusion

The high level of means square error is caused by the distance of the x and y values as the center of the cluster compared to the values of the x and y objects. In the K-Means deviation method the MSE value is much lower than K-Means because the cluster's center point is in the middle of the object and when performing the next test the cluster value is formed equal to the previous cluster value, resulting in cluster, iteration, and MSE the same as the first test. While K-Means there are objects that are too far away with the central point of the cluster selected at random and give different results when doing the next test.

References

[1] Bambang. (1994). Statistic 1 (Descriptive). Jakata:Gunadarma
[2] Han, J. & Kamber, M. (2012). Data Mining Concepts and Techniques. Elsevier: Amsterdam.
[3] Rahmawati. (2016). Clustering Analysis Using K-Means and Hierarchical Clustering Methods. Jurnal. Universitas Sebelas Maret.
[4] Rougher J. (2016). Ensemble Averaging and Mean Squared Error. Jurnal of Climate. Vol 29(4): 1-6
[5] Ismkham. (2018). I-k-means−+: An iterative clustering algorithm based on an enhanced version of the k-means. Elsevier. Vol 79:402-413

[6] Winda. (2015). Clustering Using K-Means Method To Determine Underfive Nutritional Status. Jurnal Informatika.15(2):160-174