Analysis of the effect early cluster centre points on the combination of k-means algorithms and sum of squared error on k centroid

D Selvida[1,*], M Zarlis[2], Z Situmorang[3]

[1]Student, Department of Information Technology, Faculty of Computer Science and Information Technology, Universitas Sumatera Utara, Indonesia
[2]Department of Information Technology, Faculty of Computer Science and Information Technology, Universitas Sumatera Utara, Indonesia
[3]Department of Information Technology, Faculty of Computer Science and Information Technology, Universitas Katolik Santo Thomas Medan, Indonesia

*desilia.selvida@gmail.com

Abstract. K-Means clustering is a clustering algorithm based on a partition with data only entered into one grub K, the algorithm determines the number of grub at the beginning and defines the set of K centroid. The initial determination of the cluster center is very influential on the results of the clustering process in determining the quality of clustering. The results of better clustering are often obtained after several trials. Sum of Squared Error (SSE) is a representation of homogeneity or uniformity within a cluster. In this study, the Sum of Squared Error (SSE) was used as an approach to determine the center point of the initial cluster of the K-Means algorithm. Tests were carried out on 2 datasets and the number of centroids 2,3,4,5,6,7,8, and 9 obtained values of centroids 3 and 4 in iris data had better number of iterations using a combination of K-Means and Sum of Squared Error (SSE). These results prove that the grouping with the method of determining the cluster center starting with the K-Means algorithm is based on the minimum Sum of Squared Error value that can improve clustering results and increase the value of Sum of Squared Error (SSE), compared to conventional cluster center points.

1. Introduction

Clustering is the objects (data) on the same group the each other and different objects on the another group. The greater the similarity (homogeneity) in a group and the greater the difference between groups, the better or clearer the grouping [1].

Increasing the cluster center and minimize the distance with the number of clusters that have been determined, it is difficult to predict the right K value. K-Means clustering is a clustering algorithm based on a partition with data only inserted into one grub K, the algorithm determines the number of grub at the beginning and defines the set of K centroid [2]. The initial determination of cluster centers is very influential on the results of the clustering process in determining the quality of clustering. Therefore, the determination of the initial cluster center is very important in the K-Means algorithm. There are several methods for selecting the initial cluster center, such as random methods, based on maximum and minimum distances, densities, and quadratic groupings [3].

2. Research Background

The results of Xiong, X research indicate that K-Means algorithm can increase the stability and accuracy of groupings [3]. In Zhang M research that the k-means algorithm when choosing the initial grouping must choose random. Experimental results show better accuracy and stability [4]. The Khairul U S experimental results show that the k-means algorithm can be improved efficiently in grouping speeds and good accuracy in reducing complexity [5].

3. Proposed Method

3.1. The classification data

Data grouping have to using an approach to find similarities in the data so as to be able to put data into the right groups. Data grouping will divide the data set into several groups where the similarity in a group is greater when compared to other groups [6].

3.2. Sum of Squared Error (SSE)

Sum of Squared Error (SSE) states the total sum of the squared values of the data distance with the cluster center (Rebagliati, 2013). The smaller the SSE value is the better clustering results. The Sum of Squared Error is stated by the following formula [7].

$$SSE = \sum_{i=1}^{k} d(p_i - m_i)^2 \quad (1)$$

Note:

D is minimum distance of results betwen data with clustering point center

q is feature or atribute data from data to i

n is feature or atribute of clustering point center to i

3.3. K-Means Clustering

Clustering is one of the data mining algorithms that is non-directive (unsupervised). There are two types of data clustering that are often used in the process of grouping data, namely hierarchical (hierarchical) data clustering and non-hierarchical (non hierarchical) data clustering. K-Means is one method of non-hierarchical clustering data that attempts to partition existing data into one or more clusters / groups.

In general, grouping data with the K-Means method can be done with algorithms, including: (1) determining the number of groups, (2) allocating data into groups randomly, (3) calculating the center of the group from the data in each group, (4) allocate each data to the nearest average, (5) return to step (3) if there is still data that moves groups, or if there is a change in the value of the centroid above the specified threshold value, or if the value changes in the function the objective used is still above the specified threshold value. The centroid location of each group taken from the mean of all data values for each feature must be recalculated. If M states the amount of data in a group, i declares the i i feature in a group, and p denotes the data dimension [1].

To calculate the centroid, the ke-i formula is used:

$$C_i = \frac{1}{M} \sum_{j=1}^{M} X_j \quad (2)$$

The formula is done as many as p dimensions to i starts from 1 to p. Measuring distance in Euclidean spacing using a formula:

$$D(x_2, x_1) = \|x_2 - x_1\|_2 = \sqrt{\sum_{j=1}^{p} |x_{2j} - x_{1j}|^2} \quad (3)$$

D is the distance between x_2 and x_1, and |. | is absolute point.
Measuring distance in the Manhattan spacing uses a formula:

$$D(x_2, x_1) = \|x_2 - x_1\|_1 = \sum_{j=1}^{p} |x_{2j} - x_{1j}| \quad (4)$$

Measuring distance in the Minkowsky spacing uses a formula:

$$D(x_2, x_1) = \|x_2 - x_1\|_\lambda = \sqrt[\lambda]{\sum_{j=1}^{p} |x_{2j} - x_{1j}|} \quad (5)$$

This method partition the data into clusters / groups so that data that has the same characteristics are grouped into one and the same cluster of data that has different characteristics grouped into other groups.

4. Results and Analysis

In this research to processing K-Means methode for classification data, needed center point of the cluster according to the number of desired clusters of data. In this test the author conducted a test with iris data and wine quality data of 100 data and 4 attributes with parameter points of variation in centroid values 2,3,4,5,6,7,8,9 and centroid center points randomly. The details of the data used are shown as follows.

Table 1. Dataset Iris

o	Name of Item	1	2	3	4
	IrisSetosa	1	5	4	2
	IrisSetosa	9		4	2
	IrisSetosa	7	2	3	2
	IrisSetosa	6	1	5	2
	IrisSetosa		6	4	2
	IrisSetosa	4	9	7	4
	IrisSetosa	6	4	4	3
	IrisSetosa		4	5	2
	IrisSetosa	4	9	4	2
0	IrisSetosa	9	1	5	1
	:				
	:				
	IrisVirginica				

0		2		8	6

Explanation:

X1 = Sepal length in cm
X2 = Sepal width in cm
X3 = Petal length in cm
X4 = Petal width in cm

Table 2. Dataset Wine Quality

o	Name of Item	1	2	3	4
	Wine Quality Red	4	7		9
	Wine Quality Red	8	8 8		6
	Wine Quality Red	8	7 6	. 4	. 3
	Wine Quality Red	1 2	2 8	5 6	9
	Wine Quality Red	4	7		9
	Wine Quality Red	. 4	6 6		. 8
	Wine Quality Red	9	6	. 6	6
	Wine Quality Red	. 3	6 5		. 2
	Wine Quality Red	. 8	5 8	0 2	
0	Wine Quality Red	5	5	3 6	1
	:				
	:				
0 0	Wine Quality White	. 8	. 2	. 5 9	. 9

Explanation:

X1 = Fixed acidity
X2 = Volatile acidity
X3 = Citric acid
X4 = Residual sugar

To processing clustering of classification data, we have to needed the point center according to the number of desirable clusters of data. In this study the author varied the value of centroid (k) in testing with a combination of k-means algorithm and Sum of Squared Error. The results of the tests are shown in the following table.

Table 3. Result of Testing Data Iris

Centroid (K)	K-Means Iterations	K-Means Iterations dan Sum of Squared Error
2	10	7
3	6	4
4	8	4
5	9	5
6	11	7
7	10	11
8	15	11
9	13	10

Table 4. Result of Testing Data Wine Quality

Centroid (K)	K-Means Iterations	K-Means Iterations and Sum of Squared Error
2	10	8
3	9	6
4	9	7
5	13	11
6	17	11
7	13	11
8	16	11
9	19	13

To see clearly the performance generated from each method on all data used in this study, shown in the following figure.

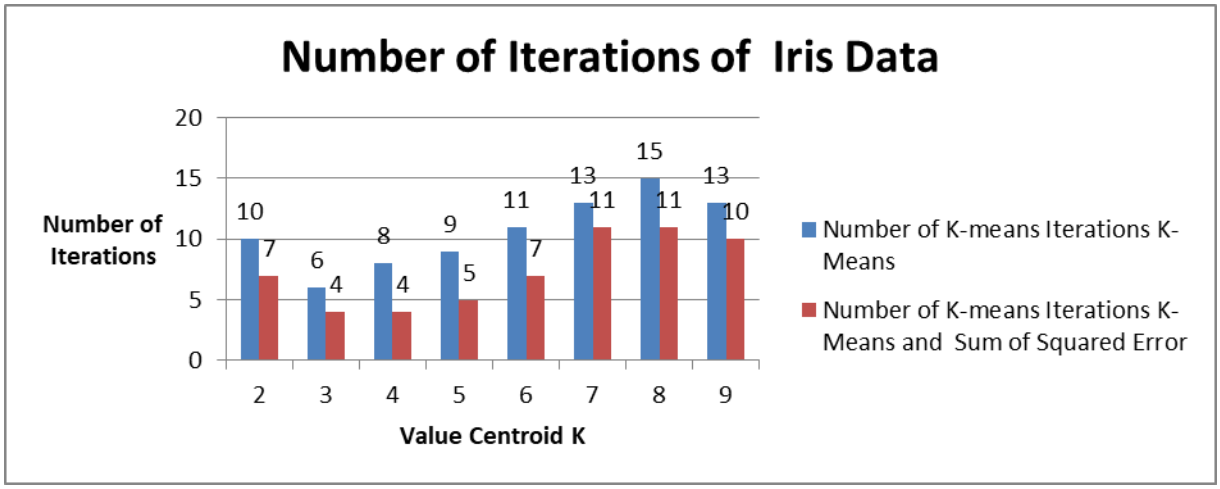

Figure 1. Number of Iterations of Iris data Graph

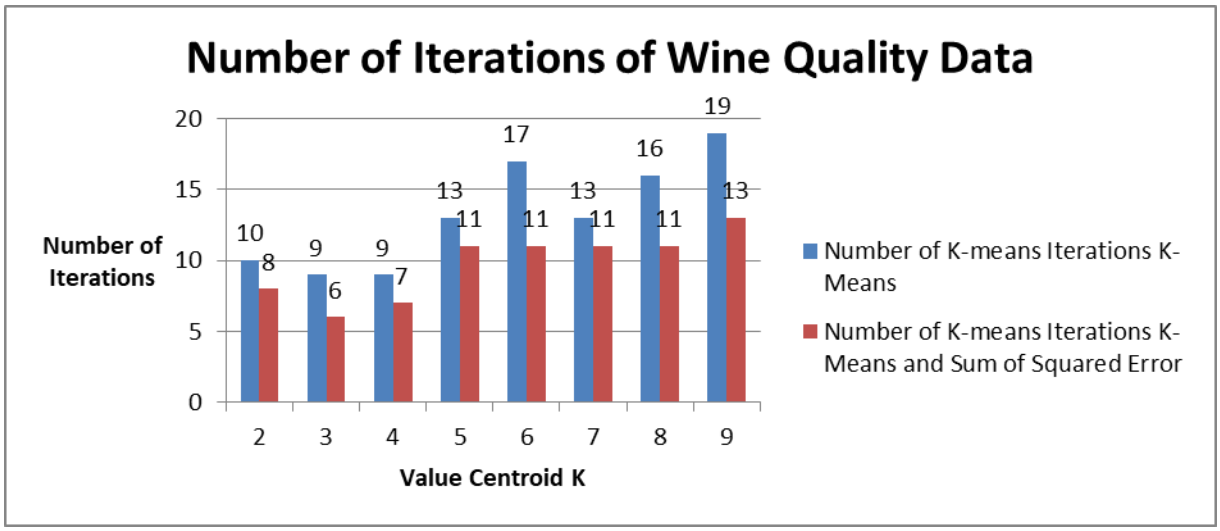

Figure 2. Graph of Number of Iterations of Wine Quality Data

Figure 1, and figure 2 shows that the proposed method is able to provide better performance than K-Means. The results of the authors conducted a test with variations in the number of centroids (K) with a value of 2,3,4,5,6,7,8,9. The authors concluded the number of centroids 3 and 4 had a better iteration of values than the number of centroids that were increasingly high and low based on iris and wine quality datasets.

5. Conclusion

Based on the results of clustering testing with the Sum of Squared Error in both datasets, the method of determining cluster centre points based on the value of minimum Sum of Squared Error (SSE) can improve the quality of clustering and show better clustering results compared to the method of determining cluster centre conventional.

References

[1] Prasetyo, E. (2014). Data Mining-Mengolah Data Menjadi Informasi Menggunakan MATLAB . Yogyakarta: ANDI.

[2] Nayak, J., Kanungo, D.P., Behera, H.S. 2016. An Improved Swarm Based Hybrid K-

Means Clustering for Optimal Cluster Centers. Advances in Intelligent Systems and Computing. 343-352.

[3] Xiong, C., Hua, Z., Lv, Ke. & Li, X. 2016. An Improved K-means text clustering algorithm By Optimizing initial cluster centers. *International Conference on Cloud Computing and Big Data* : 265 - 268.

[4] Zhang, M., Duan, K.F. 2015. Improved Research To K-Means Initial Cluster Centers. Ninth International Conference on Frontier of Computer Science and Technology. 349-353.

[5] Khairul, U S, Zulfahmi, M. & Aldi A N. 2017. Perbandingan Rapid Centroid Estimation (RCE) — K Nearest Neighbor (K-NN) Dengan K Means — K Nearest Neighbor (K-NN). InfoTekJar (Jurnal Nasional Informatika dan Teknologi Jaringan). 79-89.

[6] Gorunescu, F. 2011. *Data Mining Concept, Model and Techniques*. Springer-Verlag : Berlin.

[7] Qi, J., Yu, Y., Wang, L. & Liu, J. 2016. K*-Means: An effective and efficient kmeans lustering algorithm. Big Data and Cloud Computing (BDCloud), Social Computing and Networking (SocialCom), Sustainable Computing and Communications (SustainCom), 2016 IEEE International Conferences on IEEE, pp. 242-249.