IMPLEMENTATION OF K-MEANS ALGORITHM ELBOW METHOD AND SILHOUETTE COEFFICIENT FOR RAINFALL CLASSIFICATION

[1]Daniel Adrian Setiady, [2]Hironimus Leong
[1,2]Program Studi Teknik Informatika Fakultas Ilmu Komputer,
Universitas Katolik Soegijapranata
[2]marlon.leong@unika.ac.id

ABSTRACT

Rain is one of the hydrological cycles which is a cycle of water rotation from the earth to the atmosphere and back to the earth continuously. High Rainfall may cause some areas that are in lowlands or those with low water infiltration systems will be very susceptible to flooding. For that it is necessary to have a system to classify weather data and rainfall in each city and district so the city that has high rainfall and extreme weather can be given special attention to prevent any natural disaster like flooding. The collected data will be processed with K-Means algorithm to classify the cities or district that have low, medium, high, or very high rainfall data. In the K-Means algorithm the amount of k or cluster usually determined by randomly, on this project will be used a method that is Elbow Method to determine the value of k or cluster and Silhouette Coefficient Method will be used for testing the quality amount of a cluster. The data that will be used is the rainfall data from dataonline.bmkg.go.id at a certain period of time to be classified using the K-Means algorithm. The elbow method and the silhouette method can be used in selecting a good optimal number of clusters, and both methods mostly have the same results in determining the optimal number of clusters, it can be seen that the calculation of accuracy between using the optimal number of clusters is higher rather than not using the amount optimal number of clusters. This can be seen in the results of the clustering in Semarang on February 1 - 28, 2021, when using the amount of K = 4 produce the accuracy result 92.8571429 %, while when using the optimal number of cluster K=3 the accuracy result is higher (97.6190476 %). In the Cilacap city classification on April 1-30 2021, the elbow method and the silhouette coefficient method produce different optimal cluster results, but the accuracy obtained when using the optimal number of clusters from the silhouette coefficient (85.7142857 %) is higher than using the optimal cluster from the elbow method.(74.6031746 %), but when the data is processed with centroid on table 5.10, the elbow method and silhouette coefficient method produce the same amount of optimal number of clusters is 2. This shows that differences in the use of the initial centroid point can affect the results of the elbow method and the silhouette coefficient method

Keywords: Classification, K-Means, Elbow Method, Silhouette Coefficient, Weather

INTRODUCTION

Rain is one of the hydrological cycles which is a cycle of water rotation from the earth to the atmosphere and back to the earth continuously. High Rainfall may cause some areas that are in lowlands or those with low water infiltration systems will be very susceptible to flooding. For that it is necessary to have a system to classify weather data and rainfall in each city and district so the city that has high rainfall and extreme weather can be given special attention to prevent any natural disaster like flooding.

To make that system, then an algorithm is needed to classify rainfall and weather data. Rainfall can be classified into 6 types that is cloudy (0mm), light (0.5-20mm), moderate (20mm – 50mm), heavy (50mm-100mm), very heavy (100mm-150mm), and extreme (>150mm). With this classification, we can easily determine the district or city that need to be given special attention.

The collected data will be processed with *K-Means* algorithm to classify the cities or district that have low, medium, high, or very high rainfall data. In the *K-Means* algorithm the amount of *k* or *cluster* usually determined by randomly, on this project will be used a method that is *Elbow Method* to determine the value of *k* or *cluster* and *Silhouette Coefficient Method* will be used for testing the quality amount of a *cluster.* The data that will be used is the rainfall data from dataonline.bmkg.go.id at a certain period of time to be classified using the *K-Means* algorithm.

METHODS

The algorithm used in this study is the K-Means algorithm and will be assisted by the Elbow Method and the Silhouette Coefficient. The Elbow Method is a method that can be used to determine the number of clusters based on the WCSS (Within Cluster Sum of Squares) value, while the Silhouette Coefficient is a method that can be used to see the quality of a cluster. The Silhouette Coefficient has a range of values between -1 and 1, a good cluster quality value will be indicated by a value close to 1. To determine the distance of each data to the center of the centroid in each cluster, the Euclidian Distance formula in the K-Means algorithm will be used. The first stage is to determine the number of k or clusters using the Elbow Method. The Elbow Method uses WCSS (Within Cluster Sum of Squares) or the distance between each data and each cluster at this stage will also determine the centroid point randomly. After determining the number of k or clusters obtained from the Elbow Method, the Silhouette Coefficient method will be used to determine how much the quality of the cluster is compared to the number of other clusters. After finding a good number of clusters, the next step is to calculate the data with K-Means. The next step after determining the number of k or clusters using the Elbow Method and the Silhouette Coefficient, the K-Means algorithm will be calculated starting with randomly determining the centroid point, the first iteration process will be carried out, in the first iteration the Euclidian Distance formula will be used to determine the distance of each data with the center of the centroid.

K-Means Algorithm

The K-Means Clustering algorithm is an iterative clustering algorithm that partitions the data set into a number of K clusters that have been set at the beginning [5]. Data with similar characteristics are grouped into one cluster/group and data with different characteristics are grouped with other clusters/groups so that data in one cluster/group has a small degree of variation [6]. The proximity of two objects is determined by the distance between them. Likewise, the proximity of a data to a particular cluster is determined by the distance between the data and the center of the cluster. The closest distance between one data and a certain cluster will determine which data belongs to which cluster [1]. Calculation of the distance of all data to each cluster center point using the Euclidean distance theory which is formulated as follows:

$$d(x,y)\sqrt{\sum_{i=1}^{n}(x^i - y^i)}^{2} \quad (1)$$

$where$
$x^i = first\ point$
$y^i = second\ point$

According to Prasetyo [3], the steps for performing the K-Means Clustering Algorithm areas follows:

1. Determine the value of K as the number of clusters.
2. Select K from the dataset X as the centroid.
3. Allocate all data to centroid with distance metric using equation 1.
4. Recalculate centroid C based on the data that follows each cluster. Repeat

Elbow Method

The Elbow method is a method which is used to generate information in determining the best number of clusters by looking at the percentage of comparison results between the number of clusters that will form an angle at a point [8]. This method provides ideas by selecting the cluster value and then adding the cluster value to be used as a data model in determining the best cluster. And besides that the percentage of the resulting calculation is a comparison between the number of clusters added [10]. The results of different percentages of each cluster value can be shown by using a graph as a source of information. If the value of the first cluster with the value of the second cluster gives the angle in the graph or the value has decreased the most, then the value of the cluster is the best [11].

$$WCSS = \sum_{i=1}^{k} (x^i - c^i) \quad (2)$$

$where$
$x^i = first\ point$
$y^i = centroid\ of\ data(i)$

To get the comparison is by calculating the SSE (Sum of Square Error) of each cluster value. Because the larger the number of K clusters, the smaller the SSE value will be. SSE formula on K-Means [9].

Elbow Method Algorithm in determining the value of K on K-Means

1. Start
2. Initial initialization of K . value
3. Increase the value of K 4. Calculate the sum of square error of each value of K
4. See the result of the sum of square error of the K value which has dropped drastically
5. Set the value of K that is in the form of an angle
6. Done [11]

Silhouette Coefficient

The Silhouette coefficient is used to see the quality and cluster strength, how well an object is placed in a cluster. This method is a combination of cohesion and separation methods. Calculation stage The Silhouette coefficients are as follows:

1. Calculate the average distance from a document for example I with all other documents in one clusters (a(i))
2. Calculate the average distance from the document I with all documents in another cluster, and retrieved smallest value (b(i))
3. The Silhouette Coefficient score is:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i)} \quad (3)$$

IMPLEMENTATION AND TESTING

Below is the result for the final clustering on K=2 – K=4

Table 1. K=1 Final Centroids

Cluster	Average Temperature	Average Humidity	Rainfall	Sunshine	Average Wind Velocity
C1	25	90	173.5	4	3
C2	26.48888889	91.55555556	27.5	2.988888889	3.666666667

Table 2. K=2 Final Centroids

Cluster	Average Temperature	Average Humidity	Rainfall	Sunshine	Average Wind Velocity
C1	25	90	173.5	4	3
C2	26.48888889	91.55555556	27.5	2.988888889	3.666666667

Table 3. K=3 Final Centroids

x	Average Temperature	Average Humidity	Rainfall	Sunshine	Average Wind Velocity
C1	25	90	173.5	4	3
C2	26.16	92	43.38	21.18	3.4
C3	26.9	91	7.65	5.25	4

Table 4. K=4 Final Centroids

Cluster	Average Temperature	Average Humidity	Rainfall	Sunshine	Average Wind Velocity
C1	26.2	92.5	36.475	0.85	3
C2	26.9	91	7.65	5.25	4
C3	25	90	173.5	4	3
C4	26	90	71	2.5	5

Count WCSS to Calculate Elbow Method

Table 5. WCSS on K=1

Data Point	Distance from Centroid	Class	Distance with Centroid Squared
1	29.48361748	C1	869.2837
2	35.57228837	C1	1265.3877
3	37.20733395	C1	1384.3857
4	28.9757433	C1	839.5937
5	13.25412011	C1	175.6717
6	131.4188103	C1	17270.9037
7	4.808918797	C1	23.1257
8	5.776478166	C1	33.3677
9	9.024616335	C1	81.4437

10	36.22609143	C1	1312.3297
Sum			23255.493
WCSS			23255.493

Then the next process is to count all WCSS until K=4

Table 6. Result all WCSS on K=1-K=4

K	WCSS
1	23255.493
2	4065.6
3	1187.588
4	223.5975

The calculation process has been done and the amount of WCSS from k=1 to k=4 are 23255.493, 4065.6, 1187.588, 223.5975. and ilustration below will show the graphs of WCSS.

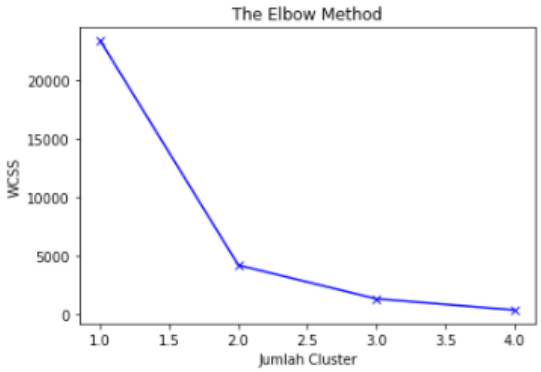

Figure 1. Elbow Method Graph

From that illustration above, the point starts to decrease linearly and form a elbow at the amount of k is equal to 2. so based on Elbow Method the amount of cluster is set to 2.

Silhouette Coefficient

Table 7. Silhouette Coefficient Score

Number of Clusters	Silhouette Coefficient
2	0.841706224
3	0.645288222
4	0.776207438

From the table above the amount of 2 cluster have the best score for silhouette coefficient, this makes the k=2 is an optimal cluster

Final Result

Table 8. Test Results

City	Start Date	End Date	Elbow Method	Silhouette Score	K used	Accuracy
Semarang	1 February 2021	28 February 2021	3	3	4	92.8571429 %
					3	**97.6190476 %**
Cilacap	1 April 2021	30 April 2021	3	2	**2**	**85.7142857 %**
					3	74.6031746 %
Cilacap (different centroid)	1 April 2021	30 April 2021	2	2	**2**	**100 %**
					3	68.2539683 %
Semarang	1 January 2021	31 May 2021	2	2	**2**	**95.3642384 %**
					3	77.4834437 %
					4	82.781457 %
Tegal	1 January 2021	31 May 2021	2	2	**2**	**94.0740741 %**
					3	74.3209877 %
					4	91.4814815 %
Cilacap	1 January 2021	31 May 2021	2	2	**2**	**95.8677686 %**
Jakarta	1 January 2021	31 May 2021	2	2	**2**	**96.4912281 %**
					3	79.5321637 %
					4	91.4814815 %
Bandung	1 January 2021	31 May 2021	2	2	**2**	**92.71523181%**
					3	79.6909492 %
					4	76.4900662%

CONCLUSION

Based on the results of the test above, can be conclude that:

1. K-Means Algorithm can be used to classify weather, specifically rainfall data
2. The cluster quality test results (determining the optimal number of K) from the elbow method and the silhouette method have the same results, but there are times when both showing a different number of optimal cluster.
3. The differences in the use of the initial centroid point can affect the results of the elbow method and the silhouette coefficient method. although different but the results are similar to each other.
4. On Cilacap 1-30 April 2021 Silhouette Coefficient method can produce the amount of optimal cluster consistently with different centroids

REFERENCES

[1] Santosa B, "Data Mining : Teknik Pemanfaatan Data untuk keperluan Bisnis," Graha Ilmu-Yogyakarta, 2007.

[2] Santoso, "Statistik Multivariat," Elekmedi Komputindo-Jakarta, 2005. [3] Prasetyo, Eko, "Data Mining Mengolah Data menjadi Informasi dengan Matlab," Andi- Yogyakarta, 2009.

[4] Dephut, " P.32/Menhut-II/2009 tentang Tata cara penyusunan rencana teknik rehabilitasi hutan dan lahan Daeah Aliran sungai". Jakarta, 2009.

[5] Rajagopal, Sankar, "Customer data clustering using data mining Technique," in International Journal of Database Management System Vol.3 No.4, 2011.

[6] Oscar, Johan Ong, "Implementasi Algoritma K-Means Clustering untuk menentukan Strategi Marketing President University", Jurnal Ilmiah Teknik Industri, Vol.12 No.1, 2013.

[7] Handoyo, Rendy, dkk. 2014. Perbandingan Metode Clustering Menggunakan Metode Single Linkage dan K-Means Pada Pengelompokan Dokumen. JSM STMIK Mikroskil, volume 15, no 2.

[8] Madhulatha, T.S., 2012. An Overview On Clustering Methods. IOSR Journal of Engineering, II(4), pp.719-725

[9] Irwanto, et. al (2012). Optimasi Kinerja Algoritma Klasterisasi K-Means untuk kuantisasi Warna Citra. Jurnal Teknik ITS, I(1), pp.197-202.

[10] Kodinariya, Trupti M. & Makwana, Prashant R., (2013). Review on determining number of cluster in K-Means Clustering. International Journal of Advance Research in Computer Science and Management Studies, I(6), pp. 90-95

[11] Bholowalia, Purnima & Kumar, Arvind, 2014. EBK-Means: A Clustering Techiniques based on Elbow Method and K-Means in WSN. International Journal of Computer Application (0975-8887), IX(105), pp. 17-24