ILKOM Jurnal Ilmiah Vol. 13, No. 1, April 2021, pp. 31-35
Accredited 2nd by RISTEKBRIN No. 200/M/KPT/2020; E-ISSN 2548-7779 | P-ISSN 2087-1716

The comparison of k-means and k-medoids algorithms for clustering the spread of the covid-19 outbreak in Indonesia

Wargijono Utomo
Universitas Krisnadwipayana, Jl. Kampus UNKRIS Jatiwaringin, Kota Bekasi 13077, Indonesia
wargiono@unkris.ac.id

Article history: Received February 7, 2021; Revised March 25, 2021; Accepted April 17, 2021; Available online April 30, 2021

Abstract

The coronavirus spreads quickly through human-to-human transmission via close contact and respiratory droplets such as coughing or sneezing. Various studies have been carried out to deal with Covid-19. However, the cure for this virus has not been found until now. Based on data from the covid19.go.id page retrieved on January 1st, 2021, which was updated by the Ministry of Health, the overall number of confirmed cases was 1,078,314 active cases reaching 175,095 or 16.2% of confirmed cases, recovered 873,221 or 81.0% of confirmed cases, and death 29,998 or 2.8% of the confirmed cases. This study compares the two algorithms of data groups to analyze clustering patterns to determine the best data processing method. The data in this study sourced from the Ministry of Health, contained 4 attributes, including confirmed cases, treatment, recovery, and death cases. In this study, only 2 attributes were used: the confirmed and death cases. From the data analysis and processing results through a comparison between the K-Means method and the K-Medoids for clustering the spread of the coronavirus in Indonesia, a conclusion is derived. With the Davies Boulden index value from K2 to K9 values, it turns out that the K-Means method gets the smallest value at the K-5 of 0.064, while K-Medoids at the k-2 value of 0.411. Thus, from the two methods used, it can be concluded that the best method for clustering the spread of the coronavirus outbreaks in Indonesia is the K-Means method.

Keywords: COVID-19; Clusterization; K-Means; K-Medoids.

Introduction

COVID-19 is a disease that causes severe acute respiratory symptoms, which first emerged in Wuhan, China, at the end of 2019 [1] - [2]. Since then, the disease has been rapidly spread by tourists to almost every country in the world and was declared a pandemic by the WHO on March 22nd, 2020. This virus is hazardous because it can cause deadly symptoms such as shortness of breath, chest pain, inability to speak and difficulty breathing, even lead to death [3] - [4]. People with weak immune systems and the elderly are very vulnerable to this viral infection and its effects. The spread of this virus is effectively very fast through human-to-human transmission via close contact and respiratory droplets such as coughing or sneezing [5]. Various studies have been carried out to deal with COVID-19. However, the cure for this virus has not been found until now [6].

Based on data from the covid19.go.id page retrieved on January 1st, 2021, from 34 provinces in Indonesia, the first highest level of Covid-19 spread was from DKI Jakarta Province, with 269,178 confirmed cases and 4,273 death cases. Then, the second is in West Java Province, with confirmed cases 150,336 and 1,9732 death cases. The third is in Central Java, with confirmed cases 125,355 and 5,043 death cases. Then, the fourth is in East Java Province, with confirmed cases 112,795 and 7,805 death cases.

Various studies have been carried out by both local and foreign researchers, including the classification of COVID-19 based on incomplete heterogeneous data using the KNN algorithm [3], Classification of Covid-19 images using the deep features algorithm and fractional-order marine predators [5], K-Means Clustering COVID-19 Data [6], Clustering the spread of Corona Virus in DKI Jakarta using the K-Means Method [7], Implementation The K-Means algorithm for determining the level of the spread of the Covid-19 outbreak in Indonesia [2] and K-Medoids method clustering for learning applications during the COVID-19 outbreak [14]. From the results of previous research, this study has a difference in the method, which uses two methods as a comparison clustering between the K-Means and K-Medoids algorithms in clustering the spread of the coronavirus disease 19 in Indonesia.

http://dx.doi.org/10.33096/ilkom.v13i1.763.31-35

This study directly compares the two algorithms from the dataset to analyze clustering patterns and determine the best algorithms in data processing. While the data used sourced from the Ministry of Health, there are 4 attributes, including confirmed, treatment, recovery, and death cases. However, only 2 attributes are used in this study namely the confirmed and death cases.

In this test, the RapidMiner application version 9.8 was used. Then, the Davies-Bouldin index value was also used as a reference in clustering. From the generated clustering pattern, then the data can be analyzed. So that it becomes new information and can help decision-makers (stakeholders) reduce the spread of the coronavirus and minimize the number of patients who have been confirmed positive for Covid-19. Other than that, it can produce the best algorithm from validity testing applied to K-Means and K-Medoids.

Method

The implementation of the method used in the study of clustering the spread of the coronavirus in Indonesia with the K-Means and K-Medoids algorithms is through several stages as can be seen in **Figure 1**.

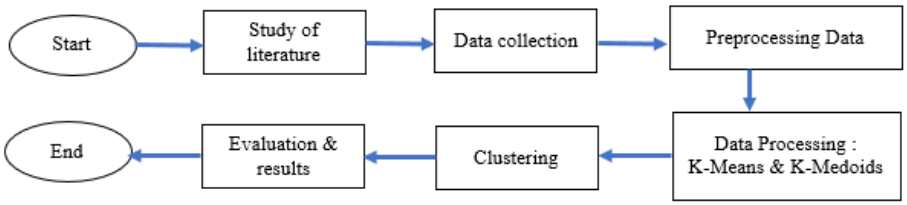

Figure 1. Research Methodology

A. Data Mining

Data mining is a series of finding the pattern of a relationship, extracting added value, both data and essential information such as knowledge, to find a relationship and simplify data. So that information can be understood and useful, and assisted through statistics, mathematics, artificial intelligence, and machine learning [7]. Clustering is the task of dividing data and vectors into several clusters according to their respective characteristics. The data with similar characteristics will be sorted into the same group. Then, the data with different characteristics will be segregated into different groups. No label is required on each processed data because it can be given after the cluster or group is formed. As it does not use class labels on each data, it is often called unsupervised learning [8]

B. K-Means Algorithm

The K-Means algorithm is an algorithm that clusters data by trying to separate data into groups so that the data that have similarities are in the same group. However, the different data are classified in other groups [9].

The following is a stage in the K-Means algorithm:

First, determine the number of groups (k), then choose the group center randomly. Second, compute the distance from each data to the cluster center. Third, cluster the data into groups with the nearest distance. Fourth, compute the new group center and then repeat the second to the fourth steps until no more data is moved to other groups [10]. In the process of clustering, it can be started by identifying grouped data, using the Euclidean Distance formula (1) (2) [11].

$$d_{(i,j)} = \sqrt{(X_{1i} - X_{1j})^2 + (X_{2i} - X_{2j})^2 + \cdots + (X_{ki} - X_{kj})^2} \quad (1)$$

Notes:

D (i, j) = The distance between the i to cluster data center j
X ki = The data to the i on the data attribute to k
X kj = The center point to j on the attribute to k

$$C = \frac{\sum m}{\text{n}} \quad (2)$$

Where C is the data center, m is a data member that enters a particular center or certain centroid, and n is the amount of data that belongs to a certain centroid.

C. K-Medoids Algorithm

The K-Medoids algorithm is part of the group clustering. The k-medoids method is quite efficient in collecting small data [12]. The first step is to determine the most representative points (medoids) in the data group by computing the distance in a cluster from all combinations of medoids so that the distance between points in a group is small, while the distance of the point between groups is large [13]. K-Medoids are objects that represent their point of reference, not taking values as the mean of an object in each group. The algorithm will take the parameter from the input of k, with the number of groups that will be segregated between one part of n objects [14] - [15].

The steps of the K-Medoids algorithm are as follows [16] - [17]:

1. Initiate a cluster center (number of clusters)
2. Allocate each data (object) to the closest cluster using the Euclidian Distance equation as in the equation formula 1:
3. Select objects randomly in each cluster as candidates for the new medoid.
4. Then compute the distance to each object in each cluster with a new medoid candidate.
5. Next, compute the total deviation (S) by computing the new total distance value with the old total distance value. When S <0, then swap objects with cluster data to form a new set of k objects as medoids.
6. Then repeat the steps 3 to 5 until there is no change in medoid so that the cluster and members of each cluster are obtained.

D. Davies-Bouldin Index (DBI)

The Davies-Bouldin Index (DBI), introduced by David L. Davies and Donald W. Bouldin in 1979, is a metric for evaluating clustering algorithms. It is an internal evaluation scheme, where the validation of how well the clustering has been done is made using quantities and features inherent to the dataset [18]. Meanwhile, the separation is based on the distance between the cluster center point to its cluster. As a measure, DBI can maximize the distance between clusters Ci and Cj while trying to minimize the distance between points in the cluster. When the distance between the clusters is maximum, that is what distinguishes each cluster significantly. Thus, the slight differences between the clusters are clearer. When the distance between the clusters is minimal, it means that each object in the cluster has a high level of characteristics. Formula (3) is the formula used in computing the Davies-Bouldin Index [19]:

$$DBI = \frac{1}{K}\sum_{i=1}^{k} Ri, qt \quad (3)$$

Where the formula; k is the number of clusters used. When the Davies-Bouldin Index value obtained is getting smaller (non-negative> = 0), then the resulting cluster will be better [10].

Results and Discussion

A. Steps in Data Pre-processing

The data in this study was sourced from the covid19.co.id website page. It was retrieved on January 1st, 2021, and was updated by the Ministry of Health on January 31st, 2021. Then the data pre-processing stage is carried out before running the clustering process of some of the attributes used. There are 34 provinces and 4 attributes from the data obtained, namely death, recovery, in care or independent isolation and confirmed cases. Furthermore, there are 2 attributes chosen as a reference in the clustering process, namely confirmed and death cases, as shown in **Table 1**.

Table 1. Dataset for the spread of covid-19 in Indonesia

No	Province	Confirmed	Death
1	DKI Jakarta	269,718	4,273
2	West Java	150,336	1,932
3	Central Java	125,355	5,043
4	East Java	112,795	7,805
5	South Sulawesi	48,261	740
6	East Kalimantan	41,212	991
…	…….	……	…….
34	North Maluku	3.452	100

B. Steps of Data Processing

In the previous stage there is a pre-processing data, which is then processed with the Rapid Miner version 9.8 application. The purpose of this research is to compare which clusters are best between the K-Means and K-Medoids

algorithms in clustering data on the spread of the coronavirus-19 disease in Indonesia in cluster 1 to cluster 10. Then there is the clustering process using these two algorithms, which in the next stage of this research is to determine the best number of clusters with Rapidminer as presented on the DBI index. This processing can be done to find the DBI index value of each algorithm in each group. Then the test process is carried out from clusters k = 2 to k = 10. Furthermore, the results of the DBI index comparison can be shown in **Figure 2**.

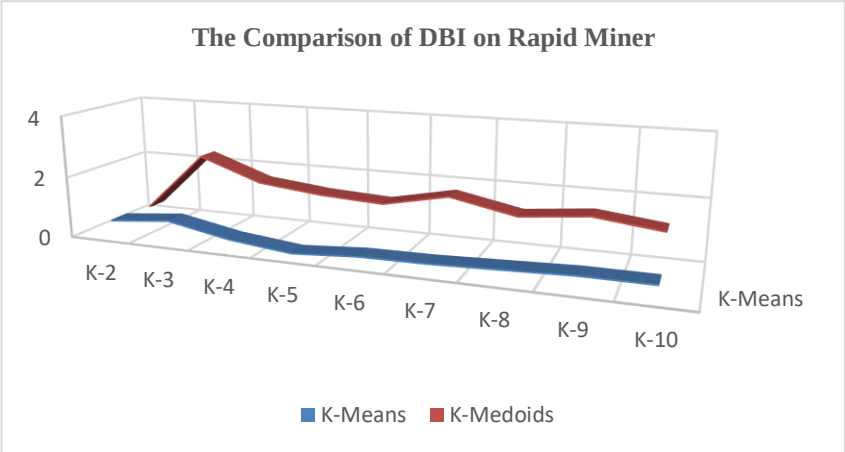

Figure 2. Comparison of the K-Means and the K-Medoids algorithm

Furthermore, it can be seen that the test results in each cluster using the K-Means and K-Medoids algorithm methods can be seen in **Table 2**.

Table 2. Results of the DBI index comparison

Klaster	*K-Means*	*K-Medoids*
K-2	0.481	0.411
K-3	0.677	2.357
K-4	0.286	1.632
K-5	0.064	1.405
K-6	0.207	1.295
K-7	0.230	1.726
K-8	0.306	1.296
K-9	0.385	1.512
K-10	0.383	1.258

After previously making comparisons through the Davies Boulden index using the K-Means and K-Medoids algorithm methods, started from K2 to K10 as presented in table 2, then, find the smallest Davies Boulden index value from K-Means, that is located in the 5th K-Means cluster with a Davies Boulden index value of 0.064. After that, find the smallest Davies Boulden index value on K-Medoids, located in the 2nd K-Medoids cluster with a Davies Boulden index value of 0.411. The analysis using the Davies Boulden index value on K-Means (k = 5) shows that the value is smaller than the K-Medoids. Therefore, it can be concluded that clustering using the K-Means algorithm is better than the K-Medoids algorithm in clustering the spread of coronavirus. Based on the analysis results, the best cluster then interprets the data clustering of the spread of coronavirus, and the results can be obtained in **Table 3**.

Table 3. The best cluster results with K-Means (K-5)

Cluster	Item	Province
1	8	South Sulawesi, East Kalimantan, Riau, West Sumatera, Banten, Bali, Special Region of Yogyakarta North Sumatera
2	1	Special Capital Region of Jakarta
3	1	West Java
4	22	South Kalimantan, Papua, South Sumatera North Sulawesi, Central Kalimantan, Lampung, Southeast Sulawesi, Aceh, Riau islands, Central Sulawesi, West Nusa Tenggara North Kalimantan, West Papua, Maluku

Cluster	Item	Province
		East Nusa Tenggara, Bangka Belitung islands, Jambi, Bengkulu, Gorontalo West Kalimantan, West Sulawesi North Maluku
5	2	Central Java, East Java

Conclusion

A conclusion is derived from the data analysis and processing results by comparing the K-Means and K-Medoids algorithms in grouping the spread of the coronavirus in Indonesia. It can be concluded that the Davies Boulden index value obtained from the two methods, namely by cluster testing from 2-k to 9-k, produces the smallest Davies Boulden index value with the K-Means algorithm, namely at a K-5 value of 0.064. Meanwhile, cluster formation using the K-Medoids algorithm method obtained the smallest K-2 Davies Boulden index value of 0.411. Therefore, from the two algorithms, it can be said that the best method for clustering the spread of coronavirus disease in Indonesia is the K-Means algorithm method.

Reference

[1] M. A. Shereen, S. Khan, A. Kazmi, N. Bashir, R. Siddique, "COVID-19 infection: Origin, transmission, and characteristics of human coronaviruses," Journal of Advanced Research 24 (2020) 91–98

[2] N. Dwitri dkk, "Penerapan Algoritma K-Means dalam Menentukan Tingkat Penyebaran Pandemi Covid-19 di Indonesia," Jurnal Teknologi Informasi, Vol. 4, No.1, Juni 2020

[3] A. Hamed et al, "Accurate Classification of COVID-19 Based on Incomplete Heterogeneous Data using a KNN Variant Algorithm, "Research Square 2020

[4] M. Rubaiyat Hossain Mondal, Subrato Bharati, Prajoy Podder, Priya Podder, "Data analytics for novel coronavirus disease," Informatics in Medicine Unlocked 20 (2020), 100374.

[5] A. T. Sahlol et al, "COVID-19 image classiication using deep features and fractional-order marine predators algorithm," Scientific Report, 10:15364 https://doi.org/10.1038/s41598-020-71294-2

[6] R. A. Indraputra , R. Fitriana, "K-Means Clustering Data COVID-19," Jurnal Teknik Industri, Volume 10 No.3. Desember 2020.

[7] A. Solichin, K. Khairunnisa, "Klasterisasi Persebaran Virus Corona (Covid-19) Di DKI Jakarta Menggunakan Metode K-Means," Fountain of Informatics Journal Volume 5, No. 2, November 2020

[8] F. Farahdinna, I. Nurdiansyah, A. Suryani, A. Wibowo, "perbandingan algoritma k-means dan k-medoids dalam klasterisasi produk asuransi perusahaan nasional," Jurnal Ilmiah FIFO, Volume 11 No. 2, November 2019.

[9] I. Parlina, A. P. Windarto, A. Wanto, M. R. Lubis, "memanfaatkan algoritma k-means dalam menentukan pegawai yang layak mengikuti asessment center untuk clustering program sdp," CESS (Journal of Computer Engineering System and Science), Volume 3, No.1, Januari 2018.

[10] P. R. Nastiti, A. B. W. Putra, "perbandingan algoritma k-means dan fuzzy c-means clustering untuk kualifikasi data kinerja dosen di jurusan teknologi informasi polnes," PROSIDING SNSebatik 2017, Vol 1 No. 1, Juni 2017.

[11] N. Agustina, P. Prihandoko, 2020, Perbandingan Algoritma K-Means Dengan Algoritma Fuzzy C-Means Untuk Clustering Tingkat Kedisiplinan Kinerja Karyawan, Jurnal RESTI, Volume 2 No.3 (2018), 621-626.

[12] I. Kamila, U. Khairunnisa, M. Mustakim, "Perbandingan Algoritma K-Means dan K-Medoids untuk Pengelompokan Data Transaksi Bongkar Muat di Provinsi Riau," Jurnal Ilmiah Rekayasa dan Manajemen Sistem Informasi, Vol. 5, No. 1, Hal. 119-125, Februari 2019.

[13] N. L. Anggraeni, "Teknik Clustering Dengan Algoritma K-Medoids Untuk Menangani Strategi Promosi Di Politeknik TEDC Bandung," Jurnal Teknologi Informasi dan Pendidikan, vol. 12 no. 2 pp. 1-7, 2019.

[14] Samudi, Slamet Widodo, Herlambang Brawijaya,2020, The K-Medoids Clustering Method for Learning Applications during the COVID-19 Pandemic, Jurnal dan Penelitian Teknik Informatika Volume 5, Number 1, Oktober 2020

[15] Iqbal Dzulfiqar Iskandar, Melisa Winda Pertiwi, Mira Kusmira, Imam Amirulloh, "komparasi algoritma clustering data media online pada proses bisnis," Jurnal IKRA-ITH Informatika Vol. 2 No. 3, November 2018.

[16] R. D. Ramadhani and D. J. Ak, "Evaluasi K-Means dan K-Medoids pada Dataset Kecil," Semin. Nas. Inform. dan Apl., no. September, pp. 20–24, 2017.

[17] H. Zayuka, S. M. Nasution, and Y. Purwanto, "Perancangan Dan Analisis Clustering Data Menggunakan Metode K-Medoids Untuk Berita Berbahasa Inggris Design and Analysis of Data Clustering Using Kmedoids Method For English News," e-Proceeding Eng., vol. 4, no. 2, pp. 2182–2190, 2017

[18] A F Khairatia , A A Adlinaa , G F Hertonoa , B D Handaria, "Kajian Indeks Validitas pada Algoritma K-Means Enhanced dan K-Means MMCA," Prosiding Seminar Nasional Matematika, PRISMA 2 (2019): 161-170.

[19] R. D. Kusumah, B. Warsito, M. A. Mukid, "perbandingan metode k–means dan self organizing map (A case study: The clustering of districts/cities in Central Java based on human development index indicators 2015)," JURNAL GAUSSIAN, Volume 6, Nomor 3, Tahun 2017, Halaman 429-437.