International Journal of Information System & Technology
Akreditasi No. 158/E/KPT/2021 | Vol. 5, No. 6, (2022), pp. 709-716

Comparison of Euclidean with Manhattan in K-Means Clustering for Grouping Palm Oil Production in the Province North Sumatra

Solikhun[1], Lise Pujiastuti[2]

[1]Information Management, AMIK Tunas Bangsa, Pematangsiantar, Indonesia
[2] Information Systems Study Program, STMIK Antar Bangsa, Tangerang, Indonesia

[1]solikhun@amiktunasbangsa.ac.id, [2] lise.pujiastuti@gmail.com

Abstract

North Sumatra is the largest palm oil-producing province in Indonesia. The region of North Sumatra has an extensive area of oil palm plantations compared to other provinces in Indonesia. To produce a good clustering of oil palm production using the K-Means Clustering method, it is necessary to compare several calculation methods to find the shortest distance in K-Means Clustering. This study focuses on comparing Euclidean Distance with Manhattan Distance on K-Means Clustering. To determine the best method of calculating the shortest distance, the researchers looked for the smallest Davies Bouldin Index (DBI). The smallest DBI value is at k=2 0.145. The result of grouping oil palm production in Sumatra province with k=2 is the high group being Asahan, Langkat and North Labuhanbatu regencies, while 30 other regencies/cities are in the low group.

Keywords: *Data Mining, Clustering, K-Means, Davies Bouldin Index.*

1. Introduction

K-Means uses the average value (mean) as the cluster's centre. The K-Medoids algorithm has the advantage of overcoming the weaknesses of the K-Means algorithm, which is sensitive to noise and outliers, where objects with large values allow them to deviate from the data distribution[1][2][3]. DBI is used to get the performance of a clustering algorithm to get optimal clustering [4]. There are several distance calculations in the K-Means clustering algorithm: Euclidean Distance and Manhattan Distance. The Euclidean Distance method is a method of finding the proximity of the distance of two variables; besides being easy, this method is also more time-efficient, and the process is fast. Euclidean Distance is a heuristic function obtained based on direct barrier-free distances, such as to get the value of the length of the diagonal line in a triangle. But before getting the results, both points must be represented in 2-dimensional coordinates (x, y)[5]. Manhattan Distance is used to calculate the absolute difference (absolute) between the coordinates of a pair of objects[6]. DBI is used to obtain optimal clustering results by comparing distance calculations using Euclidean Distance and Manhattan Distance on the K-Means Clustering algorithm by looking at the smallest DBI value.

2. Research Methodology

a) Davies Bouldin Index (DBI)

Using the Davies Bouldin Index technique, the cohesion matrix (closeness of one group) and the separation matrix will be known (differences between groups). The smaller the Davies Bouldin Index value, the more optimal the grouping model. The Davies Bouldin Index evaluation will produce an optimal grouping[7].

b) Distance Calculation

Several ways to determine the closest distance in the K-Means Clustering algorithm include [8]:

ISSN : 2580-7250

1. Calculation of Euclidean Distance
The formula to find the closest distance value using Euclidean Distance:

$$D(x,y) = \left(\sum_{k=1}^{m} (X_{ik} - X_{jk})^2 \right)^{\frac{1}{2}} \quad (1)$$

Xi = X value in training data
Yi = Y value on test data
M = Limit of the amount of data

2. Calculation of Manhattan Distance
The formula to find the closest distance value using Manhattan Distance:

$$d(x,y)^n = \sum_{i=1}^{m} |xi - yi| \quad (2)$$

Xi = X value in training data
Yi = Y value on test data
M = Limit of the amount of data

c) Research Framework

To produce an optimal grouping, it is carried out in several stages as shown in Figure 1.

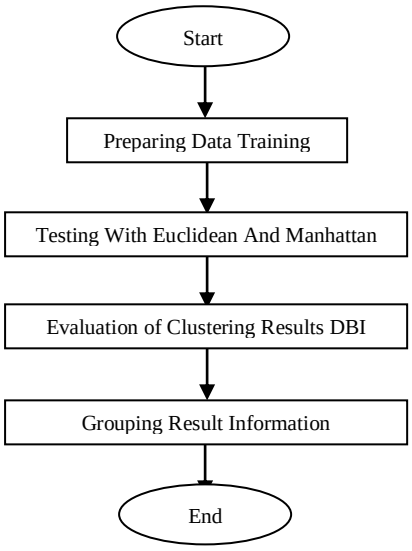

Figure 1. Research Framework

3. Results and Discussion

a) K-Means Algorithm Calculation With Euclidean Distance Using Cluster 3

This research data was taken from the Central Statistics Agency of North Sumatra in oil palm production in the province of North Sumatra from 2018-to 2020.

Table 1. Raw data

No	2018	2019	2020
1	0.00	0.00	0.00
2	73,133.70	306,172.73	315,129.09
3	16,555.44	71,677.27	78,831.82
4	8,870.45	36,390.91	42,290.91
5	42.67	263.64	331.82
6	1,846.59	11,977.27	14,000.00
7	125,775.01	505,372.73	532,600.00
8	405,538.64	1,622,468.18	1,631,013.64
9	122,341.97	512,095.45	520,518.18
10	859.05	3,559.09	3,690.91

No	2018	2019	2020
11	5,738.35	27,622.73	27,677.27
12	49,571.01	217,372.73	224,595.45
13	187,421.11	758,718.18	764,222.73
14	695.45	3,127.27	3,740.91
15	438.43	1,763.64	2,686.36
16	457.95	2,286.36	2,331.82
17	0.00	0.00	0.00
18	42,221.59	219,340.91	227,845.45
19	26,921.18	131,322.73	138,763.64
20	64,382.39	339,345.45	347,286.36
21	122,216.57	515,231.82	521,672.73
22	157,167.05	637,304.55	2.73
23	270,009.55	1,083,036.36	1,117,481.82
24	0.00	0.00	0.00
25	0.00	0.00	0.00
26	0.00	0.00	0.00
27	0.00	0.00	0.00
28	0.00	0.00	0.00
29	0.00	0.00	0.00
30	0.00	0.00	0.00
31	0.00	0.00	0.00
32	86.36	536.36	736.36
33	0.00	0.00	0.00

Before calculating the K-Means algorithm with the Euclidean Distance calculation, normalization is carried out first.

Table 2. Data Normalization Results

0.000000	0.000000	0.000000
0.180337	0.188708	0.193211
0.040823	0.044178	0.048333
0.021873	0.022429	0.025929
0.000105	0.000162	0.000203
0.004553	0.007382	0.008584
0.310143	0.311484	0.326545
1.000000	1.000000	1.000000
0.301678	0.315627	0.319138
0.002118	0.002194	0.002263
0.014150	0.017025	0.016969
0.122235	0.133977	0.137703
0.462154	0.467632	0.468557
0.001715	0.001927	0.002294
0.001081	0.001087	0.001647
0.001129	0.001409	0.001430
0.000000	0.000000	0.000000
0.104112	0.135190	0.139696
0.066384	0.080940	0.085078
0.158758	0.209154	0.212927
0.301368	0.317561	0.319846
0.387551	0.392799	0.000002
0.665805	0.667524	0.685146
0.000000	0.000000	0.000000
0.000000	0.000000	0.000000
0.000000	0.000000	0.000000
0.000000	0.000000	0.000000

0.000000	0.000000	0.000000
0.000000	0.000000	0.000000
0.000000	0.000000	0.000000
0.000000	0.000000	0.000000
0.000213	0.000331	0.000451
0.000000	0.000000	0.000000

After the data is normalized, it is continued by grouping it with the K-Means algorithm using the Euclidean Distance calculation.

1) Determination of Centroid value

Table 3. Iteration-1 Centroid Value

Centroid	2018	2019	2020
C1	0.021890	0.000000	0.000000
C2	0.462154	0.462154	0.462154
C3	1.000000	1.000000	1.000000

2) Calculating the distance from Centroid

Calculation of distance using Euclidian Distance iteration-1 with the formula:

Table 4. Iteration Centroid Distance-1

C1	C2	C3	Closest Distance	Results Cluster
0.021890	0.800473	1.732051	0.021890	C1
0.313124	0.475945	1.407463	0.313124	C1
0.068163	0.723512	1.655079	0.068163	C1
0.034284	0.759931	1.691505	0.034284	C1
0.021787	0.800201	1.731779	0.021787	C1
0.020706	0.788632	1.720207	0.020706	C1
0.535485	0.253373	1.184693	0.253373	C2
1.719505	0.931577	0.000000	0.000000	C3
0.528914	0.260146	1.191467	0.260146	C2
0.020022	0.796677	1.728255	0.020022	C1
0.025253	0.772681	1.704256	0.025253	C1
0.216751	0.573160	1.504667	0.216751	C1
0.795020	0.008427	0.924730	0.008427	C2
0.020397	0.797046	1.728624	0.020397	C1
0.020903	0.798271	1.729848	0.020903	C1
0.020858	0.798182	1.729760	0.020858	C1
0.021890	0.800473	1.732051	0.021890	C1
0.211072	0.582304	1.513484	0.211072	C1
0.125576	0.666441	1.597934	0.125576	C1
0.328353	0.467089	1.397358	0.328353	C1
0.530334	0.258864	1.190134	0.258864	C2
0.536656	0.473244	1.320524	0.473244	C2
1.153099	0.365207	0.566886	0.365207	C2
0.021890	0.800473	1.732051	0.021890	C1
0.021890	0.800473	1.732051	0.021890	C1
0.021890	0.800473	1.732051	0.021890	C1
0.021890	0.800473	1.732051	0.021890	C1
0.021890	0.800473	1.732051	0.021890	C1
0.021890	0.800473	1.732051	0.021890	C1
0.021890	0.800473	1.732051	0.021890	C1
0.021890	0.800473	1.732051	0.021890	C1
0.021685	0.799899	1.731476	0.021685	C1
0.021890	0.800473	1.732051	0.021890	C1

b) Calculation of K-Means Algorithm with Manhattan Distance Using Cluster 3

1) Determination of Centroid value

Table 5. Iteration-1 Centroid Value1

Centroid	2018	2019	2020
C1	0.021890	0.000000	0.000000
C2	0.462154	0.462154	0.462154
C3	1.000000	1.000000	1.000000

2) Calculating the distance from Centroid
Calculating the closest distance value using Manhattan

Table 6. 1st Iteration Centroid Distance

C1	C2	C3	Closest Distance	Results Cluster
0.021890	1.386461	3.000000	0.021890	C1
0.733576	0.824205	2.437744	0.733576	C1
0.159777	1.253126	2.866666	0.159777	C1
0.074305	1.316229	2.929768	0.074305	C1
0.022355	1.385989	2.999529	0.022355	C1
0.041886	1.365941	2.979481	0.041886	C1
1.252827	0.438288	2.051828	0.438288	C2
3.978110	1.613539	0.000000	0.000000	C3
1.233690	0.450018	2.063557	0.450018	C2
0.026492	1.379886	2.993425	0.026492	C1
0.058704	1.338316	2.951856	0.058704	C1
0.509727	0.992546	2.606085	0.509727	C1
1.845009	0.011882	1.601657	0.011882	C2
0.026690	1.380525	2.994064	0.026690	C1
0.025190	1.382645	2.996185	0.025190	C1
0.025030	1.382492	2.996032	0.025030	C1
0.021890	1.386461	3.000000	0.021890	C1
0.496803	1.007463	2.621002	0.496803	C1
0.295590	1.154059	2.767598	0.295590	C1
0.771874	0.805622	2.419162	0.771874	C1
1.236730	0.447686	2.061225	0.447686	C2
0.758464	0.606108	2.219648	0.606108	C2
2.681729	0.632014	0.981526	0.632014	C2
0.021890	1.386461	3.000000	0.021890	C1
0.021890	1.386461	3.000000	0.021890	C1
0.021890	1.386461	3.000000	0.021890	C1
0.021890	1.386461	3.000000	0.021890	C1
0.021890	1.386461	3.000000	0.021890	C1
0.021890	1.386461	3.000000	0.021890	C1
0.021890	1.386461	3.000000	0.021890	C1
0.021890	1.386461	3.000000	0.021890	C1
0.022911	1.385466	2.999005	0.022911	C1
0.021890	1.386461	3.000000	0.021890	C1

The K-Means algorithm stops when the cluster value is the same as the previous iteration.

c) Cluster Evaluation Based on DBI Value

To get the optimal clustering, the researcher evaluates the DBI value of the clustering using the K-Means method with the calculation of the Euclidean Distance and Manhattan Distance. The formula for finding the DBI value is as follows:

1. Finding SSW with the formula:

$$SSW_i = \frac{1}{m_i} \sum_{j=i}^{m_i} d(x_j, c_i) \quad (3)$$

Information :

m_i = The amount of data in the i-th cluster
X = Data in cluster
D(x,c) = Data distance to centroid
X_j = Data on the cluster
C_i = Centroid cluster i

2. Find SSB with the formula:

$$SSB_{ij} = d(c_i c_j) \quad (4)$$

Information :

c_i = Cluster 1
c_j = Other Cluster
$d(c_i c_j)$ = The distance between the centroid sat with each other

3. Find the ratio with the formula:

$$R_{ij} = \frac{SSw_i + SSW_j}{SSB_{ij}} \quad (5)$$

Information :

R_{ij} = Rasio Between Clusters
SSW_i = Cluster 1
SSW_j = Cluster 2
SSB_{ij} = Separation from cluster 1 and 2

4. Find DBI with formula :

$$DBI = \frac{1}{K} \sum_{i=1}^{k} max_{i \neq j} (R_{ij}) \quad (6)$$

Information :

K = Existing clusters
$R_{i,j}$ = Rasio between cluster i and j
Max = Find rasio between the biggest cluster

d) Grouping Results and DBI Values with Euclidean Distance

The following is the DBI value of K-Means Clustering with Euclidean Distance calculation with k=2:

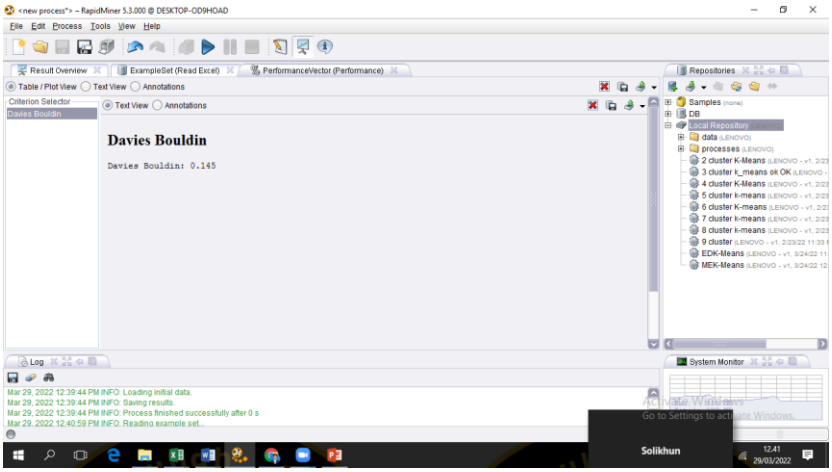

Figure 2. DBI value of K-Means Clustering with Euclidean Distance calculation with k=2

The results of grouping oil palm production using the K-Means method with Euclidean Distance with k=2:

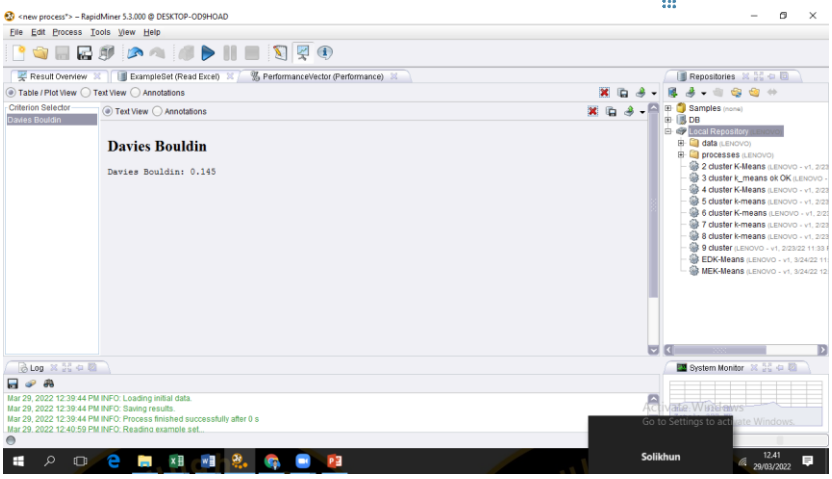

Figure 3. Grouping Results With Euclidean Distance k=2

The test is continued by finding the DBI value of K-Means Clustering by calculating the Euclidean Distance with k=3 and k=4. Then proceed with testing by finding the DBI value of K-Means Clustering by calculating the distance Euclidean Distance with k=2, k=3 and k=4.

e) Comparison of DBI Euclidean Distance Values with Manhattan Distance

From the test results using rapidminer, the DBI value from the calculation of the distance between Euclidean Distance and Manhattan Distance can be seen in the following table:

Table 7. DBI Value Comparison

No	Calculation Type	k	DBI Value
1	Euclidean Distance	k=2	0.145
2	Euclidean Distance	k=3	0.154
3	Euclidean Distance	k=4	0.176
4	Manhattan Distance	K=3	0.145
5	Manhattan Distance	k=2	0.154
6	Manhattan Distance	k=3	0.176

4. Conclusion

The result of this study is a comparison of the number of clusters based on the DBI value in the grouping of oil palm production in the province of North Sumatra. The group of oil palm production using K-Means Clustering with the calculation of Ecludiean Distance and Manhattan Distance with k=2, k=3 and k=4 show the same DBI value. The smallest DBI value is at k=2 0.145. The result of grouping oil palm production in Sumatra province with k=2 is the high group being Asahan, Langkat and North Labuhanbatu regencies, while 30 other regencies/cities are in the low group.

References

[1] H. Sulastri and A. I. Gufroni, "Penerapan Data Mining Dalam Pengelompokan Penderita Thalassaemia," *J. Nas. Teknol. dan Sist. Inf.*, vol. 3, no. 2, pp. 299–305, 2017.

[2] D. Marlina, N. Lina, A. Fernando, and A. Ramadhan, "Implementasi Algoritma K-

Medoids dan K-Means untuk Pengelompokkan Wilayah Sebaran Cacat pada Anak," *J. CoreIT J. Has. Penelit. Ilmu Komput. dan Teknol. Inf.*, vol. 4, no. 2, p. 64, 2018.

[3] F. Ramdhani and A. Hoyyi, "Pengelompokan Provinsi Di Indonesia Berdasarkan Karakteristik Kesejahteraan Rakyat Menggunakan Metode K-Means Cluster," *J. Gaussian*, vol. 4, no. 4, pp. 875–884, 2015.

[4] W. Gie and D. Jollyta, "Perbandingan Euclidean dan Manhattan Untuk Optimasi Cluster Menggunakan Davies Bouldin Index : Status Covid-19 Wilayah Riau," *Pros. Semin. Nas. Ris. Dan Inf. Sci. 2020*, vol. 2, no. April, pp. 187–191, 2020.

[5] C. A. Pamungkas, "Aplikasi Penghitung Jarak Koordinat Berdasarkan Latitude Dan Longitude Dengan Metode Euclidean Distance Dan Metode Haversine," *J. Inf. Politek. Indonusa Surakarta*, vol. 5, no. 2, pp. 8–13, 2019.

[6] M. Nishom, "Perbandingan Akurasi Euclidean Distance, Minkowski Distance, dan Manhattan Distance pada Algoritma K-Means Clustering berbasis Chi-Square," *J. Inform. J. Pengemb. IT*, vol. 4, no. 1, pp. 20–24, 2019.

[7] I. F. Ashari, R. Banjarnahor, and D. R. Farida, "Application of Data Mining with the K-Means Clustering Method and Davies Bouldin Index for Grouping IMDB Movies," vol. 6, no. 1, pp. 7–15, 2022.

[8] Yuliyanti, S. Al-Zulfa Nas, I. Jauhari Muhas, and E. Firmansyah, "Perbandingan metode pendekatan manhattan distance dengan euclidian distance pada implementasi pengenalan aksara jawa dengan menggunakan algoritma k-nearest neighbor."