Comparative Analysis of Inter-Centroid K-Means Performance using Euclidean Distance, Canberra Distance and Manhattan Distance

M Faisal1*, E M Zamzami2 and Sutarman3
1 Department of Information Technology, Universitas Sumatera Utara, Medan, Indonesia
2 Department of Computer Science, Universitas Sumatera Utara, Medan, Indonesia,
3 Department of Mathematics and Natural Sciences, Universitas Sumatera Utara, Medan, Indonesia

*E-mail: mfaisalusu@gmail.com

Abstract. Clustering is a method needed to group data or objects based on the required level between data, K-means is one of the clustering methods used that can be used easily in its implementation, there are some additions to this method according to the data center and on the weighting of the distance between data, the weighting of the distance between data on K-Means traditionally can be done using Euclidean Distance, Canberra Distance and Manhattan Distance, making this an analysis of the accuracy generated from the method produced by a combination of the Z-score and Min-Max Normalization methods, and is carried out Cluster homogeneity test using the Silhouette Coefficient method. The results of this method show that the Canberra method is superior to Euclidean and Manhattan on Iris dataset and the Canberra combination method with Z-score and Min-Max can increase the value on the glass without using the Normalization Method 37.44% to 67.46% use the Z-score and 56.52% use Min-Max and use an increase in the average value of the Silhouette Coefficient.

1. Introduction
Clustering is an activity (task) aimed at grouping data that has similarities between one data with other data into clusters or groups so that the data in one cluster has a maximum level of similarity and the data between clusters has a minimum similarity. Clustering can also be interpreted as a method of data segmentation that is implemented in several fields, including marketing, analyzing business issues and market segmentation predictions, patterns in the field of computer vision, zoning to object identification and image processing. Cluster analysis aims to find groups of objects in such a way that the objects in the group will be the same (or related) to each other and different from (or not related) objects in other groups [1]. The application of the conventional K-Means algorithm has been carried out using the Euclidean Distance measurement method. This study will analyze the distance measurement using the Euclidean, Canberra and Manhattan methods on the K-Means algorithm with a combination of Z-score and Min-Max normalization methods to see the accuracy of each method to determine the best method.
2. Related Research

There are some studies that have been done beforehand that the application of clustering using the k-means algorithm is relatively faster than clustering using other algorithms, and also provides quality clusters when using large datasets [2]. Comparison of Manhattan and Euclidean distance calculation systems in the k-means algorithm to find out the number of squared errors using the Bank dataset and tested using the WEKA tools application. The test results show that the Manhattan distance work system is better than the Euclidean work system [3]. Comparison of 3 (three) ways of working distance calculation on the k-means algorithm (Manhattan, Euclidean, and Minkowski) to find the best distance calculation method, the results of the study concluded that the workings of the Euclidean distance calculation work better than the Manhattan and Minkowski systems [4]. Other studies concerning the comparison of the workings of Manhattan, Euclidean and Chebyshev Distance calculations on the k-means algorithm to recognize the mean absolute error. The results of tests conducted using the flower dataset show that the Chebyshev Distance calculation works better than Manhattan and Euclidean [5]. In different studies, it is known that the workings of the distance calculation of Manhattan, Euclidean and Chebyshev are superior to each other depending on the data used [6], [7], [8]. In subsequent studies estimating the quality of inter-centroid distances in data clustering problems using Hybridized K-Means for classification of medical dataset data with the aim of classifying patients based on medical history. The results showed the distance function of Canberra performed better compared to Euclidean [9].

3. Research Method

Distance assessment has a very important role in ensuring the similarity or regularity between data and items. this is done to understand, with what systems the data are said to be interconnected, similar, not similar, and what distance assessments are needed to compare them [10]. In the clustering process, the level of determining or describing a quantitative score of the degree of similarity or dissimilarity of the data (proximity measure) has a very important role, so it is necessary to make a comparative analysis of some of the methods that are often used, namely the Euclidean, Manhattan, and canadian distance with a combination of methods Z-score and Min-Max Normalization to get the best method.

3.1. Euclidean Distance

Euclidean distance is one of the ways in calculating distance used to calculate the distance of 2 (two) pieces of data in Euclidean space (including fields Euclidean two dimensions, three dimensions, or more). To evaluate the degree of similarity of data with Euclidean Distance formula can be applied with the following formula [11]:

\[ d_{ij} = \sqrt{\sum_{k=1}^{n} (x_{ik} - y_{jk})^2} \]  

Where \( d \) is the distance between \( i \) and \( j \), \( i \) as the cluster data center \( j \) data on the attribute, \( k \) symbol of each data, \( n \) the amount of data, \( x_{ik} \) is the data in the cluster center to \( k \) and \( y_{jk} \) is the data on each data to \( k \).

3.2. Canberra Distance

This distance function was introduced and developed in 1966 by G.N. Lance and modified by W.T Williams in 1967. Canberra Distance is used when we need to find the distance between pairs of points, where data is around the origin in vector space [12] [13]. To evaluate the level of similarity of data with the Canberra distance formula can be applied with the following formula:
\[ d_{ij} = \sum_{k=1}^{n} \frac{|x_{ik} - x_{jk}|}{|x_{ik}| - |x_{jk}|} \]  

Where \( d \) is the distance between \( i \) and \( j \), \( i \) as the cluster data center, \( j \) data on the attribute, \( k \) symbol of each data, \( n \) the amount of data, \( x_{ik} \) is the data at the cluster center to \( k \), and \( y_{jk} \) is the data on each data to \( k \).

3.3. Manhattan Distance

Manhattan distance is used to calculate the absolute difference between the coordinates of a pair of objects. The formula used is as follows:

\[ d_{ij} = \sum_{k=1}^{n} |x_{ij} - y_{ik}| \]

Where \( d \) is the distance between \( i \) and \( j \), \( i \) as the cluster data center, \( j \) data on the attribute, \( k \) symbol of each data, \( n \) the amount of data, \( x_{ik} \) is the data at the cluster center to \( k \), and \( y_{jk} \) is the data on each data to \( k \).

3.4. Cluster Analysis Techniques

The process of grouping (clustering) data is done through the general stages of the K-Means Clustering algorithm. Before the clustering process begins, data normalization is done using two different methods, namely Min-Max and Z-score normalization.

- **Min-Max Normalization**

\[ x^1 = \frac{x - \min (a)}{\max (a) - \min (a)} \]

Where \( x \) is data per column, \( \max \) is the maximum value of data per column and the minimum value of data per column

- **Z-Score Normalization**

\[ v^1 = \frac{v - \bar{A}}{\sigma A} \]

Where, \( V1, v \) is the new and old entries of each data. \( \sigma A, A \) is the standard deviation and the average of \( A \).

The first stage, determine the number of \( K \) clusters manually, i.e. in this study the value of \( k \) given is following the category in the dataset; in the iris dataset the value \( k = 3 \), in the wine dataset the value \( k = 3 \) and the glass dataset \( k = 6 \). Then initialize the centroid value 6 times randomly simulated. Second stage, calculate the distance of the object with a centroid using several methods. This research will use the Euclidean Distance, Canberra Distance, and Manhattan Distance methods. After the distance is calculated, then the cluster homogeneity test is performed using the Silhouette Coefficient method. Cluster homogeneity testing can be determined based on the value of the Silhouette Coefficient which can be obtained through the following stages:

- **Calculate the average distance from an object to all other objects contained in one cluster with the equation:**

\[ \alpha_i = \frac{1}{|A| - 1} \sum_{j \in A, j \neq i} d(i, j) \]

Where, \( |A| \) is the amount of data in clusters \( A \), \( i \) and \( j \) as indexes of document, \( d(i, j) \) between the distance of document \( i \) to document \( j \).
• Calculate the average distance from document \( i \) with all documents in other clusters, and take the smallest score using the following formula:

\[
d(i, C) = \frac{1}{|A|} \sum_{j \in C} d(i, j)
\]

(7)

Where, \( d(i, C) \) is the average distance of object \( i \) to all objects in other clusters \( C \) where \( A \neq C \).

• Calculate the value of the Silhouette Coefficient with the following equation:

\[
s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}
\]

(8)

Third, object grouping and convergence testing are conducted between new data groups and data groups in the previous process, if they are the same (convergent), then the clustering process is complete. If not, then iterations must start from the determination of the new cluster center.

3.5. Dataset

The dataset is used to see the performance of Euclidean, Canberra and Manhattan, so this study will use 3 datasets obtained from the UCI Repository, including iris (150), wine (178) and glass (214).

**Tabel 1. Dataset.**

| Name of Dataset | Amount of data | Attribute | Classes |
|-----------------|---------------|-----------|---------|
| Iris            | 150           | 4         | 3       |
| Wine            | 178           | 13        | 3       |
| Glass           | 214           | 9         | 6       |

4. Results & Discussion

This section shows the results of simulations performed on the three datasets 6 times using the Euclidean, Canberra and Manhattan distance methods, as well as a combination using the Z-score and Min-Max methods.

**Tabel 2. Maximum number of iterations in each data.**

| Dataset | Without Normalization | Z-score | Min-Max |
|---------|-----------------------|---------|---------|
|         | Euclidean             | Canberra| Manhattan|
| Iris    | 10                    | 6       | 10      |
| Wine    | 7                     | 12      | 8       |
| Glass   | 12                    | 11      | 9       |

**Tabel 3. Level of Accuracy of Methods without using the Normalization Method.**

| No | Distance Measure | Silhouette Coefficient | Accuracy |
|----|------------------|------------------------|----------|
|    | Iris             | Wine                   | Glass    | Iris     | Wine     | Glass    |
| 1  | Euclidean        | 0.3266                 | 0.3365   | 0.0446   | 86.66%   | 84.75%   | 45.37%   |
| 2  | Canberra         | 0.3252                 | 0.3093   | 0.0120   | 98.68%   | 83.44%   | 37.44%   |
| 3  | Manhattan        | 0.3223                 | 0.2919   | 0.0535   | 84.46%   | 78.85%   | 51.08%   |
### Table 4. Level of Accuracy Using the Z-score Normalization Method.

| No | Distance Measure | Silhouette Coefficient | Accuracy |
|----|------------------|------------------------|----------|
|    | Iris             | Wine                   | Glass    | Iris       | Wine       | Glass    |
| 1  | Euclidean        | 0.3222                 | 0.3311   | 0.0408     | 86.66%     | 84.75%   | 45.91%   |
| 2  | Canberra         | 0.3362                 | 0.3375   | 0.0305     | 90.07%     | 73.18%   | 67.46%   |
| 3  | Manhattan        | 0.3235                 | 0.3095   | 0.0570     | 84.46%     | 78.85%   | 45.60%   |

### Table 5. Level of Accuracy Using the Min-Max Normalization Method.

| No | Distance Measure | Silhouette Coefficient | Accuracy |
|----|------------------|------------------------|----------|
|    | Iris             | Wine                   | Glass    | Iris       | Wine       | Glass    |
| 1  | Euclidean        | 0.3226                 | 0.3315   | 0.0397     | 86.66%     | 84.75%   | 45.37%   |
| 2  | Canberra         | 0.3314                 | 0.3470   | 0.0235     | 98.68%     | 81.69%   | 56.52%   |
| 3  | Manhattan        | 0.3284                 | 0.2919   | 0.0516     | 84.46%     | 78.85%   | 51.08%   |

### 5. Conclusion

The simulation results of the three datasets in Table 1 show that the Canberra method is superior with an accuracy of 98.68% in the iris dataset, Euclidean is superior in the wine dataset with 84.75% and Manhattan is superior in the glass dataset with 51.08%. Table 2 shows that the Canberra method also excels with an accuracy of 90.07% in the iris dataset and 67.46% in the glass dataset, Euclidean in the wine dataset with 87.75%. Table 3 shows that the Canberra method is also superior with 98.68% in the iris dataset and 56, 52% in the glass dataset, while Euclidean excels with 84.75% in the wine dataset. The accuracy produced by the Euclidean and Manhattan distance methods without using the normalization method using the Z-score and Min-Max normalization methods in the iris and wine dataset shows the clustering process produces the same results, the Euclidean method using the iris dataset produces using 86.66% and the wine dataset produces an accuracy of 84.75%, but the glass dataset shows the same results only seen in the clustering process without using the normalization method and using the Min-Max normalization method with a test value of 51.08%, seen in Table 3 in the z- method score there is a slight change in the value of accuracy. In the process of grouping using the distance Canberra method without using the normalization method using the Z-score and Min-Max normalization methods indicate changes in the value of each method used, indicating that the distance Canberra method is very sensitive to the weighting of values in each cluster, a good convincing value on the iris dataset with a value of 98.68%, on the wine set 83.44% and on the glass dataset 67.46%. From these results it can be concluded that the Canberra method is superior to Euclidean and Manhattan on the iris dataset, the Canberra combination method with z-score and Min-Max can increase the accuracy value in the glass dataset 30.02% (Z-score) and 19.08% (Min-Max).

### References

[1] P.-N. Tan, M. Steinbach, A. Karpatne, and V. Kumar, Introduction to Data Mining (2nd Edition), 2nd ed. New York: Pearson, 2018.

[2] S. Saraswathi and M. I. Sheela, “A Comparative Study of Various Clustering Algorithms in Data Mining,” vol. 3, no. 11, pp. 422–428, 2014.

[3] R. Awasthi, A. K. Tiwari, and S. Pathak, “Empirical Evaluation On K Means Clustering With Effect Of Distance Functions For Bank Dataset,” Int. J. Innov. Technol. Res., vol. 1, no. 3, pp. 233–235, 2013.

[4] A. Singh, A. Rana, and A. Yadav, “K-means with Three different Distance Metrics,” Int. J. Comput. Appl., vol. 67, no. 10, pp. 13–17, 2013.
[5] K. Kouser, “A comparative study of K Means Algorithm by Different Distance Measures,” Int. J. Innov. Res. Comput., vol. 1, no. 9, pp. 2443–2447, 2013.

[6] D. Sinwar and R. Kaushik, “Study of Euclidean and Manhattan Distance Metrics using Simple K-Means Clustering,” Int. J. Res. Appl. Sci. Eng. Technol., vol. 2, no. 5, pp. 270–274, 2014.

[7] D. J. Bora and A. K. Gupta, “Effect of Different Distance Measures on the Performance of K-Means Algorithm: An Experimental Study in Matlab,” Eff. Differ. Distance Meas. Perform. K-Means Algorithm An Exp. Study Matlab, vol. 5, no. 2, pp. 2501–2506, 2014.

[8] H. Prasetyo and A. Purwariati, “Comparison of Distance Measures for Clustering Data with Mix Attribute Types,” in International Conference on Information Technology Systems and Innovation, 2014.

[9] Ritu Vashistha and Jitender Kumar “Estimation of Inter-Centroid Distance Quality in Data Clustering problem using Hybridized K-Means Algorithm ” in Second International Conference on Electrical, Computer and Communication Technologies (ICECCT), 2017.

[10] A. Singh, J. Agarwal, and A. Rana, “Performance Measure of Similis and FP-Growth Algorithm,” Int. J. Comput. Appl., vol. 62, no. 6, pp. 25–31, 2013.

[11] H. Anton, Elementary Linear Algebra, 7th ed. New Jersey: Wiley, 1993.

[12] Nagar, S. and Khunteta, A., “A Proposed Modification Over Learning Vector Quantization and K-Means Algorithms for Performance Enhancement,” in Proceedings of the International Conference on.

[13] Al-Daoud, M. and S. Roberts (1994). New Methods for the Initialisation of Clusters. Leeds, University of Leeds: 14.
