Analysis of Euclidean Distance and Manhattan Distance in the K-Means Algorithm for Variations Number of Centroid K

Analysis of Euclidean Distance and Manhattan Distance in the K-Means Algorithm for Variations Number of Centroid K

R Suwanda[1]*, Z Syahputra[1]* and E M Zamzami[1]*

[1] Department of Information Technology, Faculty of Computer Science and Information Technology, Universitas Sumatera Utara, Medan 20155, Indonesia

*Email: rizkiswd@gmail.com, zulfahmisyahputra@gmail.com, elvi_zamzami@usu.ac.id

Abstract. K-Means is a clustering algorithm based on a partition where the data only entered into one K cluster, the algorithm determines the number group in the beginning and defines the K centroid. The initial determination of the cluster center is very influential on the results of the clustering process in determining the quality of grouping. Better clustering results are often obtained after several attempts. The manhattan distance matrix method has better performance than the euclidean distance method. The author making the result of conducted testing with variations in the number of centroids (K) with a value of 2,3,4,5,6,7,8,9 and the authors having conclusions where the number of centroids 3 and 4 have a better iteration of values than the number of centroids that increasingly high and low based on the iris dataset.

Keywords: Euclidean Distance, Manhattan Distance, K-Means.

1. Introduction

Classification is a technique used to build classification models from training data samples. The classification will analyze the input data and build a model that will describe the class of the data. Class labels from unknown data samples can be predicted using classification techniques [1]. One of the most popular classification techniques is K-Means.

Increasing the cluster center and minimizing 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 where data is only entered into one K cluster, the algorithm determines the number of grub in the beginning and defines the K centroid set [2].

Several studies have tried to develop the K-Means algorithm based on the parameter k value to improve the performance of the K-Means algorithm. Okfalisa et al (2017) also made a comparison analysis of K-Means and Modified K-Means on the classification. data. This study produced the highest accuracy value of K-Means of 93.94% and MKNN accuracy of 99.51%. This shows that K-Means has better accuracy compared to the K-Means algorithm. In addition to the K value, the distance matrix is an important factor that depends on the KNN algorithm data set. The resulting distance matrix value will affect the performance of the algorithm. The distance between two data points is determined by the calculation of the distance matrix where Euclidean Distance is the most widely used distance matrix function. There are several types of distance matrix functions besides Euclidean Distance, namely Manhattan Distance, Miskowski Distance, Canberra Distance, Braycurtis Distance, Chi-Square and others [3].

Based on the background above, the author tries to analyze and compare the accuracy of the performance of the K-Means using the Euclidean Distance and Manhattan Distance distance functions in the classification process based on the accuracy point of view. The data that will be used in this study is the Iris dataset. The source of the dataset is obtained from the UC Irvine Machine Learning Repository (UCI Machine Learning Repository).

2. Related Research

Saputra (2018) conducted research in determining the k value in the K-Means classification using the Gini Index method and Local Mean Based. This study produces the conclusion that the Gini Index method and Local Mean Based can be maximized in determining the value of k. [4].

Rulaningtyas (2015) conducted research in K-means clustering method for color image segmentation in pulmonary tuberculosis identification. This study did segmentation to separate the tuberculosis bacteria images from the background images [8].

Alamri (2016) conducted a classification study of satellite images using distance matrices by comparing Bray Curtis Distance, Manhattan distance, and Euclidean distance [7]. By producing the best classification accuracy by Bray Curtis Distance by 85% and followed by Manhattan distance (City Block Distance) and Euclidean distance by 71% [5].

Mulak and Talhar (2013) also conducted a distance analysis study using KNN on the KDD dataset. This study compares Euclidean Distance, Chebychev Distance, and Manhattan Distance using the KNN algorithm [1].

3. Proposed Method

3.1. Classification Data

Data classification must use an approach to look for similarities in data so as to be able to place data into the right groups. Grouping data will divide the data set into several groups where the similarity in a group is greater when compared to other groups [6].

3.2. Measures Distance

Distance calculation is widely used in determining the degree of similarity or not two vectors. So this method is widely used for pattern recognition [7]. Some distance methods are Euclidean Distance, Chebyshev, Angular Separation, Canberra Distance, Hamming Distance, Sorrensen Distance and so on. In the K-Nearest Neighbor algorithm, the classification process uses the Euclidean Distance method.

3.3. Euclidean Distance

Euclidean Distance is the distance between points in a straight line. This distance method uses the Pythagorean theorem. And is the distance calculation that is most often used in the process of machine learning [7]. Euclidean Distance formula is the result of the square root of the differences of two vectors.

$$d_{ij} = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2} \quad (1)$$

Information:
dij = similarity calculation distance
n = number of vectors
xik = input image vector
xjk = comparison image vector
From equation 1 the pattern of Euclidean Distance is the circle shown in figure 2.

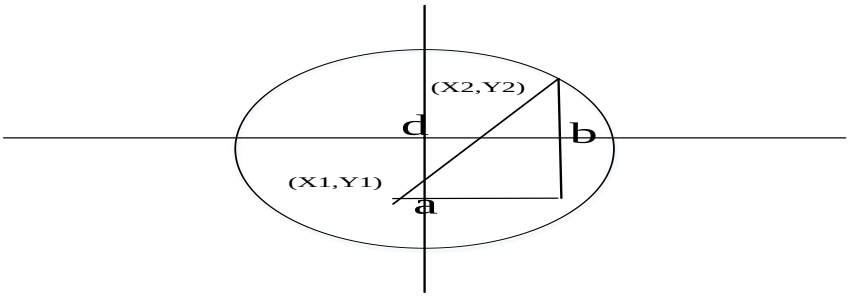

Figure 1. Euclidean distance pattern.

Information:
$a = x_2 - x_1$
$b = y_2 - y_1$
Formula Pythagoras
$a^2 + b^2 = d^2$
$d^2 = (x_2 - x_1)^2 + (y_2 - y_1)^2$

3.4. Manhattan Distance

Manhattan distance is also referred to as "city block distance" which is the sum of the distances from all the attributes. For the two data points x and y in d-space dimensions, the Manhattan distance between the points are defined as follows:

$$d_{man}(\mathbf{x}, \mathbf{y}) = \sum_{k=1}^{d} |x_j - y_j|. \quad (2)$$

4. Results and Analysis

This study analyzes the matrix of the Manhattan Distance and Euclidean Distance methods on the K-Means algorithm in grouping data. To be clear in describing the process in this study it will be explained step by step in this section. The stages, in general, can be seen in figure 2.

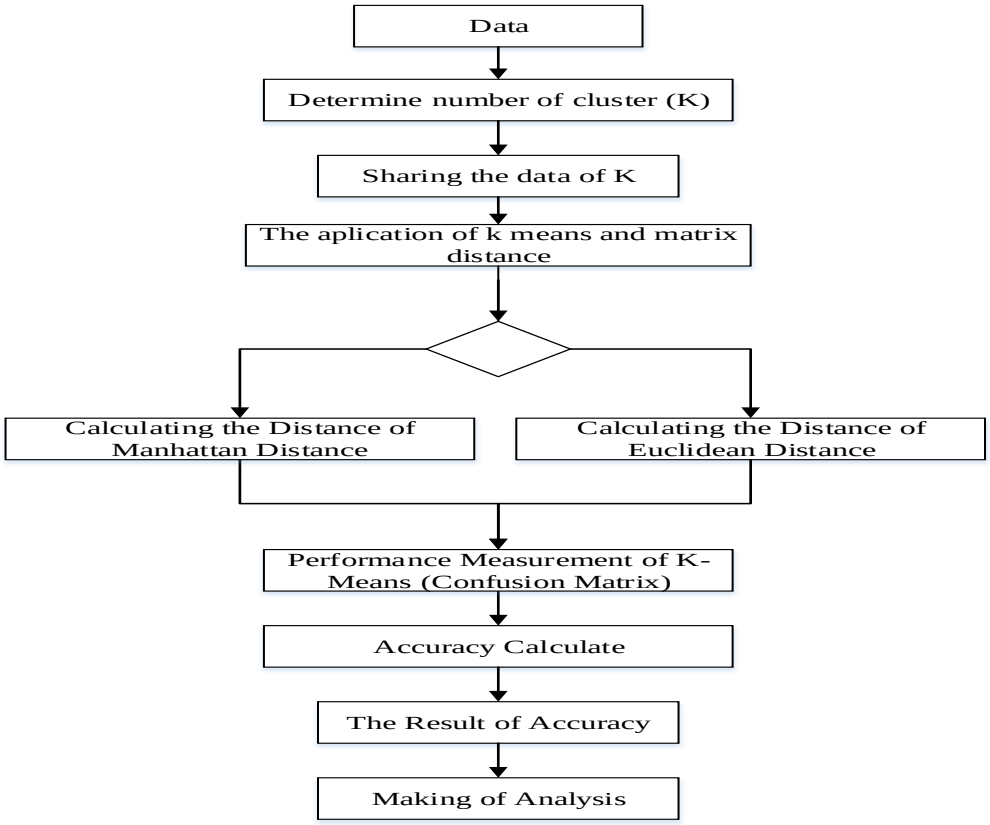

Figure 2. The research flowchart.

In this study, the Iris dataset from the UC Irvine Machine Learning Repository (UCI Machine Learning Repository) was used. This data set has 4 attributes that will be used in the classification process using K-Means. Four features of 4 features were measured from each sample length and width of sepals and flower petals in centimeters. Information on Iris data set attribute values can be seen in Table 1.

Table 1. Dataset iris.

No.	Name Item	X1	X2	X3	X4
1	IrisSetosa	5.1	3.5	1.4	0.2
2	IrisSetosa	4.9	3	1.4	0.2
3	IrisSetosa	4.7	3.2	1.3	0.2
4	IrisVersiColor	7	3.2	4.7	1.4
5	IrisVersiColor	6.4	3.2	4.5	1.5
6	IrisVersiColor	6.9	3.1	4.9	1.5
7	IrisVirginica	6.3	3.3	6	2.5
8	IrisVirginica	5.8	2.7	5.1	1.9
:	:	:	:	:	:
100	IrisVirginica	7.1	3	5.9	2.1

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

The process of clustering data classification needed the cluster center point according to the number of clusters desired from the data. The authors making testing of the centroid value (k) with a combination of the k-means algorithm and Sum of Squared Error. The results of the test are shown in the following table.

Table 2. Iris data testing results.

Value Centroid (K)	Amount of Iteration Euclidean Distance K-Means	Amount of Iteration Manhattan Distance K-Means
2	10	7
3	6	4
4	8	4
5	9	5
6	11	7
7	10	11
8	15	11
9	13	10

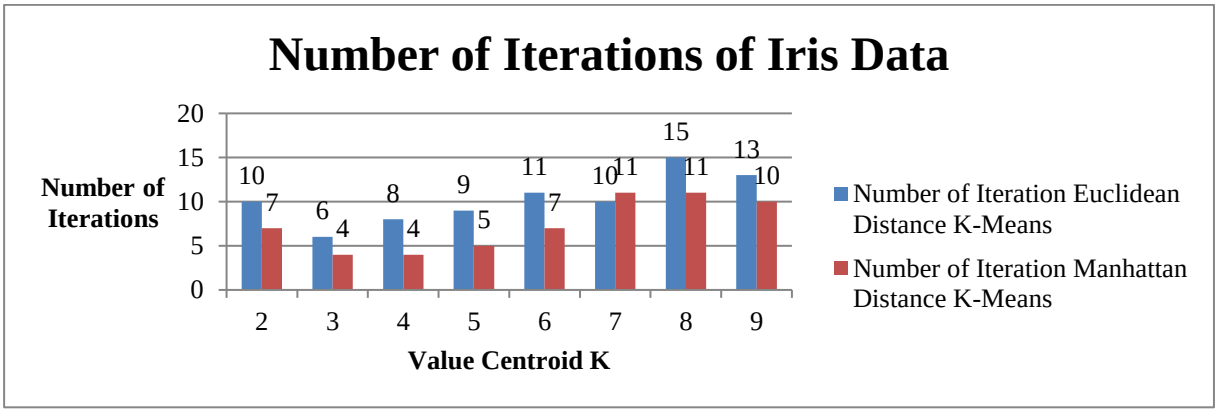

Figure 3. Test graph variation in Centroid Value (K) iris dataset.

From figure 3 it can be seen that the Manhattan K-Means method has better performance than Euclidean Distance K-Means. The results of the authors conducted testing with variations in the number of centroids (K) with a value of 2,3,4,5,6,7,8,9 the authors draw conclusions the number of centroids 3 and 4 have an iteration of values that is better than the number of centroids that are increasingly high and low based on the iris dataset.

5. Conclusions

The manhattan distance with the matrix method has better performance than the euclidean distance method. The results are centroids 3 and 4 have a better iteration of values than the variation number of centroids (K) with the value of 2,3,4,5,6,7,8,9. That is increasingly high and low based on the iris dataset.

References

[1] Mulak, P. & Talhar, N. 2015. Analysis of Distance Measures Using K-Nearest Neighbor Algorithm on KDD Dataset. International Journal of Science and Research (IJSR) 4(7): 2101-2104.

[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] Okfalisa., Mustakim., Gazalba, I., & Reza, N.G.I. 2017. Comparative Analysis of KNearest Neighbor and Modified K-Nearest Neighbor Algorithm for Data Classification. International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE), pp: 294-298.

[4] Saputra, M.E. 2018. GINI Index dengan Local Mean Based untuk Penentuan Nilai K dalam Klasifikasi K-Nearest Neighbor. Tesis. Universitas Sumatera Utara.

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

[6] Viriyavisuthisakul, S., Sanguansat, P., Charnkeitkong, P., & Haruechaiyasak, C. 2015. A comparison of similarity measures for online social media Thai text classification. 2015 12th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), pp. 1-6.

[7] Alamri, S. S. A., Bin-Sama, A. S. A., & Bin-Habtoor, A. S. Y. (2016). Satellite Image Classification by Using Distance Metric. International Journal of Computer Science and Information Security 14(3): 65.

[8] Rulaningtyas, R., Suksmono, A.B., Mengko, T. and Saptawati, P., 2015. Multi patch approach in K-means clustering method for color image segmentation in pulmonary tuberculosis identification. *4th International Conference on Instrumentation, Communications, Information Technology, and Biomedical Engineering (ICICI-BME)* (pp. 75-78).