http://infor.seaninstitute.org/index.php/infokum/index
JURNAL INFOKUM, Volume 9, No.2, Juni 2021 ISSN : 2302-9706

COMBINATION OF EUCLIDEAN DISTANCE ON X-MEANS ALGORITHM IN DATA GROUPING

Berti Sari Br Sembiring[1], Mahdianta Pandia[2], Miska Irani[3]
bertisari0@gmail.com

Program Studi Sistem Informasi, STMIK Kristen Neumann Indonesia, Jl. Jamin Ginting KM 10,5 Medan

Article Info
Received 01 June 2021
Revised 20 June 2021
Accepted 30 June 2021

Grouping can use clustering to group data based on the similarity between the data, so that the data with the closest resemblance is in one cluster while the different data is in another group. The X-Means algorithm is the development of K-Means. The weakness of X-Means is that in determining the distance matrix, the distance matrix is an important factor that depends on the X-Means algorithm data set. The resulting distance matrix value will affect the performance of the algorithm. The results of the study are: testing with variations in the number of centroids (K) with values of 2,3,4,5,6,7,8,9,10. The author concludes that the number of centroids 3 and 4 has a better iteration value compared to the number of centroids that are getting higher and lower based on the iris dataset with the jarax matrix Manhattan Distance. From the test results with the X-Means cluster point, calculate the Euclidean Distance distance with 100 iris data reaching the 9th iteration, while with 100 iris data by calculating the Manhattan Distance distance it reaches the 10th iteration. Meanwhile, in determining the cluster point using the X-Means method from 100 data iris reaches its 7th iteration.

Keywords: X-Means, Euclidean Distance

1. INTRODUCTION

Grouping can use clustering to group data based on the similarity between the data, so that the data with the closest resemblance is in one cluster while the different data is in another group. The process of grouping data into several clusters or grouping so that the data in one cluster has a maximum similarity level and between clusters has a minimum similarity is called Clustering. Clustering is divided into 2 approaches in its development, namely partitioning and hierarchical clustering. [1] X-means clustering is used to solve one of the main drawbacks of K-means clustering, namely the need for prior knowledge of the number of clusters (K). In this method, the true value of K is estimated in an undiscovered manner and is only based on the data set itself [2]

Several studies related to the distance matrix function have also been carried out by comparing the Euclidean distance with Manhattan distance, Canberra distance and Hybrid distance on the LVQ algorithm.Nakyoung Kim, Hyojin Park, Jun Kyun Choi (2017) Research modifies the method derived from the combination of Mean-Shift and X-Means. The results of the research can be time and calculation efficiency and can separately group images with the same features. [3]Latifa Greche, Maha Jazouli, et al. (2017) The results of the study by comparing the results of the classification of six facial expressions. The classification of facial features calculated using Manhattan and Euclidean methods has been realized using a neural network classifier to recognize six emotions. Both of these methods achieve the same average recognition rate of 100%, except that each reaches this level at a different stage of neural network training. [4]

Alfatih Muhammad, Ary Setijadi Prihatmanto, et al (2018) The Manhattan distance method is more appropriate for measuring syllable and phonetic distances, even if we look at the average Manhattan and Euclidean distance measurements. [5]
the distance values are almost the same. However, when the distance from the syllable and the phonetic length is far, Euclidean measurements take the midpoint of the accumulation of all parameters.

2. RESEARCH METHOD

The X-Means algorithm was developed by Dan Pelleg and Andre Moore in 2000. In this algorithm the number of clusters is calculated dynamically using the upper and lower limits provided by the user. This algorithm consists of two steps which are repeated until completion.
1. Increase-Params, in this step apply the k-means algorithm initially for k clusters until convergence. Where k is equal to the lower limit provided by the user.
2. Fix Structure, this structural repair step begins by breaking each cluster center into two children in opposite directions along a randomly selected vector. After that run k-means locally within each cluster for two clusters. The decision of each cluster center itself by comparing the BIC values.
3. If K > = kmax (upper limit) stop and report to the best scoring model found during the dance, otherwise go to step 1.

X-Means means taking advantage of Bayesian Cri Information ionized (BIC) to control the cluster separation process. In other words, if we split one cluster into two clusters ters increase the BIC score, then have two groups more likely than a single cluster. In this paper, we recommend using the Minimum Noisy Description Length(MNDL. as a cluster separation criterion, leading to for more precise predictions for the number of clusters.

X-means clustering is used to solve one of the main weaknesses of K-means clustering, namely the need for prior knowledge about the number of clusters (K). In this method, the true value of K is estimated in an unsupervised manner and based solely on the data set itself [3].

Figure 1. General Steps in X-Means Grouping

Kmax and Kmin as upper and lower bounds for the possible values of X. In the first step X-means grouping, knowing that at this time X = Xmin, X-means find the initial structure and centroid. In the next step, each cluster in the estimated structure is treated as a parent cluster, which can be divided into two groups.

This algorithm can be too slow because it needs to rerun Kmeans for each cluster split. To solve this problem, implementing kd-tree from the data set suggested in, which naturally reduce the number of nearest neighbor requests for K-means.

Distance The closest distance calculation method / similarity distance Euclidean Distance is the distance calculation method that is most often used to calculate the similarity of two vectors. Euclidean Distance is the most commonly used metric to calculate the similarity of two vectors. The Euclidean Distance formula is the root of the square of differences between 2 vectors (root of square differences between 2 vectors).

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 machine learning process (Viriyavisuthisakul et al, 2015). The Euclidean Distance formula is the result of the square root of the difference between two vectors.

$$d_{ij} = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2} \quad (2.1)$$

Information :
dij = distance similarity calculation
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.

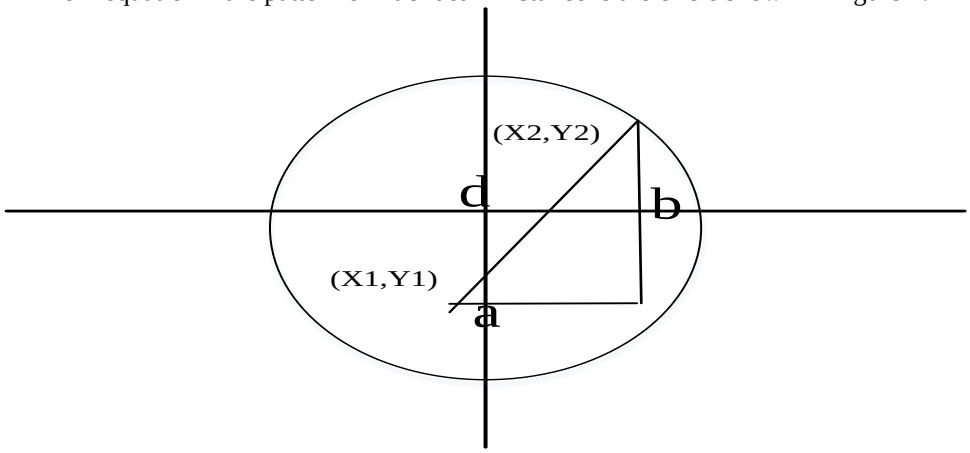


Figure 2. Euclidean Distance Patten

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

3. RESULTS AND ANALYSIS

To carry out the data classification clustering process, of course, a cluster center point is needed according to the number of clusters desired from the data. In this test, the authors conducted a test with iris data, as many as 100 data and 4 attributes with the center point of centroid 2 which were selected using the X-Means method. The following is a data classification process with a random centroid center point using the X-Means method.

Table 1. Inisialisasi Pusat Cluster X-Means Data Iris

Pusat Cluster	Nama Item	X1	X2	X3	X4
1	IrisSetosa	4.7	3.2	1.3	0.2
2	IrisVirginica	5.8	2.7	5.1	1.9

The next process is to calculate the distance of each iris data for each cluster. The following is shown in table 2.

Table 2 Data Distance to Center of Iris Data Cluster Using Euclidean Distance Method

No	Nama Item	Iterasi Ke -1 (C1)	Iterasi Ke-1 (C2)
1	IrisSetosa	0.538516481	4.208325083
2	IrisSetosa	0.3	4.33474336
3	IrisVersiColor	4.09633983	1.449137675
4	IrisVersiColor	3.686461718	1.063014581
5	IrisVersiColor	4.236744033	1.252996409
6	IrisVirginica	5.338539126	1.334166406
7	IrisSetosa	0.3	4.33474336
:	:	:	:
:	:	:	:
100	IrisVirginica	5.357238094	1.568438714

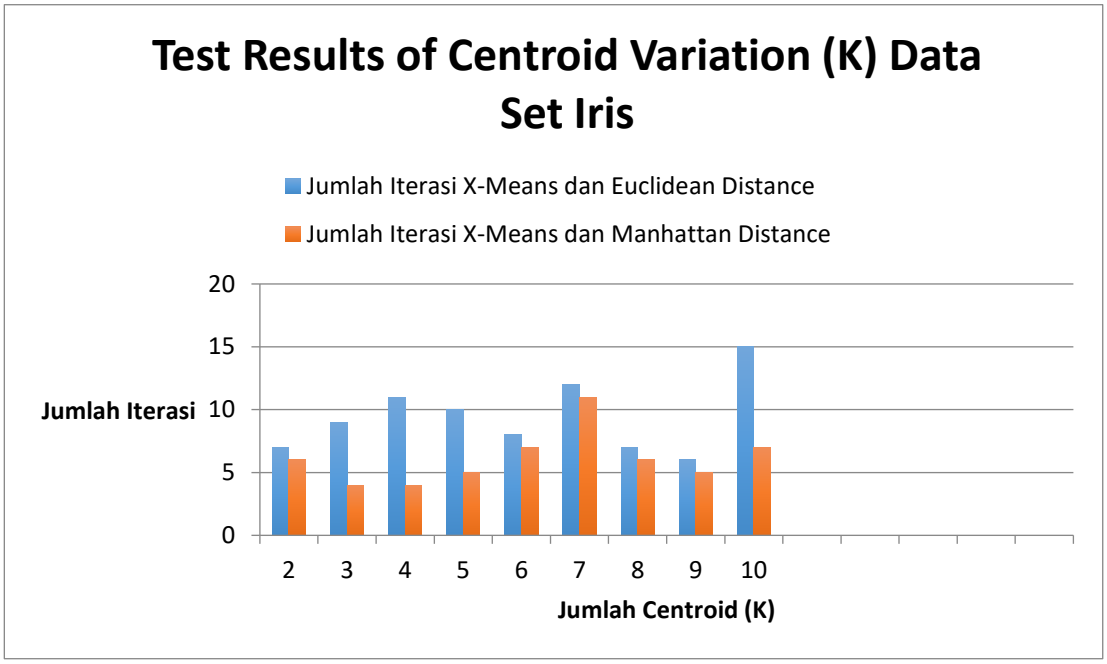
After the process of calculating the distance with the Euclidean distance to the iris data and, then the results of the final cluster center value are shown in the following table.

Table 3 Endpoint Cluster Center Point Value X-Means Euclidean Distance Data Iris

Pusat Cluster	Nama Item	X1	X2	X3	X4	Iterasi Ke-	Jumlah Data Centroid 1	Jumlah Data Centroid 2
1	IrisVersiColor	6.26	2.85	4.87	1.63	9	55	45
2	IrisSetosa	5.03	3.45	1.47	0.25			

From the results of the tests carried out on the SetIris Data as many as 100 test data with varying numbers of centroids, the accuracy of the X-Means method with Manhattan Distance is better than the X-Means method with Euclidean Distance. This is based on testing the classification of the Iris Data Set with variations in the number of centroids (K) with values of 2,3,4,5,6,7,8,9,10.Based on the Accuracy assessment of the iris data set, it was found that the Manhattan Distance matrix was better than the Euclidean Distance distance matrix, namely at the values of k=6, k=7, and k=8. The best Accuracy value of Braycurtis Distance is 96%. The best Accuracy Euclidean Distance value is 95.33%.

The following graph of the results of testing the centroid variation (K) is shown as follows.



Figur 3 Output Graph Testing the Variation of K Value Dataset Iris

In connection with the results of the description above, it can be explained that the Euclidean distance is one of the distance calculation methods used to measure the distance from 2 (two) points in Euclidean space (covering a two-dimensional, three-dimensional, or even more Euclidean plane). The test results show that the variation in the number of centroids (K) with values of 2,3,4,5,6,7,8,9,10 in the iris data set has a longer iteration result than the Manhattan Distance. Tests with X-Means cluster points calculate the distance of Euclidean Distance using 100 iris data reaching the 9th iteration, while using 100 iris data by calculating the distance of Manhattan Distance reaching the 4th iteration. 7th iteration.

4. CONCLUSIONS

Based on the testing and evaluation of the method of determining the center point of the cluster with X-Means and the Euclidean Distance and Manhattan Distance matrix, the results of the study can be drawn several conclusions, including: Output Graph Testing the Variation of K Value Dataset IrisWhere the test using the Euclidean distance calculation on the iris data has a better iteration accuracy than the Manhattan distance. Based on the results of the X-Means iteration with the Euclidean Distance and Manhattan Distance matrix test parameters, the number of iterations varies from the number of variations of K with values of 2,3,4,5,6,7,8,9,10. The author concludes that the number of centroids 3 and 4 has a better iteration value using Manhattan Distance compared to the number of centroids that are getting higher and lower based on the iris dataset. Based on the Accuracy assessment of the iris data set, it was found that the Manhattan Distance matrix was better than the Euclidean Distance distance matrix, namely at the values of k=6, k=7, and k=8. The best Accuracy value of Braycurtis Distance is 96%. The best Accuracy Euclidean Distance

value is 95.33% and the best Accuracy Canberra Distance is 94.7%. The results of the authors conducted tests with variations in the number of centroids (K) with values of 2,3,4,5,6,7,8,9,10. The author concludes that the number of centroids 3 and 4 has a better iteration value compared to the number of centroids that are getting higher and lower based on the iris dataset with the Manhattan Distance matrix distance.

Reference

[1] Poteras, C. M., Mihaescu, M .C., & Mocanu, M. (2014). *An Optimized Version of the K-Means Clustering*. Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 695–699.

[2] Latifa Greeche., Maha Jazouli., Najia Es-Sbai., Aicha Majda., & Arsalane Zarghili. (2017). IEEE. pp. 1-4.

[3]Alfatih Muhammad., Ary Setijadi Prihatmanto., Rifki Wijaya., Harits Ar Rosyid., & Hashfi Rasis Hakim. (2018). *Distance Measurements Method for The Demite Pronunciation Assessment*. IEEE 8th International Conference on System Engineering and Technology (ICSET 2018), 15 - 16 October 2018, Bandung, Indonesia. pp. 189-194

[4]Nakyoung Kim., Hyojin Park., Jun Kyun Choi., & Jinhong Yang. (2017). *Time Gap Accounted Video Scene Segmentation with Modified Mean-shift X-means Clustering*. IEEE 6th Global Conference on Consumer Electronics (GCCE 2017) pp. 1-2

[5]Mahdi Shahbaba, Soosan Beheshti. (2012). *Improving X-Means Clustering With MNDL. he 11th International Conference on Information Sciences, Signal Processing and their Applications: Special Sessions* pp.1298-1302.

[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.