http://infor.seaninstitute.org/index.php/infokum/index
JURNAL INFOKUM, Volume 10, No.1, Desember 2021 ISSN : 2302-9706

APPLICATION OF X-MEANS METHOD FOR GROUPING EARLY CHILDHOOD DISEASES

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

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

Article Info
Received : 15 September 2021
Revised : 20 October 2021
Accepted : 04 November 2021

Abstract

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, Early Childhood Diseases

1. Introduction

The purpose of clustering is that objects (data) in a group are the same (related) to each other and different (unrelated) objects in other groups. The greater the similarity (homogeneity) within a group and the greater the differences between groups, the better or clearer the grouping. One of the algorithms that can be used in grouping is X-Means (Prasetyo, 2012).

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 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.

2. Method

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.

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].

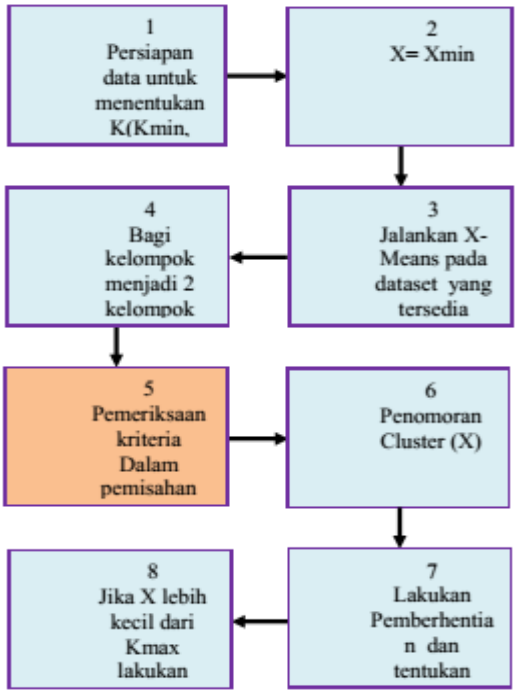

Figure 1. General Steps in X-Means Grouping

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.

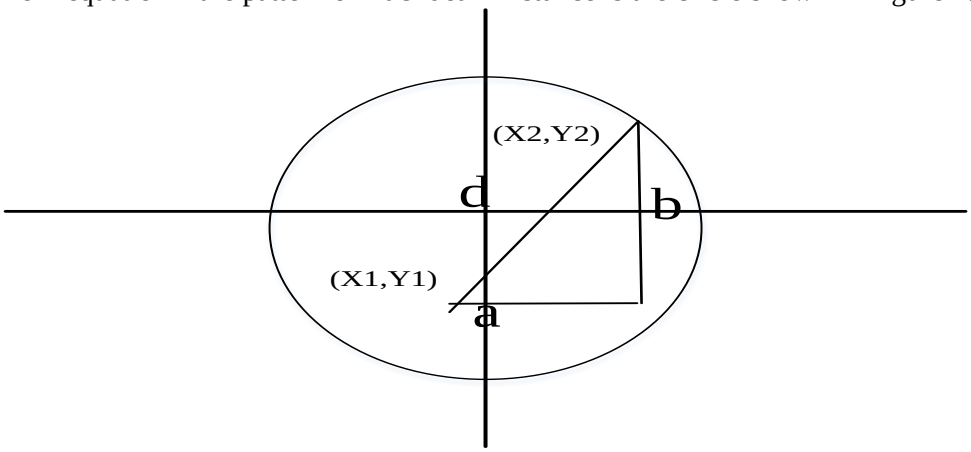

Figure 2. Euclidean Distance Patten

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

In using the clustering method, the process carried out for the formation of clusters is to transform the data into numeric form with specified codes, then determine the number of groups (K), calculate the centroid, calculate the distance of the object to the centroid and then group it based on the closest distance, if there is no object that moves the group then the iteration is complete. To determine the group of an object, the first thing to do is to measure the Euclidean distance between two object points (X and Y) which is defined as follows: Euclidean Distance: $\sqrt{((x2-x1)^2 + (y2-y1)^2}$

Table 3.1 Sample Patient Data to be Clustered

NO	NAME	AGE	GENDER	DIAGNOSIS
1	Aldi	3	Laki – laki	Demam
2	Mykayla	2	Perempuan	Pilek
3	Andini	5	Perempuan	Penyakit Diare
4	Alfi	4	Laki – laki	Mual dan Muntah
5	Alif Fadhilah	3	Perempuan	Cacar Air
6	Rudi	4	Laki – laki	Campak
7	Nining	2	Perempuan	Masalah Kulit
8	Hidayat	1	Laki – laki	Infeksi Telinga
9	Annisa	5	Perempuan	Radang Tenggorokan
10	Rendi Irawan	3	Laki – laki	Penyakit Eksim
11	Rudiansyah	4	Laki – laki	Mual dan Muntah
12	Anggun	5	Perempuan	Cacar Air
13	Wawan	4	Laki – laki	Campak
14	Endang	5	Perempuan	Masalah Kulit
15	Mega	4	Perempuan	Infeksi Telinga
16	Fahmi	3	Laki – laki	Radang Tenggorokan
17	Agung	4	Laki – laki	Penyakit Diare
18	Ayu Annisa	6	Perempuan	Mual dan Muntah
19	Sundari	2	Perempuan	Cacar Air
20	Wanda	4	Laki – laki	Penyakit Diare

1. Transformation Data

In order for the above data to be processed using the K-means algorithm, non-numeric data such as age, gender and diagnosis must be initialized in numeric form (numbers). This data can be expressed in variables, namely Age (X), Gender (Y), Diagnosis (Z) are as follows:

Table 3.2 Transformed Data

NO	NAME	AGE (X)	GENDER (Y)	DIAGNOSIS (Z)
1	A	3	1	1
2	B	1	2	2
3	C	3	2	1
4	D	3	1	3
5	E	1	2	4
6	F	1	1	5
7	G	2	2	6
8	H	2	1	7
9	I	2	2	6
10	J	3	1	8
11	K	1	1	2
12	L	1	2	4
13	M	2	1	7
14	N	2	2	7
15	O	3	2	8
16	P	3	1	3

17	Q	1	1	5
18	R	1	2	5
19	S	3	2	1
20	T	1	1	2

2. Distance Calculation Based on Euclidean Distance

The following is the process of calculating the X-Means method. In this example, 3 clusters will be formed. The centroid values are taken randomly, namely the center of cluster 1 and the center of cluster 3, the initialization of the cluster is chosen randomly with the data range from the lowest value to the highest value. The values for each cluster center are shown as follows:

Table 3.3 Center Point Sample Cluster Random

NO	Nama	X	Y	Z
1	A	3	1	1
3	C	3	2	1
7	G	2	2	6

Table 3.4 Result Iterasi 1

No	Nama	X	Y	Z	C1	C2	C3	Grup
1	A	3	1	1	0	1	5,196	1
2	B	1	2	2	2,449	2,236	4,123	2
3	C	3	2	1	1	0	5,099	2
4	D	3	1	3	2	2,236	3,316	1
5	E	1	2	4	3,741	3,605	2,236	3
6	F	1	1	5	4,472	4,582	1,732	3
7	G	2	2	6	5,196	5,099	0	3
8	H	2	1	7	6,082	6,164	1,414	3
9	I	2	2	6	5,196	5,099	0	3
10	J	3	1	8	7	7,071	2,449	3
11	K	1	1	2	2,236	2,449	4,242	1
12	L	1	2	4	3,741	3,605	2,236	3
13	M	2	1	7	6,082	6,164	1,414	3
14	N	2	2	7	6,164	6,082	1	3
15	O	3	2	8	7,071	7	2,236	3
16	P	3	1	3	2	2,236	3,316	1
17	Q	1	1	5	4,472	4,582	1,732	3
18	R	1	2	5	4,582	4,472	1,414	3
19	S	3	2	1	1	0	5,099	2
20	T	1	1	2	2,236	2,449	4,242	1

After calculating using the existing cluster formula, in iteration 1 and iteration 2 the position of the cluster does not change or there is no data moving groups again, then the calculation can be stopped. The group results obtained from the calculation of iteration 1 and iteration 2 are as follows:

Group 1: (1,2,2,1,3,3,3,3,3,3,1,3,3,3,3,1,3,3,2,1)

Group 2: (1,2,2,1,3,3,3,3,3,3,1,3,3,3,3,1,3,3,2,1)
Information :
From 20 data obtained 3 groups, it can be concluded as follows:
1. Group 1 Centroid 1: (2,2 1,4 2,2) there are 5 data. Based on the above calculations, it can be seen that in cluster 1 group 1, the patients are children.
2. Group 2 Centroid 2: (1 1,6 1,6) there are 3 data. Based on the above calculations, it can be concluded that in cluster 2, group 2 is an adult patient.
3. Group 3 Centroid 3: (1,75 1,58 6) contains 12 data. Based on the above calculations, it can be concluded that in cluster 3, group 3 is the patient is the elderly.

4. Conclusions

From the writing of the thesis entitled Grouping Diseases in Patients Based on Age with the K-Means Clustering Method (Case Study: Bahorok Health Center), the following conclusions can be drawn:

1. The results of the pattern analysis above From 20 data obtained 3 groups, it can be concluded as follows:

Group 1 Centroid 1: (2,2 1,4 2,2) there are 5 data. Based on the above calculations, it can be seen that in cluster 1 group 1, the patients are children.
Group 2 Centroid 2: (1 1,6 1,6) there are 3 data. Based on the above calculations, it can be concluded that in cluster 2, group 2 is an adult patient.
Group 3 Centroid 3: (1,75 1,58 6) there are 12 data. Based on the above calculations, it can be concluded that in cluster 3, group 3 is the patient is the elderly.

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.
[7] Purwa Hasan Putra, Et Al, Application Of The K-Means Algorithm In Identifying Types Of Skin Disease, JURNAL INFOKUM, Volume 9, No. 2, Juni 2021