JURNAL IPTEKS TERAPAN
Research of Applied Science and Education V14.i2 (152-159)

ISSN : 1979-9292
E-ISSN : 2460-5611

Analysis Of The Use Of X-Means Method In Grouping Interest And Talent Data Students

Purwa Hasan Putra[1], Muhammad Syahputra Novelan[2]
[1]Fakultas Sains dan Teknologi, Universitas Pembangunan Panca Budi, Jl. Jenderal Gatot Subroto, KM 4,5 Sei Sikambing 20122 Medan, Sumatera Utara, Indonesia
purwahasanputra@dosen.pancabudi.ac.id
[2]Fakultas Sains dan Teknologi, Universitas Pembangunan Panca Budi, Jl. Jenderal Gatot Subroto, KM 4,5 Sei Sikambing 20122 Medan, Sumatera Utara, Indonesia
putranovelan@dosen.pancabudi.ac.id

Abstract

The X-Means algorithm is an algorithm used for grouping data. The x means algorithm is the development of k-means. 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 way and only based on the data set itself. The research results using the X-Means algorithm with Davies-Bouldin Index evaluation Determination of the number of Centroid clusters is done by modifying the X-Means method. In grouping this data, clustering is performed on each student data of the collected variables. Each student's gift and interest will be matched with the college and department of what each student is interested in.

Keywords : X-Means, Decision Support System

INTRODUCTION

Grouping can use clustering to group data based on the similarity between data, so that the data with the closest resemblance are in one cluster while the data is different in other groups[1]. The process of grouping data into clusters or groupings so that data in one cluster has a maximum level of similarity and between clusters has a minimum similarity is called Clustering[2]. Clustering is divided into 2 approaches in its development namely sclustering partitioning and hierarchical approaches[3]. The purpose of clustering is that the objects (data) in a group are the same (related) to each other and different (unrelated) objects in another group[4]. The greater the similarity (homogeneity) in a group and the greater the differences between groups, the better or clearer the grouping. One algorithm that can be used in grouping is X-Means[5]. The X-Means algorithm is a development of K-Means. The weakness of X-Means is in determining the distance matrix, the distance matrix is an important factor that depends on the data set of the X-Means algorithm. The resulting distance matrix value will affect the performance of the algorithm[6].

METHOD

In doing this Framework are steps that will be taken in order to solve the problem that will be discussedIn doing this Framework are steps that will be taken in

Submitted : 06 Jul 2020 – Accepted : 29 Jul 2020 – Published : 1 Agt 2020

order to solve the problem that will be discussed. In this study studied literature relating to the problem. Then the literature studied is selected to determine which literature will be used in research. Literature sources obtained from libraries, journals, articles and other concepts that support in completing the system to be built including references. In conducting this research, data and information collection at this stage was carried out to find out about the system under study. From the data and information collected, data will be obtained to support research and data collection is done to determine the needs of users. In conducting this research, data and information collection at this stage was carried out to find out about the system under study. From the data and information collected, data will be obtained to support research and data collection is done to determine the needs of users. Conduct interviews with parties related to the flow of the problem. This interview was conducted to obtain writing material and an explanation of the observations made. System design activities carried out as a prelude to the design of the system to be built as needed. And at this stage the interface design of the system will be made. The system implementation is carried out according to the design and design of the application interface to be built. At this stage, coding or making programs so that the system designed can be used by the user. System testing is carried out to determine the feasibility test of the system that has been built as expected and by doing the test can find out the weaknesses and strengths of the system designed so that it can be repaired at a later stage.

RESULT

In this study the authors conducted an evaluation of the problem to help students make decisions that determine students' interests and talents. The data used were sourced from primary data obtained through the implementation of interest and aptitude tests with "X" Vocational School student participants. From the primary data obtained, the data used is separated from a set of operational data.

Application of the X-Means Method
Initialization of cluster center 1 and cluster center 2, initial cluster with range of data between lowest value and highest value. The values for each cluster center are shown as follows.

Table.1 Cluster Center Point

Pusat Cluster	Nama Item	X1	X2	X3	X4
1	TKJ	4.9	3	1.4	0.2
2	RPL	5.8	2.7	5.1	1.9

Inf:
X1 = Visual Reasoning
X2 = Numerical Reasoning
X3 = Verbal Analysis
X4 = Spatial Reasoning

At this stage the cluster point distance update process is carried out.

$$D(t) = d(y_j - x_i)\ {}^2\forall j,i$$

yj = i data
Number of clusters = 2
xi = jth cluster center point
1st calculation
1st data calculation (yj) with cluster center 1 (xi)

$= (5.1 - 4.9)\ ^2 + (3.5 - 3)\ ^2 + (1.4 - 1.4)\ ^2 + (0.2 - 0.2)\ ^2$
$= 0.04 + 0.25 + 0 + 0$
$= 0.29$

ISSN : 1979-9292
E-ISSN : 2460-5611

1st data calculation (yj) with cluster center 2 (xi)
$=(5.1-5.8)^2+(3.5-2.7)^2+(1.4-5.1)^2+(0.2-1.9)^2$
$= 0.49 + 0.64 + 13.69 + 2.89$
$= 17.71$

2nd data calculation (yj) with cluster center 1 (xi)
$=(4.9-4.9)^2+(3-3)^2+(1.4-1.4)^2+(0.2-0.2)^2$
$= 0 + 0 + 0 + 0$
$= 0$

2nd data calculation (yj) with cluster center 2 (xi)
$=(4.9-5.8)^2+(3-2.7)^2+(1.4-5.1)^2+(0.2-1.9)^2$
$= 0.81 + 0.09 + 13.69 + 2.89$
$= 17.48$

2nd data calculation (yj) with cluster center 2 (xi)
$=(4.7-4.9)^2+(3.2-3)^2+(1.3-1.4)^2+(0.2-0.2)^2$
$= 0.04 + 0.04 + 0.01 + 0$
$= 0.09$

3rd data calculation (yj) with cluster center 2 (xi)
$=(4.7-5.8)^2+(3.2-2.7)^2+(1.3-5.1)^2+(0.2-1.9)^2$
$= 1.21 + 0.25 + 14.44 + 2.89$
$= 18.79$

4th data calculation (yj) with cluster center 1 (xi)
$=(7-4.9)^2+(3.2-3)^2+(4.7-1.4)^2+(1.4-0.2)^2$
$= 4.41 + 0.04 + 10.89 + 1.44$
$= 16.78$

4th data calculation (yj) with cluster 2 (xi) center point
$=(7-5.8)^2+(3.2-2.7)^2+(4.7-5.1)^2+(1.4-1.9)^2$
$= 1.44 + 0.25 + 0.16 + 0.25$
$= 2.1$

5th data calculation (yj) with cluster center 1 (xi)
$=(6.4-4.9)^2+(3.2-3)^2+(4.5-1.4)^2+(1.5-0.2)^2$
$= 4.41 + 0.04 + 10.89 + 1.44$
$= 13.59$

5th data calculation (yj) with cluster center 2 (xi)
$=(6.4-5.8)^2+(3.2-2.7)^2+(4.5-5.1)^2+(1.5-1.9)^2$
$= 0.36 + 0.25 + 0.36 + 0.16$
$= 1.13$

6th data calculation (yj) with cluster center 1 (xi)
$=(6.9-4.9)^2+(3.1-3)^2+(4.9-1.4)^2+(1.5-0.2)^2$
$= 4 + 0.01 + 12.25 + 1.69$
$= 17.95$

6th data calculation (yj) with cluster center 2 (xi)
$=(6.9-5.8)^2+(3.1-2.7)^2+(4.9-5.1)^2+(1.5-1.9)^2$
$= 4 + 0.01 + 12.25 + 1.69$
$= 17.95$

7th data calculation (yj) with cluster center 1 (xi)
$=(6.3-4.9)^2+(3.3-3)^2+(6-1.4)^2+(2.5-0.2)^2$
$= 1.96 + 0.09 + 21.16 + 5.29$
$= 28.5$

7th data calculation (yj) with cluster center 2 (xi)
$=(6.3-5.8)^2+(3.3-2.7)^2+6(4.9-5.1)^2+(2.5-1.9)^2$
$= 0.25 + 0.36 + 0.81 + 0.36$
$= 1.78$

8th data calculation (yj) with cluster center 1 (xi)
$=(5.8-4.9)^2+(2.7-3)^2+(5.1-1.4)^2+(1.9-0.2)^2$
$= 0.81 + 0.09 + 13.69 + 2.89$
$= 17.48$

8th data calculation (yj) with cluster center 1 (xi)
$=(5.8-5.8)^2+(2.7-2.7)^2+(5.1-5.1)^2+(1.9-1.9)^2$
$= 0 + 0 + 0 + 0$
$= 0$

9th data calculation (yj) with cluster center 1 (xi)
$=(7.1-4.9)^2+(3-3)^2+(5.9-1.4)^2+(2.1-0.2)^2$
$= 4.84 + 0 + 20.25 + 3.61$
$= 28.7$

9th data calculation (yj) with cluster center 2 (xi)
$=(7.1-5.8)^2+(3-2.7)^2+(5.9-5.1)^2+(2.1-1.9)^2$
$= 1.69 + 0.09 + 0.64 + 0\ .04$
$= 2.46$

Submitted : 06 Jul 2020 – Accepted : 29 Jul 2020 – Published : 1 Agt 2020

ISSN : 1979-9292
E-ISSN : 2460-5611

10th data calculation (yj) with cluster center 1 (xi)

$= (6.3-4.9)^2 + (2.9-3)^2 + (5.6-1.4)^2 + (1.8-0.2)^2$

$= 1.96 + 0.01 + 17.64 + 2.56$

$= 22.17$

10th data calculation (yj) with cluster center 1 (xi)

$= (6.3-5.8)^2 + (2.9-2.7)^2 + (5.6-5.1)^2 + (1.8-1.9)^2$

$= 0.25 + 0.04 + 0.25 + 0.01$

$= 0.55$

Calculation of the update of the center point of the cluster with the attribute data has been calculated and the results obtained. The values from each calculation of the cluster 1 center point and the cluster 2 center point can be shown as follows:

Table 2 Cluster Point Distance Updates

Data Ke-	Jarak Data Dengan Titik Pusat Cluster 1	Jarak Data Dengan Titik Pusat Cluster 2
1	0.29	17.71
2	0	17.48
3	0.09	18.79
4	16.78	2.1
5	13.59	1.13
6	17.95	1.57
7	28.5	1.78
8	17.48	0
9	28.7	2.46
10	22.17	0.55

From the results of the calculation of the cluster point distance update, it is obtained that the 3rd and 10th Data become the new cluster center points to be used in the next iteration. The results of the calculation of the new cluster center points are shown as follows:

Table 3 Cluster Center Points

Pusat Cluster	Nama Item	X1	X2	X3	X4
1	IrisSetosa	4.9	3	1.4	0.2
2	IrisVirginica	6.3	2.9	5.6	1.8

Furthermore, after obtaining the cluster center point by the X-Means method. Then the next calculation is done with the euclidean distance formula. a. Calculate the distance to all data points with the euclidean distance formula. The calculation is as follows:

Euclidean Distance: √((x2-x1)² + (y2-y1)²

First iteration

Cluster 1

$= \sqrt{((5.1-4.7)^{\wedge}2 + (3.5-3.2)^{\wedge}2 + (1.4-1.3)^{\wedge}2} + (0.2-0.2)^{\wedge}2 = 0.509$

$= \sqrt{((4.9-4.7)^{\wedge}2 + (3-3.2)^{\wedge}2 + (1.4-1.3)^{\wedge}2} + (0.2-0.2)^{\wedge}2 = 0.3$

$= \sqrt{((7-4.7)^{\wedge}2 + (3.2-3.2)^{\wedge}2 + (4.7-1.3)^{\wedge}2} + (1.4-0.2)^{\wedge}2 = 4.276$

$= \sqrt{((6.4-4.7)^{\wedge}2 + (3.2-3.2)^{\wedge}2 + (4.5-1.3)^{\wedge}2} + (1.5-0.2)^{\wedge}2 = 3.849$

$= \sqrt{((6.9-4.7)^{\wedge}2 + (3.1-3.2)^{\wedge}2 + (4.9-1.3)^{\wedge}2} + (1.5-0.2)^{\wedge}2 = 4.415$

$= \sqrt{((6.3-4.7)^{\wedge}2 + (3.3-3.2)^{\wedge}2 + (6-1.3)^{\wedge}2} + (2.5-0.2)^{\wedge}2 = 5.472$

$= \sqrt{((5.8-4.7)^{\wedge}2 + (2.7-3.2)^{\wedge}2 + (5.1-1.3)^{\wedge}2} + (1.9-0.2)^{\wedge}2 = 4.334$

$= \sqrt{((7.1-4.7)^{\wedge}2 + (3-3.2)^{\wedge}2 + (5.9-1.3)^{\wedge}2} + (1.8-0.2)^{\wedge}2 = 5.529$

Cluster 2

$= \sqrt{((5.1-6.3)^{\wedge}2 + (3.5-2.9)^{\wedge}2 + (1.4-5.6)^{\wedge}2} + (0.2-1.8)^{\wedge}2 = 4.690$

$= \sqrt{((4.9-6.3)^{\wedge}2 + (3-2.9)^{\wedge}2 + (1.4-5.6)^{\wedge}2} + (0.2-1.8)^{\wedge}2 = 4.708$

$= \sqrt{((7-6.3)^{\wedge}2 + (3.2-2.9)^{\wedge}2 + (4.7-5.6)^{\wedge}2} + (1.4-1.8)^{\wedge}2 = 1.244$

$= \sqrt{((6.4-6.3)^{\wedge}2 + (3.2-2.9)^{\wedge}2 + (4.5-5.6)^{\wedge}2} + (1.5-1.8)^{\wedge}2 = 1.183$

$= \sqrt{((6.9-6.3)^{\wedge}2 + (3.1-2.9)^{\wedge}2 + (4.9-5.6)^{\wedge}2} + (1.5-1.8)^{\wedge}2 = 0.989$

$= \sqrt{((6.3-6.3)^{\wedge}2 + (3.3-2.9)^{\wedge}2 + (6-5.6)^{\wedge}2} + (2.5-1.8)^{\wedge}2 = 0.9$

$= \sqrt{((5.8-6.3)^{\wedge}2 + (2.7-2.9)^{\wedge}2 + (5.1-5.6)^{\wedge}2} + (1.9-1.8)^{\wedge}2 = 0.741$

$= \sqrt{((7.1-6.3)^{\wedge}2 + (3-2.9)^{\wedge}2 + (5.9-5.6)^{\wedge}2} + (1.8-1.8)^{\wedge}2 = 0.911$

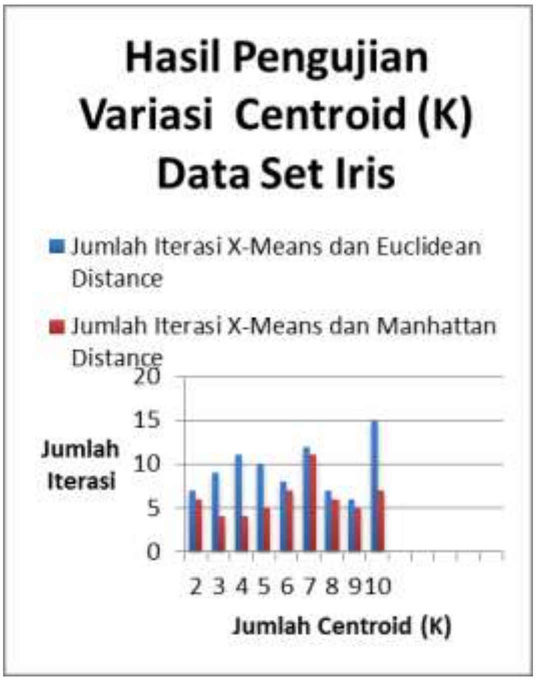
Table 4. Clusters for Each Data Group

No	Nama Item	Bakat	Kelompok Cluster
1	TKJ	Ekonomi	Cluster 1
2	TKJ	IT Komputer	Cluster 1
3	TKJ	IT Komputer	Cluster 1
4	RPL	Psikologi	Cluster 2
5	RPL	Tarian	Cluster 2
6	RPL	Akuntansi dan Keuangan	Cluster 2
7	RPL	Tarian	Cluster 2

Submitted : 06 Jul 2020 – Accepted : 29 Jul 2020 – Published : 1 Agt 2020

8	RPL	Ekonomi	Cluster 2
9	RPL	Tarian	Cluster 2
10	RPL	Ekonomi	Cluster 2

After the cluster process, from table 3 you can see the clustering of each data according to the cluster center point. The following are the results of the 1st iteration calculation, where Data 2 and 8 become the center of the cluster. For the next iteration it is calculated with the same stage until the convergence of each data found in the classification. The following graphs of centroid (K) variation test results are shown as follows.



Picture 1. Testing Graph Output

CONCLUTION

Based on testing and evaluation of the method of determining the cluster center point with X-Means and Euclidean Distance and Manhattan Distance matrices, as for the results of the study can be drawn several conclusions, Based on testing and evaluation of the method of determining the cluster center point with X-Means and Euclidean Distance and Manhattan Distance matrices, as for the results of the study can be drawn several conclusions. Based on the results of the X-Means iteration with Euclidean Distance and Manhattan Distance matrix test parameters have a number of iterations that vary from the number of variations of K with a value of 2,3,4,5,6,7,8,9,10. The author draws the conclusion that the number of centroids 3 and 4 has a better iteration of values using Manhattan Distance compared to the number of centroids that get higher and lower based on the iris dataset. Based on the Accuracy assessment on the iris data set, it was found that the Distance Distance Manhattan matrix is 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 Distances is 96%. The best Euclidean Distribution Accuracy value is 95.33% and the best Canberra Accuracy Value is 94.7%. The results of the authors conducted testing with variations in the number of centroids (K) with a value of 2,3,4,5,6,7,8,9,10. The author draws the conclusion that the number of centroids 3 and 4 has a better iteration of values compared to the number of centroids that are getting higher and lower based on the iris dataset with the distance matrix Manhattan Distance.

REFERENCE

[1] Eka Sabna, Muhardi. 2016. “Penerapan Data Mining Untuk Memprediksi Prestasi Akademik Mahasiswa Berdasarkan Dosen, Motivasi, Kedisiplinan, Ekonomi, dan Hasil Belajar.” *Jurnal CoreIT* 41-44.

Submitted : 06 Jul 2020 – Accepted : 29 Jul 2020 – Published : 1 Agt 2020

[2] Fakhroddin Noorbehbahani., Sadeq Mansoori. (2018). *A New Semi-supervised Method for Network Traffic Classification Based on X-means Clustering and Label Propagation*. 8th International Conference on Computer and Knowledge Engineering (ICCKE 2018), October 25-26 2018, Ferdowsi University of Mashhad. pp. 120-125

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

[4] 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.

[5] 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

[6] 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.

Submitted : 06 Jul 2020 – Accepted : 29 Jul 2020 – Published : 1 Agt 2020