International Journal of Psychosocial Rehabilitation, Conference Special Issue
ISSN: 1475-7192

Determine the optimal number of clusters k-means

[1]Handry Eldo, [2]Syahril Efendi,[3] Herman Mawengkang

ABSTRACT-- *K-means clustering technique has been very widely used in various fields such as academics, practitioners and so on. However, k-means itself still has some shortcomings, including the problem of the accuracy of the algorithm used to measure the similarity between objects being compared. To overcome this problem, the optimum number of clusters will be calculated in this study (euclidean distance, Manhattan distance, and Chebyshev distance) to find out the optimum number of clusters. Silhouette Coefficient test results for each distance measure, including Euclidean Distance worth 0.232149, Manhattan Distance worth 0.240016, and Chebyshev Distance worth 0.242821. Based on the results of the silhouette coefficient testing conducted, the most optimal distance measure for this case is Chebyshev Distance, that is, the silhouette coefficient value closest to 1 is 0.242821*

Keyword *-- k-means, cluster, Euclidean, Manhattan,*

I. INTRODUCTION

Clustering is an unsupervised data mining method. There are two types of clustering that are often used in the grouping process, namely: hierarchical and non-hierarchical. The purpose of the clustering process is to group data into one cluster, so that objects in a cluster have very large similarities with other objects in the same cluster, but are not similar to objects in another cluster.

The implementation of basic k-means algorithm has been done by using Euclidean Distance measurement. In this study distance measurements will be implemented using the Euclidean Distance, Manhattan and Chebyshev methods on k-means. After that the accuracy of each method will be compared to find out the optimum number of clusters in each silhouette coefficient algorithm.

II. LITERATURE REVIEW

Distance measurement is one of the most important roles in determining the similarity or regularity between data and items. This is done to find out, in what way the data is said to be related, similar, not similar, and what distance measurement method is needed to compare it. In the clustering process, the stages of determining or describing quantitative values of the degree of similarity or dissimilarity of data (proximity measure) have a very

[1]*Department of Computer Science, Faculty of Computer Science andInformation Technology, Universitas Sumatera Utara, Medan, Indonesia*.
[2]*Department of Computer Science, Faculty of Computer Science andInformation Technology, Universitas Sumatera Utara, Medan, Indonesia*.
[3]*Department of Computer Science, Faculty of Computer Science andInformation Technology, Universitas Sumatera Utara, Medan, Indonesia*.

Received: 02 Jan 2020 | Revised: 12 Feb 2020 | Accepted: 17 Mar 2020

important role, so it is necessary to do calculations to obtain the optimum number of clusters among the most frequently used methods namely euclidean, manhattan, and Chebyshev.

1. Euclidean Distance

Euclidean distance is a type of distance calculation method used to measure the distance of 2 (two) points in Euclidean space. To measure the degree of similarity of data with the euclidean distance formula the following formula is used:

$$d(x,y) = |x - y| = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

2. 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:

$$d(x,y) = \sum_{i=1}^{n} |x_i - y_i|$$

3. Chebyshev distance

Chebyshev distance (Tchebychev) is a method for measuring distance based on the absolute value of the difference in the coordinates of two points. That is if there are two different vector values for each element, the distance measured using Chebyshev is based on the absolute value of the difference in elements. The general equation used is as follows:

$$d_{ij} = {}^{max}_{k}(x_{ik} - x_{ij})$$

III. K-MEANS ALGORITHM

k-means is an algorithm for grouping data based on its attribute value into as many k clusters. The workflow of the k-means algorithm is as follows:

1. Determine the number of cluster clusters to be formed
2. Determine the number of k data to be used as a cluster center
3. Determine the membership of a cluster by collecting data to the nearest cluster center
4. Calculate the middle value of a cluster to become a new cluster center. This process iterates through step 3 if the new cluster center formed is not the same as the previous cluster center, so the cluster center does not change again. If described, then the data flow diagram of the k-means algorithm is as follows:

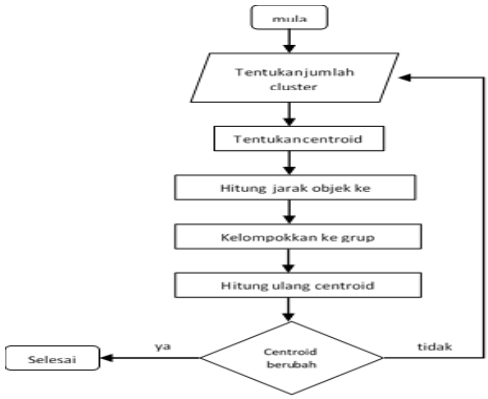

Figure 1: K-Means Algorithm

IV. SILHOUETTE COEFFICIENT

This Silhouette algorithm is used to see the quality and strength of clusters, how well an object is placed in a cluster. This Silhouette method is a combination of cohesion and separation methods. The stages of Silhouette Coefficient calculation are as follows:

1. Calculate the average distance from a document

$$a(i) = \frac{1}{|A| - 1} \sum j \in_{A, j \neq i} d(i, j)$$

where j is another document in one cluster A and d (i, j) is the distance between document i and j.

2. Calculate the average distance from document i with all documents in other clusters, and take the smallest value.

$$d(i, c) = \frac{1}{A} \sum j \in C \; d \; (i, j)$$

where d (i, C) is the average distance of document i to all objects in other clusters C where A ≠ C.

$$b(i) = \min C \neq A \; d \; (i, C)$$

3. The Silhouette Coefficient value is:

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

V. Dataset

The dataset used for testing is sales data in a self-service supermarket that has been normalized / cleaned beforehand to take some data samples related to algorithm testing.

Table 1: sales dataset

idpelanggan	usia	Pendapatan (juta rupiah)	Rating pengeluaran (1-100)
1	19	15	39
2	21	15	81
3	20	16	6
4	23	16	77
5	31	17	40

VI. RESULTS AND DISCUSSION

5.1 Clustering Results

1. Euclidean Distance

Following are the final clustering results from Euclidean Distance. Tested with 5 data:

Table 2: Clustering Results

Id pelanggan	usia	Pendapatan (juta rupiah)	Rating pengeluaran (1-100)	C1	C2	C3
1	19	15	39		x	
2	21	15	81	x		
3	20	16	6			x
4	23	16	77	x		
5	31	17	40		x	

2. Manhattan Distance

Following are the final clustering results from Manhattan Distance.

Table 3: Manhattan Distance

Ipelanggan	usia	Pendapatan (juta rupiah)	Rating pengeluaran (1-100)	C1	C2	C3
1	19	15	39	X		
2	21	15	81		X	
3	20	16	6			X
4	23	16	77		X	
5	31	17	40		X	

3. Chebyshev Distance

Following are the final clustering results from Chebyshev Distance.

Table 4: Chebyshev Distance

Ipelanggan	usia	Pendapatan (juta rupiah)	Rating pengeluaran (1-100)	C1	C2	C3
1	19	15	39	X		
2	21	15	81		X	
3	20	16	6		X	
4	23	16	77			X
5	31	17	40		X	

5.2 Silhouette Coefficient Test Results

Silhouette Coefficient testing is used to see the quality and strength of the cluster, how well an object is placed in a cluster. Tests carried out on the results of clustering each distance measure on the K-Means Clustering method. The results obtained from the overall calculation of the Silhouette Coefficien are as follows:

Table 5: Silhouette Coefficient Test Results

nomor	Jenis distance	Hasil Silhouette Coefficient
1	Euclidean Distance	0,232149
2	Manhattan Distance	0,240016
3	Chebyshev Distance	0,242821

VII. CONCLUSIONS

In this case the most optimum number of clusters using the calculation of the eucliende distance, mahattan distance and Chebyshev distance is k = 3 followed by k = 4 and k = 5. By testing Silhouette Coefficient Clustering as follows:

a. The Euclidean Distance value of the Silhouette Coefficient is 0.232149
b. Manhattan Distance's Silhouette Coefficient value is 0.240016
c. The Chebyshev Distance value of the Silhouette Coefficient is 0.242821

The most optimal distance measure to use in this case is Chebyshev Distance. Because based on the evidence done using Silhouette Coefficient that the higher the Coefficient value, the more optimal.

REFERENCES

1. Santoso,Budi.dkk, (2017). Optimasi K-Means untuk Clustering Kinerja Akademik Dosen Menggunakan Algoritme Genetika. E-ISSN:2548-964x
2. Sinwar, Deepak & Rahul Kaushik (2014). Study of Euclidean and Manhattan Distance Metrics using Simple K-Means Clustering. ISSN: 2321-9653
3. P.-N. Tan, M. Steinbach, A. Karpatne, and V. Kumar, Introduction to Data Mining (2nd Edition), 2nd ed. New York: Pearson, 2018.
4. Ong, Johan Oscar. 2013. “Implementasi Algoritma K-Means Clustering Untuk Menentukan Strategi Marketing.” (April): 10–20.
5. M. Nishom, “Implementasi Metode K-Means berbasis Chi-Square pada Sistem Pendukung Keputusan untuk Identifikasi Disparitas Kebutuhan Guru,” J. Sist. Inf. Bisnis, vol. 8, no. 2, pp. 1–8, 2018.
6. 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.
7. 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.

Received: 02 Jan 2020 | Revised: 12 Feb 2020 | Accepted: 17 Mar 2020