Analysis of determining centroid clustering x-means algorithm with davies-bouldin index evaluation

M Mughnyanti[1], S Efendi[2], M Zarlis[3]

[1]Student, Department of Information Technology, Faculty of Computer Science and Information Technology, Universitas Sumatera Utara, Indonesia
[2,3]Department of Information Technology, Faculty of Computer Science and Information Technology, Universitas Sumatera Utara, Indonesia

syahril.efendi@usu.ac.id[2], m.zarlis@yahoo.com[3]

Abstract. Clustering is a process to group data into several clusters or groups so the data in one cluster has a maximum level of similarity and data between clusters has a minimum similarity. X-means clustering is used to solving one of the main weaknesses of K-means clustering need for prior knowledge about the number of clusters (K). In this method, the actual value of K is estimated in a way that is not monitored and only based on the data set itself. The results of the study using the X-Means algorithm with the Davies-Bouldin Index evaluation to determine the number of Centroid clusters is done by modifying the X-Means method to do some centroid determination to get 11 iterations. The result is produces cluster members that have a good level of similarity with other data. In determining the number of centroids, use the Davies-Bouldin Index method where testing with 2 clusters has a minimum value with a DBI value close to 0.

1. INTRODUCTION

Clustering is a process to group data into several clusters or groups so the data in one cluster has a maximum level of similarity and data between clusters has a minimum similarity.

The center of the cluster or centroid is the starting point for the group in clusters in the K-Means algorithm. The data is done by calculating the closest distance to the initial cluster center point as a central point in the formation of each group or cluster. But in this situation this determination of the initial cluster center point is the weakness of the K-Means algorithm. This is because there is no approach used in selecting and determining the cluster center point. The cluster center point is chosen arbitrarily or randomly from a set of data. The clustering results of the K-Means algorithm are often not optimal and not optimal in every conducted experiment. Therefore, it can be said that the good and bad results of clustering depend on the center point of the cluster or the initial centroid [1].

X-means clustering is used to solving one of the main weaknesses of K-means clustering need for prior knowledge about the number of clusters (K). In this method, the actual value of K is estimated in a way that is not monitored and only based on the data set[2].

Davies-Bouldin Index (DBI) is one method used to measuring cluster validity in a clustering method. Measuring with Davies Bouldin Index maximizes inter-cluster distance and at the same time tries to minimize the distance between points in a cluster. If the inter cluster distance is maximal, it means that the characteristics of each cluster are small so that the differences between clusters are more apparent. If the minimum intra-cluster distance means that each object in the cluster has a high level of characteristic similarity [3]

The clustering results obtained from determining the proposed cluster center point are then evaluated by the DBI method. So that it can be seen that the correlation of the method of determining cluster center points is based on the Sum of Squared Error to increase cluster quality based on DBI values obtained.

2. RESEARCH BACKGROUND

According to the results of Mahdi Shahbaba, Soosan Beheshti research shown that MNDL (Minimum Noiseless Description Length) is stronger than that of BIC (Bayesian Information Criterion) in terms of increasing variance. MNDL is more accurate in predicting the correct number of clusters, yet computational complexity is the same as BIC. MNDL (Minimum Noiseless Description Length) increases variance making clusters more dispersed and consequently it is more difficult to distinguish between the number of clusters and cluster members [2].

The result of the study from Dela Arundina, Shaufiah shown that performance measurements with parameter changes in cluster range, are known that although the range of cluster values in X-means changes, BIC scores, number of clusters are produced, and Silhouette Coefficient accuracy does not change. Measuring the accuracy of the Silhouette Coefficient clustering, it is evident that X-means has a better level of accuracy than K-means. Although the difference is not too significant, the use of X-means is recommended if the desired number of clusters is unknown. Weaknesses: The use of X-means is done with the exact number of clusters that are unknown [4].

The result of Bartosz Krawczyk, Micha I Wo'zniak shown that x-means has an optimization of the clustering model with little time. Xmeans is able to provide very satisfying results in one class problems. Weaknesses: Checking different criteria in measuring the similarity of each cluster model should be focused [5].

3. PROPOSED METHOD

3.1. Clustering

Clustering is a method of grouping or partitioning data in a data set. Basically clustering is a method to finding and grouping data that has similarity between one data and another (Bhusare, 2014). Clustering is the process of separating a set of data or objects into smaller groups or clusters based on the similarity of characteristics that are possessed [6].

3.2. X-Means Clustering

X-means clustering is used to solving one of the main weaknesses of K-means clustering need for prior knowledge about the number of clusters (K). In this method, the actual value of K is estimated in a way that is not monitored and only based on the data set. Kmax and Kmin as the upper and lower limits for the possible values of X. In the first step of the X-means grouping, know that at present X = Xmin, X-means finding the initial structure and centroid. In the next step, each cluster in the estimated structure is treated as the parent cluster, which can be divided into two groups.

3.3. Davies-Bouldin Index

David L. Davies and Donald W. Bouldin introduced a method and the name of this methode using name with both of them, namely the Davies-Bouldin Index (DBI) used to evaluating clusters. Evaluation using Davies-Bouldin Index has an internal cluster evaluation scheme, where good or not cluster results are seen from the quantity and proximity between

cluster results. Davies-Bouldin Index is one method used to measure cluster validity in a grouping method, cohesion is defined as the sum of the proximity of the data to the cluster center point of the cluster followed. While the separation is based on the distance between the cluster center points to the cluster. Measurements with Davies-Bouldin This index maximizes the inter-cluster distance between the Ci and Cj clusters and at the same time tries to minimize the distance between points in a cluster. If the inter-cluster distance is maximal, it means that the similarity of characteristics between each cluster is small so that the differences between clusters appear more clearly. If the minimum intra-cluster distance means that each object in the cluster has a high level of characteristic similarity [3].

4. RESULTS AND ANALYSIS

In this study is an explanation of the use of the X-Means method with the determination of the centroid value, as well as an analysis of several methods applied that can also get results in research by clustering data. Training and testing in this study is by doing the best data grouping on the X-Means method by doing some testing on determining the number of centroids.

In the predetermined dataset, a model of the method that will be used, namely X-Means, with the graph used can be seen as follows:

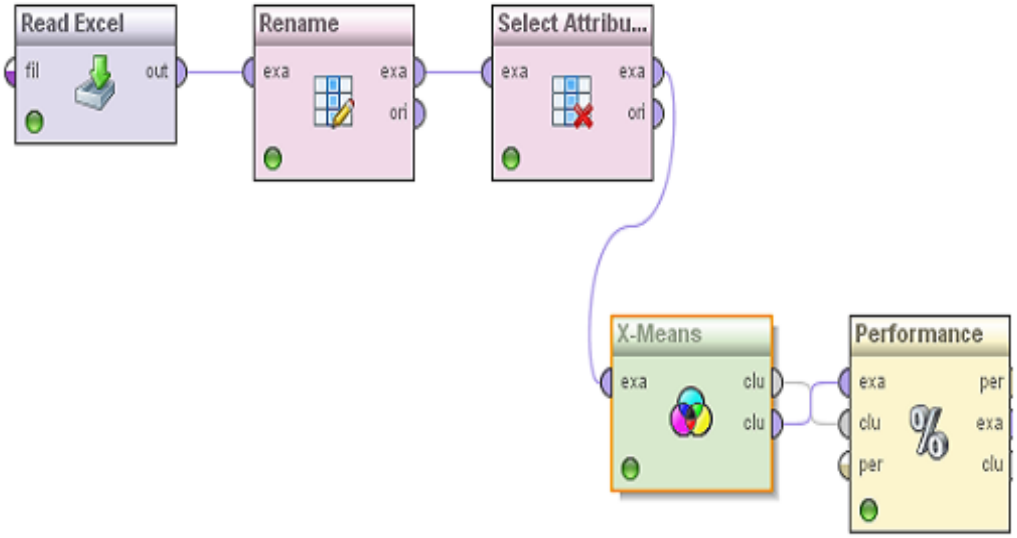

Figure 1. X-Means Model Plan.

In the testing of data clustering analysis is done by using the iris dataset as test data. The clustering process uses X-Means by initializing the number of clusters C = 2 with the center of cluster 1 and the center of cluster 2. The iris dataset is shown below.

Tabel 1. Dataset Iris

No.	Item Name	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	IrisSetosa	4.6	3.1	1.5	0.2
5	IrisSetosa	5	3.6	1.4	0.2

6	IrisSetosa	5.4	3.9	1.7	0.4
7	IrisSetosa	4.6	3.4	1.4	0.3
8	IrisSetosa	5	3.4	1.5	0.2
9	IrisSetosa	4.4	2.9	1.4	0.2
10	IrisSetosa	4.9	3.1	1.5	0.1
:	:	:	:	:	:
:	:	:	:	:	:
100	IrisVirginica	7.2	3	5.8	1.6

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

Clustering evaluation is carried out with the aim of knowing how well the quality of the clustering results. In this study, evaluation of the results of clustering used is davies-bouldin index. To get the Davies-Bouldin Index value, first calculated the value of the Within-cluster Sum of Square (SSW), Sum of Square Between-cluster (SSB) and Ratio.

The first stage of clustering evaluation using Davies-Bouldin Index is to calculate the value of Sum of Square Within-cluster (SSW).

The SSW value is obtained from calculating the distance of each data to the final cluster center point using Euclidiean Distane. The SSW value obtained from the overall SSW calculation is as follows

SSW1 = 15.3517
SSW2 = 13.3528

After the SSW value is obtained, then it is calculated the value of Sum of Square Between-cluster (SSB). To get an SSB value is to calculate the distance between the cluster center points of each cluster.

SSB1,2 = 30.0671

The next calculation after SSW and the SSB value obtained is to calculating the value of the Ratio. Value Ratio is obtained by calculating the distance between the cluster center points of each cluster.

Table 2. Value of each cluster

R	Data ke – i		R – Max
	1	2	
1	0	0.9547	0.9547
2	0.9547	0	0.9547

Then after the clustering results are obtained, the next step is to calculating the DBI value of each clustering result. Where testing uses variations in the number of clusters of 2.3 4. Then DBI values are calculated from all experiments and in each number of clusters

.

Table 3. The value of Davies-Bouldin Index

X	Avarage range from centroid	Avarage range from centroid on cluster		Davies-Bouldin Index
		C1	C2	
Min 2, Max 100	225.891	15.352	13.353	0.955

Min 3, Max 100	187.973	168.487	216.509	1.345
Min 4, Max 100	168.562	111.478	167.479	1.531

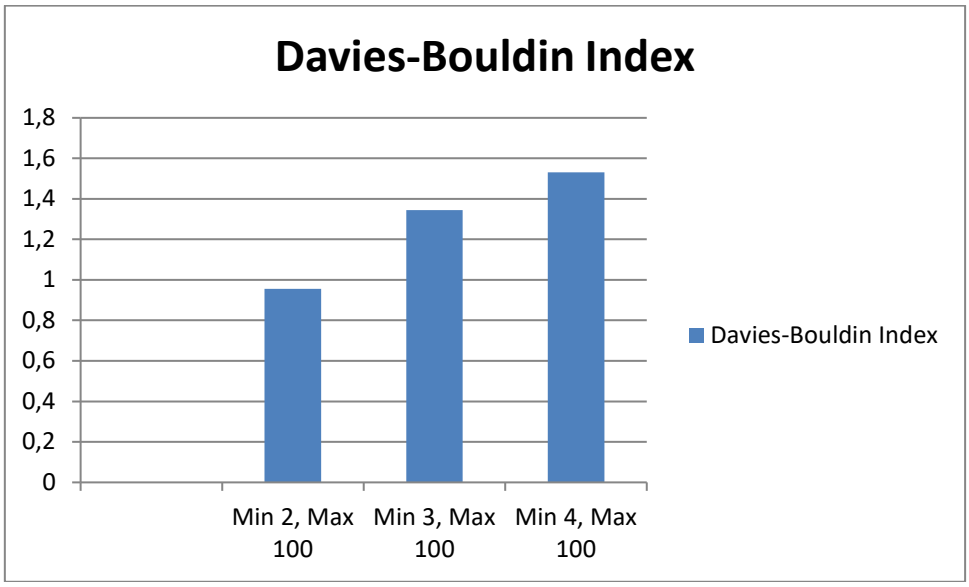

Figure 2. Davies – Bouldin Index Graph

Based on the results of the DBI obtained, the name of 2 clusters with the results of DBI = 0.955, 3 clusters DBI = 1,345 results and 4 clusters DBI = 1.531 results, it can be concluded the evaluation results from clustering in table 4.9 that the best results are in the centroid 2 clusters because the DBI value obtained is close to 0.

5. CONCLUSION

Based on the results of the study using the X-Means algorithm with the Davies-Bouldin Index evaluation determining the number of Centroid clusters is done by modifying the X-Means method to do some centroid determination to get 11 iterations. Fromthe test results produce cluster members that have a good level of data similarity with other data. In determining the number of centroids, use the Davies-Bouldin Index method where testing with 2 clusters has a minimum value with a DBI value close to 0.

Acknowledgments

The author thanks Dr. Syahril Efendi and Prof. Dr. Muhammad Zarlis for the guidance given until this research was completed well.

References

[1] Baswade, A.M. & Nalwade, P.S. 2013. Selection of initial centroids for k-meansalgorithm. International Journal of Computer Science and Mobile Computing(IJCSM) 2(7): 161-164.

[2] MahdiShahbaba, Beheshti Soosan. 2012. Improving X-Means Clustering with MNDL. International Conference on Information Science, Signal Processing and their Applications (ISSPA).

[3] Wani, M. A. & Riyaz, R. 2017. A novel point density based validity index for clustering gene expression datasets. International Journal of Data Mining and

Bioinformatics 17(1): 66–84.
[4] R. Dela Arundina, Shaufiah,Toto Suharto. 2010. Analisis Dan Implementasi Algoritma X-Means Pada Data Pelanggan Telekomunikasi. Universitas Telkom. Bandung.
[5] Krawczyk Bartosz and Micha_l Wo´zniak. 2013. Pruning Ensembles of One Class Classifiers with X-means Clustering. Wroc_law University of Technology.
[6] Serapiao, A.B.S., G.S. Correa, F.B. Goncalves, and V.O. Carvalho. 2016. CombiningK-Means and K-Harmonic With Fish School Search Algorithm for Data Clustering Task on Graphics Processing Units. Applied Soft Computing41: pp.290-304