Enhancement Clustering Evaluation Result of Davies-Bouldin Index with Determining Initial Centroid of K-Means Algorithm

Bernad Jumadi Dehotman Sitompul¹, Opim Salim Sitompul², Poltak Sihoming³

¹, ², ³Department Computer Science and Information of Technology, University of North Sumatera, Medan, Indonesia
¹bernadsitompul@yahoo.com, ²opim@usu.ac.id, ³poltakhombing@usu.ac.id

Abstract. K-Means is one of the most popular clustering algorithms because it is easy and simple when implemented. However, clustering results from K-Means are very sensitive to the selection of initial centroid. Better clustering results are often obtained after several experiments. In this study, Sum of Squared Error (SSE) was used as an approach to determine initial centroid of K-Means algorithm. If the SSE value is smaller then the data in one cluster will be more homogeneous and certainly give a good cluster result. In this study, Sum of Squared Error (SSE) was used as an approach to determine initial centroid of K-Means algorithm. Testing was performed on 3 datasets and the number of clusters 2, 3 and 4. From the test, the average value of Davies-Bouldin Index (DBI) for 3 datasets was 0.2427, while the average value of Davies-Bouldin Index (DBI) for 3 datasets was 0.2805. These results prove that clustering with method of determining initial centroid of K-Means algorithm based on Sum of Squared Error minimum able to improve clustering result and enhance DBI value obtained by simple determine initial centroid of K-Means algorithm.

Keywords—K-Means, Clustering, Centroid, SSE, DBI

1. Introduction

K-Means is one of the clustering algorithms that is included in partitioning clustering. K-Means algorithm partitions the data set into a number of cluster k that have been set up in the beginning. Partition data sets are performed to determine the characteristics of each cluster, so clusters that have similar characteristics are grouped into one cluster and that have different characteristics grouped into other clusters [11]. The advantages of K-Means algorithm is that the required execution time is relatively fast and easy to implement. But the K-Means algorithm also has weaknesses. One of them is the determination centroid of the cluster or the initial centroid randomly selected. Center of cluster or centroid is the starting point of clustering in the clustering method of partitioning clustering. The clustering results of the K-Means algorithm rely heavily on the initial centroid [1]. If initial centroid is given that a centroid is not good then it can be ascertained its clustering results are also not good [4].

Research similar has do by Nainggolan [10] with modification determining centroid algorithm K-Means based Sum of Squared Error. On early research centroid used to process clustering done and then calculated value Sum of Squared Error from results clustering on iteration last. Process the do on over and over again as many as 20 centroid to found a centroid with Sum of Squared Error minimum. While on research this, period center cluster early raised on random and more first calculated the value Sum of Squared Error to as many as 100 centroid. From 100 centroid, are calculated value Sum of Squared Error respectively. Centroid with Sum of Squared Error value is the least used as initial centroid for process clustering algorithm K-Means. For test technique determination centroid with Sum of Squared Error in this research better than research conducted by [10]. For prove that existence difference technique completion in determine centroid based on Sum of Squared Error, then do evaluation clustering with Davies-Bouldin Index (DBI).
Davies-Bouldin Index (DBI) is one methods used for measure validity cluster on something method clustering. Measurement with Davies-Bouldin Index is maximizing inter-cluster distance and on same time try for minimize distance between point in a cluster. If maximum inter-cluster distance, means similarity characteristics between each cluster a bit so difference inter-clusters visible more clear. If minimal intra-cluster distance means each object in the cluster have level similarity the characteristics of the high [16]. Clustering results obtained from proposed determining centroid then in evaluation with DBI method. So that could is known correlation from determining centroid method based on Sum of Squared Error to enhancement quality cluster based on the value of DBI obtained.

2. Related Work

Some researchers have previously conducted research on the determining centroid by using some partitioning clustering method. [9] applies improvisation of the Particle Swarm Optimization (PSO) algorithm that is used as an approach to determine centroid of Possibilistic Fuzzy C-Means (PFCM) algorithm. The results of this study obtained a pretty good result compared to conventional Fuzzy C-Means Possibilistic algorithm. Subsequent research [4] uses improvisation of the Pillar Algorithm algorithm to determine centroid of K-Means algorithm. In the experiments conducted tests conducted on the dataset Iris, Thyroid and Wine. From this research, a good accurate percentage of accuracy dataset accuracy was 89%, Thyroid 80% and 67% in Wine dataset. While the determining of centroid randomly on K-Means algorithm accuracy yielded 66% on Iris dataset, Thyroid dataset 80% and Wine dataset 67%. Then in the study [8], Genetic Algorithm (GA) was used to determine centroid of K-Means algorithm. In that study, the best centroid is determined based on the fitness value of the genetic algorithm to obtain its Mean Square Error value. The research was tested on one dataset of Iris Dataset and from the test obtained the result of Average Mean Square Error better than the determination of random centroid.

3. Proposed Method For Determining Initial Centroid

The step of proposed method for determining initial centroid in this paper are following:
1. Determine and input number of cluster
2. Determine number of centroid will be tested.
3. Choose centroid with randomly.
4. Calculate Sum of Squared Error (SSE) value centroid respectively.

\[ SSE = \sum_{i=1}^{k} d(p_i - m_i)^2 \]  \hspace{1cm} (2.1)

And then get centroid with minimum SSE from all centroid have been calculated. Back to step 2 until number of centroid is tested equal with input number of centroid. Centroid with minimum Sum of Square Error is used as initial centroid in clustering step.
5. Calculate distance between data and centroid with Euclidean distance formula.

\[ D_e = \sqrt{(x_i - s_i)^2 + (y_i - t_i)^2} \] \hspace{1cm} (2.2)

6. Classify data with minimum distance from step 5.
7. Calculate and get new centroid based on mean value from membership of cluster
8. If new centroid value equal input centroid then clustering is stopped, else back to step 5.
4. Experimental Results

4.1 Clustering Evaluation Result

Clustering results have been obtained from the 10 experiments that have been done, then the next step is to calculate the DBI value of each clustering experiment. The result of comparison of Davies Bouldin Index value from both methods in dataset 1 can be seen in the following table 1.

Table 1. Testing and comparison of Davies Bouldin Index values of both methods on the dataset 1

| Experiment | Simple K-Means | K-Means with centroid based min SSE |
|------------|----------------|-----------------------------------|
|            | k=2 | k=3 | k=4 | k=2 | k=3 | k=4 |
| 1          | 0.4206 | 0.6214 | 0.0406 | 0.1366 | 0.0403 | 0.0272 |
| 2          | 0.4206 | 0.6214 | 0.0358 | 0.2392 | 0.0540 | 0.0509 |
| 3          | 0.4206 | 0.6214 | 0.0322 | 0.2392 | 0.0505 | 0.0373 |
| 4          | 0.4206 | 0.6214 | 0.0357 | 0.2392 | 0.0493 | 0.0409 |
| 5          | 0.2845 | 0.6409 | 0.0391 | 0.2392 | 0.0493 | 0.0372 |
| 6          | 0.2845 | 0.6409 | 0.0319 | 0.2392 | 0.0493 | 0.0319 |
| 7          | 0.2845 | 0.6409 | 0.0329 | 0.2392 | 0.0493 | 0.0373 |
| 8          | 0.4206 | 0.6214 | 0.0407 | 0.2392 | 0.0491 | 0.0409 |
| 9          | 0.4206 | 0.6214 | 0.0356 | 0.1366 | 0.0403 | 0.0366 |
| 10         | 0.4206 | 0.6214 | 0.0407 | 0.2392 | 0.0493 | 0.0366 |
| Average    | 0.38607 | 0.0395 | 0.03858 | 0.21668 | 0.04987 | 0.03807 |

In Table 2, the clustering evaluation of the Davies Bouldin Index obtained from conventional K-Means is 0.38607 for the sum of k = 2. While on the proposed K-Means method, the average value of Davies Bouldin Index obtained is 0.21868. Then on the number of clusters k = 3, has an average value of Davies Bouldin Index of 0.05595. While on the proposed K-Means method, the average value of Davies Bouldin Index obtained is 0.0498. And the last experiment with a total of 4 clusters, Conventional K-Means has a ratio between the distance between the cluster and intra-cluster - it amounted to 0.03858. While the K-Means method proposed, the average values obtained Davies Bouldin Index amounted to 0.03807.

Table 2. Testing and comparison of Davies Bouldin Index values of both methods on the dataset 2

| Experiment | Simple K-Means | K-Means with centroid based min SSE |
|------------|----------------|-----------------------------------|
|            | k=2 | k=3 | k=4 | k=2 | k=3 | k=4 |
| 1          | 0.0611 | 0.1216 | 0.0834 | 0.9631 | 0.2620 | 0.0834 |
| 2          | 1.0671 | 0.2620 | 0.0083 | 0.9631 | 0.2620 | 0.0083 |
| 3          | 1.0671 | 0.2620 | 0.0083 | 0.9631 | 0.2620 | 0.0083 |
| 4          | 0.9631 | 0.2620 | 0.0083 | 0.9631 | 0.2620 | 0.0083 |
| 5          | 0.9631 | 0.2620 | 0.0083 | 0.9631 | 0.2620 | 0.0083 |
| 6          | 1.0671 | 0.2620 | 0.0083 | 0.9631 | 0.2620 | 0.0083 |
| 7          | 1.0671 | 0.2620 | 0.0083 | 0.9631 | 0.2620 | 0.0083 |
| 8          | 1.0671 | 0.2620 | 0.0083 | 0.9631 | 0.2620 | 0.0083 |
| 9          | 1.0671 | 0.2620 | 0.0083 | 0.9631 | 0.2620 | 0.0083 |
| 10         | 0.9631 | 0.2620 | 0.0083 | 0.9631 | 0.2620 | 0.0083 |
| Average    | 1.0233 | 0.3352 | 0.1058 | 0.9631 | 0.2620 | 0.0083 |

In table 3, clustering evaluation on the clustering results conventional method determination centroid and the proposed methods determination of centroid has been calculated and obtained the results. From the overall experiment, the average value of the Davies Bouldin Index obtained from conventional K-Means is 1.0255 for the sum of k = 2. While on the proposed K-Means method, the average value of Davies Bouldin Index obtained is 0.9631. Then at the number k = 3, obtained the average value of Davies Bouldin Index of 0.3352. While on the proposed method, the average value of Davies Bouldin Index obtained is 0.2620. And the last experiment with a total of 4 clusters, Conventional K-Means has a ratio between the distance between the cluster and intra-cluster amounted to 0.1058. While the K-Means method proposed, the average values obtained Davies Bouldin Index of 0.0838.
4.2 Comparison of clustering evaluation results from third dataset

Based on table 3, the result of clustering evaluation the method K-Means with determination point center conventional cluster obtained Bouldin index values Davies of 0.3592 with the number of clusters is 2 clusters. While on experiments with K-Means algorithm by method of determination point center The proposed cluster by 2 the cluster has an index of 0.3636. While experimenting on method of determining centroid by conventional with 3 cluster , has a comparison between the distance between clusters and intra-cluster nya amounted to 0.1262. While the clustering evaluation on proposed method determination of centroid, the average value obtained Davies Bouldin Index of 0.1260. Then next for the experiment with the number 4 cluster, obtained results clustering evaluation by method of conventional determination centroid is 0.0833. And clustering evaluation on the clustering result of the proposed K-Means method is 0.0794. The best clustering result improvement is on the number of cluster 2 that is by the difference of Davies Bouldin Index value of 0.0045, while on the number of other clusters there is still improvement of clustering results, but with a smaller value difference.

4.3 Comparison of clustering evaluation result

To know the effectiveness of the proposed method is the determining centroid of the cluster based on Sum of Squared Error (SSE) minimum by the conventional method determining of centroid in the K-Means algorithm to the clustering process , then the comparison of the total average of clustering evaluation results from the three datasets. The average total comparison of the clustering evaluation results of the two methods against the three datasets used, can be seen in Table 4 as follows.

Table 4. Total average value of testing and comparison results Davies Bouldin Index (DBI) from three dataset

| Number | Dataset           | Davies Bouldin Index (DBI) value |
|--------|-------------------|---------------------------------|
|        |                   | Simple-K-Means  | K-Means with centroid based minimum SSE |
| 1      | Birth and Death Rates | 0.1602           | 0.1022             |
| 2      | Wholesale Customers | 0.4888           | 0.4363             |
| 3      | Seeds Dataset      | 0.1925           | 0.1896             |
| **Total Average DBI** |                  | 0.2805           | 0.2427             |
Figure 1. Graph of the total average value of Davies Bouldin Index (DBI) of three datasets

4.4 Comparison with the last research
The clustering comparison results of a similar method used by [10] with the method proposed in this study against dataset 4, can be seen in Figure 2.

Figure 2. Graph clustering result of both research with k = 3 on the dataset 4

After the clustering results are obtained, a clustering evaluation is performed to measure the quality of the intra-cluster with the inter-cluster. Result of clustering evaluation and comparison of average result of clustering evaluation technique is used in the study [10] with the technique proposed in this study, can be seen in Table 5 the following.

Table 5. Testing and comparison of Davies Bouldin Index values from both research on the dataset 4

| Experiment | Davies-Bouldin Index |
|------------|----------------------|
|            | Nainggolan Research (Nainggolan, 2014) | Proposed Research |
| 1          | 1.4016                | 1.3714            |
| 2          | 1.4016                | 1.4016            |
| 3          | 1.4016                | 1.4016            |
| 4          | 1.4016                | 1.3714            |
| 5          | 1.4016                | 1.4016            |
| 6          | 1.4016                | 1.2807            |
| 7          | 1.4016                | 1.3572            |
| 8          | 1.4016                | 1.4016            |
| 9          | 1.4016                | 1.4016            |
| 10         | 1.4016                | 1.3714            |
| Average    | 1.4016                | 1.3715            |

5. Conclusions
Based on the results of clustering evaluation with Davies-Bouldin Index on the three datasets, the method of determining centroid based on the minimum Sum of Squared Error (SSE) value can improve in terms of improving the clustering quality and show better clustering results compared to the method of conventional determining centroid by the difference of Davies-Bouldin Index value of
0.0378 or 3.78%. And then the comparison result value of previous research with current research is 0.0301 or 3.01%. Based on this research, future work are to compare result with used categorical data, image data or determining centroid based on SSE with another clustering and then to compare with another evaluation clustering method.

References

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

[2] Bates, A. & Kalita, J. 2016. Counting clusters in twitter posts. Proceedings of the 2\textsuperscript{nd} International Conference on Information Technology for Competitive Strategies, pp. 85.

[3] Bholowalia, P. & Kumar, A. 2014. EBK-Means: A clustering techniques based on elbow method and k-means in WSN. International Journal of Computer Application 9(105): 17-24.

[4] Bhusare, B.B. & Bansode, S.M. 2014. Centroids initialization for k-means clustering using improved pillar algorithm. International Journal of Advanced Research in Computer Engineering & Technology 3(4): 1317-1322.

[5] Celebi, M.E., Kingravi, H.A. & Vela, P.A. 2013. A comparative study of efficient initialization methods for the k-means clustering algorithm. Expert Systems with Applications 40(1): 200–210.

[6] Gingles, C. & Celebi, M.E. 2014. Histogram based method for effective initialization of the k-means clustering algorithm. Proceedings of 27\textsuperscript{th} International Florida Artificial Intelligence Research Society Conference, pp. 333-337.

[7] Goyal, M. & Kumar, S. 2014. Improving the initial centroids of k-means clustering algorithm to generalize its applicability. Journal of The Institution of Engineers 95(4): 345-350.

[8] Hartono, Ongko, E. & Abdullah, D. 2015. Determining a cluster centroid of k-means clustering using genetic algorithm. International Journal of Computer Science and Software Engineering (IJCSSE) 4(6): 160-164.

[9] He, Y., Zhang, K., dan Sun, Z. 2014. A possibilistic fuzzy c-means clustering algorithm based on improved particle swarm optimization. Journal of Computational Information Systems 10(18): 7845-7857.

[10] Nainggolan, R. 2014. Modified k-means clustering algorithm on determining cluster centre based sum of squared error (SSE). Thesis. University of Sumatera Utara

[11] Nawrin, S., Rahman, M.R. & Akhter, S. 2017. Exploring K-Means with internal validity indexes for data clustering in traffic management system. International Journal of Advanced Computer Science and Applications 8(3): 264-272.

[12] Poteras, C.M., Mihăescu, M.C. & Mocanu, M. 2014. An optimized version of the k-means clustering algorithm. Proceedings of the 2014 Federated Conference on Computer Science and Information Systems, pp. 695–699.

[13] Qi, J., Yu, Y., Wang, L. & Liu, J. 2016. K*-Means: An effective and efficient k-means clustering algorithm. Big Data and Cloud Computing (BDCloud), Social Computing and Networking (SocialCom), Sustainable Computing and Communications (SustainCom), 2016 IEEE International Conferences on IEEE, pp. 242-249.

[14] Rebagliati, N. 2013. Strict monotonicity of sum of squared error and normalized cut in the lattice of clustering. Proceedings of the 30\textsuperscript{th} International Conference on Machine Learning (ICML), pp. 163-171.

[15] Rose, J.D. 2016. An efficient association rule based hierarchical algorithm for text clustering. International Journal of Advanced Engineering Technology 7(4): 751-753.

[16] Thakare, Y.S. & Bagal, S.B. 2015. Performance evaluation of k-means clustering algorithm with various distance metrics. International Journal of Computer Applications 110(11): 12-16.

[17] 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.