Determining The Appropiate Cluster Number Using Elbow Method for K-Means Algorithm

H Humaira[1], R Rasyidah[2]

Technology Information Department, Politeknik Negeri Padang,Jl. Limau Manis Kampus Unand, Padang,Indonesia [1,2]
{mira.humaira@gmail.com[1], rasyidah.faisal@gmail.com[2]}

Abstract. Several algorithms are applied in the literature of clustering such as K-means, Fuzzy C-Means, and Single Linkage, etc. K-means Algorithm is the most commonly used because of its simplicity. How ever, the best number of clusters in K-means possesses a weakness. This study was carried out in determining cluster numbers by using Elbow Method. Meanwhile, Silhouette Technique was applied for the testing methods in order to measure the quality of the clusters. The results obtained were in the form of two clusters.

Keywords: Clustering, K-means, Elbow, Silhouette

1 Introduction

Clustering is an influential process in identifying a group or a cluster in a number of the dataset. Many have carried out the clustering process in various disciplines of knowledge like learning machine, image processing, pattern recognition, data mining, bioinformatics, and decision support systems (DSS)[1], [2]. Most commonly, clustering is applied for pre-processing before processing a number of dataset.

Several clustering algorithms have been utilized in literature, such as Single Linkage and K-means [3]. This study takes a K-means Algorithm for its simplicity and low-cost computing needs. Hence, it is more appropriate to apply for large data collection.

On the K-means Algorithm, there is a partition for a number of cluster going to be inserted. It is an issue on how to determine the best number of cluster. Many methods can be taken in obtaining the best number of the cluster such as By Rule of Thumbs, Elbow Method, Information Criteria Approach, Information Theory Approach, Silhouette, and Cross-Validation [4].

2 Method

This is a preliminary study for Bidik Misi Scholarship SPK. The SPK is going to be finalized using a Fuzzy Inference System (FIS) applying Mamdani Method.

There were many inputs inserting into the system. However, only two of them becoming most influential toward the output, they were parents' income and academic achievement. Setting the distance on each term was carried out by the clustering process.

WMA-2 2018, January 24-25, Padang, Indonesia

DOI 10.4108/eai.24-1-2018.2292388

2.1 Data

The data was taken from registrants of Bidik Misi, State Polytechnic of Padang (PNP) in 2017. The data obtained was 4421 for the income category and 275 for an academic one.

2.2 Elbow Method

Whereas, Elbow Method was used to specify the number of the cluster on a set of data by using the visual technique. The graphic was obtained from Sum Square Error (SSE) calculation. The number of the cluster was determined by looking at the point position on the “elbow” arm. Seen in Fig. 1, Cluster 1 and Cluster 2 had larger value differences and similarly occurred in Cluster 2 and Cluster 3. Whereas on the cluster 3 points and so forth, small value difference was obtained developing plateau shape. This visualization indicated that the number of the best cluster was 3.

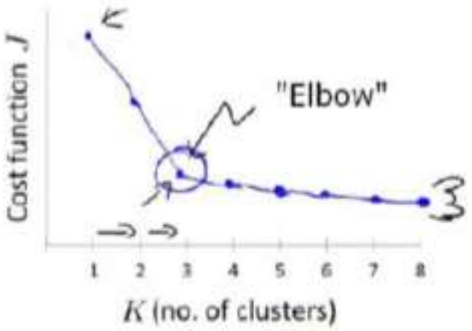


Fig. 1 Elbow Method Visualization[4]

2.3 K-Means Algorithm

K-means is a method included in the clustering partition category[3]. As its strength, K-means Algorithm is easily implemented and its complexity appropriates with a larger number of data. As its weakness, it is considered a greedy algorithm category where the solution is not optimal and the user deciding the number of the cluster beforehand [2].

2.4 Silhouette Evaluation

The silhouette is taken in order to see cluster quality and its strength, how good is one object to be put in a cluster. This is a combination of cohesion and separation methods[5]. Silhouette value ranges between 0 to 1, the smaller the obtained value is, the closer the object on the inappropriate cluster. Dataset on the appropriate cluster should have its value close to 1.

3 Result and Discussion

3.1 Academic Dataset

The academic dataset is students’ average score. The tested data consisted of two groups, the first group was 331 data, and the second one was 275 data. Group I Elbow Graph can be seen in Fig. 3 where out of four attempts was obtained random result. Attempt (1) and (3) made an elbow shape on point 4 while attempt (2) and (4) made an elbow shape on point 3.

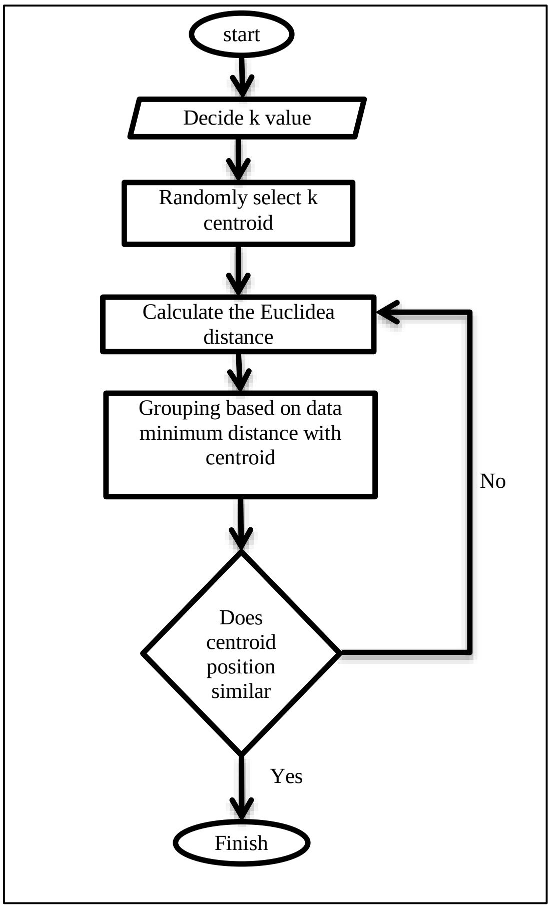
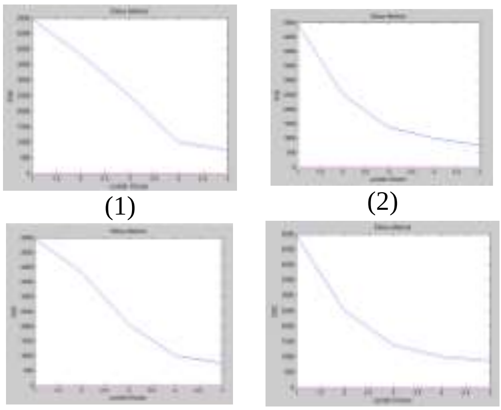

start
Decide k value
Randomly select k centroid
Calculate the Euclidea distance
Grouping based on data minimum distance with centroid
No
Does centroid position similar
Yes
Finish

Fig. 2. K-means Algorithm Flowchart [1], [4]



(1)

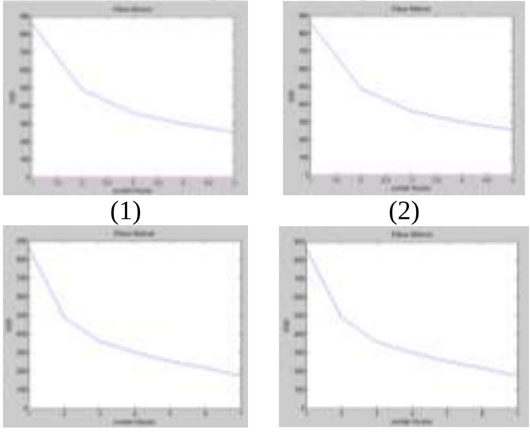
(2)

(3)

(4)

Fig 3. Attempt on Group I

The attempt was also carried out on Group II, with a 275 clean dataset. The 331 datasets had been going through a cleaning process of junk data resulting in 275 datasets. The Elbow Graph can be seen in Fig. 4, where the elbow shape was set on point 2.

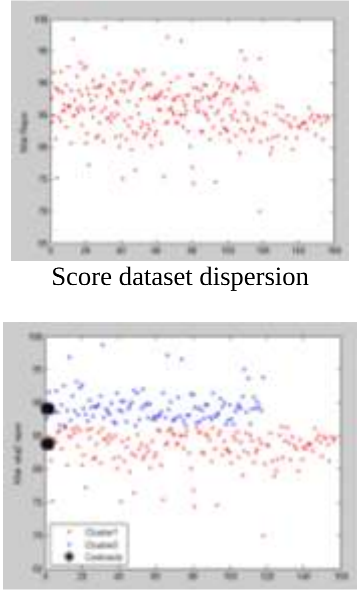


(1) (2)

(3) (4)

Fig. 4. Attempt on Group II

Academic dataset dispersion can be seen in Fig. 5, where the score was spread from 70-100. Whereas Fig. 6 contained data after clustering. The data was set for two clusters based on elbow analysis as seen in Fig. 4.

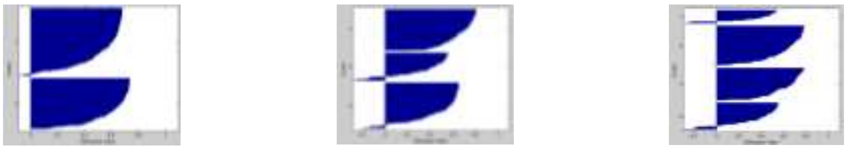


Score dataset dispersion

After clustered = 2

Fig. 5. Comparison of score dataset dispersion before and after cluster

Moreover, the Elbow Method mentioned above should be tested by using the Silhouette Technique, and the result using this technique shown in Fig. 7. Shown in the figure that clusters 2, 3, and even cluster 4 still presented Silhouette value which was smaller than 0. It can be said that the number of the cluster within this academic dataset cannot be determined yet. If it should be chosen, the approach that could be carried out on the observation result of cluster =2 was much better than cluster =3 and =4 results.



Cluster=2 Cluster=3 Cluster=4

Fig. 6. Academic Dataset using Silhouette Technique

3.2 Payroll Dataset

The payroll dataset consisted of 4421 data. Out of four attempts shown on the graph in Fig. 7, literally it is seen that the point forming elbow shape was on point 2. Nevertheless, on attempts 3 and 4 there were still points that occurred before forming the plateau, it was still tolerable as sharply decreased SSE was presented on point 2. Looking at the difference of SSE out of the four conducted attempts, the largest difference was shown in cluster 2.

Table 1. Comparison of SSE Table

cluster	SSE	Difference
K=1	1.9625	
K=2	1.165	0.7975
K=3	0.9075	0.2575
K=4	0.7375	0.17
K=5	0.5975	0.14
K=6	0.475	0.1225
K=7	0.4075	0.0675
K=8	0.39	0.0175

SSE Table on attempt 1

cluster	SSE	Difference
k1	1.9625	
k2	1.165	0.7975
k3	0.9075	0.2575
k4	0.6775	0.23
k5	0.4825	0.195
k6	0.4825	0
k7	0.395	0.0875
k8	0.3875	0.0075

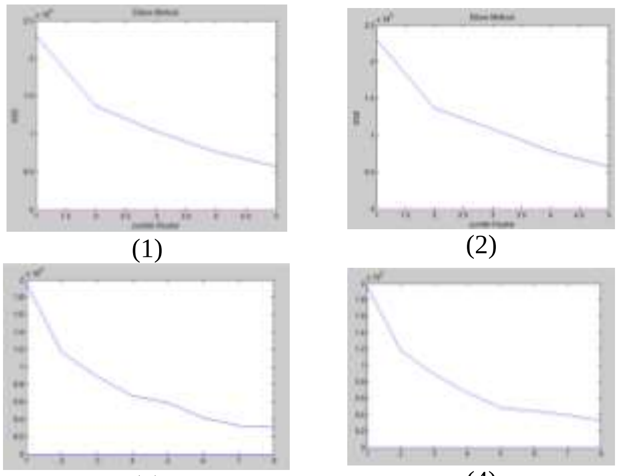
SSE Table on attempt 2

cluster	SSE	difference
k1	1.9625	
k2	1.165	0.7975
k3	0.89	0.275
k4	0.665	0.225
k5	0.5875	0.0775
k6	0.415	0.1725
k7	0.3275	0.0875
k8	0.305	0.0225

SSE Table on attempt 3

cluster	SSE	Difference
k1	1.9625	
k2	1.1825	0.78
k3	0.89	0.2925
k4	0.665	0.225
k5	0.4825	0.1825
k6	0.445	0.0375
k7	0.395	0.05
k8	0.3275	0.0675

SSE on attempt 4



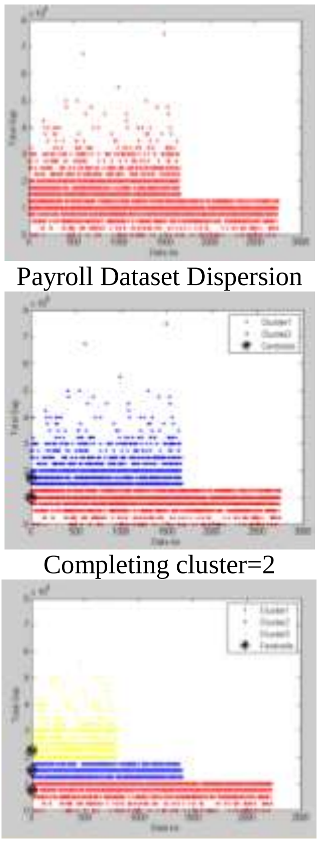
(1)

(2)

(3)

(4)

Fig. 8. Attempts on Payroll Dataset



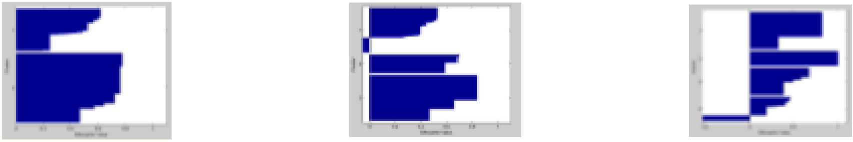
Payroll Dataset Dispersion

Completing cluster=2

Completing cluster=3

Fig. 9. Comparison of payroll dataset distribution before and after being clustered

Elbow Method generated two clusters, as confirmed by Silhouette Technique in Fig. 10. Silhouette Technique had a range of values from 0.2 to 0.8 on this cluster 2. It demonstrated dataset was appropriately being clustered. On the other hand, cluster 3 and 4 was found that some of the dataset were below 0. It indicated that this dataset was not on the appropriate cluster.



Cluster=2

Cluster=3

Cluster=3

Fig. 10. Payroll dataset using Silhouette Technique

4 Conclusion

The data set was clustered by using the K-means Algorithm, where the number of the cluster takes the Elbow Method and then is tested by using the Silhouette Technique. The payroll dataset results in two accurate numbers of clusters of Elbow Method and in accordance with Silhouette Technique testing. The academic dataset results in two accurate numbers of clusters of Elbow Method, however the testing result using Silhouette Technique is in contrast between one to another.

References

[1] N. Putu, E. Merliana, P. Studi, M. Teknik, F. T. Industri, and U. A. Jaya, “Analisa Penentuan Jumlah Cluster Terbaik Pada Metode K-Means,” *Semin. Nasionalmulti Disiplin Ilmu&Call Pap. Unisbank*, pp. 978–979.

[2] I. D. Analysis, “An overview of clustering methods An Overview of Clustering Methods,” no. October 2015, 2007.

[3] Mathwork, “Matlab 2010a,” 2010.

[4] T. M. Kodinariya and P. R. Makwana, “Review on determining number of Cluster in K-Means Clustering,” *Int. J. Adv. Res. Comput. Sci. Manag. Stud.*, vol. 1, no. 6, pp. 2321–7782, 2013.

[5] R. Handoyo, S. M. Nasution, P. Studi, S. Komputer, S. Linkage, and S. Coefficient, “Perbandingan Metode Clustering Mengggunakan metode Single Linkage dan K-Means Pada Pengelompokkan Dokumen,” *JSM STMIK Mikroskil*, vol. 15, no. 2, pp. 73–82, 2014.