Effect of Distance Metrics in Determining K-Value in K-Means Clustering Using Elbow and Silhouette Method

Danny Matthew SAPUTRA[1*], Daniel SAPUTRA[2] and, Liniyanti D. OSWARI[3]

[1]*Faculty of Computer Science, Sriwijaya University, Indonesia*
[2]*Faculty of Agriculture, Sriwijaya University, Indonesia*
[3]*Faculty of Medical, Sriwijaya University, Indonesia*
**Corresponding author: danny@ilkom.unsri.ac.id*

ABSTRACT

Clustering is one of the main task in datamining. It is useful to group and cluster the data. There are a few ways to cluster the data such as partitional-based, hierarchical-based and density based. Partitional-based clustering is a way to cluster data with non-overlapping subsets. One of the most popular partitional-based clustering algorithm is K-means. K-means is an algorithm to cluster data in to K cluster and based their distance to its centroid. Due to the pational, a few factors that must be determined before using K-means is the value of K. Determining the value of K is a big problem because there is no universal way to find the value of K. Two popular ways to determine the value of K is using elbow and silhouette method. This method is graph based. But before using this method another factor is important to determine and that is the metrics distance that will be used. This paper will show the effect of three distance metric Manhattan, Euclidian and Minkowski in finding the value of K using elbow and silhouette method. Based on this study the choice of distance matrix used has little impact in determining the value of K in K-means using elbow and silhouette. Manhattan distance has the most variant in the elbow and silhouette graph. Elbow method is difficult to use and sometimes it is unable to define the value of K in K- means based on its graph.

Keywords: *clustering, Partitional-based clustering, K-means, elbow and silhouette*

Introduction

Clustering is one of the main models in datamining because it is used in a variety of ways to cluster data into meaningful information. It is also used in a variety of fields from marketing, medicine, economics, pattern recognition, and others[1]. Because of this, clustering has benefited many fields and gave them a significant advantage when applied appropriately. But, the main problem with clustering is determining the model to cluster the data. The main models in clustering are partitional-based, hierarchical-based, and density-based.

K-means is one of the most popular method in data clustering. It is a partitional-based clustering that uses a centroid model. K-means strength lies in its simple way to cluster data based on its centroid and distance to each data. But it is also its weaknesses, because the way K-means works by initializing a random point for every centroid it has[2]. The main factor in clustering in K-means is the number of centroids (the value of K) and the initialization point of each centroid.

The number of clusters or the value of K is one of the biggest problems in clustering in general [3]. It is also a problem in hierarchical-based clustering and density-based clustering. There are no universal ways to determine the number of clusters [4]. There are a few methods that work on case-by-case basis. Two simple ways in determining the number clusters are the Elbow method and Silhouette method. These methods will produce graphics that will give general ideas of the number of clusters to be used in clustering.

Elbow and Silhouette method is one method to determine the validity of a clustering method using internal index. Internal index is a way to determine the validity of the clusters without external information. Two of the main data used to determine internal index is cluster cohesion and cluster separation. Cohesion measure how closely the data is related to each other in the same cluster. Cluster separation measure how separated each cluster is with other cluster.

The tool used to measure the cluster cohesion and cluster separation is to measure its distance. The main factor in K-means in creating a good cluster is by choosing its distance. This is also a factor that is important when using Silhouette and Elbow to determine the number of clusters. The default value of distance measurement is Euclidean, but there are also a few other metrics like Manhattan and Minkowski. This research will examine the effects of each distance metrics in determining the value of K in K-means using Elbow and Silhouette Method.

CLUSTERING

Clustering is a process of grouping a data into many groups. There are several ways to categorized clustering. But the most common categorization of clustering is partitional-based clustering, hierarchical-based clustering and density-based clustering. The main difference between partitional-based and hierarchical-based is hierarchical-based will have nested-type cluster or a cluster that is part of another cluster. Another factor when determining which type of cluster algorithm to use is if we want the clusters to be completely separated or not. For example, K-means is the simplest clustering algorithm in a partitional-based clustering where each cluster will be completely separated.

Hierarchical Clustering is data grouping based on hierarchical cluster where clusters have levels that are grouped based on distance or similarity. Hierarchical cluster will group clusters gradually depending on the strategy. If the strategy used is a top-down one, then clusters will starts from the biggest cluster which is the data itself as a whole cluster, and that big cluster will be broken down into smaller clusters. In contrast, bottom-up will start from every data in that dataset and each data will be considered as clusters and every data will be combined with another data that is closest with each other and repeats until it formed into a one whole cluster. One of the algorithm that uses this is agglomerative clustering.

PARTITIONAL-BASED CLUSTERING

Partitional-based clustering or Flat Clustering is a clustering model without hierarchical and its algorithm makes clusters based on centroids and the value of cluster specified (value of K). The main difference between partitional-based clustering and hierarchical is there are no cluster as sub cluster like hierarchical. The first step in partitional-based clustering is by determining the K value. After that, then the partitional-based clustering algorithm will form clusters based on its algorithm. Determining the value of K is one of the biggest problem in partitional-based clustering. There is no universal way to determine the value of K. Partitional-based clustering will create clusters based on the determined K value. One of the partitional-based clustering algorithm that is mostly used in general are K-means algorithm.

K-means Algorithm

K-means algorithm starts by placing centroids as a random point in the data as many as the value of K specified. Then every data distances will be measured to every centroid, then data to the closest centroid will be combined into a cluster. After all of the data combined with their respective clusters, then the centroids will change its value based on the center of all data in their respective clusters. Then every data will be measured by its distance again to newly updated centroids. These steps will keep repeating until the centroid value is too small in difference or it isn't changes at all after updated, or until a certain amount of iterations[5].

DISTANCE METRICS

Another factor in clustering algorithm is determining the distance metrics. A different equation used in measuring distance for each data would resulted in a different value and a different way on how the clusters is form. The most popular distance metrics is Euclidean distance. Two other popular distance metrics are Manhattan Distance and Minkowski Distance. But generally Euclidean and Manhattan could be categorized as varians of Minkowski where its square and square root or *p* in equation 1 is 1 for Manhattan and 2 for Euclidean. Another variant of Minkowski is supremum of chebyshev where the value *p* is nearing the limit of infinity[6].

Euclidean Distance. Mathematically, Euclidean distance is a way to measure a distance between 2 dots in a dimension that gives out results like a Pythagoras Equation. Euclidean distance could be obtain from the square root of the sum of square of differences between two dots and could be written out as shown on equation (1).

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

For equation 1 to 5 n is the number of atributes a point has and x and y are the value of that atributes for point x and y.

Manhattan Distance. Manhattan distance is known as City Block Distance. Manhattan distance is used to measure the distance from a certain data to another data. Manhattan distance reflects the distance between points on an urban road within 1 block. A mathematical equation of Manhattan distance is on equation (2) and is calculated by adding up the absolute results of the reduction between points.

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

Minkowski Distance. Minkowski distance is a metric in normed vector space that could be considered a generalization of a Euclidean distance and Manhattan distance. Minkowski distance in the order of p between 2 points could be defined by equation (3).

$$D(X,Y)=\left(\sum_{i=1}^{n}|x_i-y_i|^p\right)^{1/p} \qquad (3)$$

For p >= 1, Minkowski distance is a metric as a result of Minkowski Inequality. While p < 1, the distance between Point (0,0) and Point (1,1) are $2^{1/p} > 2$, however Point

(0,1) is at the distance 1 from those 2 points. Because this violates triangle inequality, for p <1, it is not a metric. However, a metric can be obtained to these values by removing exponent from 1/p, the resulting metric is also an F-norm. Minkowski distance is usually used with p is 1 or 2, which corresponds to Manhattan and Euclidean distance. In the case of limiting p from an infinite value, we get the chebyshev or supremum distance:

$$\lim_{p\to\infty}\left(\sum_{i=1}^{n}|xi-yi|^{p}\right)^{1/p}=\max_{i=1}|xi-yi|$$

(4)

For p that reaches negative infinite values, we get:

$$\lim_{p\to-\infty}\left(\sum_{i=1}^{n}|xi-yi|^{p}\right)^{1/p}=\min_{i=1}|xi-yi|$$

(5)

ELBOW AND SILHOUTTE METHOD

Internal measure in clustering validation is one way to determine if a cluster is created with the right value from two main value that are needed to determine the internal measure is cluster cohesion and separation. Cluster cohesion measure how closely related data is in its cluster. Separation measures how different or separated one cluster with another cluster. Both cluster cohesion and cluster separation can be measured using sum square error (SSE). Cluster cohesion can be measure using “within cluster sum of square” (WSS) and cluster separation can be measure using “between cluster sum squares”(BSS). Both of this measurement will determine how good a cluster is created without external information.

$$WSS=\sum_{i}\sum_{x\in Ci}\left(x-m_{i}\right)^{2}$$

(6)

$$BSS=\sum_{i}|C_{i}|\left(m-m_{i}\right)^{2}$$

(7)

Elbow Method. Elbow method is a way to measure the cohesion a cluster which is grouped with data that is similar to each other. But this method has one weaknesses. By increasing the value of K (number of cluster) the value of cohesion will eventuality reach a limit near 0. However this will not determine if a cluster is good or not based on this measurement alone. Elbow method suggest that each value of WSS is listed in a graph were the Y-axis label is the value of WSS and X-axis label is the value of K. Elbow method suggest that the best value of K is when the graph has a significant bend[7].

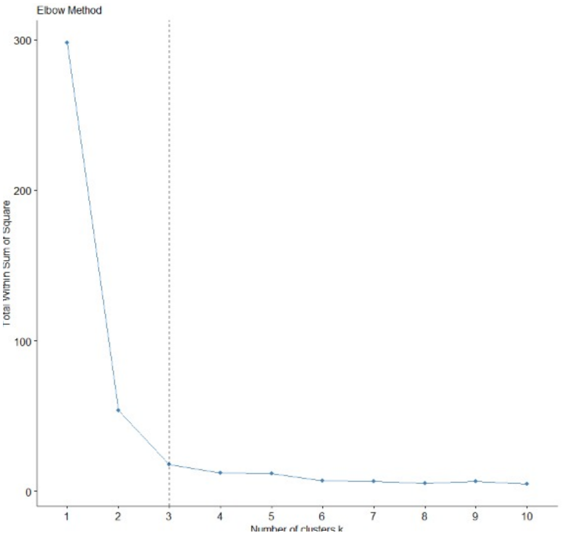

Figure 1. The result of elbow method for iris 2d using R

Silhouette Method. Silhouette Method uses a silhouette coefficient which combines both separation and cohesion. Silhouette coefficient determined by dividing separation measure by cohesion measure and subtracting that value by 1 if separation measure is bigger than cohesion measure or by 1 subtracted by the value of cohesion measure divided by separation measure if the cohesion is bigger the separation. The higher Silhouette coefficient is the better the cluster is.

$$s=1-\left(\frac{cohesion\,measure}{separation\,measure}\right)$$

If cohesion<separation

(8)

$$s=\left(\frac{separation\,measure}{cohesion\,measure}\right)-1$$ If cohesion>separation

Silhouette Method suggest creating a graph where the Y-axis is Silhouette coefficient and X-axis is the value of K, and chooses the K with the highest value of silhouette coefficient.

Determining the value of K in K-means

Data source. The data that is used comes from a variety of sources from a variety of fields such as agriculture, medical, marketing and economic. These data has a variety properties such as the number of tuples and the number of attributeses. Overall, the number of dataset used is 37. The data range from small data to big data sets and from small dimension data to large dimension data.

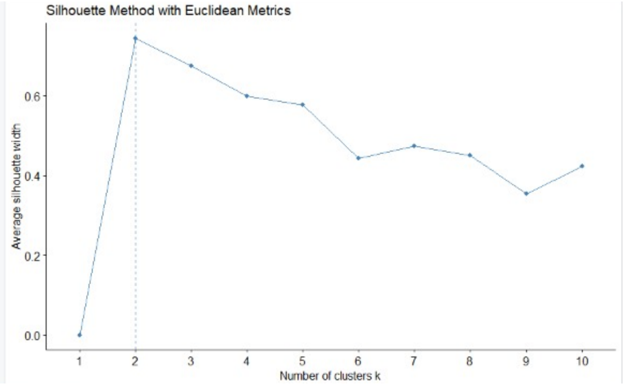

Figure 2. The result of Silhouette method for iris 2d using R

R programing. R is a programing language for statistical and mathematical problem solving. It is very powerful and easy to use and it is perfect to create elbow and Silhouette graph. In Using R, the first step is to load the data within the command line. The next step is to determine the option in creating the graph. This include determining which algorithm to use, which graph is to be created and which type of distance metrics is to be used.

In collecting data, the first step is to load a data from the data source into R and choosing K-means algorithm. K-means algorithm is chosen because it is the most common algorithm in clustering that is usually used as a benchmark to compare other clustering algorithm. The next variable that is chosen is the graph that will be outputted. The graph that will outputted is elbow method graph and the silhouette method graph. The last variable that is chosen is the distance metrics that will be used. This distance metrics that will be used are Manhattan Distance, Euclidean distance and Minkowski Distance/supremum distance.

Graph analyst. R programing is used to create elbow and silhouette method graph for each data with variance in measurement metrics, the total graph that was outputted numbered 222 graph. Each graph is analyzed to determine the value of K and the value is recorded in Table 1.

It could be seen from Table 1 that there is only one graph for elbow method where the value of K is different. There are also 2 data graph that each graph has a different value of K using silhouette method. For all the difference in the value of K, every one of them comes when comparing Manhattan distance graph to Euclidean distance and Minkowski graph. Manhattan distance graph also has the most different graph compared to the other distance graph. There are 29 graph in elbow method where the Manhattan graph is different from elbow graph that uses Euclidean and Minkowski. And there are 31 Graph in silhouette where the Manhattan graph differs from silhouette graph that uses Euclidean and Minkowski.

Another result after graph analysis is the difficulty in determining the value of K in elbow graph. Silhouette graph will automatically pick the value of K with the highest Silhouette coefficient. Elbow graph has the difficulty in determining where the bend is to determine the value of K. Out of 37 data, 15 data has elbow graph that is hard to determine the value of K and in those 15 data, 3 data has elbow graph where the value of K is undetermined. Silhouette method does not has this problem unlike elbow method.[8]

Based on the elbow graph created the value of K in the graph each K has a value of WSS for elbow graph that gets lower for each increment of K and this says that the larger the value of K, WSS will get lower. But this only shows that the cluster cohesion is getting better and if number of cluster is large. That is why elbow graph suggest we use the bend to determine the best value of K.

Silhouette graph show the highest value of Silhouette coefficient that is based on the clusters cohesion and separation. Using this value shows the internal measure of the cluster and based on this the graph shows that the value of K with the highest Silhouette coefficient is the best K to choose in determining the value of K in K-means. But using internal measure alone is the start in determining the value of K other factors should also be analyzed such as external value. The next highest value of Silhouette coefficient should also be check to see if it is a better value. One example is the data “iris 2d” that Elbow and Silhouette determine the best value of K is 2, but looking at the external data it is classified into 3 classes.

Table 1. The value of K based on elbow and silhouette with different distance metric

		K-value (Elbow)				K-value (Silhouette)			
Data Id	Number of atributes	Euclidean	Manhattan	Minkowski	diffrence	Euclidean	Manhattan	Minkowski	Diffrence
1	9	3	3	3	0	2	2	2	0
2	5	2	2	2	0	2	2	2	0
3	2	2	2	2	0	2	2	2	0
4	6	3	3	3	0	3	3	3	0
5	172	2	2	2	0	2	2	2	0
6	8	3	3	3	0	2	2	2	0
7	6	2	2	2	0	2	2	2	0
8	4	3	3	3	0	3	3	3	0
9	12	2	2	2	0	2	2	2	0
10	6	3	3	3	0	3	2	3	0
12	3	5	5	5	0	5	5	5	0
14	7	0	0	0	0	6	6	6	0
15	6	2	2	2	0	2	2	2	0
16	2	3	3	3	0	4	4	4	0
17	7	2	2	2	0	3	3	3	0
18	2	5	5	5	0	2	2	2	0
19	11	2	2	2	0	2	2	2	0
20	3	5	5	5	0	4	4	4	0
21	3	0	0	0	0	8	8	8	0
22	9	0	0	0	0	10	2	10	1
23	3	4	4	4	0	5	5	5	0
24	8	5	5	5	0	2	2	2	0
25	2	0	0	0	0	2	2	2	0
26	3	5	5	5	0	2	2	2	0
27	6	7	2	7	1	10	10	10	0
28	9	2	2	2	0	2	2	2	0
29	8	0	0	0	0	3	3	3	0
30	6	3	3	3	0	2	2	2	0
31	4	3	3	3	0	2	2	2	0
32	3	3	3	3	0	10	10	10	0
33	6	2	2	2	0	2	2	2	0
34	4	2	2	2	0	4	2	4	1
35	20	4	4	4	0	2	2	2	0
36	10	2	2	2	0	2	2	2	0
37	29	2	2	2	0	2	2	2	0
				Total	1			Total	2

CONCLUSION

Based on table 1 out of all the 37 data, 1 data has a different value of K for elbow method and 2 in silhouette method. Of all 3 that are different, all came from graph using Manhattan distance metrics. In all the graph that were analyzed Manhattan graph has the most difference compared to graph that used Euclidean and Minkowski. Elbow method has its difficulty in finding the bend to acquire the value of K. In some data it is unable to determine the value of K using elbow method.[9]

This paper shows that Elbow and Silhouette method can produce a graph that can determine the value of K and the matrix distance choices has little impact in creating variance in determining the value of K. Elbow and Silhouette method is just the first step in determining the value of K other steps such as analyzing external measure should be done to determine in the cluster is good or not. The second best and the third best value in Elbow and Silhouette method could be used as comparison to best value of K Elbow and Silhouette to see if the cluster is good and with this information it will reduce the work of analyzing each possibilities of K in K-means

ACKNOWLEDGMENT

This research was financially supported by 2019 SATEK grant of University Sriwijaya

REFERENCES

[1] C. Zhang and S. Xia, "K-means clustering algorithm with improved initial center," in *Proceedings - 2009 2nd International Workshop on Knowledge Discovery and Data Mining, WKKD 2009*, 2009, vol. 1, no. 2, pp. 790–792.

[2] S. Ray and R. H. Turi, "Determination of number of clusters in k-means clustering and application in colour image segmentation," *Proc. 4th Int. Conf. Adv. pattern Recognit. Digit. Tech.*, pp. 137–143, 1999.

[3] J. Shen, S. I. Chang, E. S. Lee, Y. Deng, and S. J. Brown, "Determination of cluster number in clustering microarray data," *Appl. Math. Comput.*, vol. 169, no. 2, pp. 1172–1185, 2005.

[4] C. A. Sugar and G. M. James, "Finding the Number of Clusters in a Dataset: An Information-Theoretic Approach," *J. Am. Stat. Assoc.*, vol. 98, no. 463, pp. 750–763, 2003.

[5] H. Singh and K. Kaur, "New Method for Finding Initial Cluster Centroids in K-means Algorithm," *Int. J. Comput. Appl.*, vol. 74, no. 6, pp. 27–30, 2013.

[6] A. Singh, A. Yadav, and A. Rana, "K-means with Three different Distance Metrics," *Int. J. Comput. Appl.*, vol. 67, no. 10, pp. 13–17, Apr. 2013.

[7] M. A. Syakur, B. K. Khotimah, E. M. S. Rochman, and B. D. Satoto, "Integration K-Means Clustering Method and Elbow Method for Identification of the Best Customer Profile Cluster," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 336, no. 1, 2018.

[8] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 2nd ed. San Francisco: Morgan Kaufmann Publishers, 2012.

[9] S. Sumathi and S. N. Sivanandam, *Introduction to Data Mining and its Applications*, Kacprzyk,J., vol. 29. Berlin Heidelberg New York: Springer, 2006.