K-means algorithm with improved initialization for clustering fruit plants

F D T Amijaya*, I Purnamasari, Wasono

1Department of Mathematics, Faculty of Mathematics and Natural Sciences, Mulawarman University, Indonesia

*Corresponding Author: fidiadta@fmipa.unmul.ac.id

Abstract. Clustering is one of data mining technique that can make a set of objects in such way that objects in the same group are more similar in some particular manner to each other than to those in other groups. K-means algorithm is one of clustering methods. Standard K-means algorithm has fast and simple computation, but it has the pitfall of randomly choosing initial the center of cluster. In this paper, we propose a mean method combined with interval index based on number cluster to choose initial the center of cluster. It can eliminate the randomness of the selection of initial the center of cluster, so it can find the optimum the center of cluster faster. The effectiveness of algorithm can be seen by maximum iteration of each algorithm. And fruit plants data will be used as data test.

1. Introduction
K-means algorithm is one of the clustering methods. This algorithm is grouping a set of datasets in such a way that items in the same cluster are more identical to each other than to those in others clusters. K-means algorithm is a partitioning algorithm which takes as input a positive integer number of grouping K and a data sets to divide in to K non-empty, non-overlapping and non-subordinated clusters [1]. K-means algorithm has fast and simple computation, but it has the pitfall of randomly choosing initial the center of cluster [2].

Previous research about improving clustering methods has been documented in the literature such as, Fabregas et. al. [1] had proposed improved initialization of K-means using weighted average for monitoring of fixed asset and it is highly applicable to accounting application, Kumar et. al. [2] had proposed density-based initialization method for K-means clustering algorithm and it has better performance over the standard K-means algorithm, Masaki et. al. [3] had proposed K-means algorithm using improved firefly algorithm and it’s more efficient than the standard K-means algorithm.

In this paper, we proposed an improved K-means algorithm called K-means algorithm with improved initialization. In our proposed algorithm, we substituted the randomness of the selection of initial the center of cluster by a mean method combined with interval index based on number cluster to choose initial the center of cluster. After that, we will used it to cluster fruit plants data and we compare the standard K-means and K-means algorithm with improved initialization by its maximum iteration to see the effectiveness of algorithm.

2. Background
Data clustering is an unsupervised learning process used to partition n objects into K groups in such a way that objects within the groups are similar to each other than those in other groups [2]. The goal of this clustering method is to find the center of cluster and each object is assigned to the closest cluster. Clustering is an optimization problem because we need to find the minimum sum of Euclidean distances between the center of cluster and the objects [3].

K-means algorithm is one of famous clustering method. The standard K-means have following steps [3]:

1) Determine the value of $K$ which is used as the number of clusters.
2) For each cluster, initialize the center of cluster randomly.
3) Calculate Euclidean distance between each objects and the center of cluster which is denoted by $d(x_i, y_j)$ using Equation (1),

$$d(x_i, y_j) = \sqrt{\sum_{a=1}^{dim} (x_{ia} - y_{ja})^2}$$

where $x_i$ is the $i$-th object ($i=1, 2, \ldots, \text{number of objects}$), $y_j$ is the center of $j$-th cluster ($j=1, 2, \ldots, K$), and dim is number of attribute object.
4) Put each object to the cluster based on minimum Euclidean distances.
5) For $t=1$, update the center of $j$-th cluster which is denoted by $y_j(t+1)$ using Equation (2),

$$y_j(t + 1) = \frac{1}{N_j} \sum_{b \in S_j} x_b$$

where $t$ is number of iterations, $N_j$ is number of objects in cluster $S_j$, $S_j$ is index of the center of $j$-th cluster, $b \in S_j$ is set of object index which include in the center of $S_j$-th cluster, $x_b$ is the $b$-th object.
6) Repeat step 3 – 5, until the position of each object no longer changes in the cluster.

The K-means algorithm requires random selection of initial cluster. As we discussed earlier, an arbitrary choice of initial the center of cluster leads to non-repeatable clustering result that may be difficult to comprehend. The result of partitional clustering algorithm are better when the initial partitions are close to the final solution [2].

3. K-means Algorithm with Improved Initialization for Clustering Fruit Plants

In this section, we first discuss about the K-means algorithm with improved initialization, followed by the application of K-means algorithm with improved initialization for clustering fruit plants, and the comparison between the standard K-means algorithm and K-means algorithm with improved initialization in term of the speed of finding optimal solution, which can be seen from maximum iteration.

3.1. K-means Algorithm with Improved Initialization

The standard K-means use randomly initial the center of cluster that may be difficult to comprehend. That is the basic idea of K-means algorithm with improved initialization. In our proposed algorithm, we substituted the randomness of the selection of initial the center of cluster by a mean method combined with interval index based on number of clusters to choose initial the center of cluster. It is expected to make the initial the center of cluster close to the optimal solution. The procedure of K-means algorithm with improved initialization can be written in following steps:

1) Determine the value of $K$ which is used as the number of clusters.
2) For each cluster, initialize the center of cluster by following steps:
   a) Calculate mean of each object which is denoted by $\bar{x}_i$ using Equation (3),

\[ \bar{x}_i = \frac{1}{N_i} \sum_{x_{i}} \text{for each } i \]
where \( \text{dim} \) is number of attribute object and \( x_{ic} \) is the \( i \)-th object \( c \)-th attribute \((i=1,2,...,\text{number of objects}, c=1,2,...,\text{number of attributes})\).

b) Sort object based on the smallest mean

c) Calculate interval index based on number of clusters which is called by \( \text{interval} \) using Equation (4),

\[
\text{interval} = \frac{n}{K-1}
\]

where \( n \) is number of objects and \( K \) is number of clusters.

d) Calculate the center of cluster index based on sorted object which is called by \( \text{indeks}_h \) using Equation (5),

\[
\text{indeks}_h = \text{ceil}(\text{indeks}_{h-1} + \text{interval})
\]

where \( h=2,3,...,K-1 \).

3) Calculate Euclidean distance between each objects and the center of cluster which is denoted by \( d(x_i,y_j) \) using Equation (1),

4) Put each object to the cluster based on minimum Euclidean distances.

5) For \( t=1 \), update the center of \( j \)-th cluster which is denoted by \( y_j(t+1) \) using Equation (2),

6) Repeat step 3–5, until the position of each object no longer changes in the cluster.

3.2. The Application of K-means Algorithm with Improved Initialization for Clustering Fruit Plants

The fruit plants data had been taken from [4]. It contains 7 sub-districts as object and 16 fruit plants as dimension. The fruit plants data will be clustered by 2, 3, 4, 5, 6 cluster using K-means algorithm with improved initialization. The result is shown by Table 1 to Table 5.

**Table 1.** Clustering fruit plants data using K-means algorithm with improved initialization with \( K=2 \)

| Cluster | Sub-district                                    | Maximum Iteration | Total Euclidean Distance |
|---------|------------------------------------------------|-------------------|--------------------------|
| Cluster 1 | Palaran, Sambutan, Loa Janan Ilir, Sungai Kunjang, Samarinda Ulu, Sungai Pinang | 2                  | 17196.53                |
| Cluster 2 | Samarinda Utara                                |                   |                          |

**Table 2.** Clustering fruit plants data using K-means algorithm with improved initialization with \( K=3 \)

| Cluster | Sub-district                                    | Maximum Iteration | Total Euclidean Distance |
|---------|------------------------------------------------|-------------------|--------------------------|
| Cluster 1 | Sambutan, Loa Janan Ilir, Sungai Kunjang, Samarinda Ulu, Sungai Pinang | 2                  | 3436.93                  |
| Cluster 2 | Palaran                                        |                   |                          |
| Cluster 3 | Samarinda Utara                                |                   |                          |
Table 3. Clustering fruit plants data using K-means algorithm with improved initialization with $K = 4$

| Cluster | Sub-district                   |
|---------|--------------------------------|
| Cluster 1 | Sambutan, Sungai Pinang       |
| Cluster 2 | Sunga Kunjang, Samarinda Ulu  |
| Cluster 3 | Palaran                       |
| Cluster 4 | Samarinda Utara               |
| Maximum iteration | 3                            |
| Total Euclidean Distance | 2310.94                      |

Table 4. Clustering fruit plants data using K-means algorithm with improved initialization with $K = 5$

| Cluster | Sub-district                   |
|---------|--------------------------------|
| Cluster 1 | Sungai Kunjang, Samarinda Ulu  |
| Cluster 2 | Loa Jalan Ilir, Sungai Pinang |
| Cluster 3 | Sambutan                      |
| Cluster 4 | Palaran                       |
| Cluster 5 | Samarinda Utara               |
| Maximum iteration | 3                            |
| Total Euclidean Distance | 1369.05                      |

Table 5. Clustering fruit plants data using K-means algorithm with improved initialization with $K = 6$

| Cluster | Sub-district                   |
|---------|--------------------------------|
| Cluster 1 | Loa Jalan Ilir, Sungai Pinang |
| Cluster 2 | Sambutan                      |
| Cluster 3 | Samarinda Ulu                 |
| Cluster 4 | Sungai Kunjang               |
| Cluster 5 | Palaran                       |
| Cluster 6 | Samarinda Utara               |
| Maximum iteration | 3                            |
| Total Euclidean Distance | 454.99                       |

From Table 1 to Table 5 we can see that K-means algorithm with improved initialization can be used for clustering fruit plants. The maximum iteration increases because the more the number of clusters, the more computational calculations needed. And Euclidean distance decreases when the number of clusters increases because the more clusters, it will represent many objects in the test data.

3.3. Simulation Results
Simulation results will show the comparison between the standard K-means and K-means algorithm with improved initialization in terms of maximum iteration to cluster fruit plants data. Number of clusters that will be used for simulation are 2, 3, 4, 5, 6 cluster. The result is can be seen in Table 6.
Table 6. The Comparison between the standard K-means and K-means algorithm with improved initialization for clustering fruit plants

| Cluster | Algorithm | Maximum Iteration | Total Euclidean Distance |
|---------|-----------|-------------------|--------------------------|
| 2       | SKA       | 2                 | 17196.53                 |
|         | SKII      | 2                 | 17196.53                 |
| 3       | SKA       | 2                 | 3436.93                  |
|         | SKII      | 2                 | 3436.93                  |
|         | SKA       | 4                 | 2310.94                  |
|         | SKII      | 3                 | 2310.94                  |

| Cluster | Algorithm | Maximum Iteration | Total Euclidean Distance |
|---------|-----------|-------------------|--------------------------|
| 5       | SKA       | 5                 | 1369.05                  |
|         | SKII      | 3                 | 1369.05                  |
| 6       | SKA       | 4                 | 914.06                   |
|         | SKII      | 3                 | 454.99                   |

SKA = Standard K-means algorithm  
KAI = K-means algorithm with improved initialization

From Table 6, the bold and underlined numbers show that the maximum iteration of K-means algorithm with improved initialization more efficient than standard K-means algorithm for clustering fruit plants data in 4, 5, 6 cluster. Because it has smaller maximum iteration to cluster fruit plants data. With 6 cluster, K-means algorithm with improved initialization has smaller total Euclidean distance than standard K-means. It means, K-means algorithm with improved initialization is better than standard K-means algorithm for clustering fruit plants data with 6 cluster. It all may be happens because the initial the center of cluster of K-means algorithm with improved initialization are close to the final solution.

4. Conclusion
In this study, we proposed a new clustering algorithm called K-means algorithm with improved initialization. In our proposed algorithm, we substituted the randomness of the selection of initial the center of cluster by a mean method combined with interval index based on number cluster to choose initial the center of cluster. From the experiment and simulation results, our proposed algorithm can be used for clustering fruit plants and it has more efficient in terms of maximum iteration than standard K-means. In future works, we try to improve our proposed algorithm in terms of updating the center of cluster and apply more complex problems.

Acknowledgements
The author would like to thank the Faculty of Mathematics and Natural Sciences, Mulawarman University, Indonesia for the support and assistance. This research was funded by BOPTN FMIPA 2018.
References

[1] Fabregas A C and Medina R P 2017 Application of Improved Initialization of K-means Algorithm for Monitoring of Fixed Asset International Journal of Learning and Teaching 3, No. 4 December 2017 (Turkey: Ankara University) pp 336 – 41

[2] Kumar A and Kumar S 2017 Density Based Initialization Method for K-means Clustering Algorithm I.J. Intelligent Systems and Applications (Hongkong: Modern Education and Computer Science Press) pp 40 – 8

[3] Masaki T, Ott T, Haruna M, Yoko U, and Yoshifumi N 2017 K-means Algorithm using Improved Firefly Algorithm International Workshop on Nonlinear Circuits, Communications and Signal Processing (Japan: Research Institute of Signal Processing) pp 225 – 8

[4] Statistics Indonesia 2017 Samarinda Municipality in Figures 2017 Statistics of Samarinda Municipality (Indonesia: Statistics Indonesia) pp 124