Improved the Performance of the K-Means Cluster Using the Sum of Squared Error (SSE) optimized by using the Elbow Method

Rena Nainggolan¹, Resianta Perangin-angin², Emma Simarmata³, Feriani Astuti Tarigan⁴
¹,²,³Department of Komputerisasi Akuntansi, Universitas Methodist Indonesia, Indonesia
⁴Department of System Informasi, STMIK TIME, Indonesia

Abstract. K-Means is a simple clustering algorithm that has the ability to throw large amounts of data, partition datasets into several clusters k. The algorithm is quite easy to implement and run, relatively fast and efficient. Another division of K-Means still has several weaknesses, namely in determining the number of clusters, determining the cluster center. The results of the cluster formed from the K-means method is very dependent on the initiation of the initial cluster center value provided. This causes the results of the cluster to be a solution that is locally optimal. This research was conducted to overcome the weaknesses in the K-Means algorithm, namely: improvements to the K-Means algorithm produce better clusters, namely the application of Sum Of Squared Error (SSE) to help K-Means Clustering in determining the optimum number of clusters. From this modification process, it is expected that the cluster center obtained will produce clusters, where the cluster members have a high level of similarity. Improving the performance of the K-Means cluster will be applied to determining the number of clusters using the elbow method.

1. Introduction

Clustering is included in an unsupervised classification. Understanding Clustering is the process of grouping or classifying objects based on information obtained from data that explains the relationship between objects with the principle of maximizing the similarity between members of one class and minimizing similarities between classes / clusters.[1]

2. K-Means Clustering Algorithm

K-means is one form the simplest grouping. The procedure simple and easy to classify data given through a number of clusters. Determination centroid is done by taking data first as the first centroid, second data as second centroid, and so on to the number of centroids required. The next step is to calculate the distance from the point to be clustered to each centroid is available and grouped accordingly with the closest distance to the centroid.[2]

The K-Means algorithm is an algorithm clustering which groups data based on cluster center point (centroid) closest to data. The purpose of K-Means is grouping data with maximize data similarity in one cluster and minimize data similarity between cluster. Similarity measures used in the cluster is the distance function. So that maximizing the similarity of data obtained based on the shortest distance between data towards the centroid point.[3]
The basic concept of the K-Means algorithm is the iterative search of the Cluster center (centroid points). The Cluster Center is set based on the distance of each data to the Cluster center. The Clustering process starts by identifying the data that will be Clustered, Xij (i = 1, ... j = 1, ... m) with n is the amount of data that will be Clustered and m is the number of variables.

At the beginning of the iteration, the center of each Cluster is set freely (arbitrary), ckj (k = 1, .... k, j = 1, .... m). Then the distance between each data is calculated with each Cluster center. the first in the center of the Cluster k (ck), given the name (dik) can be used Euclidean formula, namely:

\[ d_{ik} = \sqrt{\sum_{j=1}^{m} (x_{ij} - c_{kj})^2} \]

A data will be a member of the J-Cluster if the data distance to the center of the J-Cluster is the smallest compared to the distance to the other Cluster center. Next, group the data that is a member of each Cluster. The new Cluster center value can be calculated by finding the average value of the data that is a member of the Cluster. The one steps to Cluster with the K-Means method is as follows: Select the number of Clusters k. Initialization of the center of this Cluster can be done in various ways. But the most often done is by random. Cluster centers are given an initial value with random numbers. Allocate all data / objects to the nearest Cluster. The proximity of two objects is determined based on the distance of the two objects. Likewise, the proximity of a data to a particular Cluster is determined by the distance between the data and the Cluster center. In this stage it is necessary to calculate the distance of each data to each Cluster center. The most distance between one data and one particular cluster will determine which data is included in which cluster. To count the distance of all data to each cluster center, the Cluster can use Euclidean distance theory which is formulated as follows:

\[ D(i,j) = \sqrt{(X_{1i} - X_{1j})^2 + (X_{2i} - X_{2j})^2 + \ldots \ldots + (X_{ki} - X_{kj})^2} \]

Where:
- \( D (i, j) \) = Data distance to i to Cluster center j
- \( x_{ki} \) = Data to i on data attribute to k
- \( X_{kj} \) = Center point to j in attribute to k

Recalculate the Cluster center with the current Cluster membership. The Cluster Center is the average of all data / objects in a particular Cluster. If you wish you can also use the median of the Cluster. So the mean (mean) is not the only measure that can be used. Assign again each object using the new Cluster center. If the Cluster center does not change again then the Clustering process is complete. Or, return to step number 3 until the Cluster center doesn't change again.

The K-Means algorithm is the best algorithm in partitional Clustering algorithm and is most often used among other Clustering algorithms, because of its simplicity and efficiency

3. **Elbow Method**

Elbow method is a method used to produce information in determining the best number of clusters by looking at the percentage of the comparison between the number of clusters that will form an elbow at a point.
Elbow method is a method used to produce information in determining the best number of clusters by looking at the percentage of the comparison between the number of clusters that will form an elbow at a point. This method provides ideas / ideas by selecting cluster values and then adding the value of the cluster to be used as a data model in determining the best cluster. And besides that the percentage of the resulting calculation is a comparison between the number of clusters added. [3] Different percentage results from each cluster value can be shown using the graph as the source of the information. If the value of the first cluster with the value of the second cluster gives the angle in the graph or the value has the biggest decrease then the value of the cluster is the best. To get a comparison is to calculate SSE (Sum of Square Error) from each cluster value. Because the greater the number of cluster K, the SSE value will be smaller.

Following are the stages of the Elbow method algorithm in determining the k value in K-Means

1. Initialize the initial value of k;
2. Increase the value of k;
3. Calculating the sum of square error results from each value of k;
4. Analysis of the sum of square error results from the k value which has drastically decreased;
5. Locate and set the elbow-shaped k value.

In the Elbow method the best cluster value will be taken from the Sum of Square Error (SSE) value which has a significant and elbow-shaped decrease. To calculate SSE using a formula

4. Sum Squared Error (SSE)

SSE (Sum Square Error) is one of the statistical methods used to measure the total difference from the actual value of the value achieved [4]

\[
SSE = \sum_{i=1}^{n} (d)^2
\]

Where, d is the distance between the data and the Cluster center.

Sum of Square Error (SSE) is a formula used to measure the difference between the data obtained by the prediction model that has been done previously. SSE is often used as a research reference in determining optimal clusters.

5. Experiment testing and comparative analysis

The following are 20 data for patient data to be tested for the Modified K-Mean Clustering algorithm in determining the cluster center based on Sum of Squared Error (SSE)

| No | Nama    | Wilayah       | Pekerjaan | Umur |
|----|---------|---------------|-----------|------|
| 0  | Daniel  | Medan Perjuangan | TKI       | 33   |
| 1  | Hendrik | Medan Selayang | Peg. Swasta | 25   |
| 2  | Rinaldi | Belawan       | Pedagang  | 44   |
| 3  | Marusaha| Medan Perjuangan | TKI     | 31   |
| 4  | Ana     | Medan Perjuangan | TKI     | 37   |
| 5  | Linda   | Medan Selayang | Peg. Swasta | 40   |
| 6  | Keana   | Medan Kota    | Peg. Swasta | 24   |
| 7  | Abadi   | Belawan       | PNS       | 42   |
| 8  | Zefri   | Medan Kota    | PNS       | 38   |
| 9  | Sasan   | Medan Kota    | Therapys  | 28   |
| 10 | Ilham   | Medan Selayang | TKI       | 27   |
| 11 | Titin   | Medan Selayang | TKI       | 22   |
| 12 | Roma    | Medan Perjuangan | TKI     | 27   |
| 13 | Pdia    | Medan Perjuangan | Therapys | 30   |
| 14 | Arki    | Medan Kota    | Peg. Swasta | 44   |
5.1 Pre-processing

After all patient data is transformed into numbers, the data can be grouped using the K-Mean Clustering method. To be able to group these data into several clusters, it is necessary to take several steps, namely:

1. First determine the number of clusters desired. In this study the existing data will be grouped into three clusters.
2. Determine the starting center point of each cluster. In this study the initial center point was generated randomly.

5.2 Data Transformation

So that the data above can be processed using the K-Means Clustering method, the data that has the type of nominal data such as region and work must be initialized in the form of numbers. To initialize the area is done by sorting from the largest based on the frequency of patients coming from the region. After that the area that has the biggest frequency is given initials with number 1 and the area that has the second largest frequency is given initials with number 2.

5.3 Processing

Step 1. Determine the parameters of MK-Means Clustering (number of iterations)

The number of iterations to be tested is 100 cluster centers. The number of iterations (i) to be tested are each of the 5 cluster centers in each iteration, which is as many as 20 iterations.

| I  | SSE        | I  | SSE        | I  | SSE        |
|----|------------|----|------------|----|------------|
| 1  | 1          | 6  | 11         | 16 | 1          |
| 2  | 7          | 12 | 17         |
| 3  | 8          | 13 | 18         |
| 4  | 9          | 14 | 19         |
Step 2. The number of clusters is 3 clusters

| Pusat Cluster | Wilayah | Pekerjaan | Umur |
|---------------|---------|-----------|------|
| C1            | 2       | 1         | 37   |
| C2            | 2       | 1         | 33   |
| C3            | 3       | 1         | 27   |

Step 3. Generating initial solutions, namely cluster centers randomly.

| P | I | Pusat Cluster | SSE   |
|---|---|---------------|-------|
| 0,1,2 | 4 | [2, 1, 33][3, 1, 22][1, 3, 44] | 315.0534 |

Step 5. Cluster center evaluation based on SSE values.

Step 6. Generating a new cluster center based on the SSE value, which is the minimum SSE value in the previous cluster center.

| Iterasi | Pusat Cluster | Nilai SSE Minimum |
|---------|---------------|-------------------|
| 1       | [2, 1, 33][3, 1, 22][1, 3, 44] | 315.0534 |
| 2       | [2, 5, 30][2, 1, 33][1, 3, 41] | 314.2185 |
| 3       | [3, 2, 25][3, 2, 40][2, 5, 30] | 306.8306 |
| 4       | [3, 2, 25][3, 2, 40][2, 5, 30] | 306.8306 |
| 5       | [3, 2, 25][3, 2, 40][2, 5, 30] | 306.8306 |
| 6       | [3, 2, 25][3, 2, 40][2, 5, 30] | 306.8306 |
| 7       | [2, 1, 37][2, 1, 33][3, 1, 27] | 301.0533 |
| 8       | [2, 1, 37][2, 1, 33][3, 1, 27] | 301.0533 |
| 9       | [2, 1, 37][2, 1, 33][3, 1, 27] | 301.0533 |
| 10      | [2, 1, 37][2, 1, 33][3, 1, 27] | 301.0533 |
| 11      | [2, 1, 37][2, 1, 33][3, 1, 27] | 301.0533 |
| 12      | [2, 1, 37][2, 1, 33][3, 1, 27] | 301.0533 |
| 13      | [2, 1, 37][2, 1, 33][3, 1, 27] | 301.0533 |
| 14      | [2, 1, 37][2, 1, 33][3, 1, 27] | 301.0533 |
| 15      | [2, 1, 37][2, 1, 33][3, 1, 27] | 301.0533 |
| 16      | [2, 1, 37][2, 1, 33][3, 1, 27] | 301.0533 |
| 17      | [2, 1, 37][2, 1, 33][3, 1, 27] | 301.0533 |
| 18      | [2, 1, 37][2, 1, 33][3, 1, 27] | 301.0533 |
| 19      | [2, 1, 37][2, 1, 33][3, 1, 27] | 301.0533 |
| 20      | [2, 1, 37][2, 1, 33][3, 1, 27] | 301.0533 |
| SSE Minimum | 301.0533 |
Step 7. Update SSE values

Step 8. Criteria stop when the iteration has performed an SSE search of 20 iterations or 100 different cluster centers that are generated randomly.

Step 9. The output of a system is the minimum SSE value which is the most optimum cluster center.

| Pusat Cluster | Wilayah | Pekerjaan | Umur |
|---------------|---------|-----------|------|
| C1            | 2       | 1         | 37   |
| C2            | 2       | 1         | 33   |
| C3            | 3       | 1         | 27   |

6 Conclusions

Build a K-Means cluster that is better for finding the most optimum cluster center by modifying the K-Mean Clustering algorithm based on the minimum Sum of Squared Error (SSE) value. Implementing the K-Means cluster with the most optimal cluster center

References

[1] A. T. Rahman and R. Anggrainingsih, “Coal Trade Data Clustering Using K-Means (Case Study PT. Global Bangkit Utama),” vol. 6, no. 1, p. 8, 2017.
[2] N. H. Kristanto, “Implementasi K-Means Clustering untuk Pengelompokan Analisis Rasio Profitabilitas dalam,” vol. 02, no. 01, p. 7, 2016.
[3] E. Muningsih and S. Kiswati, “SISTEM APLIKASI BERBASIS OPTIMASI METODE ELBOW UNTUK PENENTUAN CLUSTERING PELANGGAN,” Joutica, vol. 3, no. 1, p. 117, Apr. 2018.
[4] B. Yusuf Bakhtiar, A. Bima Murti Wijaya, and H. Dwi Cahyono, “PENGEMBANGAN SISTEM ANALISIS AKADEMIS MENGGUNAKAN OLAP DAN DATA CLUSTERING STUDI KASUS : AKADEMIK UNIVERSITAS SEBELAS MARET SURAKARTA,” J. Teknol. Inf. ITSmart, vol. 4, no. 1, p. 01, Sep. 2016.
[5] Ong, J. O. 2013. Implementasi Algoritma K-Means Clustering Untuk Menentukan Strategi Marketing President University(12):10-20.
[6] Paulanda, Z. 2012. Model Profil Mahasiswa Yang Potensial Drop Out Menggunakan Teknik Kernel-K-Means Clustering Dan Decision Tree. Tesis. Universitas Sumatera Utara. 2013.