International Journal of Database Theory and Application
Vol.6, No.5 (2013), pp.13-22
http://dx.doi.org/10.14257/ijdta.2013.6.5.02

Evaluation of the Selection of the Initial Seeds for K-Means Algorithm

ShinWon Lee[1] and WonHee Lee[2]

[1]*Department of Computer System Engineering, Jungwon University, Goesan-eup, Goesan-gun, Chungbuk, 367-805, Republic of Korea*
[2]*Department of Information Technology, Chonbuk University, Baekje-daero, deokjin-gu, Jeonju, Jeonbuk, 561-756, Republic of Korea*
[1]*swlee@jwu.ac.kr*, [2]*wony@jbnu.ac.kr*

Abstract

Clustering method is divided into hierarchical clustering, partitioning clustering, and more. K-Means algorithm is one of partitioning clustering methods and is adequate to cluster a lot of data rapidly and easily. The problem is it is too dependent on initial centers of clusters and needs the time of allocation and recalculation. We compare random method, max average distance method and triangle height method for selecting initial seeds in K-Means algorithm. It reduces total clustering time by minimizing the number of allocation and recalculation.

Keywords: *Clustering, K-Means, Initial seeds*

1. Introduction

Clustering method, gathering several clusters according to special value on a large data, is divided into hierarchical clustering [1, 9], partitioning clustering [6, 10], and graph theory clustering. Mass information of modern society is limited to process data using hierarchical clustering or graph theory clustering and is inefficient to time complexity.

In this paper, we deal with K-Means algorithm which is one of the partitioning clustering methods for mass data. It is easy to implement, if the time complexity is O(n) and the number of pattern is N. But it is entirely dependent on initial centers of clusters. That is, the result of clustering is different to the initial selected centers of cluster. Generally, when K-Means algorithm processes allocation and recalculation repeatedly, centers move into proper location. If initial centers of cluster are selected and concentrated in partial area, the result is not proper for the time of allocation and recalculation is increased. So we improved the performance of K-Means to select initial centers of cluster with calculating rather than random selecting. This method maximizes the distance among initial centers of cluster. After that, the centers are distributed evenly and that result is more accurate than initial cluster centers selected at random.

In this paper, Chapter 2 describes K-Means algorithm and initial center refining method of previous study. Chapter 3 describes the method using max average distance and triangle height for initial center setting method. Chapter 4 experiments and evaluates these methods. In Chapter 5, we conclude.

ISSN: 2005-4270 IJDTA

2. Related Work

2.1. K-Means Algorithm

K-Means algorithm is the most commonly used partitioning clustering algorithm. The concept of this algorithm is to minimize the average Euclidean distance between the patterns and the center of the pattern clustering [4, 5]. The center of cluster is the means of the pattern belonging to the cluster or called center $\vec{\mu}$, and defined as equation (1).

$$\vec{\mu}(\omega) = \frac{1}{|\omega|} \sum_{\vec{x} \in \omega} \vec{x} \tag{1}$$

In this expression, ω is a set of patterns belonging to the cluster, $\vec{x}$ is a particular pattern belonging to the cluster. The pattern is represented as a vector with real values. The cluster is considered a sphere with a center of gravity.

RSS (Residual Sum of Squares) is a measure of how well center expresses patterns belonging to cluster, and represents the sum of squared distance of each pattern center for all patterns belonging to each cluster, and is shown in the following equation (2).

$$RSS_k = \sum_{\vec{x} \in \omega_k} \left| \vec{x} - \vec{\mu}(\omega_k) \right|^2$$

$$RSS = \sum_{k=1}^{K} RSS_k \tag{2}$$

RSS is the objective function of K-Means, this should be minimized. Figure 1 is K-Means algorithm.

It is terminated if the following conditions are available.

(1) It is repeated a predefined number of times. This condition limits the running time of clustering algorithm. If the number of iterations is not enough the quality of clustering can be reduced.

(2) The cluster belonging to the vector is repeated until it doesn't change. This condition is very good for quality of gathering except for when the cluster is small it is time consuming to focus on small clusters.

(3) It is repeated until the center is no longer changed.

(4) RSS is repeated until it drops below the threshold. When it is completed up to the standard, the quality of gathering is very good.

Actually, we use the end condition that combines the method of limiting repeat numbers and repeating until threshold drops below.

$K-Means\left(\{\vec{x}_1,...,\vec{x}_N\},K\right)$
1. $\left(\vec{s}_1,\vec{s}_2,...\vec{s}_K\right)$ // *Select Random Seeds* $\left(\{\vec{x}_1,...,\vec{x}_N\},K\right)$
2. *for* $k \leftarrow 1$ *to* K
3. $\quad do\ \vec{\mu}_k \leftarrow \vec{s}_k$
4. *while stopping criterion has not been met*
5. *do for* $n \leftarrow 1$ *to* N
6. $\quad do\ j \leftarrow \arg\min_j \left|\vec{\mu}_j - \vec{x}_n\right|$
7. $\quad\quad \omega_j \leftarrow \omega_j \cup \{\vec{x}_n\}$ // vector reallocation
8. $\quad$ *for* $k \leftarrow 1$ *to* K
9. $\quad do\ \vec{\mu}_k \leftarrow \frac{1}{|\omega_k|}\sum_{\vec{x}\in\omega_k}\vec{x}$ // center recalcuration
10. *return* $\{\vec{\mu}_1,...,\vec{\mu}_K\}$

Figure 1. K-Means Algorithm using Random Initial Center

2.2. Initial Value Setting of K-Means

Performance of K-Means algorithm greatly varies depending on how the initial centers are selected. The basic initial centers consist of random K selected patterns or random K coordinates within a set of patterns. This method is bound to the large variation. To solve these problems, much research has been ongoing about the setting of initial center.

Shinwon [11] has the idea that characteristics of initial center of cluster belong to a particular set of patterns with a common pattern of the property. In the selected initial cluster, it selects three documents that are represented as index and weight, instead of selecting any one random pattern. And the initial cluster center vector is set to the center. This algorithm is the following equation (3).

$$c_i^{initial} = avg\ big\left(\sum_{j=1}^{3} d_j\right) \tag{3}$$

$c_i^{initial}$ is a ith cluster vector, dj is a jth document vector.

1. Pick two center c1, c2 as in the 2-means case, that is choose $x, y \in X$ with probability proportional to $\|x-y\|^2$.
2. Pick a random point $x \in X$ with probability proportional to $\min_{j\in\{1,2,...,i\}}\|x-c_j\|_2$ and set that as c_{i+1}.
3. This procedure consists of k iterations.

Figure 2. Initial Center Refining Procedures of [8]

Rafail [8] considers the distance that is division size of each cluster. It proposes the idea that each optimal center can have an initial center. This method processes to develop division condition to find k initial centers with very close optimal center. The process of obtaining the initial center is shown in the following Figure 2.

Paul [7] studied K-Means algorithm about protocol to adapt communication security system, and it is called Two-Party K-Means clustering protocol. Before implementing Two-Party, it needs algorithm of initial center setting for single data set, and is shown in Figure 3. The basic idea is to find the center on the whole document.

1. Center of Gravity : $C = \frac{1}{n}\left(\sum_{i=1}^{n} D_i\right)$

2. Distance to Center of Gravity : $\tilde{C}_i^0 = Dist^2(C, D_i)$

3. Average Squared Distance : $\overline{C} := \frac{1}{n}\left(\sum_{i=1}^{n} \tilde{C}_i^0\right)$

4. Pick First Cluster Center :

$$\mu_1 = D_i, \Pr[\mu_1 = D_i] = \frac{\overline{C} + \tilde{C}_i^0}{2n\overline{C}}$$

5. Iterate to Pick the Remaining Cluster Centers :

$\mu_j, j = 2,...,k$

5.1 $\tilde{C}_i^{j-1} = Dist^2(\mu_{j-1}, D_i), \quad 1 \le i \le n$

5.2 $\tilde{C}_i = \min\{\tilde{C}_i^l\}_{l=0}^{j-1}, \quad 1 \le i \le n$

5.3 $\overline{C} = average\tilde{C}_i$ (*over all* $1 \le i \le n$)

5.4 $\mu_j = D_i, \Pr[\mu_j = D_i] = \frac{\tilde{C}_i}{n\overline{C}}$

Figure 3. Initial Center Refining Procedures of [7]

This procedure uses each data, DB and DA, and each center, μ_j^B and μ_j^A to send and receive between two virtual users, Bob and Alice. This algorithm used two data sets. It is identical and proposed method to security of the message that sent and received.

3. Cluster Center Setting

3.1. Using Max Average Distance

This method should be the selected when initial centers of cluster are far apart. By doing so, the initial centers of cluster randomly selected will be biased in some areas, and this phenomenon can be prevented. The clustering was to improve speed and the accuracy of clustering. Figure 4 is initial center refining algorithm. In the proposed K-Means algorithm, a set C of the initial centers of cluster is the following equation (4).

$$C = \max \sum_{i=1}^{K} \left\| c_{avg} - c_i \right\|^2 \tag{4}$$

ci is ith center of cluster, cavg is average from c1 to ck.

1. Select Random K centers

2. *for* $x \in X$

2.1 Select Candidate Cluster with the closest x

candidate Cluster←min disti=0,…,k(x,ci)

2.2 After replacing previous center by selected candidate Cluster, calculate new average distance

$newDistAvg \leftarrow avg \sum_{i=1}^{k} \left| c_{avg} - c_i \right|^2$

If ci=candidate Cluster then $\left| c_{avg} - x \right|$

2.3 if newDistAvg > oldDistAvg then $c_i \leftarrow x$

3. return {c1,…,ck}

Figure 4. Initial Center Refining Algorithm

Figure 5 describes the setting of initial centers of cluster using two-dimensional data, when K is 3. There are c1, c2, c3 centers, and new data x will look for the closest center. Comparing the distance between each center c1, c2, c3 and x, we can confirm that the result is c1. Now, put x instead of c1, calculate the distance $\{d_1', d_2', d_3'\}$ between each centers and average as following equation (5) and (6).

$$newDistAvg = \frac{1}{K} \sum_{i=1}^{k} d_j' \tag{5}$$

$$oldDistAvg = \frac{1}{K} \sum_{i=1}^{k} d_j \tag{6}$$

This distance can be compared with distance between the existing centers. Comparing the two average distance, newDistAvg value, substituted x for c1 is larger value, so x is replaced by the new c1. This process is repeated for the set X with x.

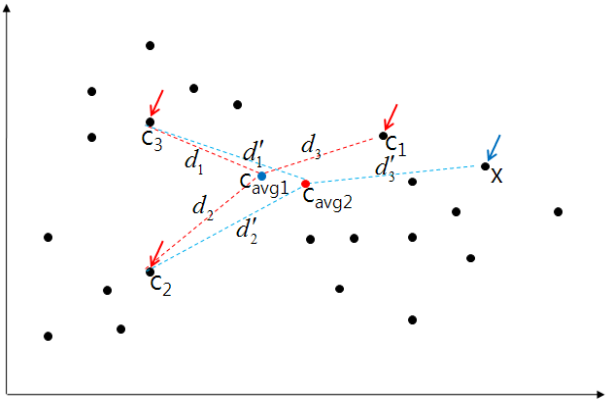

Figure 5. Initial Centers Shifting using Max Average Distance

3.2. Using Triangle Height

This method uses triangle height to replace the initial center. When we know the three lengths of the triangle, we can calculate triangle height by Heron's formula. By doing so, the initial centers of cluster randomly selected will be biased in some areas, and this phenomenon can be prevented. And the clustering was used to improve speed and the accuracy of clustering. A set C of the initial centers of cluster is the following equation (7).

$$C = \max \sum_{i=1}^{k} \left\| c_{height} - c_i \right\| \tag{7}$$

c_i is *i*th center of cluster, c_{height} is triangle height from c_1 to c_k.

1. Select Random K centers

2. *for* $x \in X$

2.1 Select Candidate Cluster with the closest x

candidate Cluster←min disti=0,…,k(x,ci)

2.2 After replacing previous center by selected candidate Cluster, calculate new triangle height

$newHeight \leftarrow \sqrt{C^2 - \left(\left(a^2 + c^2 - b^2\right)/2a\right)^2}$

$a = \overline{c_2, c_3}, b = \overline{x, c_3}, c = \overline{x, c_2}$

2.3 if newHeight > oldHeight then $c_i \leftarrow x$

3. return {c1,…,ck}

Figure 6. Initial Center Setting Algorithm

Figure 7 describes setting of initial centers of cluster using two-dimensional data, when K is 3. There are c_1, c_2, c_3 centers, and new data x_1 will look for the closest center. Comparing the triangle height between h_0, h_1, we can confirm that the result is x_1. Now, put x_1 instead of c_1, calculate the height $\{h_0, h_1\}$ between each centers as follows.

$$newHeight \leftarrow \sqrt{C^2 - \left(\left(a^2 + c^2 - b^2\right)/2a\right)^2}$$
$$a = \overline{c_2, c_3}, b = \overline{x_1, c_3}, c = \overline{x_1, c_2} \tag{8}$$

$$oldHeight \leftarrow \sqrt{C^2 - \left(\left(a^2 + c^2 - b^2\right)/2a\right)^2}$$
$$a = \overline{c_2, c_3}, b = \overline{c_1, c_3}, c = \overline{c_1, c_2} \tag{9}$$

Each height can be compared with the distance between the existing centers. Comparing the two heights, *newHeight* value substituted x_1 for c_1 the larger value, so x_1 is replaced by the new c_1. (x_1, $_{c2}$, c_3) is *newHeight*, and is then compared with x_2. On the other hand in the case of x_2, the height h_2 of (x_2, c_2, c_3) is the smaller than height h_0 of c_1,c_2,c_3, so x_2 isn't replaced by the new c_1. This process is repeated for the set X with x_i.

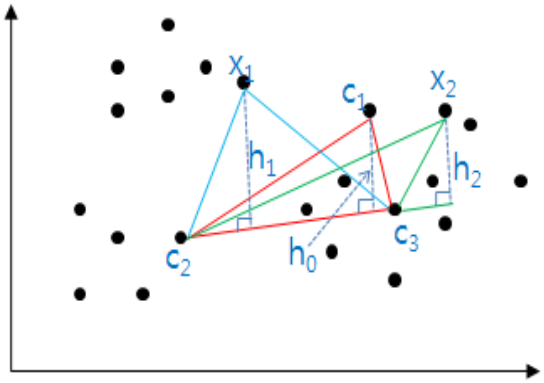

Figure 7. Initial Center Shifting using Triangle Height

4. Experiment

For evaluation of clustering results we created the 200 pieces of data and tested this clustering performance. The number of data points was a small number of restrictions to make it easy to identify with the naked eye. Clustering experiments execute 10 repetitions around each initial cluster center setting method, checking the result.

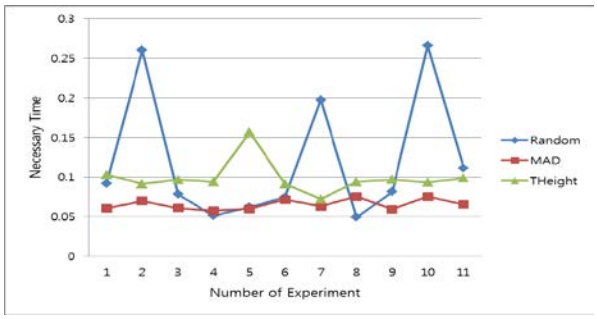

Figure 8. Necessary Time, K=5

Figure 9. The Number of Iteration, K=5

As shown in Figure 8, when using max average distance (MAD), we can see that the necessary time is reduced to 0.06546 then to 0.11118 when using random. The triangle height (THeight) method is reduced to 0.09878.

Figure 9 displays the number of iteration for each cluster number. Even though triangle height distance required modifying time, reducing the number of allocation and recalculation, the total necessary time could be reduced.

5. Conclusion

In this paper, we compared methods for the initial selection of seeds to improve the performance of K-Means algorithm. Because it is one of partitioning algorithm methods mainly used large amounts of data. However, the result of the cluster is dependent on the initial centers of cluster.

This method means to enhance the performance of clustering according to place, as far as the initial centers of cluster escaped random selection. In terms of accuracy, the standard clustering algorithm is exactly done well, but it can be not relatively. We can find that the clustering result is dependent on initial centers of cluster. It is linear for number of documents, and can reduce the time spent on total clustering. In addition, clustering result is consistent.

Clustering has been used in various fields, information retrieval, e-mail clustering, communication protocols, and clustering for medical information. Advanced K-Means algorithm using the proposed maximum average distance can be applied also in these areas.

In the future, this method can be applied to hierarchical clustering as well as partitioning clustering, and actual research can be applied to information retrieval.

References

[1] G. Adami, P. Avesani and D. Sona, “Clustering documents in a web directory”, Proceedings of the 5th ACM international workshop on Web information and data management, **(2003)**.
[2] C. D. Manning, P. Raghavan and H. Schütze, “Introduction to Information Retrieval”, Cambridge University Press, **(2008)**, pp. 331-338.
[3] A. K. Jain and R. C. Dubes, “Algorithms for Clustering Data”, Prentice-Hall advanced reference series. Prentice-Hall, Inc., Upper Saddle River, NJ, **(1988)**.
[4] S. P. Lloyd, “Least squares quantization in PCM”, Special issue on quantization, IEEE Transaction Information Theory, vol. 28, **(1982)**, pp. 129-137.
[5] J. McQueen, “Some methods for classification and analysis of multivariate observations”, Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, **(1967)**, pp. 281-297.

[6] D. A. Meedeniya and A. S. Perera, "Evaluation of Partition-Based Text Clustering Techniques to Categorize Indic Language Documents", IEEE International Advance Computing Conference(IACC 2009), **(2009)**, pp. 1497-1500.
[7] P. Bunn and R. Ostrovsky, "Secure Two-Party k-Means Clustering", Proceedings of the 14th ACM conference on Computer and communications security, Alexandria, Virginia, USA, **(2007)**, pp. 486-497.
[8] R. Ostrovsky, Y. Rabani, L. J. Schulman and C. Swamy, "The Effectiveness of Lloyd-Type Methods for then k-Means Problem", Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science, **(2006)**, pp. 165-176.
[9] N. Sahoo, J. Callan, R. Krishnan, G. Duncan and R. Padman, "Incremental hierarchical clustering of text documents", Proceedings of the 15th ACM international conference on Information and knowledge management, **(2006)**, pp. 357-366.
[10] Y. Yonghong and B. Wenyang, "Text clustering based on term weights automatic partition", Computer and Automation Engineering (ICCAE), 2010 The 2nd International Conference, **(2010)**, pp. 373-377.
[11] S. Lee, "A Study on Hierarchical Clustering using Advanced K-Means Algorithm for Information Retrieval", Doctoral Thesis, Chonbuk University, **(2005)**.

**Corresponding author: WonHee Lee, Ph.D.
Department of Information Technology, Chonbuk University,
Baekje-daero, deokjin-gu, Jeonju, Jeonbuk, 561-756, Republic of Korea
E-mail: wony@jbnu.ac.kr

