A New Algorithm for Optimization of K-Means Clustering with Determining Maximum Distance Between Centroids

Ali Ridho Barakbah
Electronic Engineering Polytechnic Institute of Surabaya - ITS
E-mail: ridho@eepis-its.edu

Abstract

K-means algorithm is very sensitive in initial starting points. Because of initial starting points generated randomly, K-means does not guarantee the unique clustering results so that it is very difficult to reach global optimum. A new algorithm for optimization of K-means clustering is proposed in this paper. It determines position of initial centroids in farthest accumulated distance among them. The accumulated distance metric is built at first in order to designate the initial centroids. A new initial centroid can be selected from a data which has maximum accumulated distance metric. The iterative process is needed so that the all initial centroids are determined. The new approach proposed in this paper can positionate all centroids far separately among them in the data distribution. The experimental results show effectiveness of the proposed algorithm to improve the clustering results of K-means clustering.

Keywords: clustering, initial centroids, K-means algorithm.

1. Introduction

Clustering is an effort to classify similar objects in the same groups. Cluster analysis constructs good cluster when the members of a cluster have a high degree of similarity each other (internal homogeneity) and are not like members of other clusters (external homogeneity) [4, 9]. It means that the process to define a mapping $f:D \rightarrow C$ from some data $D=\{d_1,d_2,..,d_n\}$ to some clusters $C=\{c_1,c_2,..,c_n\}$ on similarity between d_i. The applications of clustering is diversely in many fields such as data mining, pattern recognition, image classification, biological sciences, marketing, city-planning, document retrievals, etc.

The most well known, widely used and fast methods for clustering is K-means clustering developed by Mac Queen in 1967. The simplicity of K-means clustering made this algorithm used in various fields. K-means clustering is a partitioning clustering method that separates data into k mutually excessive groups. Through such the iterative partitioning, K-means clustering minimizes the sum of distance from each data to its clusters. K-means clustering is very popular because of its ability to cluster a kind of huge data, and also outliers, quickly and efficiently. It remains a basic framework for developing numerical or conceptual clustering systems because various possibilities of distance and prototype choice [2].

However, K-means clustering is very sensitive to the designated initial starting points as cluster centers. K-means clustering generates initial clusters randomly. If a randomly designated initial starting point close to a final cluster center, then K-means clustering can find the final cluster center. It, however is not always. If a designated initial point is far from the final cluster center, it will lead to incorrect clustering results [10]. Because of initial starting points generated randomly, K-means clustering does not guarantee the unique clustering results [12]. K-means clustering is difficult to reach global optimum, but only to one of local minima [7].

Several methods proposed to solve the cluster initialization for K-means clustering. A recursive method for initializing the means by running k clustering problems is discussed by Duda and Hart (1973). A variation of this method consists of taking the entire data into account and then randomly perturbing it k times [12]. Bradley and Fayyad (1998) proposed an algorithm that refines initial points by analyzing distribution of the data and probability of data density [3]. Penã et al. (1999) presented empirical comparison for four initialization methods for K-means clustering those are random, Forgy approach, Mac Queen approach, and Kaufman approach [5]. Barakbah and Helen (2005) presented a new algorithm, called as Optimized K-means, that spreads the initial centroids in the feature space so that the distances among them are as far as possible [14]. Barakbah et al. (2005) presented the optimization of initial starting points for K-means using Simulated Annealing [15].

In this paper we propose a new approach to optimize K-means clustering with maximum distance between centroids. This approach determines the position of centroids by calculating accumulated distance metric between each data to all previous centroids, and then, selects a data which has the maximum distance. It can positionate all centroids far separately among them in the data distribution.

2. Basic theory of K-means

Let $A=\{a_i \mid i=1,\ldots,n\}$ be attributes of n-dimensional vector and $X=\{x_i \mid i=1,\ldots,r\}$ be each data of A. The K-means clustering separates X into K partitions called

Ali Ridho Barakbah, "A new algorithm for optimization of K-means clustering with determining maximum distance between centroids", Proc. Industrial Electronics Seminar (IES) 2006, pp.240-244, November 9, 2006, Surabaya, Indonesia.

clusters $S=\{s_i \mid i=1,\ldots,k\}$ where $M \in X$ is $M=\{m_i \mid i=1,\ldots,n(s_i)\}$ as members of S. Each cluster has cluster center of $C=\{c_i \mid i=1,\ldots,k\}$.

K-means clustering algorithm can be described as follows:

1. Initiate its algorithm by generating random starting points of initial cluster centers c_k.
2. Calculate the distance $d(x,c)$ between vector x_i to cluster center c_k. Euclidean distance used to be used to express the distance.
3. Separate x_i into s_k which has minimum $d(x,c)$.
4. Determine the new cluster centers defined as:

$$c_i = \frac{1}{p}\sum_{j=1}^{p} m(s_i, j) \quad \text{where } p=n(s_i) \qquad (1)$$

5. Go back to step 2 until $Ci = Ci\text{-}1$.

It may stop in the t iteration with a threshold ε [7] if cluster center has been updated by the distance below ε:

$$\left|\frac{C^t - C^{t-1}}{C^t}\right| < \varepsilon \qquad (2)$$

3. Proposed algorithm

This section describes a basic concept of the proposed algorithm, how to determine the initial centroids to optimize K-means clustering, and the algorithm the proposed approach.

3.1. Basic concept

The lack of K-means algorithm that generates the initial centroids randomly does not consider the placement of them spreading in the feature space. It makes the initial centroids may be placed closely so that one of them can be ignored. Therefore the initial centroids generated by K-means may be trapped in the local optima. We propose in this paper how to place the initial centroids in farthest accumulated distance among them.

The proposed algorithm in this paper is inspired by from Optimized K-means [14]. Even though the ideal initial centroids reside near average gravity of the cluster members, the placement of initial centroids can also be set in the surrounding of the members as far as the members still consider the same centroid as nearest centroids. To stabilize the placement of initial centroids in data distribution, different from Optimized K-means which spreads the initial centroids based on closeness between each data to grand mean of data, the proposed algorithm determines the initial centroids based on the highest distance weights among the centroids. The experimental results present better performance in several datasets.

3.2. Determining initial centroids

First of all, the grand mean of data is calculated as center of data distribution. Then distance metric is built between each data to the grand mean. The data which has highest distance in the metric can be selected as first centroid. Figure 1 illustrates m as grand mean of data and C_1 is determined as first centroid because it has farthest distance to m.

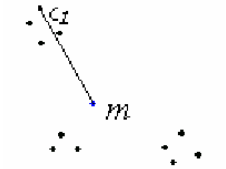

Figure 1. Determining the first initial centroid

Then, to determine the second initial centroid, distance metric is again built between each data to the first initial centroid. The second distance metric, then, is summed to the first distance metric up. It avoids the other three data near C_1 to be chosen as the second initial centroid. This approach can spread the second initial centroid far from the first one.

The iterative process is needed so that the all initial centroids are determined. The new approach proposed in this paper can positionate all centroids far separately among them in the data distribution.

3.3. Algorithm of proposed algorithm

Let $X=\{x_i \mid i=1,\ldots,n\}$ be data, and k be number of clusters. The following execution steps of the proposed algorithm are described as:

1. Set $C=\varnothing$ as set of initial centroids will be determined
2. Set $DM=[]$ as accumulated distance metric
3. Determine m as grand mean of X
4. Build $D(X,m)$ as distance metric between X to m
5. Set $i=1$ as counter to determine the first centroid
6. $DM=DM+D$
7. Select $x_j \leftarrow \text{ArgMax}(DM)$ as c_i
8. $C=C \cup c_i$
9. Set $DM(x_j,C)=0$
10. Build again $D(X, c_i)$ as distance metric between X to c_i
11. $i=i+1$
12. If $i \le k$, go back to step 6
13. C is the solution as initial centroids

4. Experimental results

To establish practical applicability of our proposed algorithm, we made a series of experiments and tested its performance on number of real datasets those are Ruspini, Fossil, Iris, New Thyroid and Wine dataset. The following variance ratio, v is defined as a performance measure in the experiments. Variance constraint [8] can express the density of the clusters with variance within cluster and variance between clusters [6, 13]. The ideal cluster has minimum variance within clusters, called as V_w, to express internal homogeneity and maximum variance between clusters, called as V_b, to

Ali Ridho Barakbah, "A new algorithm for optimization of K-means clustering with determining maximum distance between centroids", Proc. Industrial Electronics Seminar (IES) 2006, pp.240-244, November 9, 2006, Surabaya, Indonesia.

express external homogeneity [11]. Then, the variance ratio can be determined as:

$$V = \frac{V_w}{V_b} \qquad (3)$$

Therefore, the ideal cluster is expressed by minimal V because of minimization of V_w and maximization of V_b.

We conducted the comparison between the proposed algorithm and several clustering algorithms those are Single Linkage, Complete Linkage, Centroid Linkage, Average Linkage, and Fuzzy C-means (FCM). Beside we also conducted the comparison between the proposed algorithm and several approaches of initial centroids optimization for K-means, those are Forgy approach, Mac Queen approach, Kaufman approach, Refinement approach [3], Genetic Algorithm (GA), Simulates Annealing (SA) [15], Optimized K-means (Op-K) [14] and K-means using random initialization. For Forgy approach, Mac Queen approach and Refinement approach, because those approaches can not give an unique clustering result, we made 1000 times experiments and recorded the average result. For Genetic Algorithm, we used 10 individuals in the population and 100 generations. For Simulated Annealing, we used 10000 iterations with random update state in the feature space.

4.1. Ruspini dataset

The Ruspini data set represents a simple, well-known example that is commonly used as a benchmark problem in evaluating clustering methods and is widely available, incorporated as a built-in data object in both R and S-plus statistics packages. The data set consists of 75 bivariate attribute vectors. There are four classes. The data set contains 23, 20, 17 and 15 in classes 1, 2, 3 and 4 respectively.

Table 1. Comparison result between proposed algorithm and several clustering algorithms in Ruspini dataset

Clustering algorithm	V
Single	2320.260786
Complete	2990.737794
Centroid	2320.260786
Average	2320.260786
FCM	2320.260786
Proposed algorithm	2320.260786

Table 2. Comparison result between proposed algorithm and several approaches of initial centroids optimization for K-means in Ruspini dataset

Optimization of initial centroids algorithm	V
Kmeans (1000x)	4112.807717
Forgy (1000x)	5172.782753
Mac Queen (1000x)	5979.883652
Kaufman	2320.260786
Refinement (1000x)	4131.005425
GA (10i, 100g)	2320.260786
SA (10000 itr)	2320.260786
Op-K	9589.162817
Proposed algorithm	2320.260786

Table 1 performs that the proposed algorithm can reach minimal V as well as the other clustering algorithms those are Single Linkage, Centroid Linkage, Average Linkage and FCM. Table 2 performs that the proposed algorithm can also reach minimal V, better than K-means with random initialization and K-means with Forgy approach, Mac Queen approach, Refinement approach and Optimized K-means.

4.2. Fossil dataset

The Fossil data is obtained from Chernoff [1]. It consists of 87 nummulitidae specimens from Eocene yellow limestone formation of northwestern Jamaica. There are three 6 attributes with 3 classes which the distribution is 40 examples of class 1, 34 examples of class 2 and 13 examples of class 3.

Table 3. Comparison result between proposed algorithm and several clustering algorithms in Fossil dataset

Clustering algorithm	V
Single	6976.736041
Complete	7341.345452
Centroid	7049.986245
Average	7006.111129
FCM	6003.198779
Proposed algorithm	5579.178530

Table 4. Comparison result between proposed algorithm and several approaches of initial centroids optimization for K-means in Fossil dataset

Optimization of initial centroids algorithm	V
Kmeans (1000x)	6803.459233
Forgy (1000x)	6511.566477
Mac Queen (1000x)	6503.394527
Kaufman	6755.958012
Refinement (1000x)	8373.585841
GA (10i, 100g)	6773.145550
SA (10000 itr)	7033.734524
Op-K	6773.145550
Proposed algorithm	5579.178530

Table 3 performs that the proposed algorithm can reach minimal V compared with the other clustering algorithms. Table 4 shows that the proposed algorithm

Ali Ridho Barakbah, "A new algorithm for optimization of K-means clustering with determining maximum distance between centroids", Proc. Industrial Electronics Seminar (IES) 2006, pp.240-244, November 9, 2006, Surabaya, Indonesia.

can also reach minimal *V*, better than the other optimization of initial centroids for K-means.

4.3. Iris dataset

We obtained this data set from UCI Repository [16]. This data set contains information about Iris flowers. There are three classes of Iris flowers, namely Iris Setosa, Iris Versicolor and Iris Virginica. The data set consists of 150 examples with 4 attributes. One class is well separable the other two. The others have a large overlap.

Table 5. Comparison result between proposed algorithm and several clustering algorithms in Iris dataset

Clustering algorithm	***V***
Single	3050.046506
Complete	2026.395551
Centroid	1786.916492
Average	1786.916492
FCM	1776.930734
Proposed algorithm	1775.294307

Table 6. Comparison result between proposed algorithm and several approaches of initial centroids optimization for K-means in Iris dataset

Optimization of initial centroids algorithm	***V***
Kmeans (1000x)	1846.710203
Forgy (1000x)	2015.381242
Mac Queen (1000x)	2070.701539
Kaufman	1923.613081
Refinement (1000x)	2603.379798
GA (10i, 100g)	1775.294307
SA (10000 itr)	1904.553353
Op-K	1826.746787
Proposed algorithm	1775.294307

Table 5 presents that the proposed algorithm can reach minimal *V* compared with the other clustering algorithms. Table 6 shows that the proposed algorithm and Genetic Algorithm reaches minimal *V*, better than the other optimization of initial centroids for K-means.

4.4. New thyroid dataset

The new thyroid data set is also obtained from UCI Repository [16]. The data set contains information about classification whether a patient's thyroid to the class euthyroidism, hypothyroidism or hyperthyroidism. The diagnosis (the class label) was based on a complete medical record, including anamnesis, scan etc. The data set consists 5 attributes, with 215 examples. The distribution is 150 of class euthyroidism, 35 of class hypothyroidism and 30 of class hyperthyroidism.

Table 7. Comparison result between proposed algorithm and several clustering algorithms in New thyroid dataset

Clustering algorithm	***V***
Single	2336.463638
Complete	3927.459158
Centroid	2824.843089
Average	5316.100046
FCM	7316.680996
Proposed algorithm	4924.099210

Table 8. Comparison result between proposed algorithm and several approaches of initial centroids optimization for K-means in New thyroid dataset

Optimization of initial centroids algorithm	***V***
Kmeans (1000x)	5546.534675
Forgy (1000x)	6433.481167
Mac Queen (1000x)	6560.548912
Kaufman	7356.873519
Refinement (1000x)	5996.861973
GA (10i, 100g)	6485.240066
SA (10000 itr)	4924.099210
Op-K	6485.240066
Proposed algorithm	4924.099210

Table 7 shows that the proposed algorithm can not reach minimal *V* compared with Single Linkage, Complete Linkage and Centroid Linkage. Nevertheless, in Table 8, the proposed algorithm performs that it can reach minimal *V* as well as Simulated Annealing, better than the other optimization of initial centroids for K-means.

4.5. Wine dataset

We obtained this data set from UCI Repository [16]. The data is the result of a chemical analysis of wines grown in a region in Italy but derived from three different cultivars. There are three classes. The dataset consists of 178 examples each with 13 continuous attributes. The data set contains distribution 59 examples of class 1, 71 examples for class 2 and 48 examples for class 3.

Table 9. Comparison result between proposed algorithm and several clustering algorithms in Wine dataset

Clustering algorithm	***V***
Single	2696.251016
Complete	1731.856172
Centroid	1310.338496
Average	1310.338496
FCM	1734.348078
Proposed algorithm	1518.072510

Table 10. Comparison result between proposed algorithm and several approaches of initial centroids optimization for K-means in Wine dataset

Optimization of initial centroids algorithm	*V*
Kmeans (1000x)	1570.065965
Forgy (1000x)	1683.812181
Mac Queen (1000x)	1651.586974
Kaufman	1735.084480
Refinement (1000x)	1590.891917
GA (10i, 100g)	1735.084480
SA (10000 itr)	1735.084480
Op-K	1518.072510
Proposed algorithm	1518.072510

Table 9 shows that the proposed algorithm can not reach minimal *V* compared with Centroid Linkage and Average Linkage. But, it performs in Table 10 that it can reach minimal *V* as well as Optimized K-means, better than the other optimization of initial centroids for K-means.

5. Conclusion

A new algorithm for optimization of K-means clustering is proposed in this paper. It determines the position of centroids by calculating accumulated distance metric between each data to all previous centroids, and then, selects a data which has the maximum distance as new centroids. It can positionate all centroids far separately among them in the data distribution. In the experimental results, the proposed algorithm in this paper can reach minimal *V* in most of datasets. Compared with the other clustering algorithms, the proposed algorithm performed the most minimal *V* in Ruspini, Fossil and Iris dataset. Beside the proposed algorithm also performed the most minimal *V* in all datasets used in the experiment, better than the other optimization of initial centroids for K-means.

References

[1] C. Yi-tsuu, *Interactive Pattern Recognition*, Marcel Dekker Inc., New York and Basel, 1978.

[2] H. Ralambondrainy, "A conceptual version of the K-means algorithm", *Pattern Recognition Lett*, 16, 1147-1157, 1995.

[3] P.S. Bradley, U.M Fayyad, "Refining initial points for K-means clustering", *Proc. 15th Internat. Conf. on Machine Learning (ICML'98)*, 1998.

[4] G.A. Growe, "Comparing Algorithms and Clustering Data: Components of The Data Mining Process", *thesis*, department of Computer Science and Information Systems, Grand Valley State University, 1999.

[5] J.M. Penã, J.A. Lozano, P. Larrañaga, "An empirical comparison of the initilization methods for the K-means algorithm", *Pattern Recognition Lett*, 20, 1027-1040, 1999.

[6] S. Ray, R.H. Turi, "Determination of number of clusters in K-means clustering and application in colthe image segmentation", *Proc. 4th ICAPRDT*, pp.137-143, 1999.

[7] B. Kövesi, J.M. Boucher, S. Saoudi, "Stochastic K-means algorithm for vector quantization", *Pattern Recognition Lett*, 22, 603-610, 2001.

[8] C.J. Veenman, M.J.T. Reinders, E. Backer, "A maximum variance cluster algorithm", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 9, pp. 1273-1280, 2002.

[9] V.E. Castro, "Why so many clustering algorithms-a position paper", *ACM SIGKDD Explorations Newsletter*, Vol. 4, Issue 1, pp. 65-75, 2002.

[10] Y.M. Cheung, "k*-Means: A new generalized k-means clustering algorithm", *Pattern Recognition Lett*, 24, 2883-2893, 2003.

[11] A. R. Barakbah, K. Arai, "Identifying moving variance to make automatic clustering for normal dataset", *Proc. IECI Japan Workshop 2004 (IJW 2004)*, Musashi Institute of Technology, Tokyo, 2004.

[12] S.S. Khan, A. Ahmad, "Cluster center initialization algorithm for K-means clustering", *Pattern Recognition Lett*, 25, 1293-1302, 2004.

[13] W.H. Ming, C.J. Hou, "Cluster analysis and visualization", *Workshop on Statistics and Machine Learning*, Institute of Statistical Science, Academia Sinica, 2004.

[14] A.R. Barakbah, A. Helen, "Optimized K-means: an algorithm of initial centroids optimization for K-means", *Proc. Soft Computing, Intelligent System, and Information Technology (SIIT) 2005*, pp.2-63-66, Petra Christian University, Surabaya, 2005.

[15] A.R. Barakbah, A. Fariza, Y. Setiowati, "Optimization of Initial Centroids for K-means using Simulated Annealing", *Proc. Industrial Electronics Seminar (IES) 2005*, pp.286-289, Electronic Engineering Polytechnic Institute of Surabaya-ITS, Surabaya, 2005.

[16] UCI Repository, http://www.sgi.com/tech/mlc/db/