A New Clustering Validity Index based on K-means Algorithm

Xiangru Hou*

Department of Information Engineering, Heilongjiang International University, Harbin, 150025, china

*Corresponding author’s e-mail: 3306530748@qq.com

Abstract. Although cluster analysis has got great achievements, there are many questions in it. In this paper, the question on determining optimal number of clusters in cluster analysis is studied mainly. KMS (K-means Silhouette) for determining optimal number of clusters in K-means clustering algorithm are proposed. KMS (K-means Silhouette) algorithm improves the way of setting initial clustering centers in K-means clustering algorithm, and uses Silhouette validity index to determine optimal number of clusters. The experimental results on artificial datasets indicate the effectiveness of the proposed algorithms.

1. Introduction

K-means clustering algorithm is one of the most widely used algorithms in clustering analysis. The K-means algorithm, based on the determined clustering number k and the selected initial clustering center, is a clustering algorithm that obtains the minimal sum of the distances (squared) from all samples to the center of the category to which they belong. In practice, it is difficult to accurately define the k value. At present, some function indexes of clustering validity have been proposed.

For the selection methods of K-means algorithm’s initial clustering center, k samples were usually randomly selected as the initial clustering center. If this method of randomly selecting the initial clustering center was used to cluster some complex data sets, it might result in unstable and wrong clustering results. At present, researchers have proposed a lot of methods to select the initial clustering center of K-means algorithm. Reference [1] introduced a method of using multiple iterative sampling data sets to obtain initial values, which could solve K-means algorithm’s higher dependence on the initial clustering center. Reference [2] used a density-sensitive similarity measurement to calculate the density of samples and to heuristically generate the initial clustering center of samples. Based on max-min distance algorithm, the research on the improvement in the selection of K-means clustering algorithm’s initial clustering center was conducted to better determine the optimal clustering number.

2. K-means clustering algorithm

The algorithm selects the k cluster and selects k initial cluster centers, and each sample is assigned to a cluster in the k cluster according to the minimum distance principle. After that, heart cluster and the category of each sample are adjusted continuously, and finally. The sum of the squares of each sample to the center of its category is minimum. The algorithm steps are as follows:

Algorithm K-means clustering algorithm

1. For n samples, k samples are selected as the initial clustering centers \((z_1, z_2, ..., z_k)\).
2. For each sample $x_i$, find the nearest cluster center $z_v$, and assign it to the cluster $u_v$ marked by $z_v$.

3. The average method was used to calculate the centers after reclassification

4. Computing $D = \sum_{i=1}^{n} \left[ \min_{r=1...k} d(x_i, z_r)^2 \right]$.

5. If the $D$ value converges, then returns $(z_1, z_2, ..., z_k, U)$, and terminates the algorithm, or goes to the step (2).

3. The method for determining the optimal clustering number based on the improved initial center

3.1. The method for setting initial clustering center

3.1.1. Max-min distance algorithm. Max-min distance algorithm is based on heuristics in the pattern recognition field. Its basic idea is to take the object as far as possible as the clustering center, preventing the initial clustering center from excessive affinity in the initial value selection. This method not only intelligently determines the number of initial clustering centers, but also improves the efficiency of sample partitioning. The data set $S_n = \{x_1, x_2, ..., x_n\}$, select scale factor $\theta$. The algorithm steps are as follows:

**Algorithm Max-min distance algorithm**

1. Take a sample from $S_n$ as the first clustering center $z_1$

2. The largest sample of $z_1$ distance from $S_n$ as the second clustering center $z_2$

3. The distance between the sample $x_i$ and $z_1$, $z_2$ is not calculated as the cluster center, and the minimum value is $d$

4. If $D_i = \max \{d_i\} > \theta \| z_1 - z_2 \|$, the corresponding sample $x_i$ is used as the third clustering center $z_3$

5. If there are $k$ cluster centers, a sample that is not taken as a cluster center and the distance of each cluster center $d_{ij}$, Calculate out $D_j = \max \{\min(d_{i1}, d_{i2}, ..., d_{ik})\}$.

6. Repeat the same treatment, until we can't find a new cluster center that meets the requirement

7. Each sample is divided into various types according to the minimum distance principle

3.1.2. Basic setting method. This paper attempts to combine clustering center initialization with the clustering validity index. In order to determine the optimal clustering number. In the process of searching the clustering space and gradually increasing clustering number by K-means clustering algorithm, when the clustering number is $k_{\text{min}}$, $k_{\text{min}}$ samples are selected as the initial clustering center based on the principle of max-min distance algorithm; every time a clustering number is added, add an initial clustering center according to the principle of max-min distance algorithm when maintaining the same center as the last initial clustering, and try to keep the continuity of conditions and the stability of clustering results. In addition, the initial clustering center selected based on max-min distance algorithm is more likely to belong to different clusters, so that better clustering results can be obtained, in order to better determine the optimal clustering number through the validity index. Since the clustering number is unknown in the max-min distance algorithm, the clustering number needs to be obtained by selecting the proportionality coefficient $\theta$ as a constraint.

3.1.3. Setting method analysis. The K-means clustering algorithm is used for a data set. Under the known cluster number and the established initial cluster center conditions, the final clustering results are uniquely determined and thus are stable. This can be proved by the operation of replacing the known conditions into the K-means clustering algorithm. In the first iteration of the algorithm, the
number and location of the initial cluster center are determined. Each sample is divided into a cluster
according to the minimum distance principle, so the classification of each sample is uniquely
determined. In the next iteration, taking an average method to calculate the cluster centers after the
reclassification. The clustering center and clustering results will change in the two adjacent iterations.
However, because each input condition is uniquely determined, the classification algorithm based on
the minimum distance principle is stable, so the result of each iteration process is always unique. In
this paper, the time complexity of the initial clustering center is $O(nkd)$, so as long as the k-means
clustering algorithm is reduced once. The method of this paper is equivalent to the traditional method
of generating initial clustering center. If the number of iterations decreases by more than two times,
the efficiency of this method will be higher than the traditional method.

3.2. KMS algorithm for determining the optimal number of clusters

The traditional K-means clustering algorithm is used to determine the optimal number of clusters, for
each cluster number k, the cluster center should be re initialized. Because the initial cluster centers are
different, the clustering results of different cluster numbers have poor comparability of the validity
index values, which make the original algorithm to solve the optimal number of clusters unstable.
Therefore, this paper improves the existing method of random selection of the initial clustering center
method. Initial clustering centers are determined by maximum and minimum distance algorithm. With
the increase of cluster number, the original initial cluster centers remain unchanged. Based on the
principle of maximum and minimum distance, the initial cluster centers are gradually increased, so
that there is an inheritance relationship between the initial cluster centers of different cluster numbers.
The upper limit of the search range of the algorithm is changed from the existing $Int(\sqrt{n})$ to the
number of clusters $k_{opt}$ generated by the AP algorithm. The Silhouette index was used to analyze the
clustering results to determine the best clustering number. The new algorithm to determine the optimal
cluster number is recorded as KMS. The algorithm is summarized as follows:

**Algorithm KMS algorithm**

1. Select the search range of cluster numbers $[k_{min}, k_{max}]$
2. For k= $k_{min}$ to $k_{max}$.
   (1) Initialize k initial cluster centers $z_k$
   (2) Using K-means clustering algorithm, update membership matrix $U^k$ and cluster center $Z^k$
   (3) Check the termination conditions, if not satisfied, turn(2)
   (4) Using clustering results to calculate the Silhouette index value, turn (2).
3. Comparing the Silhouette index value, the k corresponding to the maximum of the index value is
   the best number of clusters $k_{opt}$
4. Output the best number of clusters, validity index and clustering results.

4. Experimental Results and Analysis

4.1. $k_{max}$ Simulation experiment and analysis

In this experiment, AP algorithm was used to estimate the $k_{max}$ cluster number for the following 12
datasets. Set the AP algorithm $p = p_m$. It is shown in Table1. The dataset used in the experiment
includes real datasets and artificial datasets, with general numerical data, gene expression data, image
data, etc. Kes2 is a two-dimensional two cluster artificial data set, Kes3 is a two-dimensional three
cluster of artificial data sets, and the rest of the data set details refer to the relevant references.
Table 1. $k_{\text{max}}$ evaluated by Affinity Propagation clustering algorithm

| Data set    | Sample number | Correct cluster number | AP cluster number | $\text{Int}(\sqrt{n})$ | Source of data set |
|-------------|---------------|------------------------|-------------------|-------------------------|---------------------|
| Face Image  | 900           | 100                    | 106               | 30                      | Literature [3]      |
| Random S    | 500           | 35                     | 38                | 20                      | Literature [4]      |
| fishers_iris| 70            | 6                      | 6                 | 8                       | Literature [5]      |
| iris        | 150           | 3                      | 6                 | 12                      | Literature [6]      |
| Wine        | 180           | 3                      | 9                 | 13                      | Literature [7]      |
| Model2      | 100           | 3                      | 3                 | 10                      | Literature [8]      |
| Kes2        | 52            | 2                      | 5                 | 7                       | artificial          |
| Kes3        | 58            | 3                      | 5                 | 7                       | artificial          |

4.2. Experiment and analysis of initial cluster center setting

In this experiment, based on the search range of clustering number and the validity index - Silhouette index, the data sets Kes2 and Kes3 were experimented according to the above initial clustering center setting method 1 and compared with method 2 randomly determining the initial clustering center. The impacts of these two clustering center setting methods on Silhouette index were mainly compared, thus affecting the determination of optimal clustering number. In addition, the random clustering accuracy rate of these two setting methods and efficiency of the clustering algorithm were compared (including running time and number of iterations). While the random clustering accuracy rate (RCAR) was defined as the given correct number of clusters, the clustering algorithm was run for w times on the data set, which was expressed by the percentage of the correct clustering number and w ratio. Here, the running time refers to the initial clustering center’s setting time and K-means clustering algorithm running time under the given correct number of clustering. The number of iterations refers to the iteration number of the K-means clustering algorithm based on different setting methods given the correct number of clusters. In order to reduce the error, the algorithm was run 20 times repeatedly, and the running time and the number of iterations took the average of 20 running results.

(a) The relation graph of Method1  
(b) The relation graph of Method2

Figure1. Clustering numbers-index relationship diagram of Kes2
Figure 1 (a) shows the cluster -Silhouette index relation graph with Method1 and Kes2. The optimal clustering number is 2, and the Silhouette index corresponding to the optimal cluster number is 0.7892. Figure 1(b) shows the cluster -Silhouette index relation graph with Method2 and Kes2. The optimal clustering number is 3, and the Silhouette index corresponding to the optimal cluster number is 0.7627. Figure 2 (a) shows the cluster -Silhouette index relation graph with Method1 and Kes3. The optimal clustering number is 3, and the Silhouette index corresponding to the optimal cluster number is 0.7797. Figure 2 (b) shows the cluster -Silhouette index relation graph with Method2 and Kes3. The optimal clustering number is 4, and the Silhouette index corresponding to the optimal cluster number is 0.7740.

Experimental results of determining the optimal number of clusters based on different initial cluster center setting method. It is shown in Table 2. From this, by Silhouette index, Method1 can get the correct optimal cluster number. Method2 cannot get the correct optimal number of clusters.

| Data set | Correct cluster number | Method1 \( k_{max} \) | Method 2 \( k_{max} \) | Method1 Optimal number of clusters | Method1 Maximum of Silhouette index | Method2 Optimal number of clusters | Method2 Maximum of Silhouette index |
|----------|------------------------|------------------------|------------------------|----------------------------------|----------------------------------|----------------------------------|----------------------------------|
| Kes2     | 2                      | 5                      | 2                      | 0.789                            | 3                                | 0.76                             |
| Kes3     | 3                      | 5                      | 3                      | 0.771                            | 4                                | 0.77                             |

The correctness and efficiency of the clustering algorithm are obtained based on different initial cluster center setting method. It is shown in Table 3. The RCAR value of the Method 1 reflects the quality of clustering. Each clustering result is correct, and it is also proved that the clustering result is uniquely determined under the condition that the initial cluster center is known. Method RCAR values of 2 reflect the data sets Kes2 and Kes3, random initial cluster centers are determined by a random method. The K-means clustering algorithm is unstable and the clustering results are poor. For the Kes2 data set, the running time of the clustering algorithm of method 1 is 0.0067 seconds. The running time of the clustering algorithm of method 2 is 0.0128 seconds, which shows that the clustering algorithm of method 1 is significantly less than method 2. Based on Method 1, the number of iterations of the K-means clustering algorithm is 2, and based on method 2, the number of iterations of the K-means clustering algorithm is 6. The number of iterative times of method 1 is obviously less than method 2.
The same conclusion is also found for the Kes3 dataset. Therefore, using method 1, the accuracy of the clustering algorithm and the efficiency of the algorithm are higher than the method 2.

Table 3. Accuracy and efficiency of the Clustering algorithm based on different setting method

| Data set | Method1 | Method2 |
|----------|---------|---------|
|          | RCAR    | Running time | Iteration times | RCAR    | Running time | Iteration times |
| Kes2     | 100%    | 0.0067    | 2               | 15.41%  | 0.12        | 6               |
| Kes3     | 100%    | 0.0075    | 2               | 14.56%  | 0.13        | 6               |

Using method 1, the clustering effects of data set Kes2 and Kes3 are shown in Figure 3 and Figure 4, respectively. It can be seen that the clustering of method 1 is very effective.

5. Conclusions
Since k-means clustering algorithm randomly selected k samples as the initial clustering center, the clustering results were greatly influenced by the initial clustering center under the condition that the
clustering structure was not easily identified, the method for solving the optimal clustering number by the K-means clustering algorithm was very unstable. Based on the max-min distance algorithm, a new method for setting the initial clustering center was proposed in this paper, in order to improve the quality of clustering. Algorithm determining the K-means algorithm’s optimal clustering number - KMS algorithm was proposed by combining with Silhouette validity index. The theoretical research and experimental results have verified the validity and good performance of the above algorithm.

References
[1] U. Fayyad, C. Reina, P. S. Bradley. (1998) Initialization of Iterative Refinement Clustering Algorithms, Proceedings of the 4th International Conference on Knowledge Discovering in Databases and Data Mining. New York: AAAI Press, 11:194-198
[2] K. Krishna, M. N. Murty. (1999) Genetic K-means Algorithm, IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics, 29(3):433-439
[3] B. Frey, D. Dueck. (2016) Clustering by Passing Messages Between Data Point, Science, 315(5814):972-976
[4] B. Frey, D. Dueck. (2008) Response to Comment on Clustering by Passing Messages Between Data Points, Science, 319(5864):72-80
[5] M. Brusco, H. Köhn. (2008) Comment on Clustering by Passing Messages Between Data Points, Science, 319(5864):22-81
[6] D. Dembélé, P. Kastner. (2013) Fuzzy C-means Method for Clustering Microarray Data, Bioinformatics, 19(8):973-980
[7] C. Blake, C. Merz. UCI Repository of Machine Learning Databases, (University of California), Available: Http://mlearn.ics.uci.edu/MLRepository.html.
[8] D. Davies, D. Bouldin. (1979) A Cluster Separation Measure, IEEE Transactions on Pattern Analysis and Machine Intelligence, 1(2):224-227