International Journal of Advanced Research in Big Data Management System
Vol.4, No.1 (2020), pp.17-26
http://dx.doi.org/10.21742/ijarbms.2020.4.1.02

Improved K-Means Clustering Algorithm Based on Dynamic Clustering

Liguo Zheng

Harbin Normal University, Harbin, Heilongjiang province, China
zhenglg@hrbnu.edu.cn

Abstract

Cluster analysis can not only find potential and valuable structured information in the data set, but also provide pre-processing functions for other data mining algorithms, and then can refine the processing results to improve the accuracy of the algorithm. Therefore, cluster analysis has become one of the hot research topics in the field of data mining. K-means algorithm, as a clustering algorithm based on the partitioning idea, can compare the differences between the data set classes and classes. We can use the K-means algorithm to mine the clustering results and further discover the potentially valuable knowledge in the data set. Help people make more accurate decisions. This paper summarizes and analyzes the traditional K-means algorithm, summarizes the improvement direction of the K-means algorithm, fully considers the dynamic change of information in the K-means clustering process, and reduces the standard setting value for the termination condition of the algorithm to reduce The number of iterations of the algorithm reduces the learning time; the redundant information generated by the dynamic change of information is deleted to reduce the interference in the dynamic clustering process, so that the algorithm achieves a more accurate and efficient clustering effect. Experimental results show that when the amount of data is large, compared with the traditional K-means algorithm, the improved K-means algorithm has a greater improvement in accuracy and execution efficiency.

Keywords: *Cluster analysis, K-means, Dynamic clustering, Data mining*

1. Introduction

Data mining is not the result of a single field, it covers multiple fields, it is a new type of interdisciplinary, and it is a crystallization of diversified knowledge fusion. Clustering is an integral part of data mining and one of the effective tools for analyzing data. Clustering, as its name implies, aggregates data into different categories. The principle is to divide the sample set into different clusters according to the characteristics (or attributes) of the sample points themselves and the corresponding criteria, so that the sample points in the same cluster are as similar as possible. The sample points in different clusters are as different as possible. In this way, we can better identify the relationship and spatial distribution of each sample point in the sample set.

Cluster analysis has been widely used in many fields, such as statistics, pattern recognition, image processing, and neural networks. As an important data preprocessing tool, it has an indispensable position in data mining. The cluster analysis method is a typical unsupervised

Article history:
Received (February 9, 2020), Review Result (March 18, 2020), Accepted (April 20, 2020)

Print ISSN: 2208-1674, eISSN: 2208-1682 IJARBMS

classification method. It is one of the effective methods for processing and analyzing data. Without any prior knowledge, it only uses the analysis and processing of the characteristics of the data objects in the data set. Get effective information. Traditional cluster analysis methods strictly classify each sample point in the data set into a certain category (that is, each sample point belongs to a certain cluster). For example, the traditional K-means clustering algorithm uses the distance between sample points as a measure of similarity, and then divides each sample point into corresponding clusters according to the principle of minimum distance. In this clustering algorithm, the division of each sample point is clear and unique, and there is no case where the same sample point belongs to two or more categories at the same time.

In recent years, the research and exploration of cluster analysis algorithms are mainly manifested in two aspects: improving traditional clustering algorithms and proposing new algorithm concepts. Aiming at the disadvantages of traditional clustering algorithms that depend on the setting of the number of clusters and the selection of the initial clustering center, many literatures have proposed improvement methods. For example, [1] proposed an algorithm for maximum and minimum distances, and selected K sample points that are the farthest from each other as the initial clustering center; [2] proposed an algorithm based on density, using the K sample points with the highest density as the initial Clustering center; Literature [3] proposed a clustering method based on the average difference degree, and the K points with the largest average difference degree were selected as the initial clustering centers. These improved algorithms reduce the instability of the algorithm caused by the random selection of the initial clustering center in the traditional clustering algorithm to a certain extent, but still need to manually determine the number of clusters, and choose different numbers of clusters for the final clustering. The results have a noticeable effect. Literature [4] proposed a method to automatically determine the number of clusters. The best K value was determined by comparing the clustering results when all possible K values were taken, but this method solved the problem of a wide range of K values. Time, it often leads to too long calculation time and excessive calculation cost. In order to avoid manually initializing the cluster center, Frey and Dueck proposed the Affinity Propagation (AP) clustering algorithm, which only needs to use the similarity between sample points. The algorithm treats each sample point equally, that is, It may become a cluster center, and at the same time, the connection between the sample points is processed into the information transfer process, and then iterative updates are repeated until the termination condition is met, and finally each sample point is assigned to the cluster to which the highest information value corresponds to the sample point.

Therefore, a more effective clustering algorithm that can dynamically determine the clustering center according to the characteristics of the data set needs to be proposed.

2. Cluster analysis

As an effective data processing tool, cluster analysis algorithm has been widely used in many fields such as high-performance computing, pattern recognition, image processing, and data visualization. It is an unsupervised learning process, does not rely on any prior knowledge, and uses the characteristics or attributes of the sample points themselves as the sole criterion for discrimination. The essence is to identify the internal connection between the sample points, and then find the characteristics of the data distribution [5]. The cluster analysis algorithm is to explore other data that has relevance value to the data object from the given data. Researchers can use this association method to achieve uniform analysis and processing of data objects in the cluster. The application of cluster analysis in the data set can

accurately identify the sparseness and denseness of the data set, so as to better grasp the overall distribution status and the value correlation between data attributes [6].

2.1. Basic steps of clustering

The process of cluster analysis generally includes the following four steps.

Feature selection (or feature extraction) [7][8]: This step should satisfy: in the feature space, the sample points in the same cluster are similar, and the sample points in different clusters are different, and it is also required to be able to effectively identify noise points.

Design of clustering algorithm (or algorithm selection): Set or select a suitable clustering algorithm based on the characteristics of the current data set and the constraints that need to be met.

Cluster confirmation: This step is to verify the clustering results.

Interpretation of results: This is a valuable step in the clustering algorithm, which provides users with a reliable explanation.

2.2. Major cluster analysis algorithms

In the actual application process, because of the differences in data types, purposes, and requirements, the requirements for cluster analysis algorithms also differ significantly. Therefore, in the actual application process, an appropriate clustering algorithm should be selected. This is also very important. important. Using multiple cluster analysis algorithms in the same data set, it can analyze the potential use value of the data and the characteristics of Sohu's buy ability, and provide a strong basis for further data mining and exploration. Typical cluster analysis algorithms mainly include basic density methods, hierarchical methods, partition methods, and grid-based methods.

Dividing method: Given a specific data set, for example, it contains 100 million data objects, the dividing method is to divide the data set into multiple clusters, such as 100 clusters, and each cluster should meet the following two condition. First, each cluster contains at least one data object; second, each data object can belong to only one cluster. In simple terms, 100 million data objects are divided into 100 clusters according to the corresponding rules, and each data can only exist in one cluster. However, in some fuzzy division methods, the degree of restriction can be appropriately relaxed. The constructed cluster should become the optimal objective division, and then the distance between objects in the same cluster should be minimized, and the distance between objects in different clusters should be as wide as possible. The degree of similarity of clusters can generally be used as a measure of the directness of the quality of the division method. An effective division can promote the high similarity of the data in the same cluster, and the lowest similarity between different clusters. Degrees, the most commonly used division methods are K-means and K-medoids algorithms. The partitioning method must be capable of processing the data set into the memory at one time, so as to limit the multi-faceted application in the large data set to the greatest extent. The division method needs to be divided into multiple data according to the user's needs. This will also cause subjective judgments to form the quality of the cluster. The division method only uses a certain fixed rule for clustering, which will cause clustering. The shape is irregular, and the accuracy of the clustering result is relatively low.

The output of the hierarchical method can form a clustering tree for the data objects. The hierarchical method is divided into a top-down and a bottom-up analysis method. However, no matter which method is used, it can obtain a multi-level clustering structure with different granularities, but there are corresponding defects, such as after splitting and merging, and

before it can no longer be traced back, this defect also has the corresponding Enthusiasm, so in the process of splitting and merging, we must consider the different options that lead to the split of the combination.

2.3. K-means clustering algorithm based on partitioning

The K-means algorithm is an iterative cluster analysis algorithm. Its operation principle is that the data is divided into K groups in advance, then K objects are randomly selected as the initial cluster center, and then each object and each seed are calculated the distance between the cluster centers, and assign each object to the cluster center closest to it.

Input: the number of clusters K and the data set containing n objects.

Output: k clusters, making the criterion function value satisfy the condition.

Step 1: Determine an initial cluster center point for each cluster;

Step 2: Distribute the data in the data set to the nearest cluster according to the Euclidean distance principle;

Step 3: Use the mean of the sample data in each cluster as the new cluster center;

Step 4: Repeat Steps 2 and 3 until the algorithm converges.

Step 5 ends, and K result clusters are obtained.

K-means algorithm is one of the most commonly used methods for cluster analysis. The first subtlety of the algorithm proposed by MacQueen is that it is simple, efficient and suitable for processing large-scale data. It has been applied to many fields, including: natural language Treatment, astronomy, ocean, soil, etc. Although the K-means algorithm is simple and efficient, it also has many disadvantages. First, using this algorithm requires determining the K value. Secondly, the initial clustering center has a great influence on the clustering results. Furthermore, it is difficult to process the data of classification attributes, and it is easy to fall into the local optimal solution. Finally, when the amount of processing data is too large, the time complexity of the algorithm is large and the data redundancy is too large. By studying the algorithm, an improved algorithm can be proposed to avoid these shortcomings. Algorithms; some propose a K-means text clustering algorithm based on density and nearest neighbors; some propose an optimized K-means text feature selection algorithm in clustering mode, which is based on the K-means algorithm for class center point initial An improved algorithm for the problem of too sensitive checkpoints; some proposed an accurate attribute weighting K-means clustering algorithm based on information entropy; and a method of plant leaf recognition based on cosine and K-means .

Aiming at the fourth shortcoming of the K-means algorithm, this paper proposes a K-means dynamic clustering method. This algorithm fully considers the dynamic changes of information in the K-means clustering process. By setting standard values for the termination conditions of the algorithm, Reduce the number of algorithm iterations and reduce the learning time; by removing redundant information generated by dynamic changes of information, reduce the interference in the dynamic clustering process, and make the algorithm achieve a more accurate and efficient clustering effect.

3. K-means dynamic clustering algorithm

3.1. The shortcomings of the K-means algorithm

In the process of clustering sample data, not only the distance between each clustered object and their center object needs to be calculated, but also the mean value of the clusters whose center objects have changed needs to be recalculated, and the calculation is repeated in

iterations. When the amount of data is too large, the performance of the algorithm will be reduced. Secondly, the dynamic k-means algorithm is a changing process during clustering, which will cause some unnecessary data redundancy and have a certain impact on the clustering results.

3.2. Improved K-means algorithm

Aiming at the above defects of the K-means algorithm, two optimization principles are proposed: 1) reducing the number of iterations in the clustering process; 2) reducing the amount of data in the clustering process. The basic idea of K-means: Since the K-means algorithm divides the data set into different categories through an iterative process, the variable at the center point is now set to a value σ1, and the initial value of σ_1 is 0. Obtain the value of σ_1 in the process as follows: Calculate the absolute error caused by replacing the original center point O_j with the new center point O_i, the formula:

$$E = \sum_1^k \sum_{p \in C_j} |p - O_j| \quad (1)$$

Among them, p is a data point in space, and $O_j$is a center point of the cluster C_j Calculate the difference between the absolute error caused by O_jnd the absolute error caused by the original center point O_j. The calculation formula is:

$$e = E_I - E_j \quad (2)$$

When e <0, the change amount of the center at this time is the set σ_1. Once the value of σ_1 is obtained, this value does not change. During the iteration, when the change amount of the center point is less than σ1, the entire cluster is added to the selected data set and deleted from the sample set, so that only the samples that are not correctly identified are retained in the original sample data set. The formula for calculating the change amount of the center point from Definition 2 is:

$$\sigma_r = \frac{1}{|T_i|} \sum_{a_i \in T_{r,j}} a_i - \frac{1}{|T_{i-1}|} \sum_{a_i \in T_{r-1,j}} a_j \quad (3)$$

Among them, r is the number of iterations of the algorithm, and Tr, j represents the j-th category of the r-th iteration. When $\sigma_r \leq \sigma_1$, the conditions are met, and then other samples are screened until all sample data is correctly identified. The algorithm flow is shown in [Figure 1]. The algorithm is described as follows:

Step 1 Determine the number of clusters K between 2 and √N according to the rules of experience, where N is the number of all data points in the data space. By selecting K values one by one in the interval [2, √N], and using the clustering validity function to evaluate the effect of clustering. Finally, the optimal K value is obtained.

Step 2: Use the K-center value method to select the initial cluster center. The so-called K-center value method is to use the average value of the cluster as the reference point in order to avoid the interference of isolated points. This method is still based on the principle of minimizing the sum of the differences between all objects and the reference point.

Step 3: Distribute the samples in the sample set to the nearest clusters according to the Euclidean distance principle.

Step 4: Calculate the centroid point of each class.

Step 5: Determine whether the change amount of the cluster center point satisfies the set conditions. If so, add it to the selected feature set and delete it from the data sample set.

Step 6: Determine whether the data sample set is empty. If it is empty, end the algorithm. If it is not empty, traverse the number of center points N. When N <K, go to step 3. When N = K, continue to the next step.

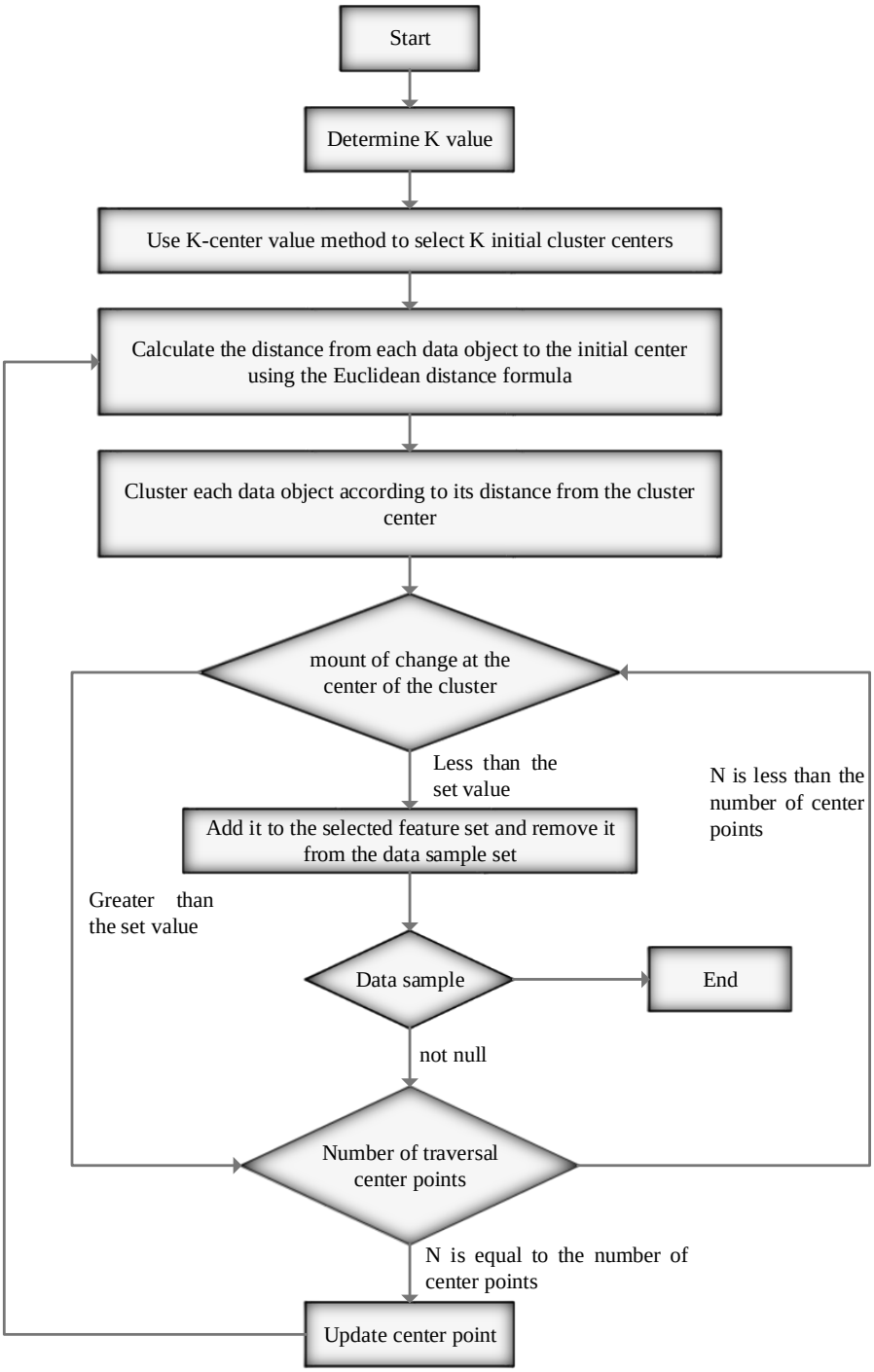

Figure 1. K-means algorithm flow

Step 7: Update the center point. Calculate the centroid of the cluster whose center point change amount is greater than the set value, and use it as the new cluster center, then go to step 3.

Step 8 ends, the data sample is the empty set, and K result clusters are obtained.

4. Experimental research

4.1. Experimental data set

In order to analyze the clustering performance of the K-means dynamic clustering algorithm, simulation experiments use five different public data sets. The data sample collections are all from the UCI machine learning database, which is a public database dedicated to data mining algorithms and testing machine learning. [Table 1] describes the summary information of the five groups of data, such as the name, number of samples, and number of categories.

It can be seen from Table 1 that the five sets of data are composed of different numbers of samples and categories, and the diversity of the data set has verified their performance under different conditions to a certain extent, ensuring that the experimental results are universal.

Table 1. Experimental data set

Serial number	Database	Number of samples / pieces	Category / piece
1	Lung-cancer	3303	27
2	Promoter	10780	16
3	Splice	30900	32
4	Coil	56233	49
5	Isolet	158000	57

4.2. Comparison of experimental results

In order to avoid the inherent defects of the K-means algorithm affecting the experimental results, the experimental data is pre-processed. First determine the number of clusters K. According to the rule of thumb, the value of K is between 2 and $\sqrt{N}$, where N is the number of all data points in the data space. By selecting K values one by one in the interval [2, $\sqrt{N}$], and using the more traditional validity function W. The index is used to evaluate the clustering effect, and then the optimal K value is obtained. The K values of the five groups of data are: 23, 18, 42, 39, 73. Then use K-center value method to select the initial clustering center. Then use Euclidean distance to calculate the distance from each data sample to the center point and cluster each data object according to its distance from the cluster center. The above is the preprocessing operation part of the data; the σ1 value obtained during the algorithm iteration is 1.31 × 19-7. Next, the entire clustering process is completed for the traditional K-means algorithm and the K-means dynamic clustering algorithm according to the algorithm flow. The accuracy and time efficiency of the clustering algorithm are compared, and the experimental results are shown in [Table 2] and [Table 3].

Table 2. Comparison of accuracy between traditional K-means algorithm and K-means dynamic clustering algorithm

Algorithm	Database	Clustering accuracy%
Traditional K-means	Lung-cancer	86.53
	Promoter	84.24
	Splice	83.87
	Coil	76.59
	Isolet	74.33
K-means dynamic clustering	Lung-cancer	66.86
	Promoter	71.32
	Splice	81.35
	Coil	86.28
	Isolet	90.98

Table 3. Comparison of execution time between traditional K-means algorithm and K-means dynamic clustering algorithm

Algorithm	Database	Execution time / ms
Traditional K-means	Lung-cancer	36
	Promoter	73
	Splice	118
	Coil	180
	Isolet	400
K-means dynamic clustering	Lung-cancer	17
	Promoter	17
	Splice	18
	Coil	36
	Isolet	57

It can be seen from [Table 2] that the K-means dynamic clustering algorithm achieves high clustering accuracy for Coil2000 and Isolet datasets with large data volumes, and for the Lung-cancer and Promoter datasets with smaller data volumes, K-means dynamic The accuracy of the clustering algorithm is lower than that of the traditional K-means algorithm, indicating that the larger the amount of data, the more advantageous the improved algorithm is. It shows that by deleting redundant information in the clustering process, the clustering process is gradually reduced. The interference can indeed improve the accuracy of clustering. This shows that when the amount of data is small, the improved algorithm is not desirable. It

can be seen from [Table 3] that the traditional K-means algorithm has a longer execution time than the improved algorithm, which is 2.19 times, 4.50 times, 6.88 times, 5.29 times, and 7.19 times respectively. The improved algorithm has a greater improvement in execution efficiency than traditional algorithms. The data shows that the larger the amount of data, the higher the efficiency.

5. Conclusions and prospects

The cluster analysis algorithm is one of the important tools for data analysis and one of the effective tools for extracting useful information from massive data. The essence of the clustering algorithm is to divide into different clusters according to the characteristics or attributes of each sample point in the data set, so that the sample points in the same cluster cluster are as similar as possible, and the sample points in different cluster clusters are as close as possible. However, this method can better discover the spatial distribution characteristics of sample points and the correlation between data objects. However, the existing clustering algorithms still have some shortcomings in some aspects, especially the commonly used K-means algorithm. In view of these shortcomings, this paper proposes an improved K-means dynamic clustering algorithm, which is verified by a series of experiments. The effectiveness of the proposed algorithm is shown. In recent years, with the continuous development of data mining technology, cluster analysis algorithms have received more and more attention. Many scholars have improved some traditional clustering algorithms or based on traditional clustering algorithms. A new algorithm concept is proposed, which solves some problems existing in traditional algorithms, such as the dependence on the initial clustering center selection, the threshold setting, and the sensitivity to the input order of sample points in the data set. The clustering algorithm is a step forward.

However, in real life, the data sets we need to process are not always standardized, and the shape of their clusters is not fixed. Therefore, further exploration and research is needed to propose a clustering method suitable for data sets of arbitrary shapes. Considering the feasibility of practical operations, when selecting the efficiency of the algorithm, more data sets of normal size are selected, so its effectiveness for large data sets cannot be confirmed, which will also be a problem that we need to continue to think and explore in the future. . In addition, due to the strong domain dependency of the clustering algorithm, how to obtain a universal method or framework requires more effort.

6. References

[1] Zhangt maf, “Improved rough k-means clustering algorithm based on weighted distance measure with Gaussian function,” International Journal of Computer Mathematics, vol.94, no.4, pp.663-675, **(2017)**

[2] Linxd maogj, “Distributed data stream clustering algorithm based on density grid,” Computer Engineering, vol.38, no.16, pp.70-73, **(2012)**

[3] Li Wu, Zhao Jiaoyan, and Yan Taishan, “Improved K-means clustering algorithm based on the average difference degree to optimize the initial clustering center,” Control and Decision, vol.32, no.4, pp.759-762, **(2017)**

[4] Zhou Shibing, Xu Zhenyuan, and Tang Xuqing, “Method for determining the optimal clustering number of K-means algorithm,” Journal of Computer Applications, vol.30, no.8, pp.1995-1998, **(2010)**

[5] Zuo Jin and Chen Zemao, “Anomaly detection algorithm based on improved K-means clustering,” Computer Science, vol.43, no.8, pp.258-261, **(2016)**

[6] Xu Dachuan, Xu Yicheng, and Zhang Dongmei, “Summary of K-means algorithm initialization methods,” Journal of Operations Research, vol.22, no.2, pp.111-114, **(2017)**

[7] Jainak, Murtymn, and Flynnpj, “Data clustering: areview,” A cm Computing Surveys, vol.31, no.3, pp.264-323, **(1999)**

[8] Jain, Du in, Mao J. “Statistical pattern recognition: are view,” IEEET rans action son Pattern Analysis & Machine Intelligence, vol.27, no.11, pp.1502-1502, **(2002)**