ISSN 2616-5775 Vol. 4, Issue 5: 37-40, DOI: 10.25236/AJCIS.2021.040506

K-means clustering algorithm: a brief review

Bao Chong

Data Science and Big Data Technology, Shanxi University of Finance and economics, Taiyuan, Shanxi, 030000, China

Abstract: *K-means clustering is a very classical clustering algorithm, and it is also one of the representatives of unsupervised learning. It has the advantages of a simple idea, high efficiency, and easy implementation, so it is widely used in many fields. However, K-means clustering also has some limitations, such as the number of clusters, the value of K is challenging to select, the selection of initial class center, the detection of outliers, and so on. This paper introduces the traditional K-means clustering algorithm and its improved method in detail. The advantages and disadvantages of the improved algorithm are analyzed, and the existing problems are pointed out. The development direction and trend of the K-means algorithm have been prospected.*

Keywords: *K-means algorithm, outliers, improved algorithm*

1. Introduction

As we all know, clustering is one of the most important tools in data science. Because the algorithm is simple and efficient, it has been widely used in various fields: for example, in bioinformatics, marketing department, computer vision, geostatistics, astronomy and horticulture [1]. The goal of clustering algorithm is to divide the data into several classes or clusters, and group the data according to the Euclidean distance according to the proximity between the data [2]. Taking neural network as an example, K-means clustering algorithm is usually called unsupervised feature learning [3].

Although the K-means clustering algorithm is widely used, it still has some defects: Firstly, if the selection of K value is inappropriate, it is easy to make the algorithm fall into local minimum [4], [5], [6], [7]; Secondly, the algorithm is also susceptible to outliers. If the outliers are used as the initial centroid, the algorithm's efficiency will be significantly reduced [8]; Thirdly, the clustering results are easily affected by noise and outliers, which makes the clustering results unsatisfactory [9].

2. Review of K-Means Clustering Algorithm

K-means algorithm is used as an iterative clustering algorithm. It takes the distance as the measurement standard, gives the K clusters in the data set, calculates the average value of the distance, and then gives the initial centroid. Each cluster is described by the centroid [4]. The goal is to form the disjoint groups of n data points $\{x1,x2,\ldots,xn\}$ into $k < n$ sets$\{S1,S2,\ldots,Sk\}$ to minimize the total average value (including the square distance from the point to the centroid). Therefore, the goal of optimization is to find:

$$\underset{S}{argmin}\frac{1}{n}\sum_{i=1}^{k}\sum_{x\in S_i}||x-\mu_i||^2$$

Where μ_i is the mean of points in S_i.

Assignment: assign each point to the same cluster as the nearest center to that point. That is: Step 1. Randomly select a set of initial sets as the initial centroid.

Step 2. Assign each point to the cluster with the same centroid closest to the point, that is, the following formula needs to be satisfied:

$$S_i^{(t)}=\{x:||x-\mu_i^{(t)}||^2\leq||x-\mu_j^{(t)}||^2\ \forall j,\ 1\leq j\leq k\}$$

$S_i \cap S_j=\Phi\ \forall i,j\leq k$. In other words, if a point is equidistant from multiple centroids, it can only be assigned to any one of them.

Step 3. The mean value of all objects in each category is used as the cluster center of the category to update the cluster center.

$$\mu_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \in S_i} x_j$$

Step 4: Judge whether the values of cluster center and objective function have changed. $\mu^{(t+1)} = \mu^{(t)}$ $\forall i \leq$ k.This means that if the maximum number of iterations is reached, the cluster allocation will not change on update. Otherwise, go to step 2.

3. Related Work

Chunhui yuan et al. [4] the selection of initialization K values. K-means clustering depends heavily on the selection of K-values, which helps to improve the efficiency of the algorithm. Unfortunately, K-values are generally difficult to define, because data clustering is closely related to the selection of K-values. Therefore, researchers first give a series of elbow calculation methods of K-value by using the square of the distance between the sampling point in each cluster and the cluster centroid, then determine the optimal number of clusters through the gap statistics algorithm, then make the model applicable and acceptable through the contour coefficient algorithm, and finally use the tree crown algorithm to optimize and accelerate the clustering algorithm.

Uri Stemmer [5] noted that solving K-means variables is also very important. Therefore, we focus on local private K-means clustering. The K-means algorithm depends on the selection of K values. The use of Euclidean helps improve the efficiency of the algorithm. Generally, the value should be $(1, \sqrt{n})$. Specify different K-values to perform any cluster analysis. Different K-values correspond to different clustering results. The traditional K-means clustering algorithm does not consider that the algorithm only converges to an inadequate local minimum. Based on the improved self-learning theory, the generalization ability of the algorithm can be effectively improved, and the clustering model can achieve optimal performance to prevent the influence of noise and outliers on the clustering results.

Peter o. olukanmi [6] and others are very sensitive to selecting initial cluster centers, resulting in local minimum values. Seed technology is introduced to prevent the selection of outliers in the initialization process, which significantly improves the practicability and efficiency of the algorithm. The core of solving the non-robustness of outliers, maintaining the structure of the K-means algorithm, and improving the calculation of cluster center is: if the distance between data points and the current cluster centroid exceeds the threshold, these data will be classified as outliers. The threshold adopts a reliable and straightforward distribution rule to produce the ability to detect and avoid anomalies automatically.

Kristina P. Sinaga [8] and others are committed to studying the cluster number of K-means clustering algorithm. They believe that the number of clusters can be found through the effectiveness index, but it should be independent of the clustering algorithm, and many clustering effectiveness indexes related to the clustering algorithm are proposed to assist. They first add a learning process to the K-means clustering algorithm, which can calculate the initial number of clusters without any initialization or parameter selection. Then, they adjust the deviation through the entropy penalty and create a new learning model to find out the actual number of clusters.

Guojun Gan et al. [9] mainly studied detecting outliers in clustering to make the algorithm more accurate. They use alternating optimization rules to ensure that the objective function of K-means clustering is optimal and extend the algorithm itself by introducing additional "clustering" to detect outliers. Numerical experiments show the effectiveness of the improved algorithm.

Mohiuddin Ahmed [10] and others are devoted to the initialization research of the K-means clustering algorithm. The improved K-means algorithm based on self-learning theory can help the K-means algorithm to establish an initial clustering model through competitive training subsets. After self-learning is added, the generalization ability of the algorithm is effectively improved, and the influence of noise and outliers on the clustering results is avoided. Another improved algorithm combines the improved cuckoo algorithm with the K- means algorithm, which not only obtains a better solution but also avoids falling into a local minimum.

Anant Joshi [11] and others based on the K-means cluster analysis of crimes to provide people with safety guarantees, help local police discover the possibility of crimes, and catch criminals in time. Through the K-means clustering data mining method, we can summarize each crime and the crime rate of each city, which helps the police evaluate the case and better respond to the next similar case. Different types of crimes have corresponding different types of data mining techniques to achieve the best results. The K-means algorithm has been well applied in this field with its unique data visualization capabilities.

Reza ghezelbash et al. [12] use GKC and traditional K-means clustering (tKMc) methods to generate a genetic K-means clustering algorithm to combine factor analysis and sample watershed modeling. The algorithm can handle complex geological features, has a reasonable prediction rate, and the prediction of the target is also very reliable. The algorithm has the characteristics of high efficiency and stability. It is a helpful tool for geological surveys.

Jothi, R. et al. [13] proposed a K-means deterministic initialization algorithm based on constrained double partition. The algorithm improves the K-means clustering algorithm, so that the initial value of the algorithm is no longer randomly selected, and effectively avoids the problem of falling into local minimum. At the same time, the improved algorithm has high efficiency and has good application prospects in genetic analysis. Comparing with several traditional clustering algorithms, it is found that the DK average runs faster, the convergence speed is better, and the clustering results are more ideal.

Chang Xia et al. [14] worried that the rapid development of big data technology would threaten the public's privacy, so they tried to solve the privacy protection problem of distributed K-means clustering through LDP. Based on the proposed basic framework, privacy enhancement technology is used to protect the user's sensitive data, which improves the practicability of the clustering results. The most important thing is to provide a distributed Kmeans algorithm according to the different levels of privacy requirements of users, which provides privacy protection and makes the algorithm more practical and reliable.

4. Tabular Comparison of Previous 5 Years Algorithms

Table 1

SL No	Year of Publication	Author Name	Algorithm	Dataset	Performance
4	2019	Chunhui Yuan Haitao Yang	Elbow Method Gap Statistic Silhouette Coefficient Canopy	Iris dataset	Elbow method shows better effect in results
6	2017	Peter O. Olukanmi Bhekisipho Twala	Classic K-means K-means ++ Improved K-means	Iris dataset Ruspini dataset	It evaluates the impact of the proposed modification on the classical K-means.
8	2020	KRISTINA P. SINAGA MIIN-SHEN YANG	Unsupervised K-means (U-K-means). X-means algorithm	RL-FCM dataset	It propose a learning framework for the k-means clustering algorithm.
9	2017	Guojun Gan Michael Kwok-Po Ng	KMOR algorithm NEO-K-means ODC algorithm	UCI machine dataset WBC dataset	KMOR algorithm is able to cluster data and detect outliers at the same time.
10	2020	Mohiuddin Ahmed Raihan Seraj Syed Mohammed Shamsul Islam	Constrained-k-means X-means algorithm	numerical and categorical dataset	It focused on the
11	2018	Anant Joshi A. Sai Sabitha Tanupriya Choudhury	Rapid miner method	Crime dataset COPLINK project dataset	Determine the location of frequent crimes to facilitate police investigation
12	2019	Reza Ghezelbash Abbas Maghsoudi Emmanuel John M. Carranza	TKMC algorithm GKMC algorithm	Geological dataset	Gkmc algorithm is an effective and robust tool for multi-element geochemical anomaly recognition.
13	2019	R. Jothi Sraban Kumar Mohanty ·Aparajita Ojha	MinMax algorithms DK-means PK-means K-means++ Bi-means	Breast, DLBCLA,ALB and Novartis dataset	K-means++ speeds up convergence of the clustering process and yields better clustering results

5. Conclusion

K-means clustering is a very classical clustering algorithm, and its application will become more and

ISSN 2616-5775 Vol. 4, Issue 5: 37-40, DOI: 10.25236/AJCIS.2021.040506

more common. This paper describes the shortcomings of traditional K-means clustering and the improved methods for these shortcomings in detail. K-means clustering has a broad application prospect and will face more challenges in the future. The improvement of K-means clustering is by no means limited to the direction in this paper. In the future, research can improve the ability of K-means clustering to deal with massive or multidimensional data sets. How to better use K-means to process the clustering of exponential data is a research direction.

References

[1] Kapoor, A. & Singhal, A. (2017). A comparative study of K-means, K-Means++ and fuzzy C-means clustering algorithms. In 3rd International conference on computational intelligence and communication technology (CICT), Ghaziabad, pp. 1–6.

[2] Arora, P.& Deepali, S. (2016).Analysis of K-Means and K-Medoids algorithm for big data. Procedia Computer Science, 78: 507–512.

[3] Liberty, E., et, al. (2016). An algorithm for online k-means clustering. In Workshop on Algorithm Engineering and Experiments (ALENEX), SIAM, 81–89.

[4] Yuan, C.&Yang, H. (2019). Research on K-value selection method of K-means clustering algorithm. Multidisciplinary Scientific Journal. 2: 226–235.

[5] Stemmer, U. (2020). Locally private k-means clustering. In Proceedings of the 2020 Symposium on Discrete Algorithms, pp.548-559.

[6] Olukanmi, P.O.& Twala, B.(2017). K-means-sharp: Modified centroid update for outlier-robust k-means clustering. In Proceedings of the 2017 Pattern Recognition Association of South Africa and Robotics and Mechatronics (PRASA-RobMech), Bloemfontein, South Africa, 30 November–1 December 2017; pp.14–19.

[7] Qureshi, M.N.& Ahamad, M.V.(2018). An Improved Method for Image Segmentation Using K-Means Clustering with Neutrosophic Logic. Procedia Computer Science, 132, 534–540.

[8] Sinaga, K. P. & Yang, M.-S. (2020). Unsupervised K-Means Clustering Algorithm. IEEE Access, vol. 8, pp. 80716–80727.

[9] Gan, G., & Ng, K. P..(2017). K -means clustering with outlier removal. Pattern Recognition Letters, 90, 8-14.

[10] Ahmed, M., et, al. (2020). The k-means Algorithm: A Comprehensive Survey and Performance Evaluation. Electronics, 9: 1295.

[11] Agarwal, J., et, al. (2013). Crime Analysis using K-Means Clustering. International Journal of Computer Application, 83(4): 1–4.

[12] Ghezelbash, R. et, al. (2020). Optimization of geochemical anomaly detection using a novel genetic K-means clustering (GKMC) algorithm. Computers & Geosciences, 134: 104335.

[13] Jothi, R. et, al. (2019). DK-means: a deterministic k-means clustering algorithm for gene expression analysis. Pattern Analysis and Applications, 22(2), 649-667.

[14] Xia, C. et, al. (2020). Distributed K-Means clustering guaranteeing local differential privacy. Computer Security, 90: 101699.