2020 IEEE 3rd International Conference on Electronics and Communication Engineering (ICECE) | 978-1-7281-9258-1/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICECE51594.2020.9353036

Review on the Research of K-means Clustering Algorithm in Big Data

Chen Jie
Institute of Modern Agricultural Science & Engineering
Tongji University
Shanghai, China
e-mail: chenjie18@yahoo.com.cn

Zhang Jiyue
Institute of Modern Agricultural Science & Engineering
Tongji University
Shanghai, China
e-mail: 1519421637@qq.com

Wu Junhui
Institute of Modern Agricultural Science & Engineering
Tongji University
Shanghai, China
e-mail: junhui_wu@163.com

Wu Yusheng*
Equipment Management Department
Xiamen Tobacco Industrial Co., Ltd.
Fujian, China
*Corresponding Author
e-mail: 21480276@qq.com

Si Huiping
Institute of Modern Agricultural Science & Engineering
Tongji University
Shanghai, China
e-mail: sihuiping@tongji.edu.cn

Lin Kaiyan
Institute of Modern Agricultural Science & Engineering
Tongji University
Shanghai, China
e-mail: ky.lin@163.com

***Abstract*—K-Means algorithm is an unsupervised learning algorithm, which is widely used in machine learning and other fields. It has the advantages of simple thought, good effect and easy realization. But with the rapid development of the Internet, the number of data collection terminals has increased rapidly, and people have entered the era of big data with information explosion. Therefore, the traditional K-Means algorithm exposes its limitations, such as: the initial value clustering number K in the algorithm is difficult to determine, the selection of the initial clustering center, the detection and removal of isolated points, etc. This article summarizes the improvement measures of the K-Means algorithm from many aspects, and analyzes the advantages and disadvantages of its improvement.**

Keywords-K-means; improvement; clustering center; K value; isolated points

I. INTRODUCTION

Clustering algorithm is an important branch of data mining algorithm [1]. Its core is to find valuable information hidden in data objects.Cluster analysis can classify samples according to their individual characteristics. There are many types of cluster analysis methods, mainly based on partition-based methods, hierarchical methods, density-based methods Model-based methods [2].

Among them, the K-Means algorithm is a classic clustering algorithm based on partitioning method, which is now widely used [3]. This algorithm was first used by MacQueen [4] in 1967.

Compared with other clustering algorithms, the K-Means algorithm has the advantages of better effects and simple ideas.

However, with the increasing number of Internet users and the increasing number of terminals for collecting data on the Internet, we have entered the era of big data with an explosion of information [5]. In addition, some big data computing frameworks are constantly emerging, such as Mapreduce and Spark. The K-Means algorithm began to expose some of its own limitations, such as the number k of clusters in the algorithm needs to be determined in advance, the initial clustering center is generated by random selection, and the effect of outliers on the clustering results [6]. In response to the above shortcomings, scholars in various fields have proposed different improved algorithms. This article first introduces the principle and calculation process of the traditional K-Means algorithm, and then presents the improvments of the K-Means algorithm based above-mentioned shortcomings.

II. K-MEANS ALGORITHM

A. The Principle of Traditional K-means Algorithm

The K-Means algorithm generally uses Euclidean distance as an indicator to measure the similarity between objects [7]. The similarity is inversely proportional to the Euclidean distance between objects. The greater the similarity between objects, the smaller the distance [8]. The algorithm needs to specify the initial cluster number k and initial cluster centers in advance, and continuously update the location of the cluster centers based on the similarity

between data objects and cluster centers. When its objective function converges, the clustering ends and the final result is obtained[9].

The Euclidean distance formula between the data object and the cluster center is:

$$d(x, C_i) = \sqrt{\sum_{j=1}^{k}(x_j - C_{ij})^2} \quad (1)$$

x is the object, C_i is the i-th cluster center, k is the dimension of the objects, x_j, C_{ij} are the j-th attribute of x and C_i.

The objective function of K-means clustering algorithm is:

$$E = \sum_{i=1}^{k}\sum_{x \in C_i}\|x - C_i\|^2 \quad (2)$$

E is the clustering effect.E value is large, it means that the final effect is poor, otherwise, it shows that the clustering effect is good.

B. The Steps of K-Means Clustering Algorithm

The specific description of the K-Means algorithm is as follows:

INPUT: Data set with n objects, the number of clusters K.

OUTPUT: K clusters that converge the objective function.

Begin:

a. Select k objects randomly from the data set as the initial center of clustering $C = \{C_1, C_2, \cdots, C_k\}$;
b. According to formula 2.1, calculate the Euclidean distance between the remaining data objects and the cluster center;
c. The data objects are assigned to clusters with cluster centers that are close to each other according to the value of Euclidean distance;
d. Calculate the mean of the data objects in each cluster as a new clustering center;
e. Calculate the E value of all clusters according to formula 2.2;
f. Until the evaluation function E value converges.

END

In the traditional clustering algorithm, its k initial center points are randomly selected. Therefore, it is sensitive to the initial value and easy to fall into the local optimal solution[10]. At present, the data set is getting larger and larger, the initial center point may not be locally representative, and there is a chance to select an isolated point as the initial center. Therefore, the cluster center will gradually deviate from the clustered dense area, and eventually the algorithm can only reach Local optimization is not possible to obtain ideal clustering results. Moreover, due to the randomness of the initial center point, it is need repeated calculations to determine the optimal clustering effect, therefore, the number of iterations increases. Therefore, many scholars continue to propose improved algorithms to overcome these shortcomings.

III. The Improvement of K-Means Clustering Algorithm

At present, the improvement methods of the K-means algorithm mainly focus on the following directions: the selection of the initial k value of the algorithm, the selection of the initial cluster center point, and the detection and removal of isolated points.

A. The Selection of Initial Value k

In the traditional K-Means algorithm, the number k of clusters must be determined in advance, but with the advent of the era of big data, the data set is gradually becoming larger.The value of k is difficult to determine, and the inappropriate selection of k will make the final clustering result into the local optimum. If the value of k is too large, the difference between different clusters will be very small. On the contrary, if the value of k is selected too small, it will cause the difference of the data objects in the same cluster to be too large.

Earlier, the literature [11] proposed that the value of k should be in the range of $(1, \sqrt{n})$. n is the number of object.But, as the data set becomes larger and larger, the range of K is still very large.

Reference [12] based on the idea of image segmentation, using the watershed algorithm divide the original data set into multiple regions to determine the optimal number of clusters k.

The algorithm first calculates the density of each data object:

$$p_i = \sum_{j=1}^{n}\left(-\beta\|x_i - x_j\|\right) \quad (3)$$

x_i is the data object, p_i is its intensity, and β is its empirical value.

Then draw the density value into a grayscale image, as shown in the figure, the density value image is M_1, the water level is M_2. The distribution of density value chart draws the density p_i from large to small as the vertical coordinate, the horizontal coordinate is the corresponding data point, and the water level line generates two regions P1 and P2 from x0 and x1. The center of each region is selected as the initial clustering center of the K-means algorithm, and then the number of optimal clusters k is determined.

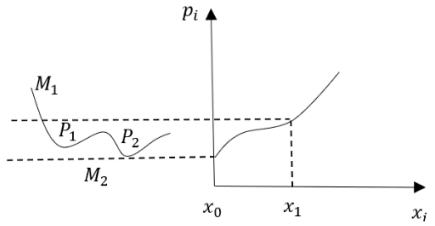


Figure 1. The distribution of density value

Reference [13] combined with the average diameter method to estimate of k. It can effectively represent the choice of initial clustering center. and can quickly obtain k value.

First calculate the average distance of all sample points, the formula is as follows:

$$Meandist = \frac{2}{n(n-1)} \times \sum_{i \neq j, i, j=1}^{n} d(x_i, x_j) \quad (4)$$

Where $d(x_i, x_j)$ is the Euclidean distance of the two samples, the formula is as follows:

$$d(x_i, x_j) = \sqrt{(x_{i1} - x_{j1})^2 + \cdots (x_{in} - x_{jn})^2} \quad (5)$$

Then select the two furthest points in data X and add them to the initial clustering center set $C, C = \{C_1, C_2\}$. Starting from $C_1$and C_2, if the distance is greater than the $Meandist$, use it as a new cluster center and join C, otherwise, do not consider looking for a new cluster center. and not continue to iteration.

After the values of the initial clustering centers *C* and *K* are set, if point *P* is closer to an initial clustering center, then point *P* belongs to the cluster,make $P.laber = I, i \leq k$. Calculated in the same cluster, also is the average of the same label point vector, a new clustering center. The iteration until all clustering center does not change, at the end of the algorithm.

Through the above steps, we can get the set *C*, the point in the set *C* is the initial clustering center point we need, and the number of the cluster center points in *C* is the size of *K*. Both of the above methods determine the *K* value by determining the initial center *C*, which requires many iterations and affects the calculation efficiency.

Reference [14] selects the *K* value corresponding to the inflection point as the optimal number of clusters according to the relationship between different *K* values and Sum of the Squared Errors (SSE). As shown in the Figure 2: when *K = 3*, the image has an inflection point. However, for the problem of unclear inflection point, literature [15] combines the exponential function, weight term, paranoid term and other parameters to determine the best *K* value.

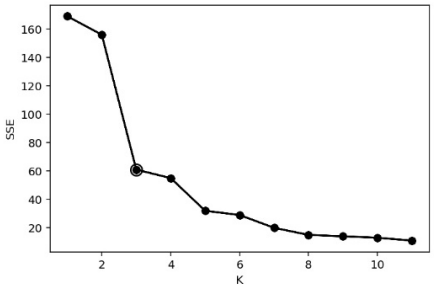


Figure 2. The relationship between SSE and K

In addition, there are still some scholars who determine the K value through the algorithm. Reference[16] sets the parameters of the AP (Affinity Propagation Clustering) algorithm, and uses the number of clusters generated by the AP algorithm as the the upper limit of the number of clusters.

B. The Selection of Initial Clustering Center

Through research, it is found that the K-Means algorithm has good local search ability, but it is easy to converge to the local optimal solution. The main reason is that the K-Means algorithm is sensitive to the selection of the initial center, which affects the stability, execution time and classification accuracy[17].

Reference [18] introduced a penalty term in the objective function to make it insensitive to the initial center. By optimizing the maximum and minimum distance algorithm, the high-density point with the largest distance product is selected as the current initial center point.
The objective function is:

$$P(U, Z) = \sum_{j=1}^{k} \sum_{i=1}^{n} u_{i,j} D_{i,j} + \gamma \sum_{j=1}^{k} \sum_{i=1}^{n} u_{i,j} \, log \, u_{i,j} \quad (6)$$

and

$$\sum_{j=1}^{k} u_{i,j} = 1, 0 < u_{i,j} \leq 1, i \in [1, n] \quad (7)$$

$U = [u_{i,j}]$ is n × k order matrix, $u_{i,j}$ represents the association degree of the ith object and the jth cluster. $Z = [z_1, z_2, \cdots, z_k]^T$ is cluster centers; $D_{i,j}$ is the dissimilarity measure between the i-th cluster center and the j-th object.

$$D_{i,\ j} = \sum_{l=1}^{m} (Z_{j,l} - x_{j,l})^2 \quad (8)$$

The relationship between the value of γ and Cluster number *K* is shown in the Figure 3. When γ increases to a certain degree, the k value is determined, and the clustering center is clear. However as γ continues to increase, the number of clusters *K* becomes smaller and smaller less than the given value of *K*. This shows that the negative entropy term plays an important role in the clustering process.

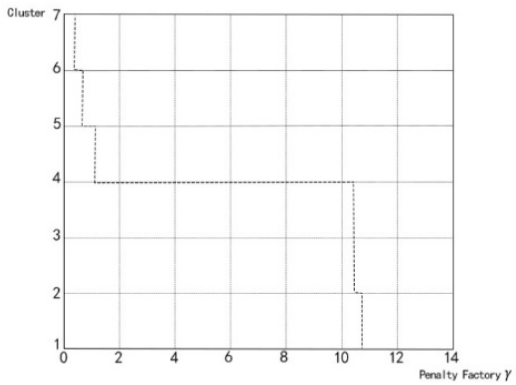
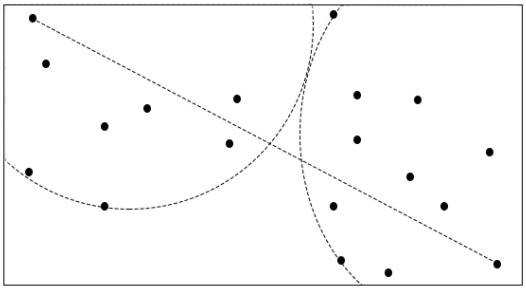


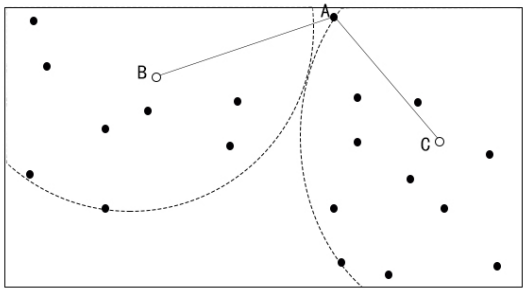
Figure 3. The relationship betweenγand K

Reference[19] first calculates the density of all data objects and finds the average density of the data set. This method treats data objects that are larger than the average density as high-density point sets. From the density point set, a data object with a high density value is selected as the first initial clustering center, and the remaining clustering centers are selected according to the principle of maximum distance from the previously selected clustering center and retrograde. Reference[20] select the center point based on the distance between the cluster center point and other center points, so as to avoid using two high-density points in a cluster as the initial cluster center. Reference[21] first select the point with the largest distance in the data set as the cluster center, and assign the remaining data objects to the corresponding clusters according to the distance of the cluster center point. Then continue to search for the point farthest from the cluster center as the next center point, and constantly update the

cluster center.As shown in Figure 4. In Figure (a), the data objects are assigned to the corresponding clusters according to the distance of the cluster center point. In Figure (b), point A is the farthest from the cluster centers B and C, so choose A for the new clustering center. In this cycle, until the number of cluster centers is *K*.



(a)



(b)

Figure 4. The clustering process of the data set

The above methods can reduce the sensitivity to the initial center point by improving the distance formula and avoid falling into the local optimal solution. But it will face the possibility of increasing the number of iterations.

Reference [22] proposes a clustering algorithm based on an improved particle swarm optimization algorithm, combining the K-means algorithm with strong local search ability and the particle swarm algorithm with strong search ability to improve the local search ability of the K-means mean algorithm. This algorithm can accelerate the convergence rate and reduce the dependence on the initial center selection.

C. The Detection of Outliers

With the huge amount of data, there are inevitably some isolated points, which are considered to be points away from the cluster center. If the outliers in the data set are used as the initial clustering center, the clustering result will fall into the local optimal value and affect the clustering effect[23].

Reference[24] proposes an improvement to the traditional K-means algorithm based on the grid density of data objects. Assuming that the data set has m-dimensional density, the length of each dimensional grid is:

$$GL_J = \frac{Jmax - Jmin}{k+1} \tag{9}$$

K is the number of clusters, and J_{max} and J_{min} are the maximum and minimum values of each dimension *J*. Then take GL_J of each dimension as the radius, and the grid density of the target object is the number of data objects within the range. If the grid density of the data object density is less than *Mins*, which is the density threshold, the point is considered as an outlier and should be eliminated from the data set. After the outliers are removed, the values in each dimension j are sequentially sorted, and the data after each sorting is divided into *K* segments, and the average value of each segment is obtained$\{p_{j1},\ p_{j2},\ p_{j3},\ \cdots,\ p_{jk}\}$,and then based on the average get initial cluster centers $\{C_1,\ C_2,\ \cdots,\ C_k\}$. Among them, $\{p_{j1},\ p_{j2},\ p_{j3},\ \cdots,\ p_{jk}\}$, $C_k = \{p_{1k},\ p_{1k},\ \cdots,\ p_{1k}\}$. Each data object is assigned to the corresponding cluster according to the distance of the cluster center, and then the cluster center of each cluster is recalculated, and the iteration is continuously updated until the cluster center no longer changes.

IV. OTHER IMPROVEMENTS OF K-MEANS ALGORITHM

With the continuous development of artificial intelligence, more and more algorithms have been proposed, and the most applied is the BP neural network. At present, many scholars combine neural networks with K-means algorithm to make them better applied to big data analysis. The BP neural network was proposed by the Rumelhart and McCelland research groups in 1986[25]. It is a multi-layer network backpropagating according to errors [26].

Reference[27] uses BP neural network to determine k value and initial clustering center. First use k-means clustering to generate k clusters and the cluster centers of each cluster, and record the distance between the cluster center and the sample point to its cluster center to get the distance matrix $d_{pi}(p = 1, \cdots, n; i = 1, \cdots, K)$, writing the initial center as $H = \{H_1, H_2, \cdots, H_k\}$ and Using this clustering to initialize the weight matrix of the neural network. Then the algorithm uses k clustering centers as hidden nodes and defines the deletion rules of hidden points. Finally, the algorithm uses the weight matrix to repeat the forward and backward calculations, modify the weights until the system error is small enough, and output the result.

Reference [28] combines genetic algorithm with K-means algorithm to improve the clustering efficiency and accuracy of K-means algorithm. The algorithm first uses Sorted-Neighborhood Method (nearest neighbor sorting algorithm) to deduplicate the original data and then normalize it, and calculate the Euclidean distance formula of the data objects in the data set. If the distance between each data object in the data set and the target point is within avg, the data object is considered to be the neighboring point of the target point. Then count the number of neighboring points and arrange them in descending order. Then, take the first *K* data objects as the initial clustering center. The genetic algorithm is used to select the roulette pair gambling area according to the size of the individual's fitness, and perform cross operation and mutation operation to eliminate the useless attribute features in the data set. If the maximum number of iterations is reached, the new population and the optimal result are

output, otherwise the genetic algorithm is used to continue the iteration.

V. CONCLUSION

Now, The K-means algorithm is widely used because of its simplicity, but how to make it more compatible with the development of the era of big data still faces very big challenges. At present, the main problem faced by clustering algorithms is that it takes a lot of time to calculate massive data. Many improved algorithms spare no effort to spend huge time costs in order to obtain accurate results, which is undesirable. In the future, how to reduce the time complexity of the K-means algorithm and improve our clustering effect still needs further optimization.

REFERENCES

[1] Ma R , Angryk R . 2017. Distance and Density Clustering for Time Series Data[C]// IEEE International Conference on Data Mining Workshops. IEEE.

[2] Liu H , Yang T . 2010. Computational Verb Clustering Algorithm and Its Applications[J]. International journal of computational cognition, 8(1-2):37-44.

[3] Mahmud M S , Rahman M M , Akhtar M N . 2012.Improvement of K-means clustering algorithm with better initial centroids based on weighted average[C]// International Conference on Electrical & Computer Engineering.

[4] Jiang S , Ferreira J , Gonzalez M C . 2017.Activity-Based Human Mobility Patterns Inferred from Mobile Phone Data: A Case Study of Singapore[J]. IEEE Transactions on Big Data.

[5] Dong-Hua Y . 2008.On the Internet and Democracy in China and the West——A Review on the Latest Literatures Related[J]. Journal of Sichuan University of Science & Engineering(Social Sciences Edition).

[6] Ma L, Gu L, Li B, et al. 2015. An improved K-means algorithm based on mapreduce and grid[J]. International Journal of Grid & Distributed Computing, 2015, 8(1).

[7] Visalakshi, N. K. , & Suguna, J. . (2009). K-means clustering using Max-min distance measure. Fuzzy Information Processing Society, 2009. NAFIPS 2009. Annual Meeting of the North American. IEEE.

[8] Abou-Moustafa K. 2016.What Is the Distance Between Objects in a Data Set?: A Brief Review of Distance and Similarity Measures for Data Analysis[J]. IEEE pulse, 2016, 7(2): 41-47.

[9] Saroj T. Kavita, R. 2016. Review: study on simple k mean and modified K mean clustering technique[J]. Int. J. Sci. Eng. Comput. Technol, 2016, 6(7): 279-281.

[10] Hung C H, Chiou H M, Yang W N. 2013.Candidate groups search for K-harmonic means data clustering[J]. Applied Mathematical Modelling, 2013, 37(24): 10123-10128.

[11] Rezaee M R, Lelieveldt B P F, Reiber J H C. 1998. A new cluster validity index for the fuzzy c-mean[J]. Pattern recognition letters, 1998, 19(3-4): 237-246.

[12] Wang X, Jiao Y, Fei S. 2015. Estimation of Clusters Number and Initial Centers of K-Means Algorithm Using Watershed Method[C]//2015 14th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES). IEEE, 2015: 505-508.

[13] Zhao Y, Zeng B. 2017.An improved k-means algorithm based on average diameter method[C]//AIP Conference Proceedings. AIP Publishing LLC, 2017, 1864(1): 020054.

[14] Weiqing C, Yanhong L U. 2015.Adaptive clustering algorithm based on maximum and minimum distances, and SSE[J]. J. Nanjing Univ. Posts Telecommun, 2015.

[15] Jianren W, Xin M A, Ganglong D. 2019.Improved K-means clustering k-value selection algorithm[J]. Computer Engineering and Applications, 2019, 55(8): 27-33.

[16] Yu S, Tranchevent L, Liu X, et al. 2011. Optimized data fusion for kernel k-means clustering[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 34(5): 1031-1039.

[17] Panapakidis I P, Christoforidis G C. 2017.Implementation of modified versions of the K-means algorithm in power load curves profiling[J]. Sustainable Cities and Society, 2017, 35: 83-93.

[18] Li M J, Ng M K, Cheung Y, et al. 2008.Agglomerative fuzzy k-means clustering algorithm with selection of number of clusters[J]. IEEE transactions on knowledge and data engineering, 2008, 20(11): 1519-1534.

[19] Xiong C, Hua Z, Lv K, et al. 2016.An Improved K-means text clustering algorithm By Optimizing initial cluster centers[C]//2016 7th International Conference on Cloud Computing and Big Data (CCBD). IEEE, 2016: 265-268.

[20] Du X, Xu N, Zhou C, et al. 2017.A density-based method for selection of the initial clustering centers of K-means algorithm[C]//2017 IEEE 2nd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC). IEEE, 2017: 2509-2512.

[21] Tanir D, Nuriyeva F. 2017. On selecting the Initial Cluster Centers in the K-means Algorithm[C]//2017 IEEE 11th International Conference on Application of Information and Communication Technologies (AICT). IEEE, 2017: 1-5.

[22] Xiaoquan C, Jihong Z. 2012.Clustering Algorithm Based on Improved Particle Swarm Optimization[J]. Journal of Computer Research and Development, 2012: S1.

[23] Lin L, Pan X, Zhang L, et al. 2016.The K-means clustering algorithm for wind farm based on immune-outlier data and immune-sensitive initial center[J]. Proceedings of the CSEE, 2016, 36(20): 461-5468.

[24] Fan Z, Sun Y. 2016.Clustering of college students based on improved K-means algorithm[C]//2016 International Computer Symposium (ICS). IEEE, 2016: 676-679.

[25] Yang Z R, Platt M B, Platt H D. 1999.Probabilistic neural networks in bankruptcy prediction[J]. Journal of Business Research, 1999, 44(2): 67-74.

[26] Ayaz E. 2014. A review study on mathematical methods for fault detection problems in induction motors[J]. Balkan Journal of Electrical and Computer Engineering, 2014, 2(3).

[27] Shi H, Xu M. 2018.A data classification method using genetic algorithm and K-means algorithm with optimizing initial cluster center[C]//2018 IEEE International Conference on Computer and Communication Engineering Technology (CCET). IEEE, 2018: 224-228.

[28] Wenjun Z. 2020. Analysis of K-means clustering algorithm based on SOM and BP networ[J]. Computer knowledge and technology 16.09(2020):24-26. doi:10.14004/j.cnki.ckt.2020.0994.