DOI: 10.3969/j.issn.1673-5188.2017.S2.007
http://kns.cnki.net/kcms/detail/34.1294.TN.20171221.1748.004.html, published online December 21, 2017

An Improved K-Means Algorithm Based on Initial Clustering Center Optimization

LI Taihao[1,2], NAREN Tuya[1], ZHOU Jianshe[1], REN Fuji[3], and LIU Shupeng[4]
(1. Beijing Advanced Innovation Center for Imaging Technology, Capital Normal University, Beijing 100048, China;
2. Flatley Discovery Lab, Boston 02129, USA;
3. Department of Information Science & Intelligent Systems, University of Tokushima, Tokushima 7708506, Japan;
4. School of Communication and Information Engineering, Shanghai University, Shanghai 200072, China)

Abstract

The K-means algorithm is widely known for its simplicity and fastness in text clustering. However, the selection of the initial clustering center with the traditional K-means algorithm is some random, and therefore, the fluctuations and instability of the clustering results are strongly affected by the initial clustering center. This paper proposed an algorithm to select the initial clustering center to eliminate the uncertainty of central point selection. The experiment results show that the improved K-means clustering algorithm is superior to the traditional algorithm.

Keywords

clustering; K-means algorithm; initial clustering center

1 Introduction

Cluster analysis is an important method of data partitioning and data grouping in data mining, so it has been widely used in statistics, machine learning, spatial databases, biomedicine, and marketing [1]–[4]. The clustering algorithms can be divided into partitioning methods, hierarchical methods, density-based methods, grid-based methods and model-based methods [5], [6]. The K-means clustering algorithm is a partitioning method proposed by Mac Queen [7], which divides the data into a predetermined number of clusters k on the basis of minimizing the error function. The algorithm is the most commonly used algorithm with high-speed clustering, easy implementation, and efficient classification of large data sets. However, the K-means algorithm has several drawbacks [8]. For example, the initial clustering center selection of the traditional K-means algorithm is random, and different initial clustering centers make the clustering results different, which results in the failure of acquiring effective clustering results [9]. So the selection of a reasonable initial clustering center, for accurate, stable and effective clustering results, is an important research topic. In this paper, based on the traditional K-means clustering algorithm, an initial center point selection algorithm is proposed based on the clustering algorithm to improve the efficiency and stability of clustering.

The K-means algorithm is based on the principle of minimizing clustering performance, usually by minimizing the sum of squares of errors for each sample point in the data set to the class center. The basic idea is to select k data objects as the initial clustering center. The iterations are used to divide the data objects into different clusters with the large similarity intra the cluster and the small similarity among the different clusters. The k value is given first and k data objects are randomly selected as the initial clustering center in data set. The algorithm takes the following steps [10]:

1) K objects are selected randomly from n data objects as the initial clustering center.
2) The distance between the remaining $n-k$ objects and the central objects are calculated, according to the mean of each cluster object; the corresponding objects are regrouped according to the minimum distance and each object is assigned to the closest class.
3) The clustering centers of k clusters are recalculated.
4) Steps 2) and 3) are repeated until the criterion function no longer changes.
5) K clusters are calculated.

The K-means algorithm is to find the k clusters that minimize the squared error function. The specific definition is as follows:

$$E=\sum_{i=1}^{k}\sum_{p\in C_i}\left\|p-m_i\right\|^2, \tag{1}$$

where E is the sum of the squared errors of all the objects in

the database, p is the point in the space, which represents the given data object, m_i is the average of the clusters C_i (p and m_i are multidimensional). This rule makes the clustering resulted as compact and independent as possible.

2 Improved K-Means Clustering Algorithm

2.1 Principle of the Improved K-Means Algorithm

We first detect the isolated points. The traditional K-means algorithm ignores isolated points. The isolated points are calculated based on distance, and the points are relatively sparsely distributed and are far apart from the cluster center. The object is those with the largest distance from the nearest neighbor data set. These points give a great impact on the clustering effect of centroid calculation. Here we scan the data set once, calculate the distance between each data object and other objects, including the total distance and the mean distance. If the total distance of one object is greater than the mean of the total distance, the data object is treated as an isolated point and removed from the data set to the isolated point set. The above process is repeated until all the isolated points are removed. Finally we get the initial collection of new clusters of isolated points.

The distance between all the sample points in the new data set is then calculated. The K-means algorithm is a clustering algorithm based on the partitioning method, and adopts Euclidean distance as the evaluation index of similarity, that is, the closer the two samples are, the better the similarity is. The Euclidean distance is as follows:

$$d\left(X_i,X_j\right)=\sqrt{(x_{i1}-x_{j1})^2+(x_{i2}-x_{j2})^2+\ldots+(x_{in}-x_{jn})^2}, \quad (2)$$

where x_{in} is the n-dimensional data of the data object X_i, and x_{jn} is the n-dimensional data of the data object X_j.

The average distance between all sample points is next calculated, defined as follows:

$$md=\frac{1}{A_n^2}\sum d\left(X_i,X_j\right),1\leqslant i\leqslant n,1\leqslant j\leqslant n, \quad (3)$$

where A_n^2 is the arranged number of 2 random points in n sample points.

We finally record the sample point X_j with the distance less than the average distance md from the sample point X_i, and call it the neighboring point and calculate all the neighboring points of X_i. When the neighboring points of all the sample points are calculated, they are sorted by the number of adjacent numbers from high to low. Then the sample point with the largest number of neighboring points is the first cluster center and then continues to look down. If the sample point with the nearest number of points 2 is the neighboring point of the existing cluster center, it will be ignored until a sample point is found not the nearest point of the existing cluster center; this point is treated as the second cluster center. This process is repeated until the initial cluster center is found.

2.2 Process of the Improved K-Means Algorithm

The input includes u data sets containing n objects and k initial cluster centers.

The output is the clustering results of the k-th cluster.

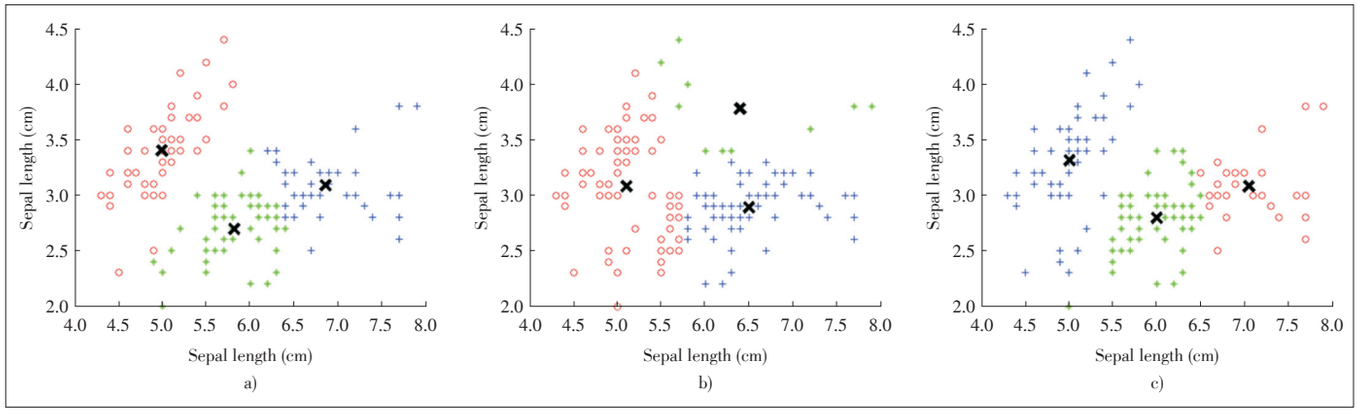
1) Initialization: t = 0, where i, j is the set of isolated points, t is the number of isolated points, m is the number of data in the original data set, and n is the number of data objects removing the isolated point.
2) Calculate the distance d between X_i and X_j of all data objects in m and calculate the total distance D of each data object in m and the mean total distance H of each data object.
3) Scan m and compare the distance D and H of each data point. The data point with $D>H$ is the isolated point $t+1$, the point is add to gl and removed from the data set U; the value of t is got until all the isolated points are removed, and $n = m- t$. Finally, a new data set is constructed: $U^{'}=U-gl$.
4) Calculate and save the distance d between every two sample points in the new data set according to (2).
5) Calculate the mean distance md between the sample points according to (2).
6) Record the sample points smaller than md from the sample points, make them the neighboring points, calculate the number of neighboring points for each sample point, and store them in the proximity table.
7) The sample points are arranged from high to low according to the number of neighboring points, and the nearest sample point is the first cluster center point.
8) In turn, the proximity point table should be found if the point is already the initial clustering center and the adjacent point is ignored; the proximity table will be obtained until a sample point that is not the center point of the existing cluster is found. This point is used as the second cluster center.
9) Repeat Step 8) until the k-th cluster center is found.
10) From the k-th clustering center, the clustering results are calculated by the K-means clustering algorithm.

3 Results and Discussion

The traditional K-means algorithm and the improved algorithm are used to test the two-dimensional data of the Iris dataset (after removing the same data, there are 132 two-dimensional data). The Iris dataset contains 4 attributes and 150 data objects. The cluster number of clusters is set to k = 3 for the three categories.

The traditional K-means algorithm randomly selects three sets of initial clustering centers to cluster the sample data. The clustering results are shown in **Fig. 1**.

The traditional K-means algorithm selects the clustering



▲Figure 1. The instable clustering results of the traditional K-means algorithm.

center randomly, so the clustering results are different and instable (Figs. 1a, 1b and 1c).

The improved algorithm selects an initial clustering center, and the data sets u equal to (6.6000, 3.0000), (6.3000, 2.9000) and (5.8000, 2.8000). **Fig. 2d** shows the clustering results.

The improved algorithm is optimized for the stochastic selection of the initial clustering center, so that the initial clustering center is unique for the same data, which eliminates the instability of the traditional algorithm results. The optimal clustering center makes the iterations number greatly reduced than that of the traditional algorithm, that is, the efficiency of the improved algorithm is better.

Fig. 2 compares the clustering results of the traditional algorithm and the improved algorithm.

4 Conclusions

The K-means algorithm has fast computation and small resource consumption, and is widely used as a clustering algorithm. The clustering results of the traditional K-means algorithm are instable with the initial clustering center selected randomly. This paper proposes an initial selection algorithm to

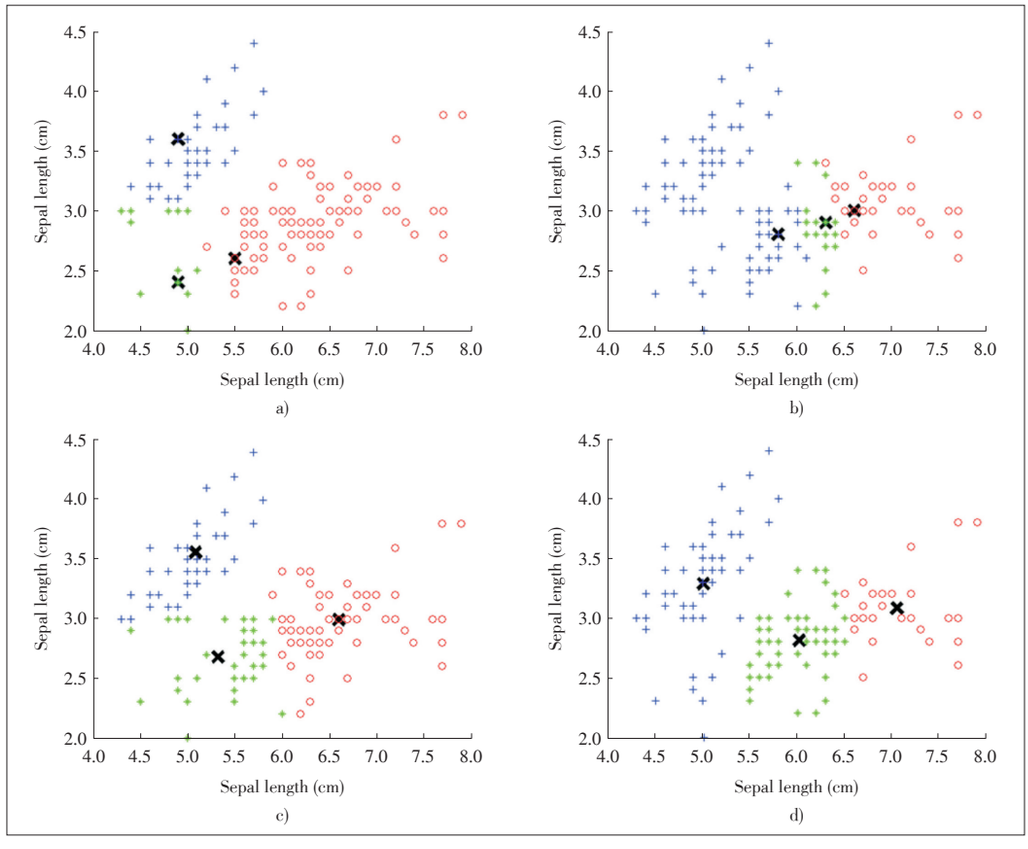


Figure 2. ▶ The clustering results of the traditional algorithm and the improved algorithm. a) The results of the traditional algorithm with first clustering; b) the results of the improved algorithm with the first clustering; c) the results of the traditional algorithm with the seventh clustering; d) the results of the improved algorithm with seventh clustering.

improve the traditional K-means algorithm. The experiment results show that the algorithm is feasible to enhance the clustering efficiency by optimizing the initial clustering center. However, the algorithm proposed in this paper will increase the time consumption in finding the distance between isolated points and calculating the sample points.

References

[1] J. Zhang, W. Jiang, R. Wang, and L. Wang, "Brain MR image segmentation with spatial constrained K-mean algorithm and dual-tree complex wavelet transform," *Journal of Medical Systems*, vol. 38, article 93, 2014. doi: 10.1007/s10916-014-0093-2.
[2] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651–666, Jun. 2010. doi: 10.1016/j.patrec.2009.09.011.
[3] T. Santhanam and M. S. Padmavathi, "Application of K-means and genetic algorithms for dimension reduction by integrating SVM for diabetes diagnosis," *Procedia Computer Science*, vol. 47, pp. 76–83, 2015. doi: 10.1016/j.procs.2015.03.185.
[4] P. Chévez, D. Barbero, I. Martini, and C. Discoli, "Application of the k-means clustering method for the detection and analysis of areas of homogeneous residential electricity consumption at the Great La Plata region, Buenos Aires, Argentina," *Sustainable Cities and Society*, vol. 32, no. 32, pp. 115–129, Jul. 2017. doi: 10.1016/j.scs.2017.03.019.
[5] J. Zhang, "K-means clustering algorithm research and application," M.S. thesis, Wuhan University of Technology, Wuhan, China, 2007.
[6] H. Khanali and B. Vaziri, "A Survey on clustering algorithms for partitioning method," *International Journal of Computer Applications*, vol. 155, no. 4, pp. 20–25, Dec. 2016.
[7] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, USA, 1965–1966, vol. 1, pp. 281–297.
[8] S. K. Popat and M. Emmanuel, "Review and comparative study of clustering techniques," *International Journal of Computer Science and Information Technologies*, vol. 5, no. 1, pp. 805–812, 2014.
[9] M. Emre Celebi, H. A. Kingravi, and P. A. Vela, "A comparative study of efficient initialization methods for the k-means clustering algorithm," *Expert Systems with Applications*, vol. 40, no. 1, pp. 200–210, Jan. 2013. doi: 10. 1016/j.eswa.2012.07.021.
[10] N. Shi, X. Liu, and Y. Guan, "Research on k-means clustering algorithm: an improved k-means clustering algorithm," in *Third International Symposium on Intelligent Information Technology and Security Informatics*, Jinggangshan, China, 2010, pp. 63–67. doi: 10.1109/IITSI.2010.74.

manuscript received: 2017-07-28

Biographies

LI Taihao (litaihao@heartdynamic.cn) received the M.S. and Ph.D. degrees in information system engineering from University of Tokushima, Japan in 2003 and 2006, respectively. During 2006–2011, he was a postdoc researcher at Harvard University, USA. He is now a professor with Beijing Advanced Innovation Center for Imaging Technology, Capital Normal University, China. His research interests include affective computing, natural language processing, and artificial intelligence.

NAREN Tuya (nrty0910@163.com) received her M.S. degree from Jilin University, China. She is a Ph.D. student at Capital Normal University, China. Her research interests include language intelligence, linguistics, and text affective computing.

ZHOU Jianshe (zhoujianshe@solcnu.net) is a board member of Chinese Linguistics Association, a director member of Expert Committee of the Linguistics Committee of Beijing, and the deputy director of Beijing Linguistics Association. He is the vice president of Capital Normal University, China and a professor with Beijing Advanced Innovation Center for Imaging Technology there. His research interests include linguistics, neuroscience, and algorithms.

REN Fuji (ren@is.tokushima-u.ac.jp) received his B.E. and M.E. degrees from Beijing University of Posts and Telecommunications, China in 1982 and 1985, respectively. He received his Ph.D. degree in 1991 from Hokkaido University, Japan. He is a professor at the Faculty of Engineering, the University of Tokushima, Japan. His research interests include natural language processing, affective computing, artificial intelligence, and language understanding.

LIU Shupeng (liusp@i.shu.edu.cn) received his Ph.D. degree in 2007 from Shanghai Jiaotong University, China. He is an associate professor with School of Communications and Information Engineering, Shanghai University, China. His research interests include signal processing, Raman spectra, and image processing.