Research on Network Public Opinion Analysis Based on Improved K-means Algorithm

Mingzhong Qi, Xiongwei Li
Shijiazhuang Campus, Army Engineering University of PLA
Shijiazhuang, Hebei, 050003, China
qkl2018@foxmail.com

Abstract. In the network public opinion analysis, the K-means algorithm is sensitive to the initial cluster center and outliers. In order to solve this problem, an improved K-means algorithm for optimizing the initial cluster center is proposed. In the improved K-means algorithm, firstly, the outliers of each data object in the dataset are calculated. And the sampling factor α is put forward to obtain the candidate initial center point set. Then, according to the idea of the maximum and minimum distance, k data objects are selected as initial clustering centers from the candidate initial center point set. Finally, a case based on the actual network text data is studied. The case study results show that the new proposed improved K-means algorithm is better than K-means algorithm and K-means++ algorithm in clustering effect. And it is more suitable for network public opinion analysis.

Keywords: Public opinion analysis, K-means algorithm, improved K-means algorithm, Initial cluster center.

1. Introduction

Public opinion is the subjective reflection of the public to the social reality in a certain period and within a certain range, and it is the comprehensive expression of the group's attitude, thought, emotion and requirements [1]. With the rise of social network platforms, such as Facebook, Weibo, WeChat and Twitter, network information has become the main field to reflect social public opinions. And network public opinion analysis has become an important part of public opinion analysis. As a result, network public opinion analysis has become an important part of public opinion analysis. For example, the famous Ceron successfully predicted the result of the French election in 2012 by analyzing the data of netizens' emotional tendency on Twitter.

Most of the network data exist in an unstructured form, such as tables, web pages, text, etc., which need to be mined to get directly available information. Clustering analysis has the advantage of no prior knowledge and it is one of the important methods of data mining. This method has attracted more and more attention in recent years.

The K-means algorithm is the most common clustering analysis method. It has many advantages, such as simple algorithm, fast convergence speed and effective processing of large data sets [2]. However, the method of randomly selecting the initial center points makes the K-means algorithm easy to fall into the local optimal solution, and the clustering result is unstable. In response to this shortcoming, many scholars have proposed some methods to optimize the selection of initial center point. For example, Arthur et al. [3] proposed the K-means++ algorithm. In K-means++ algorithm, first of all, a data object is randomly selected from data set as the first initial center point; then, the Euclidean distance between the remaining data objects and the initial center point is calculated, and the data object with the largest distance value is selected as the second initial center point. Repeat the above process until k initial center points are selected.

At the same time, the K-means algorithm is susceptible to the influence of outliers. The existence of outliers often affects the selection of initial center point and leads to poor clustering results. To solve this problem, Leng et al. [4] proposed to use the distance-based outlier detection algorithm to find outliers in the data set to avoid selecting outlier as the initial center point. In the study of this method, some of them solved the problem of randomly selecting the initial center point of the

K-means algorithm, but ignored the influence of outliers on the results, such as literature [3, 5, 6]. Although some of them avoid the influence of outliers, they still randomly select the initial center points in the end, such as literature [4].

In order to solve the two shortcomings of K-means algorithm, this paper proposes an optimized K-means initial center selection algorithm based on outlier factor and maximum and minimum algorithm, namely the improved K-means algorithm. Case study shows that the improved K-means algorithm has better clustering effect and it is more suitable for online public opinion analysis than K-means algorithm and K-means++ algorithm.

2. K-means Algorithm

The core idea of the K-means algorithm is as follows: let the data set contains *n* data objects, randomly selects *k* objects as the initial clustering centers from the *n* data objects, all remaining objects are allocated to the most similar cluster according to the similarity between each remaining data object and *k* clustering centers. Then, the average value of all objects in each cluster is taken as the new clustering center for the next iteration. The above process is repeated until the center of the cluster is no longer changed, and the convergence condition of the criterion function is met. The criterion function is as in formula (1):

$$E = \sum_{i=1}^{k} \sum_{x \in C_i} \left| x - \overline{x_i} \right|^2 \tag{1}$$

Where, E is the sum of the squared error of all objects and the clustering center, the value of E reflects the quality of clustering results. C_i is the i-th cluster, x_i is the clustering center of cluster C_i, and k is the number of clusters.

The selection of the initial clustering center of K-means algorithm is random [7], which is easy to cause unstable clustering results and fall into local optimization. In addition, because the samples may contain noise or outliers, the clustering center is sensitive to them, and it is easy to lead to poor clustering results.

3. Improved K-means Algorithm

The improved K-means algorithm is mainly divided into three steps for the initial center point selection process: a) calculating the outlier factor, b) generating a candidate set of the initial center point, c) selecting the initial center point. The three steps are detailed as follows.

3.1 Calculating Outlier Factor

When the K-means algorithm randomly selects the initial center point, if there are outliers or noise in the data set, it is easy to make the initial center point deviate from the optimal clustering center and reduce the clustering precision. The so-called outliers are data points that deviate significantly from other objects in the dataset. The outlier detection method can find that a small part of abnormal data set deviates from most data, and this method can avoid the influence of outliers on clustering [8].

In this paper, local outlier factor (LOF) algorithm [9] is used to calculate outlier factors of data objects. The value of outlier factor reflects the deviation degree of data object. The larger the outlier factor is, the higher the deviation degree is. On the contrary, the smaller the outlier factor is, the more data around the data object, the more likely the data object is to be the center point. Therefore, the improved K-means algorithm selects those data objects with small outlier factors as the initial center point.

The following is the calculation process of the LOF algorithm:

Step 1 Construct the *k*-th distance range of the object *w*.

The *k*-th distance range of the object *w* is the set of all objects with distance less than or equal to the *k*-th distance of the object *w*, the calculation formula is as follows:

$$N_k(w) = \left\{ o \in Q / \{w\} \middle| d(w,o) \le k\text{-}dis(w) \right\} \tag{2}$$

Where, Q is data set, $d(w,o)$ represents the Euclidean distance between data object w and data object o, and $k\text{-}dis(w)$ is the k-th Euclidean distance of object *w*.

Step 2 Compute the locally accessible density of the object *w*.

The *k*-th distance range of the object *w* is obtained from step 1. Thus, the locally accessible density of *w* can be calculated by formula (3) as follows:

$$lrd_k(w) = 1 \Big/ \left[\frac{\sum_{o \in N_k(w)} \max\left\{k\text{-}dis(o), d(w,o)\right\}}{\left|N_k(w)\right|} \right] \tag{3}$$

Step 3. Calculate the outlier factor of the object *w*

The outlier factor of *w* is calculated by combining formula (2) and (3). The formula is as follows:

$$LOF_k(w) = \frac{\sum_{o \in N_k(w)} \frac{lrd_k(o)}{lrd_k(w)}}{\left|N_k(w)\right|} \tag{4}$$

By analyzing formula (4), it can be seen that the value of $LOF_k(w)$ reflects the average distribution density of spatial points within $k\text{-}dis(w)$. It is easy to find that the larger the LOF value of an object, the greater the probability that the object is an outlier.

Integrating the above three steps, the outlier factor of each object in the data set can be calculated. And the distribution of outlier points can be known according to the value of outlier factor. Therefore, the influence of outliers can be avoided when selecting the initial center point.

3.2 Generating Candidate Set of Initial Central Points

According to the previous analysis, we know that data objects with small outlier factors should be considered when selecting the initial center point. The steps of the improved K-means algorithm to generate the candidate initial central point set are as follows:

Step 1 The outlier factor $LOF_k(w)$ of each data object *w* in data set Q is calculated according to section III.A.

Step 2 The data set *Q* is sorted in ascending order according to the value of outlier factor $LOF_k(w)$, and denoted as Q_L.

Step 3 On, the in front αn data objects on Q_L are selected as the candidate initial central point set, denoted as Q_{CL}. Where $0 < \alpha \le 1$ and n is the size of the sample set.

The value of the K-means algorithm is too small, which may cause the initial center point to be concentrated in one cluster, and the clustering effect is poor. If the value is too large, it is easy to select a poor initial center point. The specific value is usually determined by experiment, and this article makes it equal to 0.5.

The improved K-means algorithm introduces a sampling parameter α in the third step. If the value of α is given too small, the initial center point may be concentrated in a cluster, and the clustering effect is poor. If the value of α is given too large, it is easy to select a poor initial center point. The specific value of α is usually determined by experiments. In this paper, it is set equal to 0.5.

3.3 Determining the Initial Center Points

The improved K-means algorithm uses the data object as far as possible as the initial cluster center point, which can avoid the situation that the initial cluster center is too close. As a result, the data set can be divided is relatively good. In this paper, the maximum and minimum method [10] is used to determine the initial cluster center. The initial center point of the maximum and minimum algorithm is shown in formula (5).

$$C_i = \max\left\{Dis_j; j = 1, 2, \cdots, n\right\} \tag{5}$$

Where, Dis_j is the minimum distance of data object *j* to the center point, *n* is the number of samples. The calculation method of Dis_j is shown in formula (6):

$$Dis_j = \min_{C_i \in C}\left[d\left(C_i, x_j\right)\right] \tag{6}$$

Where, C represents the set of initial center points, $d\left(C_i, x_j\right)$ is the Euclidean distance of the data object x_j to the initial center C_i.

The improved K-means algorithm uses the maximum and minimum distance to select the initial center point, and its process is described as follows:

Input: data set Q, the number of clusters k

Output: k initial center points

Step 1 Let $C = \varnothing$, randomly select an object in Q as the first initial center C_1, $C = \{C_1\}$.

Step 2 Calculate the Euclidean distance between each remaining data objects of Q and C_1, then select the object with the largest distance as the second initial center C_2, $C = \{C_1, C_2\}$.

Step 3 Calculate the minimum distance of x_j ($x_j \in Q/C$, $j = 1,2,\cdots,n$)to the initial center as formula (6). Select the object with the maximum distance as C_i as formula (5), $C = \{C_1, C_2\} \cup \{C_i\}$. Repeat this operation if $i \le k$, otherwise stop this operation.

Step 4 Output k initial center points.

3.4 Process of Improved K-means Algorithm

The calculation process of improved K-means algorithm is shown in the Fig. 1.

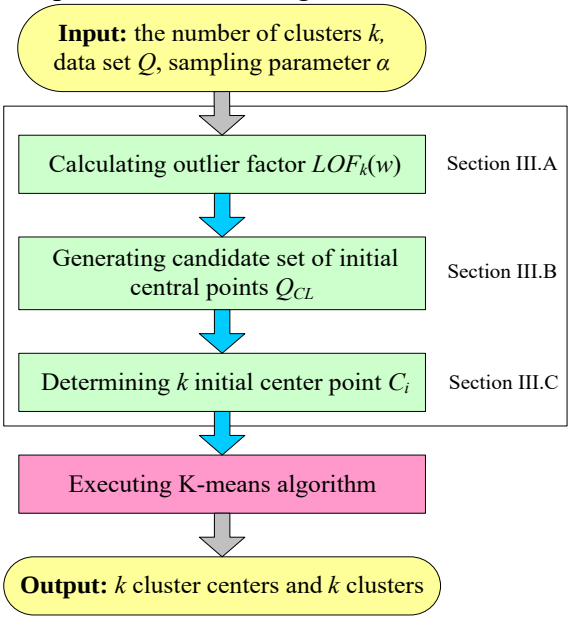

Fig. 1 The calculation process of improved K-means algorithm

In order to remove the outliers, the improved K-means algorithm sorts the data sets in ascending order according to the value of the outlier factors, and generates a candidate center points set. Then, the possibility that two initial center points are in the same cluster is excluded by the maximum and minimum distance. Therefore, the improved K-means algorithm reduces the sensitivity of the K-means algorithm to the initial center point.

4. Evaluation of Clustering Result

The effect of clustering algorithm should be obtained by evaluating the clustering results. In this paper, the clustering results are compared with the results based on manual classification, and F-measure is used as the evaluation index of the clustering results. There is a manually determined classification result on corpus Q, $A = \{A_1, A_2, \cdots, A_s\}$, the clustering results need to be evaluated by the clustering algorithm are $B = \{B_1, B_2, \cdots, B_m\}$. Where both A_i and B_j are clustering clusters, and m is not necessarily equal to s.

For any human topic A_i and cluster B_j, the precision $P(A_i, B_j)$ is shown in formula (7), the recall rate is shown in formula (8), and the F-measure value is shown in formula (9).

$$P(A_i, B_j) = \frac{|A_i \cap B_j|}{|B_j|} \tag{7}$$

$$R(A_i, B_j) = \frac{|A_i \cap B_j|}{|A_j|} \tag{8}$$

$$F(A_i, B_j) = \frac{2P(A_i, B_j) \square R(A_i, B_j)}{P(A_i, B_j) + R(A_i, B_j)} \tag{9}$$

For each human topic A_i, the optimal index value $F(A_i)$ and its corresponding cluster are selected, and the optimal index value $F(A_i)$ is used to evaluate the quality of A_i. The formula of $F(A_i)$ is shown as follow.

$$F(A_i) = \max_{j=1,2,\cdots,m} \{F(A_i, B_j)\} \tag{10}$$

The final F-measure value is:

$$F = \frac{\sum_{i=1}^{s} (|A_i| \square F(A_i))}{\sum_{i=1}^{s} |A_i|} \tag{11}$$

5. Case Study

In order to prove that the clustering result of improved K-means algorithm is better than K-means and K-means++ algorithm, the experimental data in this case study are 500 articles related to movies in WeChat and Weibo. The new films released in the 2019 Spring Festival are taken as the classification cluster, and film name, film type, actor and director are taken as key word in the clustering processing. And F-measure and running time are used to analyze the case results and compare the clustering effects between different algorithms.

In this case study, the processing of text data is the same as literature [11], the vector space model (VSM) and the weight of the feature words TF-IDF are used to represente text data. The improved K-means algorithm is more one parameter α than K-means algorithm. In this case, α is given as 0.5.

Table 1 Running time and F-measure of the three algorithms

No.	Film title	K-means		K-means++		Improved K-means	
		Running time (s)	*F-measure* (%)	*Running time* (s)	*F-measure* (%)	*Running time* (s)	*F-measure* (%)
1	The Wandering Earth	35.3	67.4	35.5	74.5	35.6	88.3
2	Crazy Alien	32.1	73.0	32.3	78.4	32.6	82.2
3	Pegasus	34.5	59.8	34.6	56.3	34.7	67.9
4	The New King of Comedy	31.4	72.5	31.4	77.1	31.5	79.0
5	The Knight of Shadows	29.6	79.4	29.9	80.2	30.2	84.1
6	Integrity	37.5	62.9	37.9	73.3	38.0	81.7
Average		33.4	69.2	33.6	73.3	33.8	80.5

The final F-measure and running time are obtained by taking the average value of the three algorithms after calculating 10 times on the case data set, respectively. The analysis results are shown in the Table I, and the last row is the average of six movies. It can be seen from Table 1, the running time of K-means, K-means++ and improved K-means are 33.4s, 33.6s and 33.8s respectively, and the F-measure is 69.2%, 73.3% and 80.5% respectively. The running time of the improved K-means is

only slightly higher than K-means and K-means++, there is almost no difference. However, the F-measure of the improved K-means is significantly higher than that of K-means and K-means++ algorithms. This phenomenon indicates that the improved K-means algorithm has enhanced the clustering analysis ability, this method is more suitable for text clustering in data mining.

6. Conclusion

Aiming at the defect that K-means algorithm is sensitive to the initial center point and susceptible to isolated points in network public opinion analysis, this paper proposes an improved K-means algorithm based on outlier factor and maximum and minimum algorithm. The paper makes an in-depth study on the selection of the initial center and the method of eliminating the outliers, and the specific operational flow of improved K-means algorithm is also given. Finally, a case based on the actual network text data is studied. The case study results show that the new proposed improved K-means algorithm is better than traditional algorithm in clustering effect. And it is more suitable for network public opinion analysis.

References

[1]. L. H. Wang, Introduction to Public Opinion Research, Tianjin: Tianjin People's Publishing House, 2007.

[2]. Z. Y. Xiong, R. T. Chen, and Y. F. Zhang, "Effective Method for Cluster Centers's Initialization in K-means Clustering," Application Research of Computers, vol. 28, no. 11, 2011, pp. 4188-4190.

[3]. D. Arthur, and S. Vassilvitskii, "K-means++: the Advantages of Careful Seeding," Eighteenth Acm-Siam Symposium on Discrete Algorithms, Society for Industrial and Applied Mathematics, 2007, PP. 1027-1035.

[4]. Y. L. Leng, Q. C. Zhang, L. Zhao, et al, "K-means Algorithm Based on Outliers Detection," Journal of Bohai University (Natural Science Edition), vol. 35, no. 1, 2014, pp. 34-38.

[5]. L. Goel，N. Jain, and S. Srivastava, "A Novel PSO Based Algorithm to fFind Initial Seeds for the K-means Clustering Algorithm," The International Conference on Communication and Computing Systems, 2016, pp. 159-163.

[6]. C. Pizzuti, and N. Procopio, "A K-means Based Genetic Algorithm for Data Clustering," International Conference on EUropean Transnational Education, 2016, pp. 211-222.

[7]. G. Q. Duan, "Auto Generation Cloud Optimization Based on Genetic Algorithm," Computer and Digital Engineering, vol. 43, no. 3, 2015, pp. 379-382.

[8]. H. P. Kriegel, M. Schubert, and A. Zimek, "Angle-based Outlier Detection in High-dimensional Data," Proceedings of the 14th ACMSIGKDD International Conference on Knowledge Discovery and Data Mining, 2008, pp. 444-452.

[9]. M. M. Breunig, H. P. Kriegel, R. T. Ng, et al, "LOF: Identifying Density Based Local Outliers," Acm Sigmod Record， vol. 29, no. 2, 2000, pp. 93-104.

[10]. I. Katsavounidis, K. C. Jay, Z. Zhang, "A New Initialization Technique for Generalized Lloyd Iteration," Signal Processing Letters, vol. 1, no. 10, 1994, pp. 144-146.

[11]. Y. M. Dai, M. H. Wang, M. Zhang, et al, "Optimizing Initial Cluster Centroids by SVD in K-means Algorithm for Chinese Text Clustering," Journal of System Simulation, vol. 30, no. 10, 2018, pp. 3835-3842.