Improved K-means Algorithm Based on the Clustering Reliability Analysis

Hong Zhang, Hong Yu, Ying Li, and Baofang Hu

Shandong Women's University, Jinan 250300, China

zhh6856@126.com, hongting220@163.com, cherry_jn@126.com, hbf0509@126.com

Abstract

Clustering analysis is the basic of data mining, and K-means algorithm is the simplest clustering algorithm. However, traditional K-means algorithm has many defects-instable K value determinations, non-universal applicable *SSE* etc. Consequently, we introduced an improved K-means algorithm basing on the clustering reliability analysis. The algorithm effectively solves the problem on uneven density and large differences in the amount of data clustering.

Keywords: clustering analysis, k-means algorithm, reliability analysis

1 Introduction

As the developing of data mining technique, an extreme important and basic theory – clustering attracts more and more researchers. The idea of clustering is "Birds of a feather flock together" [1,2]. The items within the same cluster is similar while between the clusters are far more different. Cluster analysis is an unsupervised learning algorithm. It is the foundation and core data mining, which has been widely applied. K-means algorithm [3] is a basic cluster analysis classification method. Because of its theoretically reliability, simplicity, convergence speed, and effectively handling large data sets, it has been widely used in data mining technique. However, the traditional k-means clustering algorithm has some obvious problems. For example, when selecting random number *K*, different *K* value can produce totally different clustering result; and traditional clustering criterion function *SSE* will cause the algorithm more suitable for spherical, homogeneous clusters. In view of this, we introduced an improved k-means clustering algorithm.

2 K-means algorithm

The basic idea for k-means algorithm is as follows[4]. First specify a group number, and select K items randomly as the clustering center. For the rest ($n-K$) items, calculate their similarity (distance) to each selected K items. Then cluster all items into K groups. Next, calculate the center for each group and set them as the clustering center. The process will to be done recursively until it meets the clustering criteria. The function of *SSE* (Sum of the Squared Error) shows as follows,

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

Specifically, the k-means algorithm shows as follows,

(1) Select k items randomly as the initial clustering center $c_1 c_2, \ldots, \ c_k$;

(2) For the rest ($n-k$) items, if $d_{ij}(x_i, c_j) < d_{im}(x_i, c_m)$, then $x_i \in C_i$;

(3) Calculate the centroid for each cluster $c_i^* = \frac{1}{n} \sum_{x \in C_i} x$

(4) If $c_i^* = c_i$ for all $i \in [1, k]$, algorithm terminates and $c_1^*, c_2^*, \ldots, c_k^*$ is the result. Otherwise let $c_i = c_i^k$ and GOTO (2)

3 Improved K-means algorithm

Currently, research about improved K-means algorithm [5] is mainly concentrated in the number of clusters determined, selecting of the cluster center and improving of the clustering and other criteria. Yue-Qin Zhang etc.[6] used genetic algorithm optimization clustering numerical; Lei Xiaofeng [7] made use of the fact that K-means algorithm is sensitive to the initial cluster centers, constructed K-Means Scan algorithm, and improved the clustering efficiency and the quality of clustering results; Zhangxue Feng [8] improved the clustering criterion into a weighted standard deviation of the cluster, and improved the quality of the cluster. After learning of research and analysis and comparison of existing algorithms, this paper presents a new improved k-means algorithm.

3.1 Model description

3.1.1 Clustering validity

Cluster analysis is the process of evaluating the effectiveness of clustering results merits. Good clustering should be as complete internal division structure reflects data set, so the class is similar to the sample as possible, to maximize the difference between the sample classes. In order to reflect the effectiveness of the clustering results, we propose indicators based BWP sample geometry, the data

collection of a sample for the study, then take the validity of the results of clustering analysis.

Let $K = \{X, R\}$ is the clustering space, where $X = x_1, x_2, \cdots, x_n$.

Definition: Minimum inter-cluster distance $b(j,i)$ is as follow, where k and j are cluster indexes.

$$b(j,i) = \min_{1 \le k \le c, k \ne j} \left(\frac{1}{n_k} \sum_{p=1}^{n_k} \left\| x_p^{(k)} - x_i^{(j)} \right\|^2 \right) \tag{2}$$

$x_i^{(j)}$ is the j-th item in cluster i. $\|\bullet\|^2$ is the Squared Euclidean distance.

Definition: Within cluster distance is

$$w(j,i) = \frac{1}{n_j - 1} \sum_{q=1, q \ne i}^{n_j} \left\| x_q^{(j)} - x_i^{(j)} \right\|^2 \tag{3}$$

Our goal is to ensure the effectiveness of making inter-cluster spare and incluster tight. For the viewpoint of tightly within the class, the more similar items within a cluster, the better result; and for the view of inter-cluster, the more different items between two clusters, the better result. Taking these two values, to evaluate the use of clustering results, obviously the larger the value, the better description sample clustering is. To put these two standard together, we can say $bsw(j,i) = b(j,i) - w(j,i)$ can evaluate the efficiency for cluster result. The bigger bsw, the better result is.

Definition: we define Between-Within Proportion, or BWP as follows,

$$BWP(j,i) = \frac{bsw(j.i)}{baw(j,i)} = \frac{b(j,i) - w(j,i)}{b(j,i) + w(j,i)}$$

$$= \frac{\min\limits_{1 \le k \le c, k \ne j} \left(\frac{1}{n_k} \sum\limits_{p=1}^{n_k} \left\| x_p^{(k)} - x_i^{(j)} \right\|^2 \right) - \frac{1}{n_j - 1} \sum\limits_{q=1, q \ne i}^{n_j} \left\| x_q^{(j)} - x_i^{(j)} \right\|^2}{\min\limits_{1 \le k \le c, k \ne j} \left(\frac{1}{n_k} \sum\limits_{p=1}^{n_k} \left\| x_p^{(k)} - x_i^{(j)} \right\|^2 \right) + \frac{1}{n_j - 1} \sum\limits_{q=1, q \ne i}^{n_j} \left\| x_q^{(j)} - x_i^{(j)} \right\|^2} \tag{4}$$

When the sample is properly clustering, distance between the sample and the sample within a minimum distance classes can be negligible, compared to the value of approximately *BWP* 1; if the error samples are clustering, the minimum inter-class samples distance and sample distance can be negligible compared to the value of approximately *BWP* -1. Therefore, the range of indicators is for [-1, 1]. And because the number of clusters between the minimum distance required

minimum class into two groups, so *BWP* indicators do not apply to the case when the number of clusters is 1.

From the definition of *BMP*, the index reflects the concentrated sample clustering validity of any single sample within a certain category. The value of *BMP* indicates whether the clustering result is good or bad. And if you want to represent the clustering effect of a data set, usually by obtaining a data set index value of the average of all the samples to analyze the effect of clustering data set, the greater the average, indicating that the clustering effect of the data set, the more Okay. Thus we introduce the following equation to represent the number of clusters k for data clustering effect.

$$\mathrm{avg}_{BWP}(k) = \frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n_i} BWP(j,i) \tag{5}$$

3.1.2 Modified *SSE*

The criterion function of k-means algorithm is the sum of squared errors within the cluster *SSE*, as given by the formula (1). As mentioned, the criteria that leads to k-means clustering algorithm is applied to spherical clusters of isolated nodes insensitive. If the density of date within the cluster is uneven or the difference is large, the number of large clusters containing the data errors will be spliced. So we modify the clustering criteria into a weighted one to get better clustering result. Set the ratio of clustering size and total sample size m/M as the weight, we can get the new clustering criteria as following, where M is data set size, k is the number of clusters, σ is the standard deviation of i-th cluster.

$$\varepsilon = \sum_{i=1}^{k}\frac{m_i}{M}\sigma_i \tag{6}$$

When applying the clustering criterion function in the algorithm, we have to make the cluster and the standard deviation of the weighted minimum and convergence. Criteria function ε contains a cluster's the standard deviation σ_i, and its role is to make class data objects as close to the cluster center class. The weights are m_i/M, their roles are to occupy in the within-class standard deviation criterion function comprises a greater proportion of large number of samples. In order to make the convergence criterion function, the probability of a data object is assigned to a small number of classes as the increase in. Therefore, if the two are not the same size and density of the class exists, which spacing is small and the adjacent samples in the less dense relatively large number of classes, and when the density is less the larger the number of samples in the class, the edge of large clusters is assigned to the sample with a small error probability in the cluster will be reduced, so that the effect has been optimized clustering.

3.2 Basic steps for improved algorithm

From the upper analysis, the best cluster count is

$$k_{opt} = \arg\max_{2 \le k \le n} \{avg_{BWP}(k)\} \quad (7)$$

We knew that in order to use the avgBMP, two or two more clusters are required. Moreover, many researchers [2] [7] had showed that $k_{\max} \le \sqrt{n}$ .Thus, we set $2 = k_{\min} \le k \le k_{\max} = \sqrt{n}$.

Now we intrude the modified k-means algorithm as follows:

(1) determine the cluster count $[k_{\min}, k_{\max}]$

(2) while $k_{\min} < k < k_{\max}$

(2.1) call the traditional k-means algorithm

(2.2) get the BMP for each sample

(2.3) calculate the *avgBMP(k)*

(3) get the optimal cluster count k_{opt}

(4) output the result

Also, the clustering criterion weighting applies to k-means algorithm, the optimal number of clusters combined. We obtain a new k-means clustering algorithm. The algorithm is as follows:

(1)call the *search-k( )* function, get the k_{opt}

(2)select k_{opt} initial cluster center randomly $c_1, c_2, \cdots, c_{k_{opt}}$

(3)if $w_j d_{ij}(x_i, c_j) < w_m d_{om}(x_i, c_m)$, then $x_i \in c_j$,where $w_i = 1/\sqrt{\sigma_i}$

(4)get $c_1^*, c_2^*, \cdots, c_k^*$ by $c_i = \frac{1}{n_i} \sum_{x \in C_i} x$

(5)if $c_i^* = c_i$ for all $i \in [i,k]$,algorithm terminate. Otherwise, let $c_i = c_i^*$,GOTO (2)

(6)if ε converges, output the result and BMP

4 Experiment and analysis

Our experimental platform is Matlab. We randomly selected two-dimensional data [0,100] as data collection and we used to the traditional k-means algorithm and an improved algorithm to generate data clustering algorithm iterations and compare properties analysis. Detailed steps are as follows: (1)20 sets of random data, initial cluster count is 3

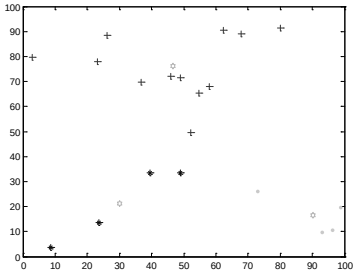

Figure 1-a K-means

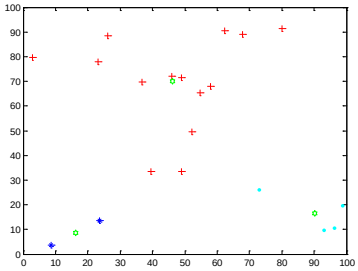

Figure 1-b Improved K-means

(2)50 sets of random data, initial cluster count is 4

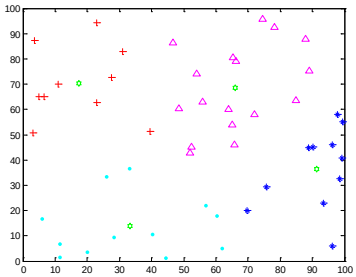

Figure 2-a K-means

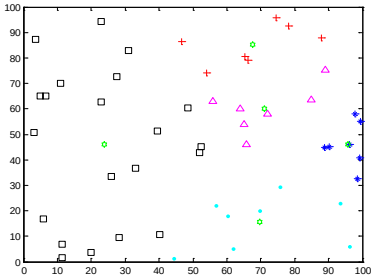

Figure 2-b Improved K-means

(3)100 sets of random data, initial cluster count is 6

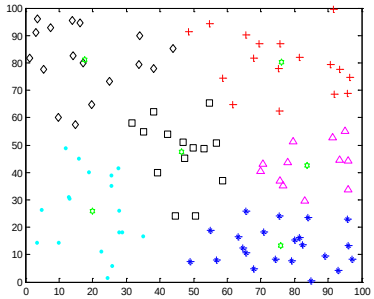

Figure3-a K-means

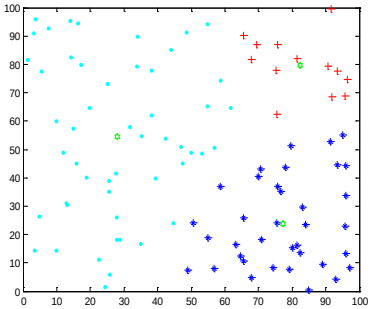

Figure 3-b Improved K-means

Figure 1-3, the center of each hexagon cluster type, the rest of the several shapes, such as star, solid circles, open circles, plus, diamond, square, triangle, respectively, represent a class.

From Figure 1-a and Figure 1-b, when the total number of sample data set for 20, the same number of clustering algorithms to improve the number of clusters with traditional algorithms. However, even under the same initial cluster centers in the selection situation, as the clustering criterion to select different functions, the final clustering result is different; See Figure 2-a and Figure 2-b, the total number of data is 50, to improve the clustering algorithm to select the optimal number is 5, compared with k-means algorithm and more, we can see, k-means clustering algorithm preset number, can not guarantee the correctness of the

index; Figure 3-a and Figure 3-b, the improved algorithm of the number of clusters is 3, and can be seen from the figure, the improved algorithm, the density has not been wrong category uneven split.

After the initial cluster centers is found, the strike to iterative cluster center until the clustering criterion function converges. Meanwhile, in the case of the same number of samples, the smaller the number of iterations, the lower the complexity of the algorithm is. In the [0,100] range were randomly selected 50,100,150,200,250,300,350,400,450,500 two-dimensional data, these data sets were executed 200 times and improved the traditional k-means algorithm, the number of iterations to get the curve, which is showed in Figure 4.

Improved algorithm iterations are less than the traditional k-means algorithm. The complexity of the k-means algorithm is *O(nkt)*, which indicates that the improved algorithm complexity than traditional algorithms.

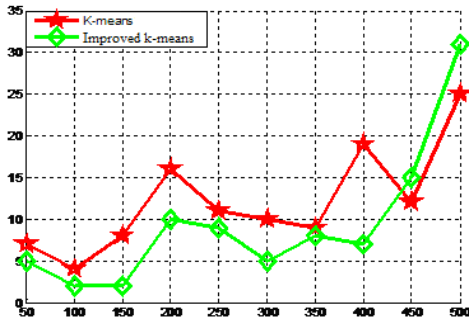

Figure 4 Recursion average for two algorithms

5 Conclusions

In K-means clustering algorithm, k is the number of options, which can be selected randomly. This lead to the un-stability of the algorithm; clustering criterion function and SSE will lead to its more suitable for spherical, homogeneous clusters of clusters. To solve these problems, we introduced clustering reliability analysis indicators BWP in this paper, and use the index to determine the optimal number of clusters of a data set. Meanwhile, we introduced the weight to be reunited with class criterion function; using the Criteria to solve the k-means algorithm does not suite to the problem of non-uniform data set. Experiments results showed that the improved algorithm can solve these problems, also reduce the number of iterations, then reduce the complexity of the algorithm.

6 Acknowledgements

This work was financially supported by the Key Laboratory of Computer Application of Shandong Women's University, the national statistics scientific research program of China (2013LZ37).

References

[1] Yuxia Lai and Jianping Liu. Optimization study on initial center of k-means algorithm. *[J] COMPUTER ENGINEERING AND APPLICATIONS*, 44(10).p.147-149, 2008

[2] Shenghui Liu Tao Huang and Yanna Tan. Research of clustering algorithm based on k-means. *[J] COMPUTER TECHNOLOGY AND DEVEL*, 21(7).p.54-57, 2011.

[3] .K Krishna and M Narasimha Murty. Genetic k-means algorithm. *Systems, Man, and Cybernetics, Part B: Cybernetics*, 29(3).p.433-439, 1999.

[4] Jim ZC Lai and Yi-Ching Liaw. Improvement of the k-means clustering filtering algorithm. *Pattern Recognition*, 41(12).p.3677-3681, 2008.

[5] Marius Muja and David G Lowe. Fast approximate nearest neighbors with automatic algorithm configuration. In *VISAPP (1)*, p. 331-340, 2009.

[6] Yueqin Zhang and Jing Liu. Application of an advanced clustering algorithm in intrusion detect- ion. *[J] JOURNAL OF TAIYUAN UNIVERSITY OF TECHNOLOGY*, 39(special issue).p.74-76, 2008.

[7] Fan Lin Xiaofeng Lei, Kunqing Xie and Zhengyi Xia. An efficient clustering algorithm based on local optimality of k-means. *[J] Journal of Software*, 19(7).p.1683-1692, 2008.

[8] Guizhen Zhang Xuefeng Zhang and Peng Liu. Improved k-means algorithm based on clustering criterion function.computer engineering and applications. *[J] Computer Engineering and Applications*, 47(11).p.123-127, 2011.