DOI: 10.1051/itmconf/20171104001

An Improved K-means Clustering Algorithm Applicable to Massive High-dimensional Matrix Datasets

Dong-Yuan LI[1] and Cai-Feng CAO[1,a]

[1] *Wuyi University, School of Computer, 529020 Jiangmen, China*

Abstract. Since K-means clustering algorithm is easy to implement and high efficient, it has been widely used in cluster analysis of massive datasets. The value of k is difficult to determine in advance and the randomness of choosing initial centers leads to a series of social problems, such as instability, local optimal solution sensitivity to outliers. Results from hierarchical clustering are more natural than those from K-means clustering, but its high time complexity and space complexity makes it difficult to be applied to a large data set. In this paper, through combination of hierarchical clustering and K-means clustering, we have proposed an improved K-means clustering algorithm, and have done experiments using datasets provided by MovieLens.

1 Introduction

As a method of unsupervised learning, clustering analysis is an important means of data mining. Without knowledge of the data distribution beforehand, clustering is to classified data into groups which are driven by data [1].

In all cluster analysis methods, K-means clustering algorithm has application usability, fast convergence and ability to handle large datasets [2]. However, K-means clustering algorithm also has some disadvantages: 1) Because of data distribution, the value of K is difficult to estimate; 2) Randomness of initial centers cause the clustering results often fall into local optimum which is not stable, but not the global optimum; 3) Sensitivity to outliers; 4) Unable to deal with non-spherical clusters and clusters of different sizes and densities.

Results from hierarchical clustering can reflect the hierarchical structure of the dataset. But the time complexity and the space complexity of hierarchical clustering are so high that hierarchical clustering is unsuited for high-dimensional massive dataset.

In this paper, we have integrated principal component analysis and sampling into combination of hierarchical clustering and K-Means clustering, and have proposed a new clustering algorithm named SPHK-means clustering algorithm.

[a] Corresponding author: cfcao@126.com

2 Associated knowledge and working

2.1 K-means clustering algorithm

K-means is a clustering algorithm based on partition.

2.1.1 Basic algorithm

Basic steps of K-means clustering algorithm are as follows:

1) Selecting initial cluster centers: randomly select k data objects from dataset containing m data objects;
2) Cluster object classification: calculate similarities between every data object and every center, and divide each data object into the group whose center is most similar to it;
3) Calculate new centers: for each group, calculate new center which is mean value of all member data objects in the group;
4) Determine whether the process is to terminate: if continuously twice iteration results in equality, the process stop, or else, proceed to the next iteration and go to step 2).

Space complexity of K-means clustering algorithm is $\mathrm{O}\left(\left(m+k\right)\cdot n\right)$ [4], and time complexity of it is $\mathrm{O}\left(t\cdot k\cdot m\cdot n\right)$ [4], m is the number of data objects, n is the number of properties each data object has, t is the number of iterations.

2.1.2 Shortcoming and solutions

There are four main disadvantage of K - Means clustering algorithm: 1) How to determine the optimal value of K; 2) Randomness of initial centers; 3) Sensitivity to outliers; 4) It is difficult to process clusters which is aspherical or with uneven density and size. According to the fact, "the distance from the sample points in the most unlikely point to the same cluster ", D.H.Zhai, etc, used the maximum distance method [5] to select initial cluster centers. Y.Qin, etc [6], generated initial centers on the basis of detecting populated area. J.P.Zhang, etc [7], optimally partitioned data sample space basing on histogram method, and determined value of k and initial centers for k-mean algorithm according to the characteristic distribution of the data sample.

2.2 Agglomerative hierarchical clustering

Hierarchical clustering whose results are closer to the natural classification of the data object is an old clustering technology. It include two basic approaches-- cohesion and splitting, the latter is more commonly used.

2.2.1 Basic algorithm

Agglomerative hierarchical clustering begins from the initial state that each data point individually becomes a cluster. In each iteration, it merge two clusters between which the distance is shortest as a new cluster. It controls the number of result clusters by setting the threshold of inter-cluster distance, once distances between all pairs of clusters are larger than the threshold set, it stop the process of merging. There are three methods of measuring the distance between clusters: 1) single-linkage; 2) Complete-linkage; 3) Group average. In this paper, we have used the last. Let $\mathbf{u}$ and $\mathbf{v}$ are respectively two clusters, $\mathbf{u}$ and $\mathbf{v}$ are respectively the i^{th} member of $\mathbf{u}$ and the j^{th} member of $\mathbf{v}$, then Distance between clusters using group average is:

$$d\left(\mathbf{u},\mathbf{v}\right)=\sum_{i,j}\frac{d\left(\mathbf{u_i},\mathbf{v_j}\right)}{\left(\left|\mathbf{u}\right|*\left|\mathbf{v}\right|\right)} \quad (1)$$

2.2.2 Shortcoming and solutions

As P.N. Tan, etc[4] mentioned, the space complexity of agglomerative hierarchical clustering algorithm is $\mathrm{O}\left(m^2\right)$ [4], and the time complexity of it is $\mathrm{O}\left(m^2\log m\right)$ [4], m is the number of data points. Relatively high time complexity and space complexity makes it difficult for hierarchical clustering algorithm to process datasets whose volume is very large.

2.3 Principal component analysis

PCA can reduce dimensionality of highly dimensional data and to some extent remove the noise. Supposing there are p characteristics: $X_1, X_2, \ldots, X_p$. PCA transform original characteristics into new characteristics which is linear combinations of original ones: $F_1, F_2, \ldots, F_k, \left(k \le p\right)$ [8],

$$\begin{cases} F_1 = \mu_{11}X_1 + \mu_{21}X_2 + \cdots + \mu_{p1}X_p \\ F_2 = \mu_{12}X_1 + \mu_{22}X_2 + \cdots + \mu_{p2}X_p \\ \cdots \\ F_p = \mu_{1p}X_1 + \mu_{2p}X_2 + \cdots + \mu_{pp}X_p \end{cases} \quad (2)$$

Meeting these conditions: 1) For each principal component, sum of squares of its factor is 1; 2) all principal components are independent of each other; 3) Variances of principal components must be in descending order, namely, their importance must be in descending order [8]:

$$\begin{cases} \mu_{1i}^2 + \mu_{2i}^2 + \cdots + \mu_{pi}^2 = 1 \\ Cov\left(F_i, F_j\right) = 0, i \ne j, i, j = 1, 2, \ldots, p \\ Var\left(F_1\right) \ge Var\left(F_2\right) \ge \cdots \ge Var\left(F_p\right) \end{cases} \quad (3)$$

3 Model design of sphk-means clustering algorithm

Let row coordinates represent the objects to be scored, and column coordinates represent users. We have built a scoring matrix Mo :

$$Mo = \begin{bmatrix} p_{0,0} & p_{0,1} & \cdots & p_{0,M-2} & p_{0,M-1} \\ p_{1,0} & \ddots & & & \vdots \\ \vdots & & \ddots & & \vdots \\ p_{N-2,0} & & & \ddots & \vdots \\ p_{N-1,0} & p_{N-1,1} & \cdots & \cdots & p_{N-1,M-1} \end{bmatrix} \tag{4}$$

The j^{th} user scored the i^{th} object $p_{i,j}$ points, $p_{i,j}$ is float, $i \in \{0,1,\ldots,N-1\}$, $j \in \{0,1,\ldots,M-1\}$. According to the above analysis, when N and M are large enough, either K-means clustering or agglomerative hierarchical clustering is unfavorable to solve the problem. Model of SPHK-means clustering algorithm summarily proceed in two phases: 1) Pre-cluster by means of integration of sampling, principal component analysis and hierarchical clustering to determine the value of and the initial centers; 2) K-means clustering with and initial centers determined in phase 1) get the final result. Its working process is shown in Figure. 1.

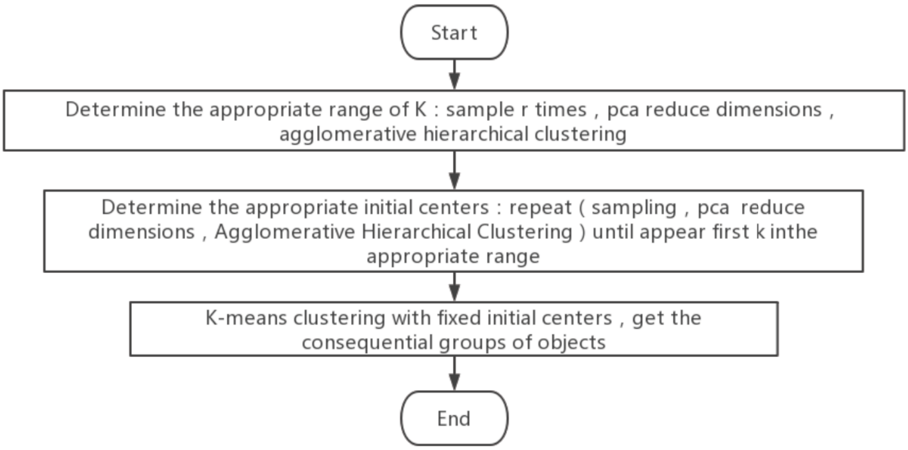
Start
Determine the appropriate range of K : sample r times , pca reduce dimensions , agglomerative hierarchical clustering
Determine the appropriate initial centers : repeat (sampling , pca reduce dimensions , Agglomerative Hierarchical Clustering) until appear first k inthe appropriate range
K-means clustering with fixed initial centers , get the consequential groups of objects
End

Figure 1. The process of SPHK-means clustering algorithm.

3.1 Sampling PCA hierarchical clustering

As pre- cluster, this stage in order to determine the value of and the initial centers.

3.1.1 Determine the appropriate value of K

Totally r samples is extracted. The process of dealing with the i^{th} $(i = 1,2,\ldots,r)$ sample is as follow:

$\frac{1}{5}N$ objects is extracted as sample from the whole including N objects. PCA reduces dimensions of sample to $\frac{1}{6}M$. Scoring matrix of sample after dimensional reduction is as follow:

$$Mp(i) = \begin{bmatrix} P_{0,0}(i) & P_{0,1}(i) & \cdots & P_{0,\frac{1}{6}M-1}(i) \\ P_{1,0}(i) & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ P_{\frac{1}{5}N-1,0}(i) & P_{\frac{1}{5}N-1,1}(i) & \cdots & P_{\frac{1}{5}N-1,\frac{1}{6}M-1}(i) \end{bmatrix} \tag{5}$$

Use formula (1) to cluster $\frac{1}{5}N$ objects by means of agglomerative hierarchical clustering algorithm. Record the numbers of consequential groups as $k(i)$. Calculate average $\bar{k}$ and standard deviation $\sigma(k)$ of $k(i)\left(i=1,2,\ldots,r\right)$:

$$\begin{cases} \bar{k} = \frac{1}{n}\sum_{i=1}^{n} k(i) \\ \sigma(k) = \frac{1}{n}\sqrt{\sum_{i=1}^{n}\left(k(i)-\bar{k}\right)^2} \end{cases} \tag{6}$$

Let $\left(\bar{k}-\frac{\sigma(k)}{2},\bar{k}+\frac{\sigma(k)}{2}\right)$ be appropriate range of K.

3.1.2 Determine the appropriate Initial centers

Repeat the above process of "sampling--reducing dimensions--hierarchical clustering", until first appear a k_0 in $\left(\bar{k}-\frac{\sigma(k)}{2},\bar{k}+\frac{\sigma(k)}{2}\right)$, then, the same time, the ultimate centers $\left\{v_1,v_2,\ldots,v_{k_0}\right\}$ are the appropriate initial centers.

3.2 K-means clustering

After reduce dimensions of N objects to $\frac{1}{6}M$ using PCA, cluster N objects by means of K-means clustering algorithm with initial centers $\left\{v_1,v_2,\ldots,v_{k_0}\right\}$, get consequential groups.

3.3 Analysis of time complexity and space complexity

In the phase of sampling PCA hierarchical clustering, samples n times, its space complexity is $\mathrm{O}\left(10\cdot\left(m/5\right)^2\right)$. Supposing the first k in appropriate range in the C^{th} sampling, its time complexity is $\mathrm{O}\left(\left(10+C\right)\cdot\left(m/5\right)^2\log\left(m/5\right)\right)$. In the phase of K-means clustering, the space complexity is $\mathrm{O}\left(\left(m+k\right)\ n/6\right)$, the time complexity is $\mathrm{O}\left(t\cdot k\cdot m\cdot n/6\right)$.

When finish pre-clustering, memory of all variables except for the appropriate initial centers is cleared. So if m ≫ n, space complexity of SPHK-means clustering algorithm is:

$$\max\left(\mathrm{O}\left(10\cdot\left(m/5\right)^2\right),\mathrm{O}\left(\left(m+k\right)\cdot n/6\right)\right)=\mathrm{O}\left(10\cdot\left(m/5\right)^2\right)=\mathrm{O}\left(\frac{2}{5}\cdot m^2\right) \tag{7}$$

However, space complexity of classical agglomerative hierarchical clustering algorithm is $\mathrm{O}\left(m^2\right)$. Space complexity of SPHK-means clustering algorithm is lower than that of classical agglomerative hierarchical clustering algorithm.

Time complexity of SPHK-means clustering algorithm is:

$$\mathrm{O}\left(\left(10+C\right)\cdot\left(m/5\right)^2\log\left(m/5\right)+t\cdot k\cdot m\cdot n/6\right)=\mathrm{O}\left(\left(\left(10+C\right)/25\right)\cdot m^2\log\left(m/5\right)\right) \tag{8}$$

However, time complexity of classical agglomerative hierarchical clustering algorithm is $\mathrm{O}\left(m^2\log m\right)$, If C is small enough, time complexity of SPHK-means clustering algorithm is lower than that of classical agglomerative hierarchical clustering algorithm.

4 Experiments

Due to restriction of article length, we have put experimental process and results analysis in another paper for this conference, *Discovering Movie Categories Based on SPHK-means Clustering Algorithm*, which has proved that SPHK-means clustering algorithm has better classification accuracy and find more movie categories than classical K-means clustering algorithm.

Conclusion

We have synthetically applied sampling, principal component analysis and agglomerative hierarchical clustering to proposing an improved clustering algorithm named SPHK-means clustering algorithm, and have analyzed that it is better than classical K-means clustering algorithm in aspect of classification accuracy and number of categories found, and that it is better than classical agglomerative hierarchical clustering algorithm in aspect of time complexity and space complexity.

MovieLens[9] provides rating data involving sufficiently large number of pairs of movie and user and category labels of movies. We have designed experiments which has been described in another paper for this conference in detail to verify the superiority of SPHK-means clustering algorithm compared to classical K-means clustering algorithm on performance.

Recommendation algorithm based on neighbourhood uses KNN algorithm which is with high time complexity and high space complexity looking for neighbours, facing the problem of low classification accuracy. Next, we could apply SPHK-means clustering algorithm to improving recommendation algorithm based on neighbourhood.

Acknowledgment

This paper is supported by the Features Innovative Project in Guangdong Province "Research and Application of Mass Media Data Mining Technology" (2015KTSCX145).

References

1. Y. Zhang, L. He, H.D. Zhu, J. Chongqing Normal Uni(Nat Sci), **33**, 97-101 (2016)
2. Z. Y.Xiong,R.T.Chen, Y.F. Zhang, Appl. Res.Comput, **28**, 4188-4190 (2011)
3. J. Gong, J. Hunan. Uni. Technol, **22**, 52-54 (2008)
4. P. N. Tan, M. Steinbach, V. Kumar, *Introduction to Data Mining* (Posts &Telecom Press, Beijing, 2011)
5. D.H.Zhai, J.Yu, F Gao, L.Yu, F Ding, Appl. Res.Comput, **31**, 713-719 (2014)
6. Y.Qin, J.W.Jing, J.Xiang, J.Grad.Sch.Chin. Acad.Sci, **24**, 771-777 (2007)
7. J.P.Zhang, Y.Yang, J.Yang, J. Sys.Sim, **21**, 2586-2590 (2009)
8. Shlens, Jonathon, *A Tutorial on Principal Component Analysis*, CoRR abs/1404.1100 (2005): n. pag.
9. http://grouplens.org/datasets/movielens/