Modified K-means Algorithm for Clustering Analysis of Hainan Green Tangerine Peel

Ying Luo, Haiyan Fu

Modified K-means Algorithm for Clustering Analysis of Hainan Green Tangerine Peel

Ying Luo[1], Haiyan Fu[2]✉

[1,2] School of Information Science Technology, Hainan Normal University, Haikou, China
yanzi78@163.com

Abstract. K-means is a classic, the division of the clustering algorithm, apply to the classification of the globular data. According to the initial clustering center, this paper comprehensive consideration the characteristics of various Hierarchical cluster algorithms and choose the appropriate Hierarchical cluster algorithm to improve K-means, and combined with Hainan Green Tangerine Peel cluster analysis of data which is compared experiments. The results indicate that the improved algorithm have increasing the distance between classes with each others, get a stable of cluster results and better implementation data mining. Finally to summary the two algorithms and the further research direction.

Keywords: clustering; Modified K-means algorithm; initial clustering center; average link.

1 Introduction

The K-means is a classic clustering algorithm which is a data exploration technique that allows samples with similar characteristics to be clustered together in order to facilitate their further processing and usually has applications in identification of patterns hidden in data. Although this algorithm has high flexibility and efficiency, in some case, it has not a good performance for random initialization centre. Owing to trial and error to achieve good representation it spend more memory and CPU time.

To address this issue, some scholar suggested improving algorithm based on other clustering algorithm that has better performance with speed, efficiency and clustering results than previous traditional clustering algorithms when solve some specified problems.

In this paper, a modified K-means algorithm based on two stages strategy is introduced. Modified K-means employed a hierarchical clustering algorithm- average linkage technology- to initialize the centre for the second stage and then uses the traditional K-means for further classification.

2 Related Research

In the traditional K-means, process depends on a near-optimal solution. For the first iteration, the algorithm is initialized randomly. Then, the clustering centers from the last iteration are used to initialize in subsequent iterations till there is no change in the average value. In worse case, a group unavailable initial clustering center lead to the K-means stops to a local minimum, that it to say, their true distribution in the problem domain will not be reflected by the distribution of objects in the result [1].

Hierarchy clustering is famous because of easy to read, but it not flexible enough due to cannot be operated with samples reversibly. Initially, all genes are considered as individual clusters followed by sequential merging of the two closest clusters in each subsequent step based on their distance, the final step has only one clustering left with all the genes in it [2]. So, this paper argues that, hierarchical clustering is suitable for used to preliminary classification. In lots of representative algorithm, there are three different technology; single linkage, average linkage and complete linkage.

The single linkage technology uses the distance of samples with most similar in different groups. Sometimes, a clustering distribution is non-homogeneous that has a bad impact on the performance of the algorithm. The complete linkage technology, also called furthest neighbor, uses the distance of samples with most dissimilar in two groups and as a complete vision, they have a better performance. However, this technology is susceptible to noise, and unable to get satisfactory results due to few samples which away from the center of group. The third technology is the average linkage technology that uses the average distance between all pairs in different clustering. This algorithm considers the structure of dataset, the most similarity group tend to be merged.

$$d(r,s) = \frac{1}{n_r n_s} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} dist(x_{ri}, x_{sj}) \tag{1}$$

where r and s are group, there are n_r samples in group r, and n_s samples in group s. Therefore, employed average linkage technology in subset of dataset for achieve initial clustering center to the second stage is available.

3 Algorithm Design

The modified K-means algorithm works in two stages: The first stage makes preliminary classification with the average linkage technology. The second stage uses the clustering centers from the first stage as initial clustering center and employed the traditional K-means algorithm on the further classification.

K is the expected number of clusters, in practice, it is empirically chosen by the user depending on characteristics of the dataset [3].

Algorithm process is as follows:

The first stage:

Step 1: Choose appropriate function calculate distance metric used for similarity of samples. The details of function choosing are explained in later.
Step 2: The similarity samples x_i and x_j are merged, the first clustering named D1, D1= $\{x_i, x_j\}$.
Step 3: Another sample x_p is added to the D1 if the average distance between x_p and samples that in the D1 less than Threshold, and then D1= $\{x_i, x_j, x_p\}$.Else, x_p is taken out and a new clustering is made with it.
Step 4: Step 3 are repeated till the number of cluster is equal to k.

The second stage:

Step 1: clustering centers from the first stage are initialized.
Step 2: Samples in dataset are assigned to the cluster that the most similar initial centers are.
Step 3: Recalculated clustering average value as centers for the sequential iterative.
Step 4: Step 2 and Step 3 are repeated till there is no change in clusters.
In the first stage, some function are common used for computes the distance metric that defined dissimilarity between samples, such as Euclidean, Hamming distance and so on. The Hamming distance between two samples is the percentage of coordinates that differ.

$$d(x_i, x_j) = \frac{m - \sum_{\alpha=1}^{m} \left(\sum x_i[v_\alpha] \oplus x_j[v_\alpha] \right)}{m} \tag{2}$$

where x_i and x_j are two samples in sets $X = \{x_j \mid j = 1, 2, \cdots n\}$, $[v_\alpha]$ is the α-th attribute, and each samples have m attributes and they are discrete.

Clustering performance criterion function

Clustering is aimed to divide a set of samples $X = \{x_1, x_2, \cdots, x_n\}$ into K disjointed subsets X_1; X_2; …; X_K so that points in the same subset share common properties while points which belong to different subsets do not share these properties. We evaluate the clustering performance using error sum of squares criterion function, it is defined as:

$$ESS = \sum_{i=1}^{K} \sum_{p \in X_i} \| p - m_i \| \tag{3}$$

where average vector $m_i = \frac{1}{n_i} \sum_{p \in X_i} p \quad i = 1, 2, \cdots, k$.

Algorithm steps are as follows:

input: $X = \{x_1, x_2, \cdots, x_n\}$ samples sets
K numbers of clustering
output: $S = \{X_1, X_2, \cdots X_K\}$ the results of clustering
the dissimilarity function: hamming distance
criterion function: error sum of squares criterion function ESS
initial state:

$d = 0$ %initial threshold

$S = \{\{x_1\},\{x_2\},\cdots\{x_n\}\}$ %each sample is a group

*step*1： compute similarity of samples, and obtained adjacent matrix C.

*step*2： examine every samples in X, and emerge the most similar x_i and x_j, the get new group: $D_1 = \{x_i, x_j\}$.

*step*3： calculate average distance of sample x_p outside of group D_1 and samples in D_1

if $average \le d$

then

$D_1 = \{x_i, x_j, x_p\}$

else

$D_2 = \{x_p\}$

$d = d + 1$

until the number of clustering is equal to K.

*step*4： obtain the centriod of K group $W = \{\mu_i \mid i = 1, 2, \cdots K\}$.

*step*5： update centriod, recalculated clustering average value as centriod for the sequential iterative.

*step*6： *step*5 are repeated until the criterion function - error sum of squares criterion function is constriction.

The first stage is important to the whole process of avoid some worse case which dependent on random initialization of the traditional K-means. While the average linkage technology is completed, lots of samples are in true distribution in the problem domain and reduce the number of iterative in the second stage.

Two parameters are compared during the testing: clustering quality and stability.

4 Comparison for Green Peel Cluster Analysis of Hainan Island

The performance of the modified K-means algorithm is compared with traditional K-means algorithms using our dataset that comes from the experiment of Green Tangerine Peel clustering Analysis of Hainan. We select 100 leaves evenly with DNA mo-

lecular markers method at five regions (group) in Hainan Island: ShiMeiWan (SMW), SanYa (SY), BaWangLing (BWL), JianFengLing (JFL), and WenChang (WC).

The dataset has 100 arrays, 206 columns (each sample has 206 characteristics) and 15 times are run for each test. All data as follow are the average value.

Two algorithms are run with Inter(R) Core(TM) 2Duo CPU T6500@2.1GHz. and 2.0GB RAM. The operation system is Vista32 and the software is written in MATLAB7.0.

The centroid of the cluster in the second stage of modified K-means is obtained by average linkage technology, and traditional K-means randomly chooses K points to be the initial centroids, we use hamming distance to calculate the dissimilarity.

4.1 Evaluation of clustering quality

The goal of clustering is to minimize the intra-cluster distances and to maximize the inter-cluster distances [3]. In order to evaluate the quality of the results of the two algorithms, the distance between centers in two clusters are used to intra-cluster distances while distances between a clustering center and the objects belonging to it are inter-cluster distances. We clustered dataset and get the results of 2-10 numbers clustering.

Table 1. the intra-cluster distances of cluster used in two algorithms

cluster	2	3	4	5	6	7	8	9	10
K-means	0.500	0.599	0.621	0.565	0.606	0.624	0.635	0.635	0.633
AK-means	0.602	0.603	0.605	0.593	0.601	0.609	0.615	0.620	0.633

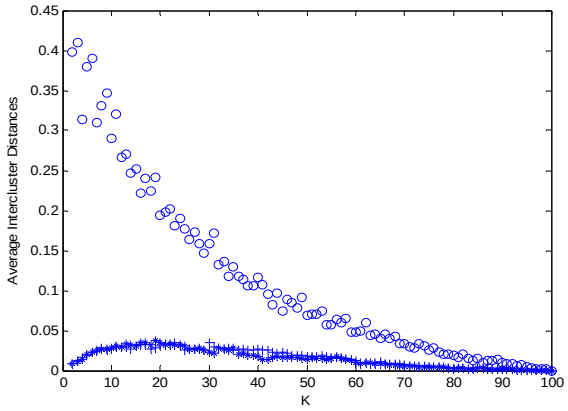

Figure 1. the "o" is on behalf the average intercluster distances of the K-means algorithm, and the "*" is represent the corresponding of the modified K-means.

On the whole, the performance of modified K-means are superior to the traditional K-means algorithm (the average inter-cluster distances are smaller).

4.2 Stability of the clustering

For compare the stability which is important that reflects the similarity between different runs using the same program of each algorithm, we need a statistic. In the K-means algorithm, based on iterative till there is no change in each cluster, at the same time, the value of the total amount of distortion wills no change more. The distortion of a cluster which describes the dispersion degree in the distribution of the samples is defined as the sum of the squared Hamming distances between its centre and the objects belonging to it.

$$S = \sqrt{\frac{\sum \left(ESS_i - \overline{ESS} \right)^2}{m-1}} \qquad (4)$$

Standard deviation of distortion in test using the same program could distribution the range of distortion. If vary of standard deviation of distortion are wildly, the clustering results are unstable, and vice versa. For comparison, we have run 12 times for each and recorded the result.

Table 2. Comparison of standard deviation of distortion to complete analysis of two clustering algorithms

cluster	2	3	4	5	6	7	8	9	10
K-means	2.499	1.929	2.334	1.606	1.255	0.654	0.358	0.178	0
AK-means	0	0	0	0	0	0	0	0	0

The table 2 shows, as we can see, the first stage of the modified K-means benefited from the average link technique performs (the standard deviation of distortion =0) better than the corresponding (which is>0). Namely, the stability of cluster is improved.

Significantly, the modified K-means clustering algorithm is better than the traditional K-means in time taken, clustering quality, and stability of algorithm.

5 Summary

The modified algorithm employed a hierarchical clustering algorithm- average linkage technology- to initialize the centre avoid a bad clustering results in finally by samples which are in the edge of the group as center in the first iterative in the traditional K-means algorithm. At the same time, thanks to reducing the influence by noise, the modified algorithm having improve the Evaluation of clustering quality in a certain degree. Besides, random selection of the initial clustering center lead to the algo-

rithm stops in a local minimum and achieve different results in many tests and modified K-means is an available strategy in this solution.

However, as the K-means, the modified K-means control the border of clustering with diameter of semi-sphere, if the clustering is not spherical, two algorithms will not work very well.

Aknowledgements.
This work is supported by NSF of China (70940007 and 71461008).

References

1. W. Jin, H.P. Chen, A k-means algorithm based on Hierarchical clustering. Journal of Hohai University-Changzhou, 1, 7-10(2007)
2. J.B. Lin, M.D. Liu, X. Chen, Data mining and OLAP theory and practice. Tsinghua university press, Bei jing(2003)
3. R.F. Hu, G.F. Yin, Y. Tan, A Hybrid Clustering Algorithm and It's Application. Journal of Sichuan University: Engineering Science Edition, 5, 68-73 (2006)
4. H.R. Wang, L.M.Zhao, J.Pei, Equilibrium Modified K-Means Clustering Method. Journal of Jilin University(Information Science Edition), 2, 41-44(2006)
5. T. Velmurugan, Performance based analysis between k-Means and Fuzzy C-Means clustering algorithms for connection oriented telecommunication data, Applied Soft Computing, 19, 134-146(2014)
6. Wan Maseri Binti Wan Mohd, A.H.Beg*, Tutut Herawan, K.F.Rabbi, An Improved Parameter less Data Clustering Technique based on Maximum Distance of Data and Lioyd k-means Algorithm. Procedia Technology, 1, 367-371(2012)