http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 12 (20): 5665-5668, 2013
ISSN 1812-5638 / DOI: 10.3923/itj.2013.5665.5668

An Adaptive Initial Cluster Center Selection K-means Algorithm and Implementation

Ma JiMing, Li XiaoJiao, Su RiJian and Zhang Xiang Mei
Zhengzhou University of Light Industry, 450002, Zhengzhou, China

Abstract: Traditional k-means algorithm randomly select initial cluster center and the quality of clustering results depends on the selection of initial cluster center. If isolate points are selected, the algorithm iterations will increase significantly; if k points in the same class are selected, the algorithm will fall into local optimum. An adaptive method to select initial cluster center for k-means algorithm is proposed, so that initial cluster centers are located in high density area and have a certain distance with each other. Experiments show that the method has higher stability and accuracy than the traditional k-means algorithm and some other similar improved algorithms.

Key words: Adaptive initial cluster center, density of sample points, k-means algorithm

INTRODUCTION

Clustering analysis is one main task of data mining and it is widely used in biology, information retrieval, climate, medicine, business and information security, etc. Clustering analysis group the data object depends on descriptions of data and relations. Its objective is to make sure objects in the same group are similar with each other and objects in different groups are quite different. The more similarity in one group and the greater difference between groups, the clustering is better (Tan *et al*., 2011). K-means algorithm is a simple but important clustering analysis technology. K-means algorithm randomly select k sample points from the data set as the initial center, the rest sample points are assigned one by one to their nearest center and the point set assigned to the same central point is a cluster. Then redistribution cluster center, according to the rule that the error sum of squares of Euclidean distance from sample points to the center is the minimum, repeat until the cluster center does not change. But the quality of the clustering results depends on the selection of the initial cluster center, if isolate points are selected, iterations will significantly increased; If k points in the same class are selected, the algorithm will fall into local optimum. Therefore, it is important to reasonably determine a set of initial cluster center according to the distribution of sample points, to improve the stability of the k-means algorithm.

Based on the problems above, An adaptive method to select initial cluster center for k-means algorithm is proposed, so that initial cluster centers are located in high density area and have a certain distance with each other. Experiments show that the method has higher stability and accuracy than the traditional k-means algorithm and some other similar improved algorithms.

K-MEANS ALGORITHM AND ITS RESEARCH STATUS

Suppose $X = \{X_1, X_2,...,X_n\}$ is the data set to do clustering analysis, X_i is a d dimension vector, n is the number of sample points, k is the number of categories. k-means algorithm steps are as follows:

- Selected k sample points at random from X as the initial center set $A = \{A_1, A_2,...., A_k\}$
- Calculate the distance each sample point X_i to each cluster center point d (X_i, A_v), $X_i \in X$, $A_v \in A$, so that each sample point is assigned to the nearest center point
- Calculate the average distance of each cluster
- Repeat 2 and 3 until the adjacent two cluster center set does not change

Because k-means randomly select initial center and initial cluster center has a great influence on the clustering results, which means that different initialization of cluster center may obtain different clustering results. In order to overcome this shortcoming, some researchers have put forward a series of improved methods from different views. To make the performance improved by adjusting the iteration (Dhillon *et al*., 2002), To choose the farthest k points in high density area as the initial cluster center (Fu and Chen, 2011). To divide the data using two farthest

Corresponding Author: Ma JiMing, Zhengzhou University of Light Industry, 450002, Zhengzhou, China

data object as the start of initial center set (Chen *et al*., 2012), so repeatedly, until they get k initial cluster center, but this method may take isolate points which can increase the number of iterations. To get k centers according to the distribution density of data objects and calculate the vertical midpoint of two nearest points (Zhou and Shi, 2012). To choose points whose distance to the initial center set are greater than a threshold value as initial cluster center and then adjust threshold according to number of the initial clustering center that meet the condition (Liu and Zhang, 2011).

ADAPTIVE INITIAL CLUSTER CENTER SELECTION K-MEANS ALGORITHM

Thoughts of the algorithm: In order to get better clustering results, location of initial points should conform to the distribution characteristics of the sample points. And in order to avoid algorithm falling into local optimum, any two initial points should be a certain distance apart.

So, first calculate the point density of all sample points in sample point set X and make the maximum density point as the first initial cluster center, then add this point to the initial point set.

Second, randomly select one point from the rest of the sample points, if its distance to each initial cluster center is greater than the limit value w, we also add this point to the initial point set, else delete it from X. Repeat screening, until the sample point set X is empty. If the value of w is appropriate, the number of initial cluster center we get is just. Compared with the algorithms in Ref. De-sheng Fu and this proposed algorithm, except for the first initial center, the rest k-1 initial point selection method is not the same.

Assumes that the collection of sample points is X = $\{X_1, X_2, ..., X_n\}$

Definition 1: Euclidean distance formula between two d dimension sample points X_i and X_j is:

$$d(X_i, X_j) = \sqrt{(X_i - X_j)^T (X_i - X_j)} \quad (1)$$

Definition 2: Average distance formula between twodimension sample points X_i and X_j is:

$$MeanDist = \sum_{\substack{i=1:n \\ j=(i+1):n}} d(X_i, X_j) / C_n^2 \quad (2)$$

Among them, n is the total number of sample points, C_n^2 is the number of permutation and combination to pick any two points from all sample points.

Steps of the algorithm

Input: Collection of sample points X to do clustering analysis

Number of categories k

Output: k classes

The following is execution process of the algorithm, the corresponding flowchart is shown as Fig. 1:

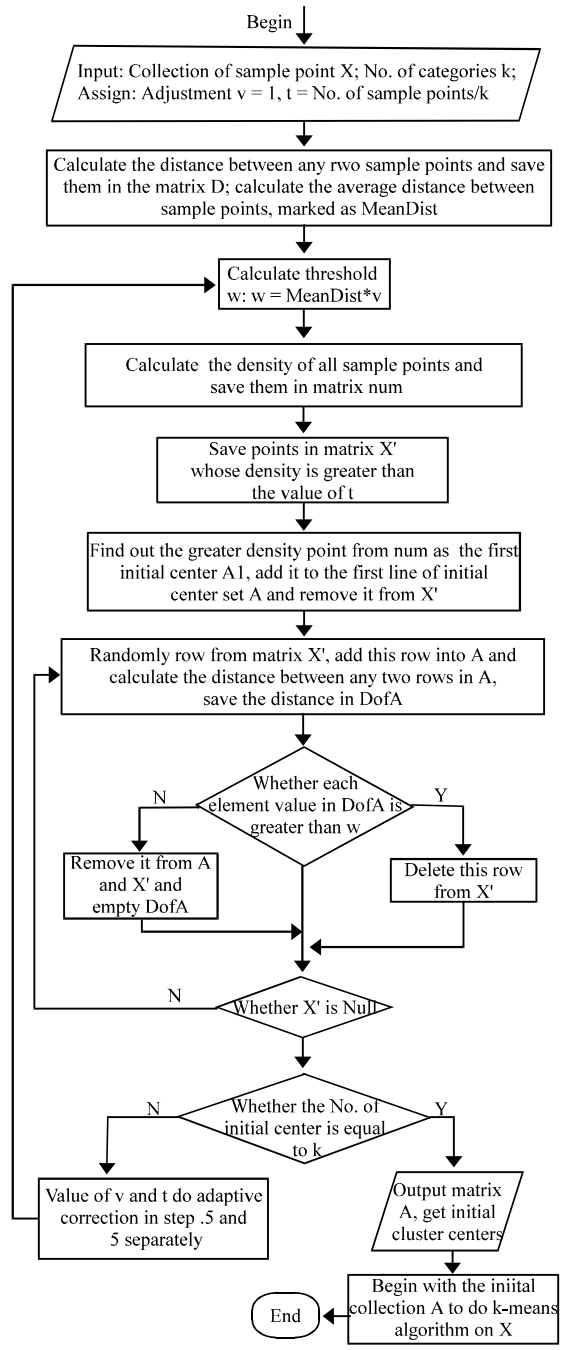

Fig. 1: Flowchart of the algorithm

- Calculate the distance between any two sample points, set the distance from each point to their own to be infinite and save them in the matrix D
- Calculate the average distance between sample points, marked as *MeanDist*, set threshold w = MeanDist*v, v is the adjustment coefficient and its initial value is 1
- Calculate the density of all sample points and save them in the matrix num. If one point's density is greater than the value of t, save the point in matrix X'. The initial value of = Total number of sample points/number of category
- Find out the greatest density point from X' as the first initial center, add it to the first line of initial center set A and remove it from X'
- Randomly select one row from X', add this row into A and calculate the distance between any two rows in A, save the distance in matrix DofA. If each element value in DofA is greater than w, then deltete this row from X', or remove it from A and X' and empty DofA
- Repeat 5 until X' is empty. If the number of initial center is greater than k, then jump to 7); If the number of initial center is less than k, then jump to 8); If the number of initial center is equal to k, then jump to 9);
- v value is incremented by 0.5 and modify t value slightly, generally keep the number of isolate points to be about 10-20% of total sample points. Return 2).
- v value is decremented by 0.5 and modify t value slightly. Return 2)
- Output matrix A, get k initial cluster centers
- Begin with the initial collection A to do k-means algorithm on X

Time complexity analysis on the algorithm: The new proposed algorithm is divided into two parts: Adaptive initial center selection optimization and the traditional k-means algorithm. Suppose r is the times to adjust the value of w, n is the total number of sample points, k is the category number and the time complexity of initial center selection optimization algorithm is:

$$O(\frac{1}{2}n^2 + (kr + r)n)$$

Traditional k-means algorithm's time complexity is O (dkln), where d is the dimension of the sample points, I is the number of iterations required for convergence. So, this study's overall time complexity is:

$$O(\frac{1}{2}n^2 + (kr + r + dkI)n)$$

bigger than traditional k-means algorithm, but to use the selected initial cluster center as the starting point can reduce the number of iterations of k-means algorithm, significantly improve the stability of the k-means algorithm, increase the accuracy of k-means algorithm.

EXPERIMENTAL RESULTS AND ANALYSIS

The new algorithm implementation platform is Windows XP, programming environment is MATLAB 7.0, Processor is 1.87 GHZ Intel Pentium and memory capacity is 2.0 G. Use four groups of data set which are Iris, Wine, New-thyroid and Glass from standard test data set UCI as test data. In which, Iris has 150 sample points, divided into three classes and each sample point has four attributes values:

- Wine has 178 sample points, divided into three classes and each sample point has 13 attribute value
- New-thyroid has 215 sample points, divided into three classes and each sample point has 5 attribute value
- Glass has 214 sample points, divided into seven categories and each sample point has 9 attribute values

Each data point in these data sets has been accurately classified, thus it is intuitive to calculate accuracy of clustering results.

According to experimental experience, in Iris the value of is 1.85, the value of is 36; in Wine the value of is 1.9, the value of is 45; in New-thyroid the value of is 2, the value of is 45; in Glass the value of is 0.5, the value of is 6; In order to verify the effect of the algorithm proposed in this study, firstly it is compared with traditional k-means algorithm that randomly select initial cluster center.

In the whole experiment process, for a particular data set, do clustering analysis using traditional k-means algorithm and improved algorithm in this article each for 15 times, record their run time, number of iterations and the accuracy of clustering results. The test results are shown in Table 1. For each set of comparative experiments, the highest average accuracy, the shortest execution time and the least number of iterations are marked with bold italics.

Experimental results indicate that for each data set the highest and lowest accuracy vary from 9-20%, using k-means clustering algorithm, while by the algorithm presented in this study the lowest and highest accuracy are the same except the data set Glass. For each data set, the minimum clustering accuracy by the algorithm in this study is higher than the average accuracy of k-means algorithm and its iterative times are less than k-means algorithm. So, the improved algorithm in this study has

Table 1: Comparison of k-means and the new algorithm

Data set	Algorithm	Clustring accuracy (%) Highest	Lowest	Average	Average time (ms)	Average iterations
Iris	K-means	89.33	56.67	82.79	54.33	7.2
	This study	89.33	89.33	89.33	65.67	4
Wine	K-means	64.61	56.74	63.04	65.13	9.13
	This study	70.22	70.22	70.22	646	8.73
New-thuroid	K-means	86.05	65.12	77.49	71.8	12
	This study	86.05	86.05	86.05	484	9
Glass	K-mean	54.21	39.25	44.39	41.6	14.33
	This study	54.67	47.2	52.03	592.25	13.4

Table 2: Clustring accuracy of similar algorithums

Improved algorithum	Clustring acuracy (%) Iris	Wine	New-thyroid	Glass
A	83.33	96.63	Not involved	Not involved
B	86.27	71.35	Not involved	Not involved
C	88.40	94.72	Not involved	50.28
D	90.00	59.63	78.35	Not involved
This study	89.33	70.22	86.05	52.30

better stability, higher accuracy and less number of iterations. But the execution time of the algorithm in this study increases obviously, especially for Wine data set with the highest number of sample points and attribute, most of the time is spent in calculating point density and the distance between point and point.

In addition, the references of Han's algorithm (Han *et al*., 2010) marked as A, Zhou's algorithm (Zhou and Shi, 2012) marked as B, Liu's algorithm (Liu and Zhang, 2011) marked as C and Cao's algorithm (Cao *et al*., 2009) marked as D are the optimization of initial cluster center selection about k-means, their comparison results are shown in Table 2.

As Table 2 shown, the clustering accuracy on Glass and New-thyroid data set is the highest by using new proposed algorithm.

CONCLUSION

A new adaptive method to select initial cluster center for k-means algorithm is proposed, multiple sets of experimental results show that this new method can overcome k-means algorithm's defect of instability, obtain better clustering effect and compared with other improved algorithms this algorithm has certain advantages.

But the new algorithm has two defects: One is that its time complexity is bigger and the clustering accuracy of Wine data set has a gap with some literatures; the other is that algorithms involved is static algorithms, which only can analyze current data and have a limit on dealing with time-series data.

Thus reducing time complexity, further improve the accuracy rate and the ability to deal with different kinds of data is the next research content. In follow-up studies, we plan to combine dynamic clustering algorithm with application problems in the related fields of medical diagnosis and communication, so that the algorithm proposed can be improved and perfect.

ACKNOWLEDGMENTS

The authors would like to greatly thank Haibin Zhu, PhD (Professor and Coordinator of Computer Science Program Director, Collaborative Systems Laboratory Nipissing University, North Bay, Canada) for examined the full content and put forward many valuable suggestions and to the anonymous reviewers for their constructive advices that have helped in improving this study.

This study is supported by the National Natural Science Funds (No. 61374014), China.

REFERENCES

Cao, Z.Y., Z.L. Zhang and Y.T. Li, 2009. Quickly find the initial cluster center in K-means algorithm. J. Lanzhou Jiao Tong Univ., 28: 15-18.

Chen, G.P., W.P. Wang and J. Huang, 2012. An improved K -means algorithm to choose initial cluster center. Mini-Micro Syst., 33: 1320-1323.

Dhillon, I.S., Y. Guan and J. Kogan, 2002. Refining clusters in high dimensional data. Proceedings of the 2nd SIAM ICDM, Workshop on Clustering High Dimensional Data, April 2002, Arlington, VA, USA.

Fu, D.S. and Z. Chen, 2011. Improved k-means algorithm and realization based on density. Comput. Appl., 31: 432-434.

Han, L.B., Q. Wang and Z.F. Jiang, 2010. An improved initial cluster center selection method in K-means algorithm. Comput. Eng. Appl., 46: 150-152.

Liu, Y.M. and H.X. Zhang, 2011. Initial center selection method in K-means algorithm based on variable threshold. Comput. Eng. Appl., 47: 56-58.

Tan, P.N., M. Steinbach and V. Kumar, 2011. Introduction to Data Mining. Posts and Telecom Press, Beijing.

Zhou, W.B. and Y.X. Shi, 2012. K-means cluster center selection optimiza ion algorithm Based on density. Comput. Appl. Res., 29: 1726-1728.