Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 2, No. 1, 2022

Initial Cluster Centers Based on Moving Two Lines Approximation in K-means Algorithm

Wenyue Feng, Shuangxia Xuan

Anhui University of Finance and Economics, Bengbu 233030, China

Abstract: The main shortcoming of K-means clustering algorithm is its great dependence on the initial cluster center point. Based on the moving two lines approximation model, this paper gives a method to pick the initial cluster center of k-means clustering. Numerical experiments and comparison criteria show that this method can get better clustering effect.

Keywords: Initial cluster centers; Moving two lines approximation; K-means algorithm.

1. Introduction

K-means algorithm is a hard-clustering algorithm and a representative of typical prototype-

based clustering methods. It is a certain distance between the data points and the prototype as the objective function of optimization, and uses the method of finding the extreme value of the function to get the adjustment rules of iterative operation. K-means clustering maximizes the distance between objects within a class, while minimizes the distance between classes.

K-means clustering algorithm is highly dependent on the selection of initial values. An

inappropriate initial value causes the algorithm to converge to a local minimum, so a lot of

work has been done on the selection of initial clustering centers.

Fayyad et al. [1] give a fast and efficient algorithnm operated over small subsamples of a given database for refining an initial starting point. Using the global optimization ability of genetic algorithm to improve the traditional K-means algorithm to prevent only local optimal solutions is described in [2,3]. Likas et al. present the global k-means algorithm which is an incremental approach to clustering that dynamically adds one cluster center at a time through a deterministic global search procedure [4]. Kkan and Ahmad propose an algorithm to compute initial cluster centers for k-means clustering. This algorithm is based on two observations that some of the patterns are very similar to each other and that is why they have same cluster membership irrespective to the choice of initial cluster centers [5]. In order to get the initial cluster centers for k-means algorithm, cells are partitioned one at a time until the number of cells equals to the predefined number of clusters, k. The centers of the k cells become the initial cluster centers [6]. Later a new algorithm for initial cluster centers in k-means algorithm is given in [7]. Two principal variables are selected according to maximum coefficient of the variaton and minimum absoute value of the correlation.

The rest of this paper is organized as follows. Section 2 introduces the Initial cluster centers computing based on moving two lines approximation. Section 3 introduces the comparision criteria such as error percentage and rand index. Section 4 elaborates the implementation of

the algorithm and presents experiments and results better. Section 4 concludes the paper.

2. Initial cluster centers computing based on moving two lines approximation

Moving multiple curves approximation is based on [8], while moving two lines approximation is a particultiple curves approximation of moving multiple curves approximation [9].

There using two lines in the moving multiple curves approximation model and modify the model as follows.

$$min \sum_{i=1}^{n}(a_1x_i + b_1y_i + c_1)^2(a_2x_i + b_2y_i + c_2)^2\theta(\|P_i - O_x\|)$$

$$s.t.\ \ {a_1}^2 + {b_1}^2 = 1 \qquad {a_2}^2 + {b_2}^2 = 1 \tag{1}$$

Where o_x is a projected point of Xon an underlying curve and X is a fixed point called a reference point near the underlying shape which is acquired according to partitioning the space occupied by $\{P_i\}_{i=1}^n$. (x_i, y_i) is the coordinate of , $P_i, i = 1,2,\cdots,n$. $a_1x+b_1y+c_1=0,\ \ a_2x + b_2y + c_2 = 0$ are algebraic equations of two lines. $a_1x_i + b_1y_i + c_1$ and $a_2x_i + b_2y_i + c_2$ are two algebraic distance from P_i. And

$$\theta(\|P_i - O_x\|) = \frac{e^{-\frac{||P_i - O_x||^2}{\rho^2}}}{\sum_{P_i \in \{P_i\}_{i=1}^n} e^{-\frac{||P_i - O_x||^2}{\rho^2}}}$$

$\rho$is a adjustment parameter.

To reduce the instability and expensive computation, moving two lines approximation (1) is rewrite as (2).

$$min \sum_{i=1}^{n}(a_1x_i + b_1y_i + c_1)^2(a_2x_i + b_2y_i + c_2)^2\theta(\|P_i - X\|) \tag{2}$$

see [9] for more details.

3. Comparision criteria

To compare the clustering results, two clustering criteria are presented here. One is error percentage [7] and another is the rand index [10].

Error percentage is defined as follows

$$Error = \frac{\varepsilon}{n} \times 100 \tag{3}$$

where $\varepsilon$is misclassified observations and $n$is the total number of observations in datasets.

The higher the value of Error, the better the result.

Suppose $Q_1 \subseteq \{P_i\}_{i=1}^n, Q_2 \subseteq \{P_i\}_{i=1}^n$ represent partitions of $\{P_i\}_{i=1}^n$ by k-means algorithm and real cluster memberships respectively. For each object pair $\{P_i, P_j\}$ there

are four possible scenarios:

a:P_i and P_j are in the same cluster in Q_1 and in the same cluster in Q_2.

b:P_i and P_j are in different cluster in Q_1 but in the same cluster in Q_2.

c:P_i and $P_j$are in the same cluster in Q_1 but in different cluster in Q_2.

d:P_i and P_j are in different cluster in Q_1 and in different cluster in Q_2.

Rand index is given as follows

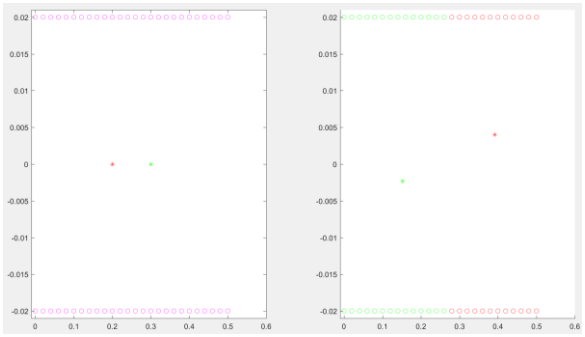
$$rand = \frac{a+d}{a+b+c+d} \tag{4}$$

And the lower the value of rand, the better the result.

4. Experiments and results

4.1. Random initial cluster centers

We consider the sampling points $\{P_i\}_{i=1}^n$ from two curves $C_1 : y\,1 = 0.02$ and $C_2 : y\,2 = -0.02$, see the pink circles in Figure 1, Figure 2 and Figure 3. We pick the initial centers $(0.2,0), (0.3,0)$ in Figure 1 and $(0.4,0.02), (0.1,-0.02)$ in Figure 2 and use k-means algorithm to cluster $\{P_i\}_{i=1}^n$ into two classes, see Figure 1 (b) and Figure 2 (b). Both of them are unreasonable results. We want the ones with the ordinate 0.02 to be one class and the ones with the ordinate -0.02 to be the other class.



(a) (b)

Figure 1. (a) the pink circles are original data and the red star and green star are (0.2,0),(0.3,0) respectively, which are random initial centers. (b) Based on (0.2,0), (0.3,0) , the result is given by k-means algorithm, there is a class in red circle and a class in green circle, the red star and green star are the final cluster centers.

4.2. Initial cluster center given by moving two lines approximation

It can be seen from the above results that if the initial clustering center is not appropriate, no good clustering result can be obtained. There the initial cluster center was obtained by moving two lines approximation. We use all the points $\{P_i\}_{i=1}^n$ in the moving two lines approximation model (2) without partition. $X$is the mean of $\{P_i\}_{i=1}^n$, denoted a pink star in Figure 3. And then from (2), two blue lines are obtained showing also in Figure 3. The computed two blue target points (0.25,0.02), (0.25,-0.02) according to (2) are considerd two initial centers of k-means algorithm, see Figure 4(a). Depend on the two initial centers from (2), we get a reasonable clustering result, see Figure 4 (b) . Besides, with the help of formula (3) and (4), we can compute error percentage and rand index for diffrent initial cluster centers, see Table 1. We found that the initial value from the moving two lines approximation corresponds to the smallest Error, and the maximum rand index. Its clustering effect is the best.

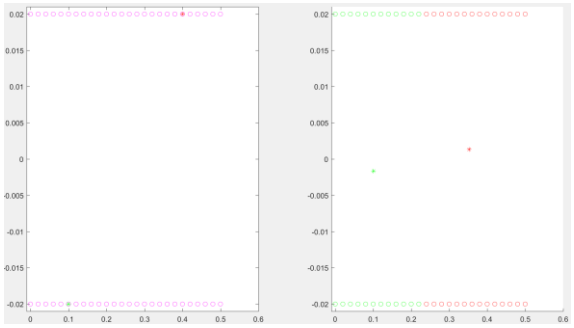


Figure 2. (a) the pink circles are original data and the two blue stars are (0.4,0.02),(0.1,-0.02) respectively, which are initial centers. (b) Based on (0.4,0.02),(0.1,-0.02) , the result is given by k-means algorithm. There is a class in red circle and a class in green circle. The red star and green star are the final cluster centers.

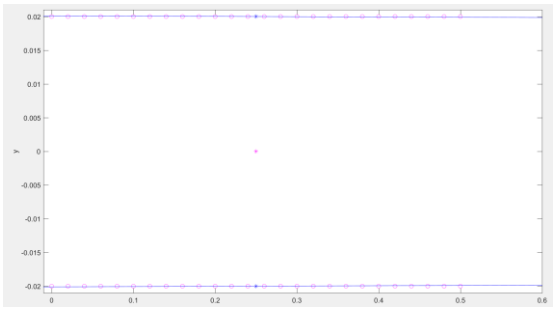


Figure 3. Based on Moving multiple curves approximation, we compute two lines and the two cluster initial centers for k-means algorithm. The pink star is X in (2).

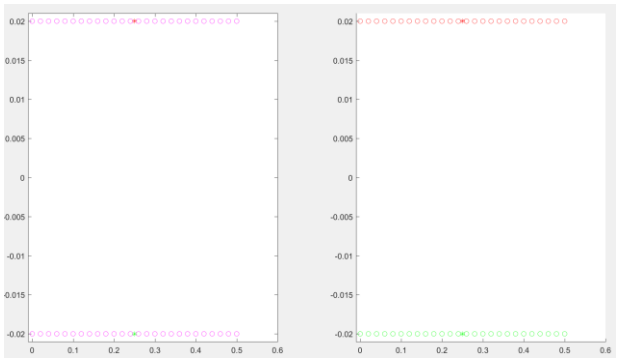


Figure 4. (a)the pink circles are original data and the two blue stars are (0.25,0.02),(0.25,-0.02), which are initial centers obtained from Fig.3; (b) The results given by k-means algorithm using the initial luster centers from (a), there is a class in red circle and a class in green circle. The red star and green star are the final cluster centers.

Table 1. Error percentage and rand index for diffrent datasets

Initial cluster centers	Error percentage	Rand index
(0.2,0), (0.3,0)	50	0.4902
(0.4,0.02), (0.1, -0.02)	50	0.4902
(0.25,0.02), (0.25, -0.02)	0	1

5. Conclusion

A new method to select the initial cluster center of k-means algorithm, that is, moving two lines approximation have provided in this paper. Moving two lines approximation is a particular case of moving multiple curves approximation. We just choose two algebraic equations of two lines in the model. For given dataset, by moving two lines approximation, we can find better initial center in k-means algorithm for given observation data and achieve good clustering effect than

random initial cluster center. Here, only moving two approximation is used to deal with binary classification problems. In the future, moving multiple curves approximation can be considered to deal with multi-classification problems, and it may be used to deal with practical dataset.

Acknowledgements

This work is supported by the University-level Projects (ACKYC20050).

References

[1] Fayyad U, Reina C, Bradley PS: Initialization of iterative refinement clustering algorithm. In： Proc of the Fourth International Conference on Knowledge Discovery and Data Mining. AAAI，Menlo Park. 1998:194-198.

[2] L. X. Bang, J. H. Yang, M.G. Wang: Using genetic algorithm to improve K-means clustering algorithm in clustering analysis[J]. Mathematical Statistics and Applied Probability, 1997, 1 2(4):350—356. (In Chinese)

[3] Krishma K, Murty MN: Genetic k-means algorithm[J]. IEEE Transactions on System, Man and Cybemetics, Part B. 1999. 29(3): 433-439.

[4] Likas, A., Vlassis, N., Jakob, J.V.: The global k-means algorithm algorithm. Pattern Recognition 2003, 36, 451–461.

[5] Khan, S.S., Ahmad, A.: Cluster center initialization algorithm for k-means clustering. Pattern Recognition Lett. 2004, 25, 1293–1302.

[6] Deelers, S., Auwatanamongkol, S.: Enhancing k-means algorithm with initial cluster centers derived from data partitioning along the data axis with the highest variance. Internat. J. Comput. Sci. 2007, 2, 247–252.

[7] Murat Erisoglu, Nazif Calis, SadullahSakallioglu: A new algorithm for initial cluster centers in k-means algorithm. Pattern Recognition Lett. 32, 1701–1705.

[8] Yang Z, Kim T: Moving parabolic approximation of point clouds[J]. Computer-Aided Design, 2007, 39(12): 1091-1112.

[9] W. Y. Feng, Z. W.Yang, J. S. Deng: Moving multiple curves/surfaces approximation of mixed point clouds[J]. Communications in Mathematics and Statistics, 2014, 2(1): 107-124.

[10] Rand, W.M.: Objective criteria for the evaluation of clustering methods. J. Amer. Statist. Assoc. 1971, 66, 846–850.