IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 4 Issue 3, March 2017
ISSN (Online) 2348 – 7968 | Impact Factor (2016) – 5.264
www.ijiset.com

# Public Bicycle Site Area Division Based on Improved K - Means Algorithm

**Jing Zhang*, Yan Liang*, Wenjun Wei***
* Chongqing Key Laboratory of Signal and Information Processing, Chongqing University of Posts and Telecommunications,Chongqing 400065, P. R. China

**Abstract**:
In order to solve the problem that the scheduling path is too long during the process of public bicycle scheduling, this paper proposes an improved k-means clustering algorithm. The k-means algorithm is used to estimate the k-center points as the initial center point. After the k-means algorithm clustering, the scheduling requirements parameters are introduced, and the k-means algorithm is obtained by the edge site to obtain the new regional partition result .The example shows that the algorithm has good global convergence, which can effectively improve the problem that scheduling path is too long and the scheduling efficiency is low.

**Key words**: *Clustering analysis, region division, k - means algorithm*

## 1.Introduction

With the development of the city, the density of the population and the motor vehicle, the pressure of urban traffic is becoming more and more serious so it will cause traffic jams in each major cities. Urban public bike rental system can effectively reduce the pressure of the traffic jam on the road, but it develops relatively late in our country, so the bike scheduling system is not yet perfect. In order to solve the difficulty in borrowing and returning that appears in the bike scheduling system, the efficiency of vehicle scheduling needs to be improved. People begin to pay close attention to how to divide the site according to its geographical location, to shorten the scheduling path and its period. Therefore, the site clustering division has drawn public attention and been wildly studied. Clustering algorithm is a series of no label objects which is divided into several groups with similar characteristics of the process in the collection. And in the process of dividing, we should ensure the similarity as large as possible as in the same site meanwhile the similarity as small as possible in the different site.

In the clustering algorithm based on region partition, k - means is a kind of iterative clustering algorithm based on partition, because of its practicability, versatility, low computational complexity, as well as it is not sensitive to data input order, fast convergence, k-means become one of the most popular and useful classic algorithms in daily life and production practice. But in the process of clustering, k - means algorithm has such shortcomings as it is sensitive to outliers, due to the difference of initial center point selection and the result is different, and it is easy to fall into regional optimal solution, we need to compute the distance of the center of each data point. With the development of the technology of data, and the increasing amount of clustering analysis data, highly complexity of the time, make the algorithm cannot adapt to the mass data processing. As for traditional k - means algorithm, many scholars put forward the improving measures. Li Xiaoyu proposes an improving k - means algorithm based on Hadoop, by introducing the Canopy algorithm to initialize the initial clustering center of k - means algorithm, to overcome the problem of the

uncertainty of initial center as it is easy to plunge into a regional optimal solution. Sina Khanmohammadi proposes a combined method (KHM - okm) to overcome these limitations by combining the harmonic mean with the overlapping k-mean. The main idea of khm - okm method is to use the output of KHM method to initialize clustering center of okm method.

On account of the above algorithm, based on the traditional K means clustering algorithm as the foundation, this paper proposes a new optimization based on clustering center, improve scheduling demand for K - means algorithm. In the case of cluster number K is known, first of all, through the statistics of each site's operating conditions and the distance with other sites to choose K initial clustering centers, secondly by optimization of the K - means algorithm to cluster results, and then introduce every site's demand for scheduling, fixed edge point in the clustering results to get the final clustering results. Theoretical analysis and experimental results show that due to improving the initial clustering center in advance, the algorithm of clustering center is more optimal in all regions than the traditional k-means algorithm, and the number of scheduling car is also reduced. So the algorithm can guarantee the accuracy of clustering and improve the efficiency of clustering by reducing the amount of data at the same time.

## 2.K -means Clustering algorithm

Clustering algorithm is to automatically divide a set of no label data automatically into several categories, and ensure that each type of elements with similar characteristics after the division. K -means algorithm is a kind of indirect clustering method based on similarity measure among the samples. This algorithm takes K as a parameter, divide N data object into K classes, making the highest similarity of the elements in each class while it is lowest similarity among different category. K -means algorithm is a kind of typical point by point iterative dynamic clustering algorithm, the point is the error sum of squares as the objective function.

Specific algorithm is as follows:

Set the data set to

$$X = \{x_i \mid x_i \in R, i = 1,2,3,\ldots, n\},$$

K Clustering centers as $m_1, m_2, \ldots, m_k$,

The Euclidean distance between the data points

$$d(x_i, x_j) = (x_i - x_j)^T (x_i - x_j)$$

The objective function is defined as

$$J = \sum_{k=0}^{k} \sum_{j=0}^{n_j} d(x_j, m_k)$$

Algorithm process is as follows:

Input: data object set, cluster number K.

Output: tend to be stable and meet the objective function of K data object classes.

Process:

(1)Randomly selected K objects as the initial clustering center from a data object set;

(2)Calculate the distance respectively from other object sets to the K center, based on the principle that which one is closest to the center, and then assign other data object to the corresponding category;

(3)Respectively calculate mean in every category, form a new cluster center;

(4)Respectively calculate the distance between other data objects and the new clustering center, based on the principle that which one is closest to the center, reassign to other data object to the new category;

(5)Return to step 3, until no longer generate a new clustering center, and the data object contained in each category no longer change or minimum objective function tend to be stable.

The original k - means algorithm exists two major defects: (1) k value is given in advance, but the original data is a pile of no label data, it is difficult to determine how many classes are divided into, so it is difficult to determine k value; (2) the coupling between clustering result and the initial clustering center is too strong, the results of the selection of different initial values may be different.

Algorithm designed in this paper is suitable for the public bicycle intelligent scheduling system, due to the determination of scheduling a commuter, K value is fixed, so the first defects does not affect this algorithm in this paper, we improve the second defect, put forward a new method to determine the initial value, moreover, we also add scheduling demand as a new parameter to correct division results.

## 3.A quadratic algorithm k-means based on the demand of scheduling

### 3.1 Basic idea

For the problem that k - means algorithm selects different initial clustering center and gets different computing results, a new k - means clustering algorithm based on scheduling demand firstly find the clustering center, with the help of public bicycle operation rule, to ensure the distance between initial clustering center and the finial clustering center is most close ; then use the k - means algorithm to divide the rest of the data points to different areas; and then introduce the scheduling demand parameters, statistics regional scheduling demand; Finally use the k - means algorithm to modify every edge sites in different regions once again which can reduce the amount of calculation in the process of clustering, so it improves the efficiency of clustering, and the absolute value of scheduling demand of every regional is minimum.

### 3.2 The correction of the initial center

The result of K -means algorithm makes a great difference with the initial center, in order to guarantee the correctness of the result, we need to choose appropriate site as the initial center. According to its own characteristic of the public bike rental system, we give the following definition:

1. Set all sites of one region A $S = \{S_1, S_2, \cdots, S_n\}$ , each site and the corresponding operation parameters $RR = \{S_{1r}, S_{2r}, \cdots S_{nr}\}$

2. Set the standard of A sites. Set the site of the data in accordance with the operating parameters and then divided them into class A sites and the no class A sites, standards of classification are as follow $S_{ir} \geq RR_A, i \in [1, n]$, the class A sites are $S_A = \{S_{A1}, S_{A2,} \cdots S_{Aj}\}$ , $RR_A = \alpha \times \sum_{i=1}^{n} S_{ir}$ , $\alpha$ as configurable parameters, it is generally determined by the size of the parameter set, namely $\alpha \approx n/5$

3. The distance between two different sites in class A. Dis (X, Y) is the distance between the site of X and Y, and then calculate the distance between each site in class A site with other site $D_{Ai} = \sum_{t=1, t\neq i}^{j} dist(S_{Ai}, S_{At})$ and generate A sites $D_A = \{D_{A1}, D_{A2,} \cdots D_{Aj}\}$.

4. The initial clustering center. Statistics and compare the size of class A site in centralized

data, regard largest site corresponding to centralized data as the first cluster center, namely $M_1 = \max(D_A)$ ; regard the second largest site corresponding to the data as the second cluster center M2, the rest can be done in the same manner, until find K clustering centers, forming the initial clustering center set $M = \{M_1、 M_2 \cdots M_K\}$.

### 3.3 The new process of the algorithm

Input: site data set D, cluster number K, initial center point set $\{M_1、 M_2 \cdots M_K\}$ Demand for scheduling of every site $\{DP_1, DP_2, \cdots DP_n\}$, scheduling bike volume V

Output: the stability of K data classes

Process:

(1)Regard focus point of initial center $\{M_1、 M_2 \cdots M_K\}$as "the center" of each class, dividing into K classes respectively.

(2)Based on the principle that which one is most close to the center, assign the original data to each corresponding class.

(3) Statistic the demand of scheduling of each class in total.

(4)Place the edge of the site that whose maximum absolute value of total scheduling demand and greater than V area $A_i$ into the data set C.

(5)Place the edge of the site which is adjacent to area $A_i$ as well as its maximum absolute value of total scheduling demand is greater than area V or its scheduling demand is opposite to the area $A_i$ into the data sets C. Using k-means algorithm to correct the edge sites in data set C again, and then redistribution the edge sites in data set C to these k areas, the principle is as follow: $d(x_i, x_j) = 0.6 * (x_i - x_j)^T (x_i - x_j) + 0.4 * |DP_i + DPj|$

(6) Recalculate the demand of scheduling of the regional class.

(7) Repeat step (4) - (6) until the process does not meet the condition of step (5).

(8Regionalism plan is finally finished.

## 4. Experimental results and analysis

### 4.1 Experimental analysis

This section is aiming at testing the function of the improved k - means algorithm. Experiment platform's computer is configured by Intel core i5 A848 CPU, 6GB of memory, a 64 - bit operating system, algorithm programs are written by MATLAB 2014 A. In the experiments, we choose data set of the real public bicycle sites as test data set. Figure 1 is the figure when we use k - means algorithm directly to divide the original data set into four categories (k value); Figure 2 is the classification figure when we use the improved k-means algorithm mentioned in this paper, we can see that some parts of the edge of the sites has changed its color, namely changed the class they belong to. Figure 3 is the figure that compares the differences of regional scheduling demand when we use the improved k-means algorithm, we can see from the figure 3, the improved algorithm can effectively improve the scheduling demand of each region.

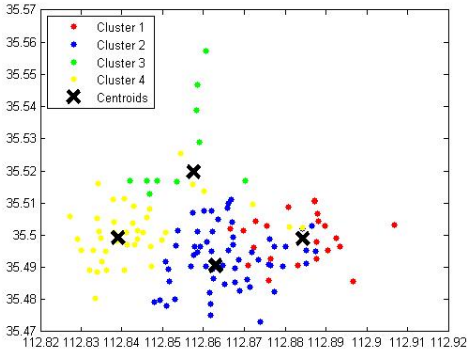


Figure 1：the classification of the data set after using k-means algorithm firstly

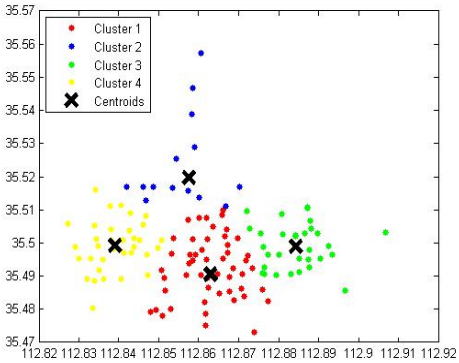


Figure 2：the classification of the data set after using the improved k-means algorithm

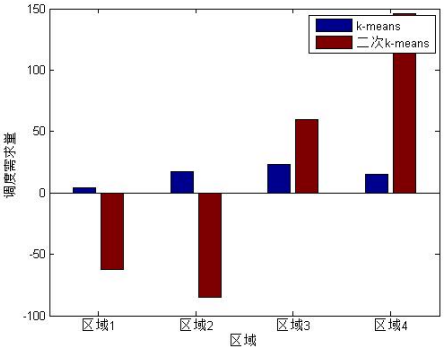


图 3：the contrast of scheduling demand of each region after using two different k-means algorithm

### 4.2 Instance analysis

In order to verify the effectiveness of the algorithm that mentioned above, we choose the city public bike scheduling as an example to analyze. The data includes 130 sites in total in one region, the latitude and longitude of each site and the scheduling demands. If dispatch vehicles used for distribution is 4, and its maximum capacity V is 60, using the improved k-means clustering algorithm to divide 130 sites into four classes, figure 1 shows the difference of the scheduling before and after the improvement from the aspects such as time, distance, car consumption.

Figure 1 the contrast after using the improved k-means algorithm

| **algorithm** | **The quantity of scheduling bikes** | **scheduling frequency** | **scheduling distance (km)** | **scheduling time (h)** |
|---|---|---|---|---|
| **k-means** | 8 | 390 | 79.82 | 2.5 |
| **quadratic k-means** | 3 | 42 | 21.23 | 0.5 |

We can see from the figure 1 that it can effectively reduce the long distance caused by cross-regional scheduling when we applied the improved k-means algorithm into public bicycle system, it can also solve the problem of large time consumption and save the cost of the scheduling.

## 5. Conclusion

As for the traditional k means clustering algorithm, clustering effect largely depends on the choice of initial clustering center, this paper regard it as a starting point, obtaining the initial clustering center and correct it through collecting operating rule of the public bicycle

system. In classifying site, we not only take the site's own geographical location into consideration but also introduce the site scheduling demands parameters of the site to correct every edge of the regional sites again, in order to achieve a result that the absolute value of scheduling demand in each area is as small as possible and its similarity is as high as possible. By compared with the traditional clustering algorithm, the improved algorithm can solve the problem that the scheduling of the bicycle is sensitive to the initial value, and it improves similarity of the regional data set. The result of the example shows that the algorithm used in the public bicycle scheduling system can effectively improve the efficiency of the scheduling, reducing excessive scheduling and cost of scheduling.

**Acknowledgments**

This work was supported in part by the Program of 2016 Chongqing Graduate Scientific Research Innovation Project (CYS16171) and special fund of Chongqing key laboratory (CSTC).

## References

[1] Li Tingting. Study on Planning of Urban Public Bicycle Lease Point Location [D] Beijing: Beijing Jiaotong University，2010.

[2] Zhou Yi-wu, Cui Dandan, Pan Yong. A K-means clustering algorithm for optimizing initial clustering centers [J]. Microcomputer and Application. 2011(13)

[3] Wang Zhong, Liu Gui-quan, Chen En-hong. A K-means Algorithm for Optimizing Initial Center Point [J]. Pattern Recognition and Artificial Intelligence.2009.4.15:299-304

[4] Zhang Wen-ming, Wu Jiang, Yuan Xiao-jiao. K-means text clustering algorithm based on density and nearest neighbor [J]. Journal of Computer Applications. 2010(07)

[5] Sun Rongzong. A fast KNN text categorization algorithm [J]. Computer Knowledge and Technology.2016,6(1):174-175,178

[6] Celebi M E,Kingravi H A,Vela P A. A comparative study of efficient initialization methods for the K-means clustering algorithm[J]. Expert Systems with Applications,2013,40 (1) :200-210.

[7] Li Xiao-yu, Yu Li-ying. .Implementation and Application of a K-means Improved Algorithm Parallelization [J] Journal of University of Electronic Science and Technology of China.2017,46(1):61-68.

[8] Sina Khanmohammadi,Naiier Adibeig, Samaneh Shanehbandy.An Improved Overlapping k-Means Clustering Method for Medical Applications[J]. Expert Systems With Applications. 10.1016/j.atmosenv.2016.09.025