Optimization of Density-Based K-means Algorithm in Trajectory Data Clustering

Mei-Wei Hao[1(✉)], Hua-Lin Dai[2], Kun Hao[1], Cheng Li[1], Yun-Jie Zhang[1], and Hao-Nan Song[3]

[1] College of Computer and Information Engineering, Tianjin Chengjian University, Tianjin 300384, China
angelsamle@126.com, littlehao@126.com, licheng.mum@gmail.com, zhangyunjietj@163.com

[2] Computing Center, Tianjin Chengjian University, Tianjin 300010, China
99871382@qq.com

[3] Department of Electrical Engineering, Tsinghua University, Beijing 10000, China
shn14@mails.tsinghua.edu.com

Abstract. Since the amount of trajectory data is large and the structure of trajectory data is complex, an improved density-based K-means algorithm was proposed. Firstly, high-density trajectory data points were selected as the initial clustering centers based on the density and increasing the density weight of important points, to perform K-means clustering. Secondly the clustering results were evaluated by the Between-Within Proportion index. Finally, the optimal clustering number and the best clustering were determined according to the clustering results evaluation. Theoretical researches and experimental results showed that the improved algorithm could be better at extracting the trajectory key points. The accuracy of clustering results was 24% points higher than that of the traditional K-means algorithm and 16% points higher than that of the Density-Based Spatial Clustering of Applications with Noise algorithm. The proposed algorithm has a better stability and a higher accuracy in trajectory data clustering.

Keywords: K-means algorithm · Based on density
Characteristics of vehicle activity · Weighted density · Initial clustering center
Between-Within Proportion (BWP) index

1 Introduction

With the advent of the era of big data and the rapid development of mobile location services, trajectory data has become an important digital resource. Since the amount of trajectory data is very large and the quality of trajectory data is usually poor, it is becoming a hot issue that how to obtain the deep semantics of trajectory data by data mining and visualization analysis. Clustering algorithm as an effective technique for trajectory data feature extraction is widely used in trajectory data mining.

The K-means algorithm [1] as a typical partition-based clustering algorithm is widely used because of its simplicity and high efficiency. However, it requires users to determine the clustering number and the initial clustering centers based on the relevant

C. Li and S. Mao (Eds.): WiCON 2017, LNICST 230, pp. 440–450, 2018.
https://doi.org/10.1007/978-3-319-90802-1_39

experience or background field, and performs sensitively to the selection of the initial clustering centers and the order of the data input, which may result in unstable and inaccurate clustering results. Aiming at these problems, many scholars have proposed different improvement programs. There are density-based improvements: the literature [2] proposed the cluster center initialization algorithm (CCIA), which clusters according to the density-distribution information of sample data points to obtain the initial clustering centers, the literature [3] proposed selecting the sample data points with farthest from the global sample data center in the high-density area as the initial clustering centers. However, these algorithms are not applicable in the sample datasets of which the data distribution is relatively uniform. When filtering the sample data points according to the density, they need to add other parameters for auxiliary judgements. There are distance-based improvements: the literature [4] proposed the method of dynamically adjusting the selection of the initial clustering centers based on the principle that the distance within the cluster is less than the distance between clusters, the literature [5] proposed the algorithm which clusters by giving every sample data point the specific weighting factor according to the distance between the data and the initial clustering center to cluster, the literature [6] proposed determining the initial clustering centers based on minimal maximum algorithm and partitioning the sample dataset based on Euclidean distance. However, these algorithms when used in larger datasets increase the algorithm time complexity and reduce the operation efficiency. In addition, the clustering results are susceptible to interference of outliers. There are also density-distance-based improvements: the literature [7] proposed the algorithm which is local clustered by dividing the sample dataset into several small subsets with the centroid and weight, but the computational complexity increases in high-dimensional sample datasets, the literature [8] proposed selecting the initial clustering centers based on the square error criteria, but the efficiency of the algorithm obviously decreases in large-scale datasets.

Since the amount of trajectory data is very large and the structure of trajectory data is relatively complex, this paper proposed an improved density-based K-means algorithm. The improved algorithm this paper proposed can automatically obtain the clustering number and the initial clustering centers, and has a strong anti-noise interference ability. In addition, because of filtering the sample trajectory data points by characteristics of vehicle activity, the accuracy of the clustering results is effectively improved. This algorithm not only has important significance for the research of regular changes in road traffic, but also has positive significance for the exploration of human activity hotspots area.

2 Related Description and Definition

Assume the set of trajectory data points $P = \{p_1, p_2, p_3, \ldots, p_n\} \in R^{d*n}, d = 3$, d means the number of the sample trajectory dataset dimension, which is three in this paper. The three dimensions include spatial longitude coordinates, spatial latitude coordinates and time dimension. n means the total number of the trajectory data points.

The following is some of the important concepts proposed in this paper:

Definition 1 (Density of a trajectory data point). The density [9] $Dens_r(p_i)$ of any one of the trajectory data points p_i is defined as follows:

$$Dens_r(p_i) = \sqrt{\frac{1}{N-1}\sum_{j=1,j\neq i}^{N}(r - Dist^2(p_i, p_j))} \tag{1}$$

r means the effective density radius; N means the total number of the trajectory data points contained within the radius; p_j means the j-th trajectory data point in the circle with the center p_i and radius r; $Dist(p_i, p_j)$ means Euclidean distance between p_i and p_j, that is, the real distance between p_i and p_j in the map, $Dist^2(p_i, p_j)$ means the square of Euclidean distance $Dist(p_i, p_j)$.

Definition 2 (Weighted density of a trajectory data point in turning state). After recognizing the straight or turning state of the trajectory data points (that is, the straight or turning activity state of the corresponding vehicle in this point), it is necessary to increase the filtering probability of the turning state of the trajectory data points by increasing the weight density. Thus, the concept of the weighted density of a trajectory data point in turning state is introduced. The weighted density [5, 9] $WDens_r(p_i)$ of any one of the trajectory data points in turning state p_i is defined as follows:

$$WDens_r(p_i) = \sqrt{\frac{1}{N-1}\sum_{j=1,j\neq i}^{N}(r - Dist(p_i, p_j))^2(1 + \frac{r - Dist(p_i, p_j)}{r})} \tag{2}$$

Definition 3 (Average distance of the trajectory data points). The average distance of the trajectory data points is defined as follows:

$$avgDist = \frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{i=1,i\neq j}^{n}(Dist(p_i, p_j)) \tag{3}$$

i and j means subscript, having. The average distance of the trajectory data points can effectively reflect the overall degree of the discretization of the trajectory dataset, and provide a valid basis for the better determination of the neighborhood radius.

Definition 4 (Neighborhood radius). The neighborhood radius as the effective density radius is not only directly involved in the calculation of the density value, but also determines how many points of the trajectory data points may be contained within the radius. Therefore, the appropriate neighborhood radius is critical to the density calculation. The neighborhood radius γ is defined as follows:

$$\gamma = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}\sum_{i=1,i\neq j}^{n}(Dist(p_i, p_j) - avgDist)^2} \tag{4}$$

The smaller the neighborhood radius is, the closer the trajectory data points in the neighborhood are.

Definition 5 (Length of trajectory step). The vehicle trajectory data are usually sampled at even time intervals. The length of trajectory step is used to reflect the average length of each sub-trajectory segmented by the trajectory data points. It also can be used to indirectly reflect the rate and other attributes of the vehicle's activity in the trajectory. The trajectory step ε is defined as follows:

$$\varepsilon = \frac{\sum_{i=1}^{m} L_i}{\sum_{i=1}^{m} P_i - m} \tag{5}$$

m means the total number of the trajectory; L_i means the length of each trajectory; P_i means the total number of the trajectory data points in each trajectory. The length of trajectory step is shorter, the rate of the vehicle is lower, and the density of the trajectory data points contained within the length is higher. It is generally considered that when the vehicle in the turning process the density of trajectory data points is high, and the rate of the vehicle is low.

3 Improved Density-Based K-means Algorithm

3.1 Selection of Initial Clustering Centers Based on Density

The improved algorithm was described as follows:

Input: The trajectory dataset X containing n objects, the predicted clustering number K, the minimum density threshold $minDen$, the neighborhood radius γ, the length of trajectory step ε

Output: The set T of the initial clustering centers containing K objects, the set P of the trajectory data points without noise

Begin:

1. Make sets $D = \{\}$, $D' = \{\}$, $W = \{\}$, $T = \{\}$, $P = \{\}$;
2. According to the formula (3) and the formula (4), calculate the neighborhood radius γ; according to the formula (5), calculate the length of trajectory step ε;
3. ***For*** $x_i \in X$, ***Do*** Give x_i a density label, and taking $r = \gamma$, according to formula (1), calculate the initial density $Dens_r(x_i)$ of x_i;
4. Calculate the minimum density threshold $\min Den = \frac{1}{n}\sum_{i=1}^{n} Dens_r(x_i)$; select the first 2K objects with the large initial density, which are put into the set D;
5. ***For*** $x_i \in X$, **Do**
   5.1 Taking $r = \varepsilon$, according to the formula (1), calculate the density $Dens_r(x_i)$ of x_i;
   5.2 ***If*** $Dens_r(x_i) \geq \min Den$, ***THEN*** Put x_i into the set D';

6. Take the objects in the set D into the set W (These objects only with the spatial and time attributes are the same as those of the set D' in the spatial dimension and the time dimension);
7. ***For*** $x_i \in W$, ***Do*** Give x_i a turning state label, and taking $r = \gamma$, according to the formula (2), calculate the weight density $WDens_r(x_i)$ of x_i, updating its density label in the set X and D;
   Repeat:
8. Take the first object d_i with no center point or boundary point label in the set D;
   8.1 Let d_i be a center of a new cluster, and give d_i and the corresponding x_i a center point label;
   8.2 ***For*** $x_i \in X$ and $x_i \neq d_i$, ***Do***
   If x_i is the point of direct density-reachable of d_i and x_i has no center point or boundary point label, ***Then*** Give x_i a boundary point label, and put it in this cluster;
   8.3 Update d_i's density label, ***If*** d_i has the turning state label, ***THEN*** Taking $r = \gamma$, according to the formula(2), calculate the weight density $WDens_r(d_i)$ of d_i, Else THEN Taking $r = \gamma$, according to the formula(1), calculate the density $Dens_r(d_i)$ of d_i;
   Until: Traverse all the objects in the set D, until each object in the set D has a central point or a boundary point label;
9. In the set D, select the first K objects with the center point label and the large density, which are put into the set T;
10. ***For*** $x_i \in X$, ***Do If*** x_i has no center point or boundary point label, ***THEN*** Give x_i a noise label, ***Else THEN*** Put x_i into the set P;
11. Output the results;

End

3.2 Optimized Selection of the Number K of Clusters

A lot of experiments and experience show [10] that the best clustering number K should be in the $[2, \sqrt{n}]$, n meaning the total number of sample data. Based on this, many scholars proposed a simple and effective way to select the clustering number K. The basic idea of it is searching for the K optimal value in the $[2, \sqrt{n}]$. That is, the sample dataset is clustered for each determined K_{opt} value, then evaluate the clustering results corresponding to the K_{opt} value by clustering validity function. When the clustering results are optimal, the corresponding K_{opt} value is the best clustering number K.

This paper used the BWP index [11] $BWP(j, i)$ as the clustering validity function to reflect the tightness within the cluster and the separation between clusters by the ratio of the clustering deviation distance $bsw(j, i)$ and the clustering distance $baw(j, i)$. j and i meaning the i-th sample data of the j-th cluster. The specific description of $BWP(j, i)$ is as follows:

$$BWP(j,i) = \frac{bsw(j,i)}{baw(j,i)} = \frac{b(j,i) - w(j,i)}{b(j,i) + w(j,i)} \tag{6}$$

b(j, i) means the minimum distance between clusters of the *i*-th sample data of the *j*-th cluster, that is, the minimum value of the average distance between the sample data and each other clusters. The specific description of *b(j, i)* is as follows:

$$b(j,i) = \min_{1 \le k \le C, k \ne j} \left(\frac{1}{n_k} \sum_{p=1}^{n_k} Dist(x_p^{(k)}, x_i^{(j)})\right) \tag{7}$$

C means the number of clustering; *k* and *j* means the cluster index; n_k means the total number of the sample data of the *k*-th cluster; $x_p^{(k)}$ means the *p*-th sample data of the *k*-th clusters; $x_i^{(j)}$ means the *i*-th sample data of the *j*-th clusters; $Dist^2(x_p^{(k)}, x_i^{(j)})$ means the square of Euclidean distance between $x_p^{(k)}$ and $x_i^{(j)}$. From the viewpoint of the degree of separation between clusters, the minimum distance between clusters *b(j, i)* is the bigger the better.

w(j, i) means the distance within the cluster of the *i*-th sample data of the *j*-th cluster, that is, average distance between the sample data and each other sample data in this cluster. The specific description of *w(j, i)* is as follows:

$$w(j,i) = \frac{1}{n_j - 1} \sum_{q=1}^{n_j} Dist^2(x_q^{(j)}, x_i^{(j)}) \tag{8}$$

From the viewpoint of the degree of tightness within clusters, the minimum distance between clusters *b(j, i)* is the smaller the better.

The higher the average BWP of all the sample data points is, the higher the quality of clustering results is. When the average BWP is maximum, the best clustering number *K* is K_{opt} (the clustering number corresponding to the clustering results). The specific description of the best number *K* of clusters is as follows:

$$K = K_{opt} = \max_{2 \le k \le \sqrt{n}} \left(\frac{1}{n} \sum_{j=1}^{k} \sum_{i=1}^{n_k} BWP(j,i)\right) \tag{9}$$

n means the total number of the sample data; n_k means he total number of the sample data of the *k*-th cluster.

3.3 Improved Density-Based K-means Algorithm with BWP Index

The basic idea of the algorithm this paper proposed was as follows. Firstly, successively selected the number K_{opt} of clusters in the $[2, \sqrt{n}]$, *n* meaning the total number of the trajectory data. Secondly, according to the selected clustering number K_{opt}, the improved density-based clustering algorithm selected the initial clustering centers by clustering iteration. Then the selected clustering number K_{opt} and the obtained initial clustering

centers were used for the improved K-means algorithm, to cluster the trajectory data and evaluate the clustering results by BWP index. Finally, when the BWP maximum, the corresponding clustering number K_{opt} was taken as the best clustering number K, and the corresponding clustering results were taken as the optimal clustering results.

The improved algorithm was described as follows:

Input: The trajectory dataset X containing n objects, the minimum density threshold *minDen*, the neighborhood radius γ, the length of trajectory step ε

Output: The best clustering number K, the set T of optimal clustering results

Begin:

1. Make sets $X = \{\}$, $T' = \{\}$, $T = \{\}$, $S = \{\}$;
2. ***For*** $k = 2$ ***To*** $[\sqrt{n}]$;
   - 2.1 According to the improved density-based clustering algorithm, select the k initial clustering centers;
   - 2.2 According to the improved K-means algorithm, the trajectory dataset without noise is clustered and the clustering results are put into the set T';
   - 2.3 According to the formula (6), calculate average BWP value *avgBWP* of the clustering results, and put (k, *avgBWP*) into the set S;
   - 2.4 Put the set T' into the set X, then make set $T' = \{\}$;
3. Comparing the *avgBWP* values of the objects in the set S, take the object *target* where the max *avgBWP* is located and record the subscript *index* of the object;
4. K is k of *target*; $T = X_{index}$, X_{index} means the *index*-th object of the set X;
5. Output the results;

End

The algorithm this paper proposed was optimized based on the traditional K-means algorithm. It ensured that the initial clustering center could be generated in the high-density trajectory data points, reduced the interference of outliers in the trajectory dataset on the clustering results, avoided the clustering results falling into the local optimal solution, and improved the accuracy of the clustering results. In addition, under the BWP index intervention, this algorithm could automatically obtain the best clustering number and highly effectively select the high-quality clustering results. It solved the big problem that when the structure of the trajectory dataset unknown, the best clustering number was difficult to be determined.

4 Experimental Results and Analysis

4.1 Experimental Platform and Data

The paper was selected from 8:00 to 10:00, 12:00 to 14:00, 16:00 to 18:00 and 20:00 to 22:00 from 4th to 14th August 2015, in Beijing Dongcheng District taxi GPS data provide by DataTang as four sets of experimental data. Each sample experiment data points included vehicle ID number, latitude and longitude coordinate point and time information. A detailed description of the data set is shown in Table 1.

Table 1. Trajectory dataset used in the experiment

Trajectory dataset	Period	Dimension	Total number of sample data	Number of trajectory path
Data1	8:00–10:00	3	21078	619
Data2	12:00–14:00	3	25162	675
Data3	16:00–18:00	3	23179	634
Data4	20:00–22:00	3	19764	596

4.2 Comparison and Analysis of Three Kinds of Clustering Algorithms

To verify the feasibility of the algorithm this paper proposed, it was compared with the traditional K-means algorithm and DBSACN algorithm, including comparison of clustering result and algorithm efficiency.

The analysis of clustering results included five aspects: the accuracy of clustering results, the distance within the cluster, the distance between clusters, the best number of clusters and the number of clustering iterations. To guarantee the reasonable validity of the traditional K-means algorithm, each dataset would repeat the experiment 50 times and take the average value as the final clustering result, when clustering experiments using the traditional K-means algorithm. The results of the three algorithms are shown in Table 2.

Table 2. Comparison of clustering results of three algorithms

Algorithm	Trajectory dataset	Accuracy	Distance within the cluster	Distance between clusters	Best number of clusters	Number of clustering iterations
Traditional K-means algorithm	Data1	58.17%	0.46	1.23	81	13
	Data2	64.83%	0.35	1.10	80	10
	Data3	63.02%	0.33	1.20	85	11
	Data4	62.17%	0.42	1.25	83	11
DBSCAN algorithm	Data1	62.08%	0.43	1.25	73	12
	Data2	74.50%	0.30	1.13	85	8
	Data3	73.11%	0.37	1.30	63	9
	Data4	70.02%	0.39	1.29	65	9
The algorithm this paper proposed	Data1	83.07%	0.41	1.32	75	10
	Data2	90.52%	0.30	1.15	85	8
	Data3	86.03%	0.36	1.30	61	9
	Data4	85.53%	0.37	1.33	63	9

It could be seen from Table 2 that the accuracy of clustering results of the proposed algorithm was 24.45% points higher than that of the traditional K-means algorithm and 16.36% points higher than that of the Density-Based Spatial Clustering of Applications with Noise algorithm. To further verify the clustering results of the three algorithms in the trajectory dataset, choose the clustering results of the trajectory dataset Data3 as shown in Figs. 1, 2 and 3.

Fig. 1. Heat map of clustering results based on the algorithm this paper proposed

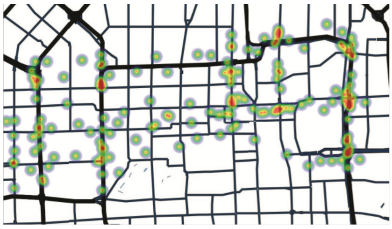

Fig. 2. Heat map of clustering results based on the DBSCAN algorithm

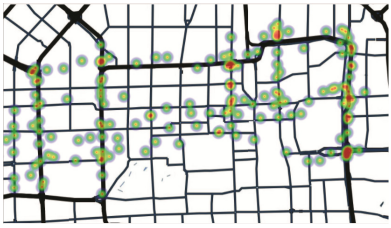

Fig. 3. Heat map of clustering results based on the traditional K-means algorithm

Compared with Figs. 1, 2 and 3, the proposed algorithm and DBSCAN algorithm were more reasonable in clustering than traditional K-means algorithm, and could better highlight the path of traffic in the trajectory distribution. In addition, the proposed algorithm had better anti-noise interference ability than the other two algorithms, and the clustering results of the trajectory data were more suitable for the actual road.

The analysis of three algorithm efficiency was analyzed by comparing the time (including the shortest, longest and average time) spent in a single cluster iteration CPU in different trajectory data sets. The results of the three algorithms were shown in Table 3.

Table 3. Comparison of operational efficiency of three algorithms

Algorithm		Traditional K-means algorithm	DBSCAN algorithm	The algorithm this paper proposed
Data1	Best/s	1.032	0.650	0.545
	Worst/s	2.998	1.695	1.572
	Average/s	2.579	1.312	1.246
Data2	Best/s	1.352	0.853	0.818
	Worst/s	4.013	3.720	3.712
	Average/s	3.188	2.641	2.575
Data3	Best/s	1.201	0.809	0.722
	Worst/s	3.913	3.263	3.38
	Average/s	3.062	2.372	2.346
Data4	Best/s	1.175	0.672	0.678
	Worst/s	3.113	2.086	2.033
	Average/s	2.473	1.492	1.498

The experimental results showed that at each iteration, the average consumption time of the proposed algorithm was similar to that of the DBSCAN algorithm. Although the time complexity was higher when searching for the initial clustering centers, in the complex trajectory datasets, the proposed algorithm could effectively reduce the number of iterations and improve the operating efficiency to a certain extent.

5 Conclusion

The improved density-based K-means algorithm was proposed in this paper. Combining the BWP index, the proposed algorithm selected the key trajectory data points in high-density area as the initial clustering centers, to ensure obtaining the high-quality clustering results which were tightness within the cluster and separation between clusters. The experimental results showed that the proposed algorithm was more accurate and highly efficient, and the simulation results showed that the proposed algorithm could extract the key points of the trajectory data well. In this paper, the time complexity is increased when calculating the sample spacing and selecting the initial clustering center. In the future research, the computational efficiency of this part will be improved.

Acknowledgments. This research was supported by the Fundamental Research Funds for the Universities in Tianjin, Tianjin Chengjian Universities (2016CJ11)

References

1. Wang, Z.C., Yuan, X.R.: Visual analysis of trajectory data. J. Comput.-Aided Des. Comput. Graph. (1), 9–25 (2015)
2. Khan, S.S., Ahmad, A.: Cluster center initialization algorithm for K-means clustering. Expert Syst. Appl. **25**(11), 1293–1302 (2004)

3. He, Y.B., Liu, X.J., Wang, Z.Q., et al.: Improved K-means algorithm based on global center and nonuniqueness high-density points. J. Comput. Eng. Appl. **52**(1), 48–54 (2016)
4. Zhu, M., Wang, W., Huang, J.: Improved initial cluster center selection in K-means clustering. Eng. Comput. **31**(8), 1661–1667 (2014)
5. Zhang, T., Ma, F.: Improved rough K-means clustering algorithm based on weighted distance measure with Gaussian function. Int. J. Comput. Math. 1–17 (2015)
6. Zhang, S.Q., Huang, Z.K., Feng, M.: An optimized K-means algorithm. Microelectron. Comput. **32**(12), 36–39 (2015)
7. Capó, M., Pérez, A., Lozano, J.A.: An efficient approximation to the K-means clustering for massive data. Knowl.-Based Syst. **117**, 56–69 (2017)
8. Zhang, S.J., Zhao, H.C.: Algorithm research of optimal cluster number and initial cluster center. J. Appl. Res. Comput. **34**(6), 1–5 (2017)
9. Rodriguez, A., Laio, A.: Machine learning. Clustering by fast search and find of density peaks. Science **344**(6191), 1492 (2014)
10. Rezaee, M.R., Lelieveldt, B.P.F., Reiber, J.H.C.: A new cluster validity index for the fuzzy c-mean. Pattern Recogn. Lett. **19**(3–4), 237–246 (1998)
11. Zhou, S.B., Xu, Z.Y., Tang, X.Q.: Method for determining optimal number of clusters in K-means clustering algorithm. J. Comput. Appl. **46**(16), 27–31 (2010)