Movement Pattern Extraction Based on a Non-parameter Sub-trajectory Clustering Algorithm

He Ailin, Liu Zhong, Zhou Dechao
College of Electronic Engineering,
Naval University of Engineering,
Wuhan, China
e-mail: heailin@yeah.net

***Abstract*—This paper proposed a movement pattern extraction method based on trajectory clustering, including three phases: trajectory partitioning, clustering and movement extracting. Firstly, trajectories are partitioned at characteristic points whose accumulated turn angle exceeds the threshold value which is determined by Minimum Description Length (MDL) principle. Secondly, a non-parameter sub-trajectory clustering based on Density Peak Clustering (DPC) is proposed to acquire trajectory line segment clusters. Lastly, trajectory segments are used to extract the overall trajectory motion trend，that is, trajectory movement pattern. Experiment results demonstrate that our algorithm can correctly extract the movement pattern from real trajectories data.**

Keywords—movement pattern; trajectory clustering; decision graph; automatic determination; clustering center; non-parameter

I. INTRODUCTION

In recent years, with the development of wireless positioning and the wide application of GPS technology, real-time tracking of moving objects has become a reality, which produces a large number of moving object trajectory data, analyzing and mining the potential value of these data has become a quite important work. Trajectory clustering, as one of the trajectory data mining techniques, can be used to discover the similar movement patterns in trajectory groups [1, 2]. The clustering results also can be used as trajectory anomaly detection [3], trajectory prediction [4], interesting place discovering [5] and other data mining task.

The object of a trajectory clustering algorithm can be an entire trajectory or a part of trajectory. When clustering the whole trajectory, the overall similarity can be discovered. However, in most applications, the overall similarity of trajectories is not obvious, but showing more local similarity. In order to find out the local similarity of trajectories, LEE et al. [1] proposed a trajectory clustering framework TRACLUS, which consists of two stages: trajectory partitioning and grouping. Because of the noise-insensitive DBSCAN[6] (Density-Based Spatial Clustering of Applications and Noise) used in the trajectory grouping stage, it is necessary to set the values of two clustering parameters in advance, and require a repeated adjustment of the predefined parameters with expert prior knowledge to get the final two parameters. At the same time, the clustering result is sensitive to the parameters. Yuan Guan et al.[2] proposed a similar trajectory clustering method, using DBSCAN in grouping stage, but in partitioning stage, this method adopts a partitioning strategy based on turn angle, which result in one more parameter than TRACLUS.

When partitioning according to turn angle, trajectory partition method is the one based on characteristic points where the behavior of a trajectory changes rapidly [2, 7-9], which is more intuitive and simple when compared with the MDL principle. These trajectory points whose the turn angle (the change of the direction) exceeds the predefined angle threshold are chose as characteristic points. However, there are two problems: (1) these papers have not shown how to determine this predefined angle threshold, when different thresholds result in different line segment sets; (2) the partition algorithm may also lose some valuable information of direction change if the accumulated change of the direction is ignored especially when the threshold is a large one, as the situation shown in Fig. 1.

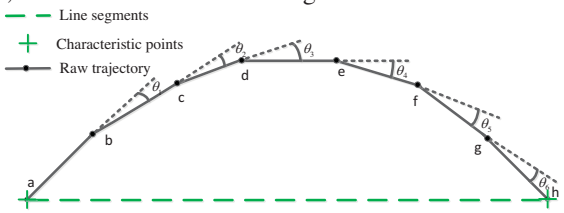

Figure 1. An example of trajectory partition based on turn angle

Considering the situation in Fig. 1, both point *a* and point *h* are the only characteristic points when all the turn angles $\{\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6\}$are less than the angle threshold. Then, the trajectory partition (*abcdefgh*) will be represented by the direct line segment between point *a* and *h*. Obviously, the segment *ah* failed to express the total details of the change of trajectory direction.

The clustering method proposed in [10] is a density-based clustering method, which requires only one trajectory clustering parameter, and it is a non-iterative process when building clusters. However, there are still some limitations: firstly, the only clustering parameter is determined by the empirical value, which will induce inaccuracy; The second one occurs when verifying the clustering noise, the determination of noise is under the condition that there

978-1-7281-1282-4/19/$31.00 ©2019 IEEE

exists neighborhoods belonged to another cluster within a given distance for an element in a cluster, which is impossible for a single cluster far away from other clusters, resulting in the failure to detect the noise in the relatively isolated clusters. Lastly, when determining the clustering center, it is necessary to judge from the decision graph manually. But for some data sets whose centers are not obvious on decision graph, this manual selection will result in non-objective results.

In this paper, we try to get the trajectory movement pattern automatically. Trajectories are partitioned at characteristic points whose accumulated turn angle exceeds the threshold value, which is determined by Minimum Description Length (MDL) principle. In order to make the trajectory clustering method get rid of the restriction of prior knowledge in the relevant domain, this paper proposes a trajectory clustering method which can automatically determine clusters based on literature [10]. Aiming to the limitations mentioned above, we put forward a method to automatically determine the only parameter with the entropy minimum principle [11], and we also provide a numerical detection strategy to automatically determine the number of clustering centers and a new method to determine trajectory outlier segments.

II. Trajectory Partition

In this section, a trajectory partitioning algorithm is proposed. Before partitioning, we need to define the distance measure between segments:

$$dist(L_i, L_j) = d_{start} + d_{center} + d_{end} \quad (1)$$

Where d_{start}, d_{center}, d_{end} are shown in Fig.2.

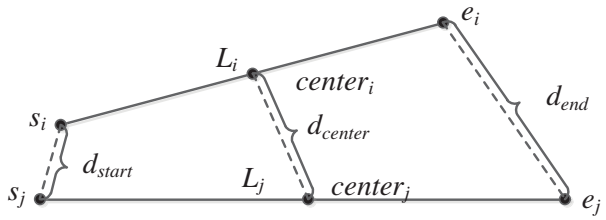

Figure 2. Illustration of distance function

After that, we determine the optimal accumulated turn angle threshold with MDL principle [1], which can minimize $L(H)+L(D|H)$ when we partition the trajectory at characteristic points whose accumulated angle is not smaller than that optimal one, and we show the calculation of $L(H)+L(D|H)$ in Fig.3.

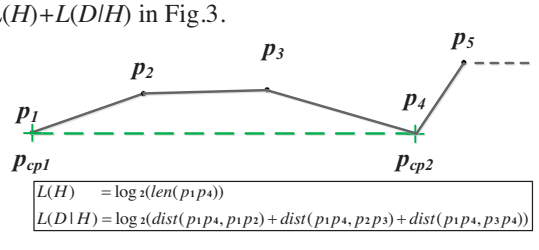

Figure 3. An example of the MDL cost

When the characteristic points are determined, the trajectory can be partitioned into trajectory segments at every characteristic point, as shown in Fig.4, trajectory segment *abc* is replaced by line segment *ac*, *ced by ce*, *efg* by *eg* and *gh* by itself.

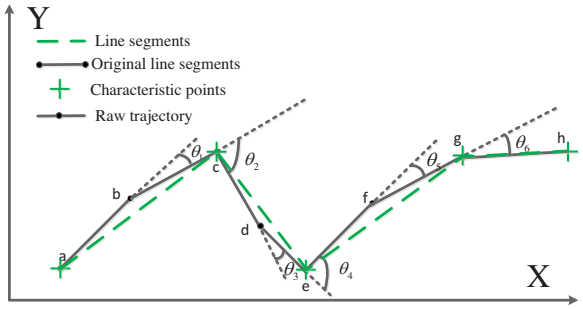

Figure 4. Trajectory partition based on accumulated turn angle

III. Trajectory Clustering

In this section, we show the details about the generation of trajectory clustering, which can be divided into three steps: firstly, we use the Entropy Minimum Principle to get the optimal d_c, which is essential when calculating the line segment potential described below; then, the process of the generation of trajectory clustering is given; the last one is the determination of outlier.

Before the generation of trajectory clusters, some necessary definitions are given as follows.

A. Relative Definitions

Definition 1. The potential of trajectory segments: The data field of the point data in the literature [12] is introduced into the calculation of the trajectory segment potential. For a given set of trajectory segments, each trajectory segment is in the data field of the other trajectory segments. When giving the set of trajectory segments $\{L_1, L_2, ..., L_n\}$, the potential for the trajectory segment L_i can be calculated as follow:

$$\varphi_i = \varphi(L_i) = \sum_{j=1}^{n} \left(m_j \times e^{-\left(\frac{d_{ij}}{d_c}\right)^k} \right) \quad (2)$$

Where m_j is the mass of the line segment L_j, and when no special semantic background, each trajectory segment is treated as equal importance, that is, $m_j = 1$. d_{ij} is the distance between the trajectory segments L_i and L_j, and d_c is the interaction range of trajectory segments, $k \in N$. As pointed in [12], the distribution of the data field depends mainly on d_c, and is independent of the specific potential function. Therefore, in this paper, we set $k = 2$, and then the potential of the trajectory segment can be expressed by the Gaussian function:

$$\varphi_i = \varphi(L_i) = \sum_{j=1}^{n} \left(e^{-\left(\frac{d_{ij}}{d_c}\right)^2} \right) \quad (3)$$

Definition 2 Delt-distance: The minimum distance between the trajectory segment L_i and the trajectory segment with a higher potential than L_i , which can be calculated as:

$$\delta_i = \min_{j:\varphi_j > \varphi_i} \left(d_{ij} \right) \tag{4}$$

For the trajectory segment L_i with the highest potential, we set δ_i = max (d_{ij}), and this can ensure the trajectory segment with the largest potential must be the clustering center.

Definition 3 Decision graph: The concept of decision graph was first proposed by Rodriguez in 2014[10], the cluster center are those points with high potential and high delt-distance, that is, those points in the upper right corner of the decision graph. According to [13], the position and number of the cluster centers on the decision graph are inevitable affected by the distribution of the data set and the subjective judgment. To this limitation, we redefine the process of generating the decision graph: firstly, normalizing the trajectory segment set $\{ \varphi_1, \varphi_2, \varphi_3, ... \varphi_n \}$ and the delt-distance set $\{ \delta_1, \delta_2, \delta_3, ... \delta_n \}$, and multiplying them respectively, $\gamma_i = \varphi_i \bullet \delta_i$. Secondly, sorting γ_i in a descending order, and plotting them as the one in Fig.5 below.

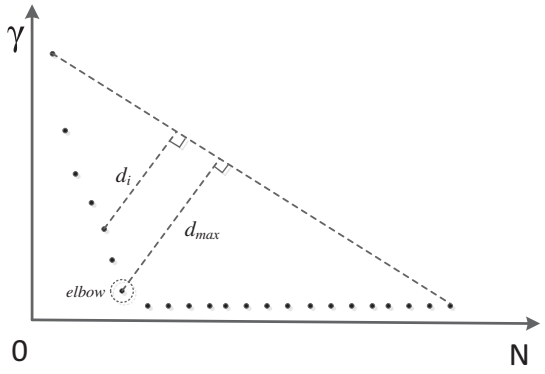

Figure 5. Illustration of decision graph

During our experiment, we find that there exist an 'elbow' between clustering center and the successive non-clustering center, and we can find that 'elbow' where d_i achieves the maximum, as shown in Fig.5.

B. The Determination of Optimal d_c

Given a set of trajectories, for each of d_c, the potential of each trajectory are defined, and the literature [10] only gives the empirical formula for d_c, which is not applicable to all data sets. In this section, a method of determining optimal d_c by using the minimum entropy principle [11] is proposed. From the perspective of information theory, the distribution of the potential of trajectory segments can be measured from the point of view of entropy, and the d_c resulting in the minimum entropy is the optimal. The entropy

$$H = -\sum_{i=1}^{n} \frac{\varphi_i}{Z} \log\left(\frac{\varphi_i}{Z} \right) \tag{5}$$

Where $Z = \sum_{i=1}^{n} \varphi_i$, is the normalization factor.

C. The Generation of Clusters

Firstly, the potential and delt-distance of a trajectory segment are calculated according to the definition 1 and 2 respectively, and then the trajectory clustering center is determined according to the method in definition 3. After that, the cluster flag set{0,1 ..., $N_{cluster}$} is assigned to the trajectory clustering center in turn, and the determination of the cluster flag for the rest of line segments is based on the following principle: the cluster number of a given line segment is the same as the one having a higher potential and the shortest distance from that given line segment. Therefore, after the determination of these clustering centers, sorting all the line segments by potential in descending order, and only scanning the sorted line segment set once, the cluster flag for all line segments can be determined.

D. The Determination of Outlier Segments

After the generation of clusters, we can see that the trajectory noise segments are mostly in the sparsely regions of each cluster where trajectory potential is relatively small. Sorting these line segments by potential in ascending order, we can find that there is an obvious jump between the noise and the successive normal one and no obvious jump between outlier or two normal ones. For this obvious jump, the specific detection strategy is given as follows:

A new potential set $\{ \varphi_1', \varphi_2', \varphi_3', ... \varphi_n' \}$ is got after sorting the original set $\{ \varphi_1, \varphi_2, \varphi_3, ... \varphi_n \}$ in ascending order, and then the slope between two successive potential can be calculated as follow:

$$slope_i = \varphi_{i+1}' - \varphi_i' \tag{6}$$

The position of the jump point in the set $\{ \varphi_1', \varphi_2', \varphi_3', ... \varphi_n' \}$ is the i where the slope is the local maximum, that is, in the set $\{ \varphi_1', \varphi_2', \varphi_3', ... \varphi_n' \}$, trajectory segments set relative to the set$\{ \varphi_1', \varphi_2', \varphi_3', ... \varphi_i' \}$ are trajectory outlier segments.

IV. MOVEMENT PATTERN EXTRACTON

In this section, we take the representative line segments calculated with sweep line approach in TRACLUS as trajectory movement pattern. Due to the limitation of paragraph, the details of sweep line approach can be seen in TRACLUS[1].

V. EXPERIMENT RESULTS AND DISCUSSION

In this section, a visual simulation system is developed to verify our method with MFC in Microsoft Visual Studio 2015.

A. Trajectory Clustering with Synthetic Data

A synthetic data set with 5 clusters in Fig.6 is used as ground truth to verify our algorithm.

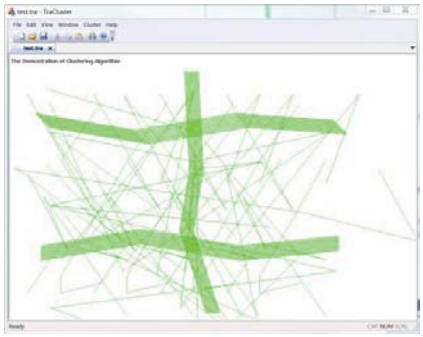

Figure 6. Illustration of synthetic data with 20% noise

Using the entropy minimum principle mentioned above, and plotting with MATLAB, we can get the variation of entropy as d_c in Fig.7：

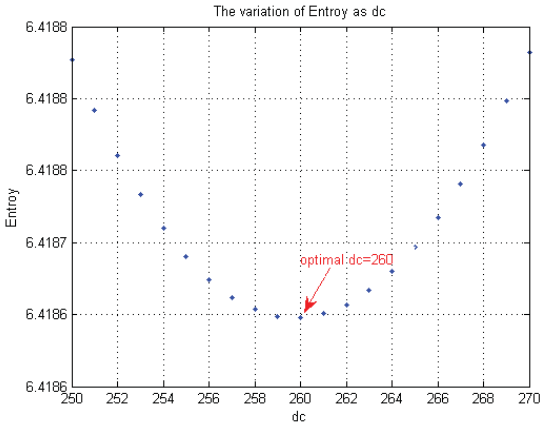

Figure 7. Variation of entropy as dc

After the determination of the optimal d_c and the calculation of trajectory potential and delt-distance, we can get the decision graph, which can be used to get the 'elbow' when d_i in Fig.5 reach to the maximum, as shown in Fig.8.

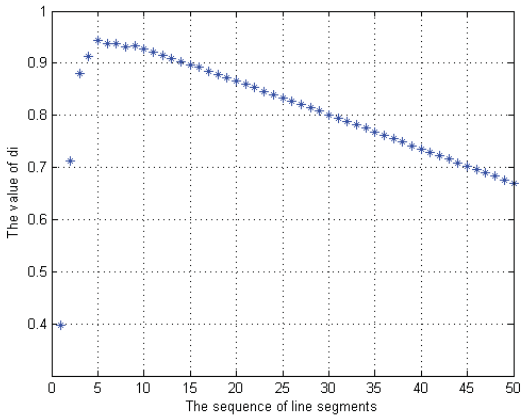

Figure 8. Variation of d_i

From Fig.8, we can see that the d_{max} is achieved at the fifth point in the decision graph, that is, there are 5 clusters in the synthetic data, which is consistent with the ground truth in Fig.6.

In order to verify the validity of the proposed detection strategy when determining these trajectory outlier segments, after the generation of each cluster, we have sorted all the line segments in one cluster by potential in ascending order as shown in Fig.9.

From the result in Fig.9, there indeed exists an obvious jump between the noise and its successive normal one which can be also proved by the result of the detection strategy as shown in Fig.10

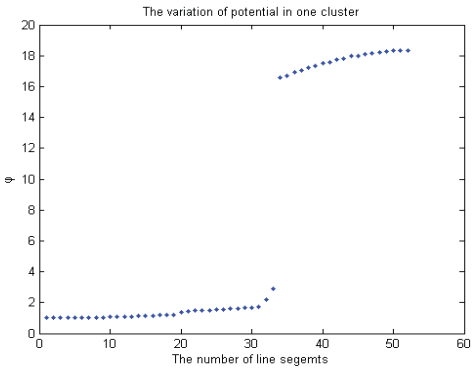

Figure 9. Variation of potential in one cluster

In Fig.9, the obvious jump proves the validity of the proposed numerical detection strategy.

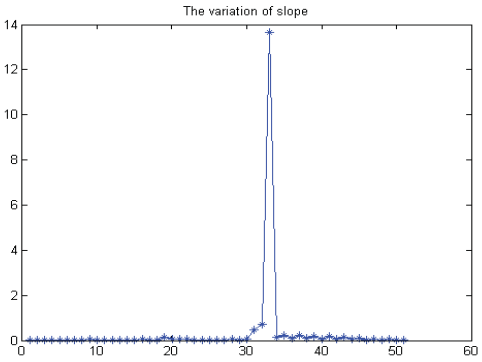

Figure 10. Variation of slope

Finally, the clustering result is shown in Fig.11, and these red line segments are the movement pattern.

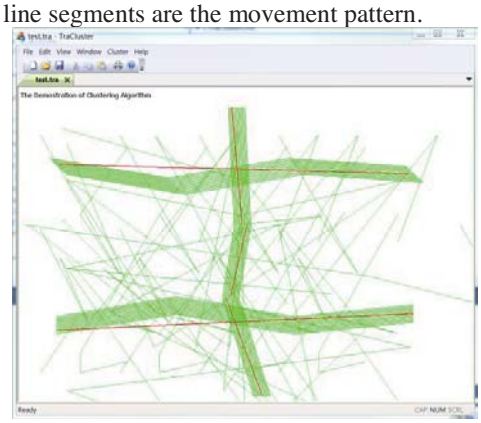

Figure 11. Clustering result of synthetic data with noise

B. *Trajectory Clustering with Deer1995 Data*

In this section, we compare our clustering result with the optimal one in TRACLUS with Deer1995 data generated by the Starkey project which can be accessed in http://www.fs.fed.us/pnw/starkey/data/tables/, and has 32 trajectories and 20065 points. We present the clustering results in Fig.12 and 13 respectively.

From the comparison between Fig. 12 and 13, we can see that our method has found not only two similar movement patterns in TRACLUS, but also two another shorter movement patterns. In TRACLUS, using two global clustering parameters results in the missing of these smaller clusters, which can be avoid in our algorithm.

Figure 12. Clustering result for Deer1995 with TRACLUS

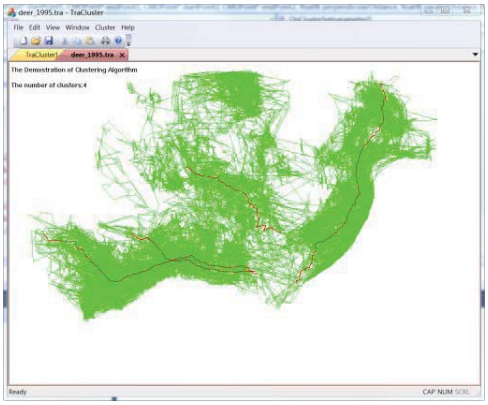

Figure 13. Clustering result for Deer1995 with our algorithm

VI. CONCLUSION

This paper has developed a method to extract trajectory movement pattern based on trajectory clustering, including trajectory partitioning, trajectory clustering and trajectory movement pattern extracting. When partitioning, we have determined the optimal accumulated turn angle threshold by using MDL principle, and then all the trajectories are partitioned into line segments with that optimal threshold. When clustering, we have proposed a non-parameter clustering algorithm based on DPC algorithm[10], and clusters are determined successfully. Finally, trajectories pattern are extracted by using sweep line approach on original line segments. Experiment results on synthetic data and real data have verified the reasonability and validity of our algorithm, and our method is insensitive to noise to some degree.

Our work above is just the first step and our next step will attempt to check the effect of different segmentation strategies and different trajectory segment distance measurement methods on clustering results.

REFERENCES

[1] J.-G. Lee, J. Han, and K.-Y. Whang, “Trajectory clustering: A partition-and-group framework, ”Proceedings of the ACM SIGMOD International Conference on Management of Data. pp. 593-604.

[2] G. Yuan et al., “An efficient trajectory-clustering algorithm based on an index tree,” Transactions of the Institute of Measurement and Control, vol. 34, no. 7, pp. 850-861, 2012.

[3] X. Ying, Z. Xu, and W. G. Yin, "Cluster-based congestion outlier detection method on trajectory data." pp. 243-247.

[4] C. Sung, D. Feldman, and D. Rus, "Trajectory clustering for motion prediction." pp. 1547-1552.

[5] A. T. Palma et al., "A clustering-based approach for discovering interesting places in trajectories." pp. 863-868.

[6] M. Ester et al., "A density-based algorithm for discovering clusters in large spatial databases with noise," In Proc. 2nd Int'l Conference on Knowledge Discovering and Data Mining. pp. 226-231, 1996.

[7] J. Gonzalez et al., "Semi-automatic extraction of ship lanes and movement corridors from AIS data," IEEE International Geoscience and Remote Sensing Symposium (IGARSS). pp. 1847-1850, 2014.

[8] J. Chen et al., "Trajectory clustering for people's movement pattern based on Crowd souring data," International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives. pp. 55-62, 2014.

[9] C. Chang, and B. Zhou, "Multi-granularity visualization of trajectory clusters using sub-trajectory clustering," ICDM Workshops 2009 - IEEE International Conference on Data Mining. pp. 577-582, Year. http://dx.doi.org/10.1109/ICDMW.2009.24

[10] A. Rodriguez, and A. Laio, “Clustering by fast search and find of density peaks,” Science, vol. 344, no. 6191, pp. 1492-1496, 2014.

[11] D. Li, and Y. Du, Artificial intelligence with uncertainty: CRC press, 2007.

[12] D. Li et al., “Data field for hierarchical clustering,” Developments in Data Extraction, Management, and Analysis, pp. 303, 2012.

[13] Z. Liang, and P. Chen, “Delta-density based clustering with a divide-and-conquer strategy: 3DC clustering,” Pattern Recognition Letters, vol. 73, pp. 52-59, 2016