Comparison study of sub-trajectory clustering in data mining

Guodong Yang, Zhitao Huang, Xiang Wang

School of Electronic Science and Engineering , National University of Defense Technology, Changsha 410073,China

Abstract. Trajectory clustering is an important method to achieve moving object data mining, multi-sensor information fusion and trajectory knowledge discovery. Sub-trajectory clustering is an important method to extract useful information from a large number of trajectory data in trajectory analysis. In this paper, comparative experiments are made on the time consumption, similarity measure and clustering performance based on the existing sub-trajectory clustering methods. Based on the comparisons, the advantages and disadvantages of different methods are presented and an improved method is proposed for dealing with trajectories with low positioning accuracy and correlating tracklets from asynchronous sensors. Besides, a general framework of trajectory data mining is discussed.

1 Introduction

Nowadays, with the popularity of mobile terminal equipment, a large number of systems based on geographic location information service, such as GPS service[15], RFID service and AIS(Automatic Identification System)service[4][8], have attracted great attention. How to make use of the geographical location information efficiently has become increasingly urgent.

Many of the known studies are considered the research on trajectory analysis as a procedure of data mining. With the analysis from trajectory data, patterns of moving object can be found. At present, the studies based on trajectory data mining are concentrated on trajectory clustering, location prediction, behavior analysis and abnormal trajectory detection.

As the basis of trajectory analysis, trajectory clustering acts as a link to high-level analysis. It is important to achieve trajectory clustering efficiently. Existing work on trajectory clustering methods can be divided into global trajectory clustering and sub-trajectory clustering. The former is regarding the sampling points of a moving object from a movement as a whole, called global trajectory. Under the circumstance, the similarity of trajectories should be measured pairwise trajectories and stored in distance similarity matrix. To measure the similarity, some distance functions such as Euclidean distance[11],Longest Common Sub-sequence(LCSS)[2], Hausdorff distance[10], Dynamic Time Warping (DTW)[3], Fréchet distance[13] can be chosen. The usage of global similarity distance measure of moving objects can grasp the moving objects with the same motion by clustering methods, but it is difficult to find the local similarity of moving objects in some important regions such as narrow waters and convergence regions.

Considering the limitation of global trajectory clustering, the clustering of sub-trajectory is proposed, which can discover the hidden behaviors. The research on sub-trajectory clustering mainly focuses on extracting feature points, similarity measure method and the clustering method. Lee, Han, and Wang[7] proposed a partition-and-group framework which defined a geometric distance function to measure similarity and clusters the sub-trajectory by DBSCAN. Shrikant, Sujor, and Lee[12] proposed a method by mapping sub-trajectory to a feature space and clustering in the feature space.

This paper focuses on the comparative analyses of the performance of these two mentioned methods and a combination method using the similarity distance measure in method 2 with DBSCAN. With the comparisons and considering the practical problems, a general architecture of trajectory data mining is discussed. The structure of =this paper is as follows. Section 2 introduced the basic concepts of trajectory definitions and proposed a detailed description of trajectory data mining framework. Section 3 focuses on the procedure of data mining and comparisons of these mentioned methods above. Section 4 discussed the results and performance of different datasets and methods.

2 Preliminaries

2.1 Basic Concepts

The trajectory of a moving object is a series of sampling position points generated by the sensors. The time interval of the sampling point sets is determined by the sampling device.

Trajectory could be described as follows:

$$TR = \{P_1, P_2, \cdots, P_i, \cdots, P_n\}$$

TR represents a spatio-temporal track of an object, $i \in [1, n]$ denotes the sampling position points which observed by sensors. For each point $P_i(lat_i, lon_i, sat_i, t_i)$, it contains the location lat_i , lon_i (indicate the latitude and longitude of an object), sat_i (indicates the status such as velocity, acceleration of an object) and t_i which denotes the timestamp. Limited by the type of sensors, the sat_i of object can be missing.

Figure 1 shows a trajectory of a moving object. $P_1, \cdots, P_7$ is the sampling points of a track.

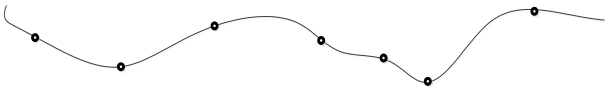


Figure 1: Sampling trajectory points diagram

Definition 1. Feature point is a set of characteristic points extracting from pre-processing trajectory which can respect the trajectory concisely.

Definition 2. A Sub-trajectory is a line segment whose start and end point are the feature points ordered by temporal index.

Figure 2 shows the feature points and sub-trajectories. The $P_1, \cdots, P_7$ are the sampling points and $P_{c1}, \cdots, P_{c4}$ are the feature points. $STR = \{P_{c1}P_{c2}, P_{c2}P_{c3}, P_{c3}P_{c4}\}$ represents the Sub-trajectory set.

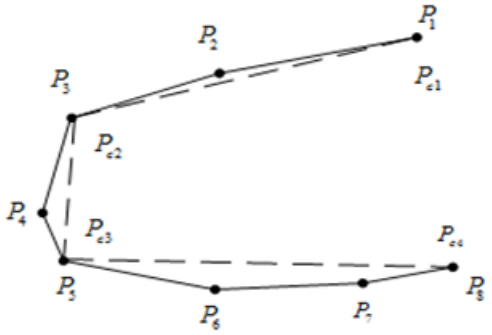


Figure 2: An Example of a trajectory sampling points and its sub-trajectories

2.2 Trajectory Mining Framework

As for the passive location system, the observations of moving objects are gathered and stored in the central database with time and object observation batch number index for each location. Considering the characteristic of passive location, a general trajectory mining framework is designed as figure 2.

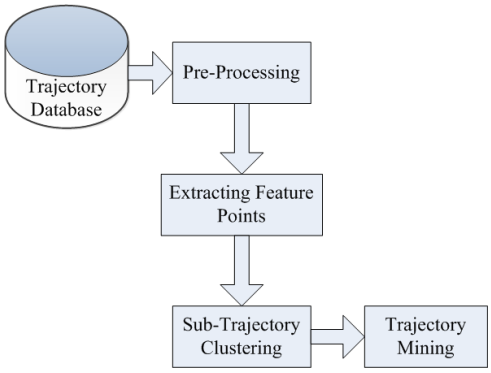


Figure 3: Trajectory Data Mining Framework

There are amount of records from different objects and distributed sensors in the database. Data quality analysis (e.g, Redundancy ration, Abnormal records and inconsistent values) should take into account firstly.With the assist of quality analysis, the data Pre-processing can be executed. By using the SQL statement, the basic statistics characteristics can be obtained and redundant records can be eliminated. Due to the deviation of passive location system, the trajectory smoothing filtering is also needed to consider. Secondly, the number of sampling point records is enormous even though the data pre-processing. The reduction of sampling point records is urgent. By using some criteria, the feature points can be extracted to represent the whole. After that, With the help of sub-trajectory clustering, the behaviors hidden from the trajectories can be discovered.

3 Mining Trajectories

3.1 Data Pre-processing and Feature Points Extracting

For most of trajectory mining methods, the step of pre-processing sampling point records is missing due to assume the high quality of sampling (both on position accuracy and low redundancy and error on the records) and the continuous observation of a moving object in a movement. Nevertheless, many practical scenarios do not satisfy this assumption (e.g, passive location in battlefields, occlusion of buildings leading to lack of sampling points and equipment shutdown due to low power). As mentioned above, the procedures of pre-processing include the correlation of tracklets in an movement, redundant records decreasing.

Facing the interruption of tracking, a trajectory of a moving object is composed of many tracklets [14]. By using the spatio-temporal constrains, tracklets can be associated. The spatio-temporal constrains include: 1) Exclusivity: The moving object can only appear at one position at the same time by single sensor; 2) Continuity: The moving object position can not be mutated out of location accuracy and its status constrain.

The exclusivity can be represented as follows:

$$P(T_i, T_j) = \begin{cases} 1, & if \quad t_{ei} < t_{sj} \\ 0, & otherwise \end{cases} \qquad (1)$$

Where $P(T_i, T_j)$ is the correlation of tracklets, t_{ei} represents the end time of tracklet T_i , t_{sj} represents the start time of tracklet T_j .

In this article, the continuity of a trajectory can be recursive by status like velocity and location using least squares estimation. The redundance can be reduced by using suitable design of SQL statement.

With the data pre-processing accomplished, Feature Points extracting should take into account to reduce the huge amount of sampling points with low information. In previous study, there are two strategies including the construction of kinematics filter and compressing the sampling points using minimum description(MDL) in information theory.

The construction of kinematics filter based on the principle that points with more information content are those points which abruptly change their status in a trajectory, including the direction and speed. Under the criterion, a heuristic method of re-sampling the points is proposed by leaving out the points whose direction difference is less than the threshold in FARM. In the other side, The MDL principle was adopted in TRACLUS method with formulas as follows:

$$L(H) = \sum_{j=1}^{par_i - 1} \log_2 \left(len\left(p_{c_j} p_{c_{j+1}} \right) \right) \tag{2}$$

$$L(D \mid H) = \sum_{j=1}^{par_{i-1}} \sum_{k=c_j}^{c_{j+1}-1} \{ \log_2 (d_\perp (p_{c_j} p_{c_{j+1}}, p_k p_{k+1})) + \log_2 (d_\theta (p_{c_j} p_{c_{j+1}}, p_k p_{k+1})) \} \tag{3}$$

The $L(H)$ is the length of a trajectory consisted by a set of feature points. The $L(D \mid H)$ represents the difference between the sampling points and feature points in a trajectory. In order to representing the trajectory more accuracy and concise, the sum of $L(H)$ and $L(D \mid H)$ needs to minimum.

Considering the passive location system, the extracting methods mentioned above are not appropriate due to position jumping between two points. A new combinatorial extracting feature method can be proposed.

The new method is contributed as follows: firstly, a grid divide method is used to divide the area contained trajectories with reference to geohash[1]. The width of grid lies on the position circle probability error. The mean value of points in the same grid can represent a normalized point in the divided area. It can reduce the impact of position jumping. Then the MDL principle can be used to extracting the feature points.

3.2 Similarity measure

Unlike the similarity measure of points, the distance function like Euclidean distance, Manhattan distance and chebyshev distance, can not be used directly. The improving distance function should be used to measure similarity between the pairwise line segments. There are two kinds of distance functions.

The first one[7] is composed of three components: the perpendicular distance($d_\perp$), the parallel distance($d_\parallel$) and the angle distance(d_θ). These components are shown in figure 5.

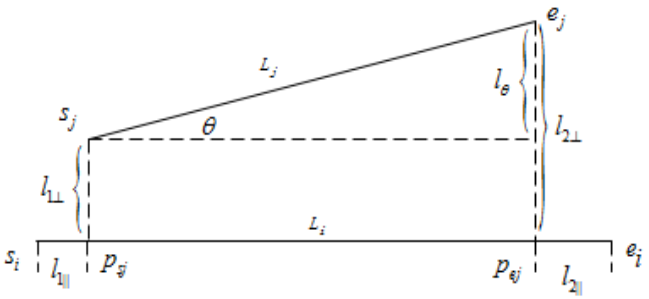


Figure 4: trajectory measure in method 1

We define two sub-trajectories $s_i e_i$, $s_j e_j$ in figure 4. $p_{sj} p_{ej}$ is the projection of $s_j e_j$ on line segment $s_i e_i$.

The perpendicular distance is defined as follows:

$$l_{\perp} = \frac{l_{\perp 1}^2 + l_{\perp 2}^2}{l_{\perp 1} + l_{\perp 2}} \quad (4)$$

$l_{\perp 1}$ is the Euclidean distance between s_j and its projection p_{sj}. $l_{\perp 2}$ is that between e_j and p_{ej}.

The parallel distance is defined as follow:

$$l_{\parallel} = \min\left(l_{\parallel 1}, l_{\parallel 2}\right) \quad (5)$$

$l_{\parallel 1}$ is the minimum Euclidean distance of p_{sj} to s_j, e_j. Likewise, $l_{\parallel 2}$ is the minimum Euclidean distance of p_{ej} to s_j, e_j.

The angle distance is defined as follow:

$$l_{\parallel} = \begin{cases} \left\| s_j e_j \right\| \sin\theta, & if \quad 0 \le \theta \le \dfrac{\pi}{2} \\ \left\| s_j e_j \right\|, & if \quad \dfrac{\pi}{2} \le \theta \le \pi \end{cases} \quad (6)$$

θ represents the angle between two sub-trajectories.

In this way, an overall evaluation of the similarity measure of the sub-trajectories can be generated, denoting the overall similarity between the two sub-trajectories as follow:

$$S_{ij} = \omega_{\perp} * l_{\perp} + \omega_{\parallel} * l_{\parallel} + \omega_{\theta} * l_{\theta} \quad (7)$$

In general, the weight coefficient of the there components $\omega_{\perp}$, $\omega_{\parallel}$ and ω_{θ}, depends on the practical problems.

The second distance function[12] maps a line segment $s_i e_i$ to a 2-tuple elements $f_i = \{P_i(lat, lon), (\theta_i)\}$, where the $P_i(lat, lon)$ is the start point of $s_i e_i$ and θ_i is the direction of the segment. Then, the set of sub-trajectories STR is represented by F, where $F = \{f_1, \cdots, f_i, \cdots, f_n\}$. The similarity of two segments is defined as follow:

$$S_{ij} = \exp - \frac{\left\| f_i - f_j \right\|^2}{2\sigma_{ij}^2} \quad (8)$$

We sum the two norm of the 2-tuple elements in each dimension, that is to say:

$$\left\| f_i - f_j \right\|^2 = (dist(P_i - P_j))^2 + (dist(\theta_i, \theta_j))^2 \quad (9)$$

The σ_{ij} is a scaling parameter which is defined as follow:

$$\sigma_{ij} = \frac{1}{c + \sigma_{\theta_{ij}}} \quad (10)$$

Where $\sigma_{\theta_{ij}}$ is the standard deviation of θ_i, θ_j.

REMARK: Notice that the order of magnitude of $dist(P_i, P_j)$ and $dist(\theta_i, \theta_j)$ is different. The matrix of $dist(P_i, P_j)$, $dist(\theta_i, \theta_j)$ should be normalized.

3.3 Sub-trajectory clustering

After the similarity measure of pairwise sub-trajectories, clustering method is needed to cluster the multiple sub-trajectories. Here, two clustering methods used in sub-trajectory clustering are discussed.

3.3.1 Clustering Based on Density. The difference between the sub-trajectories and points using the DBSCAN algorithm[9] is the definition of similarity matrix discussed in section 3.2.

Supposing D is a set of sub-trajectories, we can define as follows:

Definition 3. $\varepsilon - neighborhood$: $N_{\varepsilon}(L_i)$ is a set of sub-trajectories defined by $N_{\varepsilon}(L_i) = \{L_j \in D \mid dist(L_i, L_j) \le \varepsilon\}$, $L_i \in D$.

Definition 4. Core Sub-trajectory： A sub-trajectory $L_i \in D$ is called the core sub-trajectory if $|N_\varepsilon(L_i)| \geq MinLns$.

Definition 5. Directly density reachable: A sub-trajectory $L_i \in D$ is directly density reachable from a sub-trajectory $L_j \in D$ if $L_j \in N_\varepsilon(L_i)$ and L_i is a core sub-trajectory.

Definition 6. Density Reachable: A sub-trajectory $L_i \in D$ is density-reachable from sub-trajectory $L_j \in D$, if there is a sub-trajectory chain $L_i, L_{i+1}, \cdots, L_j$, and L_k is directly density reachable from L_{k+1}.

Definition 7. Density Connected: A sub-trajectory $L_i \in D$ is density connected to a sub-trajectory if there is a sub-trajectory $L_k \in D$ which is directly density reachable by L_i and L_j.

Based on the above definitions, sub-trajectories can be clustered by DBSCAN.

3.3.2 Spectral Clustering. The spectral clustering[6] regards clustering as a graph partitioning problem. In this practical situation, A hierarchical approach is used by dividing the data into two clusters at each steps until the standard deviation of slope θ in a cluster is less than a threshold τ.

Each sub-trajectory data is a vertex v of the graph $G = \langle V, E \rangle$ and the weight of edge e between two vertexes v_i and v_j is the similarity of L_i and L_j defined in section 3.2. The graph is divided to two sub-graph at each hierarchical tree node with the Minimum Cut Principle.

As the similarity matrix calculated, the degree matrix is deduced by summing each row in similarity matrix as a diagonal matrix D. The normalized Laplacian matrix is defined as follows:

$$L_{nom} = D^{-\frac{1}{2}}(D-S)D^{-\frac{1}{2}} = I - D^{-\frac{1}{2}}SD^{-\frac{1}{2}} \qquad (11)$$

In order to clustering the data into two clusters, two maximum eigenvalues of $D^{-\frac{1}{2}}SD^{-\frac{1}{2}}$ should be calculated. The corresponding eigenvectors can be constructed as a matrix $\mathrm{E} = [e_1, e_2]$. Using the k-means method (k=2), the data of sub-trajectories can be clustered.

3.4 Representative Trajectory Generate

Representative trajectory is mainly used to describe the hidden patterns from the cluster of sub-trajectory. It is the transformation from data to knowledge.

To transform the cluster of sub-trajectory to representative trajectory, the sweeping line algorithm is introduced. The slope of a sweeping line is the vertical average direction of a cluster, called major axis defined in (12)[7]. The sweeping line moves along the major axis and counts the number of intersections when the sweeping line meets a start point or an end point of a cluster. If the number is equal to or greater than *MinLns*, the average coordinates of the sub-trajectories intersected by the sweeping line need calculating to respect the points of a representative trajectory. In order to facilitate the computation, rotate the axes so that the X axis is made to be parallel to the major axis of the cluster. The rotation matrix is used in Formula (13).

Definition 8. Suppose a set of unit vectors $V = \{\vec{v}_1, \vec{v}_2, \cdots, \vec{v}_n\}$. The average direction vector $\vec{V}$ of V is defined as Formula (12).

$$\vec{V} = \frac{\vec{v}_1 + \vec{v}_2 + \cdots + \vec{v}_n}{|V|} \qquad (12)$$

The rotation matrix is as follows:

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \cos\phi & \sin\phi \\ -\sin\phi & \cos\phi \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \qquad (13)$$

The angle of ϕ represents the direction of major axis. The x, y is the coordinate before rotation.

4 Experiments and Comparison

4.1 Datasets and Environment

In this paper, the synthetic data and the Atlantic hurricane data sets from1950s to 2010s were used. The test contents included the comparison of clustering results and time consumption. By the combinations of similarity measures and clustering methods, we chose three combinations to compare which use the second similarity measure in both of the clustering methods and the first similarity measure with DBSCAN, i.e. TRACLUS. The algorithms were executed in Matlab 2011a. Experiments ran on an Inter core i3 3.60GHZ machine with 4.00GB in windows 7.

4.2 Results

Figure 5 shows the time consumption of the three combinations. The second similarity measure with spectral clustering is called FARM and with DBSCAN is called DB-FARM. The dataset used in the experiment is the Atlantic hurricane data. The X axis is the number of hurricane trajectories and the Y axis is the time consumption.

We ran each combination algorithm for 5 times with reducing 70 trajectories for each time. The results show that the TRACLUS has a better performance on time consumption. And the main time-consuming spends on measuring the similarity of sub-trajectories.

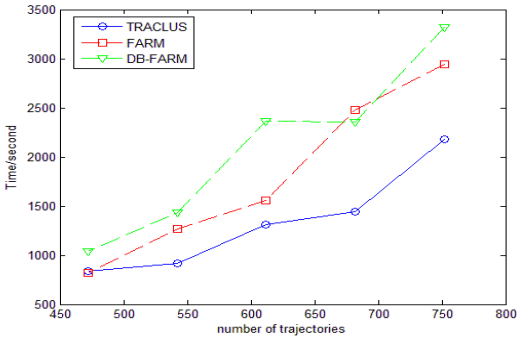
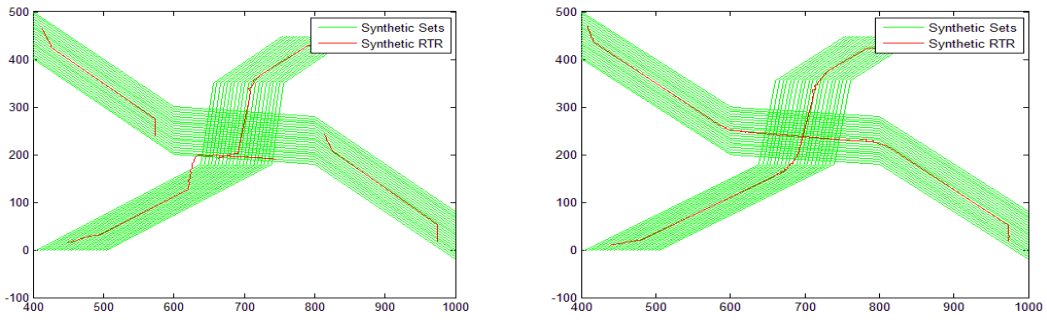
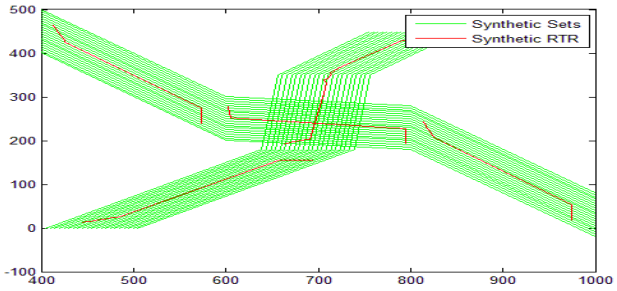


Figure 5: comparison on time consumption of sub-trajectory clustering

Figure 6 and figure 7 are the clustering performance of the three combinations. Figure 6 used the synthetic data to compare the performance of crossing trajectories on clustering. The parameters in TRACLUS are $\varepsilon = 1.6$ and $MinLns = 25$. The parameter in FARM is $\tau = 0.4$. The parameters are $\varepsilon = 1.6$ and $MinLns = 25$ in DB-FARM.



(a) Synthetic data clustering using TRACLUS (b) Synthetic data clustering using FARM

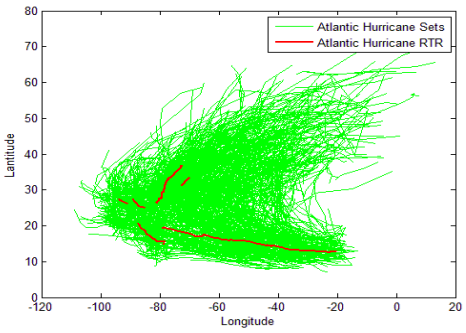


(c) Synthetic data clustering using DB-FARM

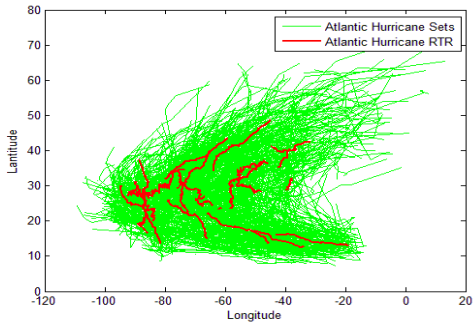
Figure 6: comparison on clustering performance with synthetic data

The Results of synthetic data with crossing trajectories of these methods show that the representative trajectories obtained by FARM are most intuitive in figure 6. Comparing with TRACLUS and DB-FARM methods, we can come to the conclusion that the second distance function is more suitable for overlapping trajectories. Since the second distance function considers the direction component whose weight is the same as location information.

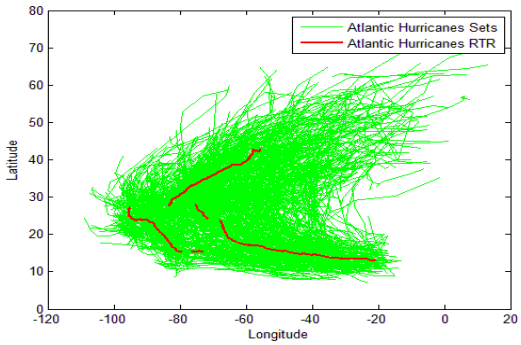
By using the Atlantic hurricane dataset, the DB-FARM and TRACLUS methods show a representative and concise result than FARM in figure 7. It is also easy to conclusion that the DBSCAN method is more appropriate for sub-trajectory clustering than spectral clustering. Since the DBSCAN method can adapt the noise in a dataset and do not need to know the number of clusters in a dataset.



(a) Hurricane data clustering using TRACLUS



(b) Hurricane data clustering using FARM



(c) Hurricane data clustering using DB-FARM

Figure 7: comparison on clustering performance with Hurricane data

Comparing the three combination methods, there are some conclusions can be draw as follows:
1) The TRACLUS shows a less spending time in sub-trajectory methods. The spending time in these methods mainly spends on calculating the similarity matrix. The spectral clustering method in this application spends more time than the DBSCAN method for the binary Hierarchical approach.
2) The similarity distance function used in FARM method is most suitable for dealing with overlapping trajectories.
3) The DBSCAN method can generate the most representative trajectories. Besides, it is robust to noise.

5 Conclusion

In this study, we propose a general framework on trajectory data mining including pre-processing, feature points extracting and trajectory clustering which is suitable for passive location system. And by comparing the methods on similarity measure and clustering, the advantages and disadvantages in different methods have been shown.

Based on this study, there are several future research directions including improving similarity measure method in clustering, prediction and online trajectory classification base on trajectory data mining.

References

[1] Jing A, Cheng C Q, Song S H, Chen B. "Regional query of area data based on Geohash", *Geogr. Geoinf. Sci*, **29**, pp.31-35, (2013)

[2] Morris B T, Trivedi M M. "Trajectory learning for activity understanding: Unsupervised, multilevel, and long-term adaptive approach", *IEEE transactions on pattern analysis and machine intelligence*, **33**, pp.2287-2301, (2011).

[3] Sankoff D, Kruskal J B. "Time warps, string edits, and macromolecules: the theory and practice of sequence comparison", *Reading: Addison-Wesley Publication*, (1983).

[4] Mazzarella F, Vespe M, Damalas D, Osio G. "Discovering vessel activities at sea using AIS data: Mapping of fishing footprints", *Information Fusion (FUSION), 2014 17th International Conference on*, pp. 1-7, (2014).

[5] Jeung H, Yiu M L, Jensen C S. "Trajectory pattern mining", *Computing with spatial trajectories*, *Springer New York*, pp. 143-177, (2011).

[6] Dhillon I S, Guan Y, Kulis B. "Kernel k-means: spectral clustering and normalized cuts", *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 551-556, (2004).

[7] Lee J G, Han J, Whang K Y. "Trajectory clustering: a partition-and-group framework", *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*, pp. 593-604, (2007).

[8] Cazzanti L, Pallotta G. "Mining maritime vessel traffic: Promises, challenges, techniques", *OCEANS 2015-Genova, IEEE*, pp. 1-6, (2015).

[9] Ester M, Kriegel H P, Sander J, Xu X. "A density-based algorithm for discovering clusters in large spatial databases with noise", *Kdd*, **96**, pp. 226-231, (1996).

[10] Pan X, He Y, Wang H, Xiong W, Peng X. "Mining regular behaviors based on multidimensional trajectories", *Expert Systems with Applications*, **66**, pp.106-113, (2016).

[11] Agrawal R, Faloutsos C, Swami A. "Efficient similarity search in sequence databases", *International Conference on Foundations of Data Organization and Algorithms*, *Springer Berlin Heidelberg*, pp. 69-84, (1993).

[12] Kashyap S, Roy S, Hsu W. "Farm: Feature-assisted aggregate route mining in trajectory data", *ICDMW'09*, pp. 604-609, (2009).

[13] Eiter T, Mannila H. "Computing discrete Fréchet distance", Tech. Report CD-TR 94/64, Information Systems Department, Technical University of Vienna, (1994).

[14] WANG J F, ZHANG M J, XIONG Z H. "Object tracking using spatio-temporal tracklet association", *Application Research of Computers*, **28,** pp.1165-1167, (2012).

[15] Yin P, Ye M, Lee W C, Li Z. "Mining GPS data for trajectory recommendation", *In Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 50-61, (2014).