Route Mining from Satellite-AIS Data Using Density-based Clustering Algorithm

Wang Yitao, Yang Lei, Song Xin
College of Aeronautics and Astronautics Science, National University of Defense Technology, Changsha 410073, China
wanyitao_nudt@163.com

Abstract. Automatic Identification System (AIS) provides a large number of vessel navigation data for Marine traffic research. In this paper, we proposed a route mining method based on satellite AIS. The method includes data preprocessing, structure similarity calculation, clustering and route extraction. The validity of the route mining method was verified by satellite AIS data from seas near Australia and Warrior Strait. This method is helpful to understand Marine traffic pattern. It provides reliable basis for route planning and lay a foundation for ship abnormal behavior detection.

1. Introduction
As the most important mode of transport in international trade, maritime transport accounts for 90% of the global trade volume[1]. In order to cope with the increasingly complex maritime traffic environment and guarantee the maritime traffic safety, the automatic identification system came into being. AIS is a new navigational aid system, which is used for ship recognition, positioning and collision avoidance. It can automatically broadcast and receive ship-related information. In recent years, AIS is widely used in many aspects, such as maritime traffic characteristics found, ship behavior research and situational awareness[2].

Using clustering method in route mining is a research hotspot in AIS data analysis. Liu[3] added constraints to COG (course over ground) and SOG (speed over ground) and adopted an improved density clustering algorithm named DBSCANS D to realize stationary areas detection and route extracting. Wang[4] took port of call as clustering feature and adopted hierarchical clustering algorithm to extract shipping routes. Ma[5] used clustering and Dijkstra algorithm to obtain key nodes of ship navigation path and extract inland river routes according to the characteristics of inland navigation. Xiao et al.[6] simplified trajectory and divided it into sub-trajectories, then applied DBSCAN to obtain typical routes. Jin[7] segmented ship track according to speed and adopted OPTICS algorithm to obtain the results of the track clustering which highlighted the speed attribute. Song[8] applied C-OPTICS algorithm with course constraint to conduct route mining for AIS data from Jiaozhou Bay in Qingdao. And the research results were consistent with the real ship traffic pattern. Existing trajectory clustering studies mostly use inland rivers and shore-based AIS data. The ship type and movement pattern are relatively simple and the space span of data is limited. There is a lack of research on large-scale data analysis using satellite based AIS.

The TianTuo-1(TT-1) and TianTuo-3(TT-3, works normally in orbit till now[9]) micro-nano satellites developed by National University of Defense Technology have received a large amount of AIS data during the mission. We can obtain many valuable maritime traffic information by analyzing these
data. This paper proposes a route mining method for satellite-AIS data based on density clustering algorithm. Firstly, satellite-AIS data need to be preprocessed and divided into some sub-trajectories. Secondly, DBSCAN algorithm is used for clustering according to the structure similarity distance. Finally, a three-dimensional sweep line method is used to realize route mining for satellite-AIS data.

2. Trajectory Clustering Method

The clustering method takes sub-trajectories as clustering objects. It including data preprocessing, structural similarity calculation, clustering and route extraction. In this paper, MMSI (unique identification), latitude and longitude, SOG and COG are used in data analysis. The trajectory of a ship whose MMSI is m can be written as \( T_m = (P_{t_1}, P_{t_2}, \ldots, P_{t_n})_m \), where \( P_{t_n} = [\text{Lng}, \text{Lat}, \text{SOG}, \text{COG}]_n \), contains the location and motion state of the ship at time \( t_n \).

2.1. Data preprocessing

Raw AIS data should be pre-processed before analysis. Those data which are invalid, incorrect, duplicate, or inconsistent can be cleaned as shown in Table 1.

| Type               | Explanation                                      | Operation |
|--------------------|--------------------------------------------------|-----------|
| Invalid data       | The state of message navigation is abnormal, such as anchoring, mooring, grounding, etc. | delete    |
| Error data         | Not conform to the specification of AIS message | delete    |
| Duplicate data     | Duplicate message contents                       | delete    |
| Inconsistent data  | Message contents are contradictory                | delete    |

Then, according to the conciseness and preciseness principle proposed by Lee[10], feature points are found to compress trajectories. Considering the conciseness principle, the location points with obvious changes of SOG and COG are selected as the feature points. And the changing rate of SOG and COG at \( P_{t_{vi}} \) is defined by equation (1). If \( R_{\text{SOG}} \) or \( R_{\text{COG}} \) at \( P_{t_{vi}} \) is greater than the average value of ship track in the research areas, it will be selected as a feature point. Moreover, the change of ship velocity and position can be not obvious in small-scale compared with the large-scale trajectory. Therefore, lower limit of the longitude and latitude span of sub-trajectory named \( LL_{\text{Lng}} \) and \( LL_{\text{Lat}} \) are set to further filter the feature points.

\[
\begin{align*}
R_{\text{SOG}} &= \frac{\Delta \text{SOG}_{t_{vi}}}{\Delta t} = \frac{\text{SOG}_{t_{vi}} - \text{SOG}_{t_n}}{t_{n+1} - t_n} \\
R_{\text{COG}} &= \frac{\Delta \text{COG}_{t_{vi}}}{\Delta t} = \frac{\text{COG}_{t_{vi}} - \text{COG}_{t_n}}{t_{n+1} - t_n}
\end{align*}
\]  

Considering the principle of preciseness to reduce the trajectory shape loss, trajectories are segmented into some sections by feature points. Then D-P (Douglas-Peucker) algorithm is applied to each section and in each section one point will be added at most. The distance threshold \( \lambda \) takes the smaller order of the mean value of longitude and latitude span of sub-trajectories. The sub-trajectory is defined as the track formed by a pair of time-adjacent feature points which belong to the same trajectory. Then the set of sub-trajectories of satellite-AIS data is as follow:

\[ S = \{s_1, s_2, \ldots, s_{all}\} = S_{m_1} \cup S_{m_2} \cup \cdots \cup S_{m_k} \]
Where $S_{m_k} = \{P_{c_1}, P_{c_2}, P_{c_3}, \ldots, P_{c_{m_k}}\}$ is the sub-trajectories of the ship whose MMSI is $m_k$, $P_{c_i}$ is the feature point, and $P_{c_{i-1}}P_{c_i}$ is the sub-trajectory represented by the spherical arc.

2.2. Trajectory clustering

Clustering is an unsupervised learning method. It is often used to analysis unlabeled samples. We used DBSCAN to classified the sub-trajectories. DBSCAN is an algorithm based on a set of neighborhood parameters $(\zeta, \text{MinPts})$. This method clusters samples according to the density-reachable relationship.

Fig. 1 is the illustration of DBSCAN algorithm when take $(\zeta, 3)$ as the neighborhood parameters. $C = \{x_1, x_2, x_3, x_4, x_5\}$ is a cluster obtained by DBSCAN. Red circles represent the $\zeta$ -neighborhood. $x_1$ is a core object and $x_2$ is directly density-reachable from $x_1$. $x_3$ is density-reachable from $x_1$ as well as density-connected with $x_4$.

2.3. Structural similarity calculation

The similarity measurement is the basis of density clustering. We used structural similarity to describe the similarity of sub-trajectories. Structural similarity is the weighted sum of spatial distance, velocity distance and directional distance. Spatial distance describes the spatial density of sub-trajectories distribution, which is define by Hausdorff distance. Velocity distance is used to measure the difference of average velocity between different sub-trajectories. Directional distance is used to measure the difference in the direction of the quantum trajectory.

For two sub-trajectories $s_m = \{p_{m_1}, p_{m_2}\}$ and $s_n = \{p_{n_1}, p_{n_2}\}$ from different ships, spherical Angle $\alpha$ is used to represent the length of sub-trajectories. Equations (3), (4) and (5) are used to calculate $d_{hd}$, $d_v$ and $d_{\theta}$.

\[
d_{hd}(s_m, s_n) = \max\{h(s_m, s_n), h(s_n, s_m)\}
\]

\[
d_v = \left|v_{s_m} - v_{s_n}\right|
\]

\[
v_{s_m} = \frac{\alpha_{s_m}}{\Delta t_{s_m}}
\]

\[
d_{\theta}(s_m, s_n) = \begin{cases} 
\min(\alpha_{s_m}, \alpha_{s_n}) \times \sin(\theta), & 0^\circ \leq \theta \leq 90^\circ \\
\min(\alpha_{s_m}, \alpha_{s_n}), & 90^\circ < \theta \leq 180^\circ 
\end{cases}
\]
In equation (3), \( h(s_m, s_n) \) is a one-way Hausdorff distance. In equation (5), \( \theta \) stands for the angle between the normal vectors of two sub-trajectories. Using the methods above, we can obtain three normalized distance matrices, including spatial distance matrix \( D_{h-norm} \), velocity distance matrix \( D_{v-norm} \) and directional distance matrix \( D_{\theta-norm} \). The structural similarity distance \( D_s \) is calculated by equation (6):

\[
D_s = w_h * D_{h-norm} + w_v * D_{v-norm} + w_{\theta} * D_{\theta-norm}
\]

Where \( w_h, w_v \), and \( w_{\theta} \) are weights and \( \omega_h + \omega_v + \omega_{\theta} = 1 \).

2.4. Route extracting

Previous studies mostly focused on shore-based AIS data and adopted two-dimensional sweep line method to extract representative routes\[6\], \[11\]. As shown in Fig.2, the basic idea of this method is to set a sweep line to scan along the cluster direction. When the number of intersection points of sweep line and sub-trajectories is greater than the threshold, the average coordinate should be calculated and set as a route node, and the representative routes are composed of all route nodes. In this paper, satellite-AIS data are used in route extracting. Considering that the span of sub-trajectories from satellite-AIS is relatively larger, so the sub-trajectories should be treated as arcs rather than lines. The difference between three-dimensional sweep line method and the two-dimensional sweep line method is that the new method takes sub-trajectories as arcs and rotates the sweep line around the axis instead of translating it in the sub-trajectories clusters.
Fig. 3 shows the realization of 3D sweep line method. The green arcs in Fig.3(a) is the sub-trajectories of cluster \( c = \{s_1, s_2, \cdots, s_i\} \) and the blue arc is the direction of cluster \( c \) whose normal vector \( N_c \) can be calculated by equation (7). The yellow arc in Fig.3(b) is a sub-trajectory \( s_i \) and the length of \( s_i \) is \( \alpha(s_i) \). \( N(s_i) \) is the normal vector of \( s_i \). \( N_{sl} \) is the normal vector of the red sweep line. The brown vector \( p_{sl} \) is the radius vector from the center of earth to the intersection of the red arc and the blue arc. \( N_{sl} \) is initialized by \( p_{sl} \). The sweep line rotates around \( N_c \) to scan cluster \( c \). Equation (8) is used to calculate the coordinates of the intersection point between the sweep line and the sub-trajectory. Then we can obtain the route nodes and get the representative routes.

\[
\begin{align*}
    N_c &= (x_c, y_c, z_c)^T = \frac{N_{c\_temp}}{|N_{c\_temp}|} \\
    N_{c\_temp} &= \frac{\sum_{s_i \in c} \alpha(s_i) \times N(s_i)}{\sum_{s_i \in c} \alpha(s_i)} \\
    OP_\theta(s_i) &= \pm N_{sl}(\theta) \times N(s_i)
\end{align*}
\]

3. Results and Discussion

The satellite-AIS data from seas near Australia and Warrior Strait is used to test the effectiveness of the trajectory clustering method.

3.1. Cluster analysis of ship tracks in seas near Australia

The seas near south Australia (30°-40° S, 115.8°-140° E) was studied. There is 4596 message of category 1,2 and 3 received by TT-3 from 18th February 2018 to 26th February 2018 (149 trajectories in total). Fig.4 shows the original data. Fig.5 and Fig.6 are the feature points and their trajectories. After data preprocessing with parameters in Table 2, there left 2221 position points and the data compression rate is 48.32%. Satellite-AIS data have the disadvantage that some sub-trajectories may be too long. These long trajectories cannot describe ship behavior as well as measure the space distance correctly. So we limited the span of sub-trajectories composed of feature points. In this example, the latitude and longitude span of sub-trajectories were limited within \((0.3684°, 3°)\) and \((0.0541°, 3°)\). Fig.7 shows 319 sub-trajectories selected from Fig.6. Then we used these sub-trajectories for clustering analysis.
Fig. 6 Feature points trajectories

Table 2 Data preprocessing parameters (seas near south Australia)

| Parameter         | Value   |
|-------------------|---------|
| $R_{SOG}$ (deg/min) | 2.1219  |
| $R_{COG}$ (n mile/5min) | 9.6290  |
| $LL_{Lon}$ (deg)   | 0.3684  |
| $LL_{Lat}$ (deg)   | 0.0541  |
| $\lambda$ (deg)    | 0.01    |

The experiments shows that when we take 0.8, 0.1, 0.1 as $\omega_h$, $\omega_v$, $\omega_\gamma$ and take (0.059, 6) as the neighborhood parameters, the clustering result (shown in Fig. 8) is good. Region A and B in Fig.8 mark the abnormal sub-trajectories due to the heading. Region C is recognized as noise because it is far from other sub-trajectories. Comparing the route extracted from satellite-AIS in Fig. 9 with the real route, the route in this example is in coincide with a section of route from Melbourne port to Fremantle Port in AUS3.

3.2. Cluster analysis of ship tracks in Warrior Strait
Located in Papua New Guinea, the Warrior Strait connects the Bismarck and Solomon Sea and is busy with shipping. The satellite-AIS data from warrior Strait ($3^\circ \sim 10^\circ$ S, $145^\circ$-153$^\circ$ E) were selected for research. There is 2554 message of category 1,2 and 3 received by TT-3 from 18th February 2018 to 26th February 2018 (281 trajectories in total). Fig.10 shows the original data. Fig.11 and Fig.12 are the feature points and their trajectories. After data preprocessing with parameters in Table 3, there left 1352 position points and the data compression rate is 42.9%. Then limit latitude and longitude span of sub-trajectories within $(0.3751^\circ, 3^\circ)$ and $(0.3264^\circ, 3^\circ)$. Fig.13 shows 309 sub-trajectories selected from Fig.12.
Fig. 10   The original trajectory

Fig. 11 Comparison of original data and feature points data

Fig. 12 Feature points trajectories

Fig. 13 Feature point trajectories (selected)

Table 3 Data preprocessing parameters (Warrior Strait)

| Parameter   | Value   |
|-------------|---------|
| $R_{COG}$ (deg/min) | 4.2835  |
| $R_{SOG}$ (n mile/5min) | 0.3803  |
| $LL_{Lon}$ (deg) | 0.3751  |
| $LL_{Lat}$ (deg) | 0.3264  |
| $\lambda$ (deg) | 0.1     |

The experiments shows that when we take 0.7, 0.1, 0.2 as $\omega_h$, $\omega_v$, $\omega_y$ and take (0.058, 6) as the neighborhood parameters, the clustering result (shown in Fig.14) is good. Considering sailing direction of the two clusters in Fig.14, cluster 0 (in blue) is ships crossing the strait from north to south, and cluster 1 (in green) is the ships crossing the strait from south to north. The clustering result indicates that speed distance in structural similarity can well distinguish ships with different sail direction. Fig.15 shows the route extraction results and it enlarges the nearest region between the two routes, where there is an interval of not less than 1km. According to the traffic separation scheme, in ports, straits and other areas with heavy traffic, the channel should be divided into two navigation channels. And one way navigation is implemented for each channel. It can be seen from Fig.15 that when ships crossing Warrior Strait, most of them obeyed the traffic separation scheme. Specifically, they sailed to the right. Comparing the routes extracted from satellite-AIS in Fig.15 with the real route, the routes in this example are in coincide with a section of route from Brisbane Port via Warrior Strait to the southeast coast of China, Japan or Korea.
4. Conclusion
In this paper, we have proposed a route extraction method based on satellite-AIS. The method includes data preprocessing, structural similarity calculation, clustering and route extraction. The real data from seas near south Australia and Warrior Strait collected by TT 3 satellite is used to verify the effectiveness of this method. However, the route extraction method performs not good enough in distinguishing different routes in high-density area, this may be because the clustering algorithm classifies some sub-trajectories wrongly. In order to solve this problem, the measurement method of structural similarity and the clustering algorithm need to be modified in the future work.

Acknowledgments
The authors appreciate the AIS data support provided by College of Aeronautics and Astronautics Science of NUDT.

References
[1] Rodrigue, J.-P. The Geography of Transport Systems, 4th ed.; Routledge: New York, NY, USA, 2017; p. 200
[2] Fournier M, Hilliard R C, Rezaee S, et al. Past, present, and future of the satellite-based automatic identification system: areas of applications (2004–2016)[J]. WMU journal of maritime affairs, 2018, 17(3): 311-345.
[3] Liu B, De Souza E N, Matwin S, et al. Knowledge-based clustering of ship trajectories using density-based approach[C]. international conference on big data, 2014: 603-608.
[4] Wang S, Gao S, Yang W, et al. Ship route extraction and clustering analysis based on automatic identification system data[C]. international conference on intelligent control and information processing, 2017: 33-38.
[5] Ma S H. Route Generation based on AIS Data [D]. Dalian Maritime University,2017.(In Chinese)
[6] Xiao X. Study on Ship’s Trajectory Clustering Model Based on AIS Data [D]. JiMei University,2015. (In Chinese)
[7] Jin X Y. Analysis and Prediction Method of Vessel Trajectory [D]. Beijing University of Posts and Telecommunications,2018. (In Chinese)
[8] Song P. Vessels Trajectory Clustering and Its Application based on C-OPTICS Algorithm [D]. Dalian Maritime University,2017. (In Chinese)
[9] Li S , Chen X , Chen L , et al. Data Reception Analysis of the AIS on board the TianTuo-3 Satellite[J]. Journal of Navigation, 2017, 70(4):1-14.
[10] Lee J G , Han J , Li X . [IEEE 2008 IEEE 24th International Conference on Data Engineering (ICDE 2008) - Cancun, Mexico (2008.04.7-2008.04.12)] 2008 IEEE 24th International Conference on Data Engineering - Trajectory Outlier Detection: A Partition-and-Detect Framework[J]. 2008:140-149.
[11] Pan S , Jingbo Y . Extracting Shipping Route Patterns by Trajectory Clustering Model Based on Automatic Identification System Data[J]. Sustainability, 2018, 10(7):2327-.