sciendo

POLISH MARITIME RESEARCH 1 (109) 2021 Vol. 28; pp. 136-148
10.2478/pomr-2021-0013

RESEARCH ON SHIP TRAJECTORY EXTRACTION BASED ON MULTI-ATTRIBUTE DBSCAN OPTIMISATION ALGORITHM

Xiaofeng Xu
Deqaing Cui*
Yun Li
Yingjie Xiao
Shanghai Maritime University, Haigang streat, ShangHai

* Corresponding author: *cuideqiang0213@163.com (D. Cui)*

ABSTRACT

With the vigorous development of maritime traffic, the importance of maritime navigation safety is increasing day by day. Ship trajectory extraction and analysis play an important role in ensuring navigation safety. At present, the DBSCAN (density-based spatial clustering of applications with noise) algorithm is the most common method in the research of ship trajectory extraction, but it has shortcomings such as missing ship trajectories in the process of trajectory division. The improved multi-attribute DBSCAN algorithm avoids trajectory division and greatly reduces the probability of missing sub-trajectories. By introducing the position, speed and heading of the ship track point, dividing the complex water area and vectorising the ship track, the function of guaranteeing the track integrity can be achieved and the ship clustering effect can be better realised. The result shows that the cluster fitting effect reaches up to 99.83%, which proves that the multi-attribute DBSCAN algorithm and cluster analysis algorithm have higher reliability and provide better theoretical guidance for the analysis of ship abnormal behaviour.

Keywords: clustering algorithm,abnormal route,DBSCAN,Feature trajectory extraction,fitting analysis

INTRODUCTION

In the continuous development of maritime traffic, AIS (automatic identification system) data plays an extremely important role in the process of extracting ship tracks, analysing ship behaviour [1], and ensuring course safety. An AIS is an automatic identification system that can realise global coverage and send ship position information to the competent department and other ships every few minutes, so as to track the ship's trend and monitor its heading. It is an important resource for studying maritime activities [2]. AIS can be used in ship trajectory detection, ship trajectory outlier analysis and other aspects [3]. It can efficiently excavate typical sections of each characteristic pattern in the water area, and effectively use the AIS data of ship navigation to provide great help for ship trajectory extraction [4][5], so as to guarantee the navigation safety of ships.

Scholars around the world have done a lot of research on AIS data application. Zhang et al. propose a multi-state ship trajectory reconstruction model. The model is processed in three steps, including (i) removal of outliers, (ii) estimation of the ship navigation state and (iii) ship trajectory fitting. The model allows the reconstruction of the ship's trajectory under different navigational conditions, such as berthing, manoeuvring and normal speed navigation. It is concluded that the performance of the model is better than that of the linear regression model [6]. Yan et al. propose a ship traffic route extraction method based on automatic ship history identification system (AIS) data. In this method, the ship trajectory with rich position information is transformed

into a ship navigation semantic object (STSO) with semantic information, and each ship navigation is abstracted into a stop-waypoint-stop navigation object. In addition, based on graph theory, STSO is further integrated into nodes and edges of directed maritime traffic charts to realise the extraction and expression of routes [7]. Wei et al. propose a new AIS trajectory compression algorithm based on ship behaviour characteristics, which takes into account the spatial and motion characteristics of the trajectory. The algorithm is mainly composed of two parts: one is to simplify the trajectory by using the Douglas–Peucker (DP) algorithm according to the spatial characteristics; the other is to simplify the trajectory by using a sliding window based on the motion characteristics. In addition, statistical theory is applied to determine the threshold of motion characteristics in the sliding window algorithm. The two results are combined to form a trajectory simplification algorithm considering ship behaviour [8]. Li et al. put forward the problem that AIS redundant information can reduce the accuracy of trajectory clustering. In order to improve the calculation accuracy and reduce the amount of calculation, the merged distance can be used to measure the similarity between tracks, and the low-dimensional spatial expression of the similarity between tracks can be used in multi-dimensional zoom (MDZ). Lin M.L used language to design and realise a prototype system of ship track clustering, so as to mine and analyse the important information in the AIS data and obtain the behavioural patterns of ships [22]. Sheng and Yin put forward a clustering model based on the track of the AIS data applied to the analysis of transport routes, where the entire model includes four main parts: data pre-processing, similarity measurement structure and typical path extraction, and clustering. The model considers the ship trajectory through geospatial information where different transport routes can be automatically classified. It allows experimental verification through specific waters. The results show that the model is effective, and helps to further understand the route model [10].

However, due to the impact of objective factors such as the environment, climate and the crew's subjective behaviour when the ship is underway, abnormal data will appear in the process of AIS data generation, and the accuracy and navigation under direct AIS data processing will be improved. Therefore, the detection of outliers in AIS data can better ensure navigation safety and make the processing of AIS data more accurate.

In AIS data processing research, a clustering algorithm is often used to process the ship trajectory, equipment and other data. Clustering algorithms have been widely used in various aspects such as ship trajectory extraction, ship anomaly detection and ship evaluation. Scholars around the world have conducted a large number of studies on this. Xiao et al. designed a ship track clustering algorithm based on AIS information. This clustering algorithm used the change of heading to obtain a candidate set of feature points, and determined the final feature points by the Minimum Description Length (MDL) criterion, so as to classify ship track class clusters, clustering large ships in specific waters and obtaining the typical representative trajectories of ships [11]. Zhou et al., based on AIS data of a large number of ships, measured the trajectory similarity by fusion distance (MD). Aiming at the problem that the traditional DBSCAN algorithm needs to query the neighbourhood of each sample repeatedly, an improved DBSCAN algorithm is proposed to reduce the number of regional queries, thus improving the time efficiency of the algorithm and completing the clustering of the existing trajectory [12]. Cui has done much work and in his paper the characteristics and structure of the ship AIS trajectory data are summarised, a ship trajectory prediction model is established by using the related method of machine learning, and the future trajectory of the ship is predicted. The main research work includes data completion and exception handling methods. Based on the original AIS data, data completion and abnormal data processing were carried out. The work also includes a clustering and regression method for ship trajectory prediction. Combined with the classification idea, the K-medoids method is used to cluster the trajectory samples, and a regression prediction is made in each class to effectively reduce the difference between the trajectory samples. The experimental results show that the clustering regression method can improve the accuracy of prediction [13]. Liu and Shi use a new way to solve the problem. The skeleton extraction technology used for model reconstruction is used to carry out trajectory clustering analysis on the historical data of ships, which lays a foundation for studying the behaviour pattern of ships, and then provides a new method for regional navigational goods supervision. In view of the problem that the current trajectory clustering algorithm consumes a lot of computing resources and cannot process the trajectory quickly, the trajectory is converted into images for gradient compression and extraction clustering. The thermal surface of the track line was constructed by relying on the thermal distance field, and then the Laplacian operator was used to iteratively shrink the meshing thermal surface, and the profile skeleton line was obtained as the clustering effect picture [14].

In numerous studies, many extended algorithms have been developed based on clustering algorithms, among which the DBSCAN algorithm is widely used in ship trajectory extraction. Jiang, Xiong and Tang proposed an improved DBSCAN clustering algorithm. The ship trajectory was segmented by taking the angle and speed change as information measures, and the discrete Frechet distance was used as a trajectory similarity measure. The DBSCAN algorithm was used to cluster the trajectory segments, and the typical trajectory of ship movement was obtained [15]. Zhao, Shi and Yang proposed an adaptive hierarchical clustering method for the ship trajectory based on DBSCAN. By analysing the characteristics of the DBSCAN algorithm, the parameters are determined according to the internal distribution law of the data set and the change law of the quasi-clustering effect. Hierarchical clustering is carried out with statistical theory to adapt to ship trajectory data with an uneven density distribution [16]. Peng et al., based on the AIS data, used cloud computing and a clustering algorithm

to carry out trajectory clustering analysis on the historical data of ships and build the normal trajectory model of ship navigation, which lays a foundation for real-time detection of abnormal ship trajectories, and then provide a new method for improving the intelligent level of water traffic supervision. Aiming at the low efficiency of the current trajectory clustering algorithm, an improved parallel DBSCAN of sub-trajectory clustering algorithm SPDBSCANST (Parallel DBSCAN of sub-trajectory based on Spark) was proposed, based on Spark memory computing technology and data partitioning in order to mine the important AIS information [17]. Jiang, Xiong and Tang used an improved DBSCAN algorithm based on the turning angle and speed change rate of ships, combined with the Frechet Distance (FD) to measure the distance, and divided the track sections of ships. After realising the clustering results, typical representative tracks of ships' navigation were extracted, using a particular body of water for experiments. The experimental results show that the algorithm can improve the clustering effect and accuracy, and lay a foundation for the detection of ship trajectory anomalies [15]. Li et al. proposed that the redundant information of the AIS would reduce the accuracy of trajectory clustering. In order to improve the accuracy and reduce the amount of calculation, the similarity between trajectories was measured by combining the distance, and the low-dimensional space expression of the similarity between trajectories used multi-dimensional zoom (MDZ). The fusion between MDZ and the improved DBSCAN algorithm can identify the trajectory route well, and through the sampling data of specific waters, the DBSCAN algorithm is used to cluster the spatial points to verify the effectiveness and accuracy of the algorithm [9].

The existing clustering algorithms are mostly K-means, Spark, etc., compared with which DBSCAN has higher accuracy and a wider application range. However, the existing DBSCAN algorithm still needs to divide the ship trajectory, which will cause the loss of the ship trajectory and decrease the accuracy. Therefore, this paper proposes a multi-attribute DBSCAN extension algorithm, which vectorises the ship trajectory and avoids having to divide it, thus ensuring the complete ship trajectory and improving the calculation accuracy.

MULTI-ATTRIBUTE DBSCAN OPTIMISATION ALGORITHM FOR SHIP TRAJECTORY CLUSTERING

SHIP TRAJECTORY CLUSTERING BASED ON MULTI-ATTRIBUTE DBSCAN OPTIMISATION ALGORITHM

Currently, there are two main two research ideas about trajectory analysis and trajectory extraction: one idea is to target the trajectory as a whole for cluster analysis, which can better dig out the trajectory of the key path, but its existence may lose some general sub-tracks defects as shown in Fig. 1.

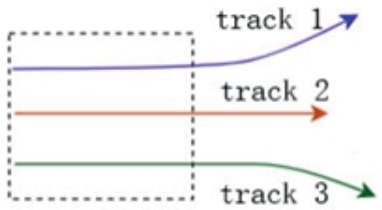

Fig. 1. Locally similar trajectories

Track 1, track 2 and track 3 have short movements with a similar general trajectory, as shown in the dotted box of the figure. If the clustering trajectory is treated as a whole, it could cause the loss of some similar sub-tracks, leading to a lack of important information. Another research idea is to divide the whole ship track into several sub-tracks. This method can effectively avoid missing sub-track information, but it also damages the integrity of the track.

SHIP TRAJECTORY CLUSTERING MODEL

Formula and similarity measure of distance between ship track points

The division of data needed to calculate the similarity between samples and the similarity calculation method between samples can be abstracted as a sample distance function. In order to calculate the distance between samples and the similarity between samples from a distance matrix, assuming that the extracted sample is X, the number of samples is n, the dimensions p, the distance matrix can be expressed in the following form:

$$\begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix}$$

where Xij represents the j-th dimension data of the i-th sample.

Design of structural similarity distance formula

In the distance calculation, the following formulas are needed: Euclidean distance [18], Manhattan distance [19], Chebyshev distance and Minkowski distance [21].

1) Euclidean distance

Used to represent the distance between two points. In n-dimensional space, the Euclidean distance formula can be expressed as

$$d_{ij} = (P_i, P_j) = \left(\sum_{k=1}^{m} \left|x_{ik} - x_{jk}\right|^2\right)^{1/2}$$

2) Manhattan distance

Used to represent the sum of the absolute values of the wheelbase of two points in the standard coordinate system. In n-dimensional space, the Manhattan distance calculation formula is expressed as

$$d_{ij} = (P_i, P_j) = \sum_{k=1}^{m} \left| x_{ik} - x_{jk} \right|$$

3) Chebyshev distance

Used to represent the maximum value of each coordinate. In n-dimensional space, the Chebyshev calculation formula is expressed as

$$d = \mathrm{Max}(\left| x_{1_1} - x_{1_2} \right|, \left| x_{2_1} - x_{2_2} \right|, , , \left| x_{n_1} - x_{n_2} \right|)$$

4) Minkowski distance

When n=1, the Minkowski distance is the Manhattan distance; when n=2, the Minkowski distance is the Euclidean distance. The calculation formula is as follows:

$$d_{ij} = (P_i, P_j) = (\sum_{k=1}^{m} \left| x_{ik} - x_{jk} \right|^n)^{1/n}$$

The Manhattan distance, also known as the "taxi distance", depends heavily on the coordinate system. The distance between points changes as the coordinate axis changes. The Euclidean distance refers to the distance between two points.

In order to solve the problem of trajectory loss caused by trajectory similarity, the concept of structural distance is introduced, and different weights are assigned to different attributes of the trajectory according to the actual scene, so as to comprehensively judge the similarity between trajectories. The distance between ship tracks is abstractly defined as the distance between vector points by using the position, speed and heading attributes contained in the ship track data. Thus, the formula of the structural similarity distance between two vector points is given as follows:

$$D_{dist}(P1, P2) = w1 * d1(P1, P2) + w2 * d2(P1, P2) + w3 * d3(P1, P2)$$
$$w1 + w2 + w3 = 1$$

where:

D_{dist} (P1, P2) represents the structural distance between the vector points P1 and P2;

d1 (P1,P2) represents the spatial distance between the vector points P1 and P2;

d2 (P1, P2) represents the speed deviation between P1 and P2;

d3 (P1, P2) represents the course deviation between the vector points P1 and P2;

wn represents the weight of the attribute N of the trajectory over the structural distance.

DBSCAN OPTIMISATION ALGORITHM

DBSCAN is one of the most widely used and referenced clustering algorithms. In density-based clustering, the phenomenon of clustering is that high-density regions are separated by low-density regions. Compared with other types of clustering, the density-based clustering algorithm is an unsupervised clustering algorithm, which is not sensitive to noise data and can find clusters of any shape and size without setting the number of clusters in advance. It is very suitable for clustering in the case of AIS trajectory data with many noise points and a highly uncertain trajectory pattern. The following describes the relevant definitions involved in density-based clustering:

1 Eps fields

Given an object P, with P as the centre, the area with a radius of eps (epsilon) becomes the eps domain of object P. The expression of the definition is as follows:

$$N_{eps}(P) = \{d \in D | dist(P, Q) \leq eps\}$$

where D is the given data set, and dist(P, Q) represents the distance between P and Q in the data set.

2 Core objects

If the eps domain of an object P includes at least minPts objects (minPts represents the minimum number of points), then P is called the core object.

3 Direct density can be reached

In a given object data set D, if P exists in the eps field of Q, where P is a core object, the object P to object Q is directly denser.

4 The indirect density is achievable

If there is an object chain P1, P2...,Pn, P1=q, Pn= P, for PI Î D, 1≤ I ≤n, if PI +1 and PI from eps and minPts are directly accessible, then the density of object P to object Q is reachable, and it is an indirect density.

5 Densities are connected

If there is an object O in the given object set D, for objects P and Q, from object Q about eps and minPts are density-accessible, then it means that objects P and Q are connected about the eps and minPts density.

The DBSCAN algorithm requires two parameters (eps and minPts) and works by differentiating between core points, border points and noise. The concept relationship of the DBSCAN algorithm is shown in Fig. 2. The operation effect of the DBSCAN algorithm is shown in Fig. 3.

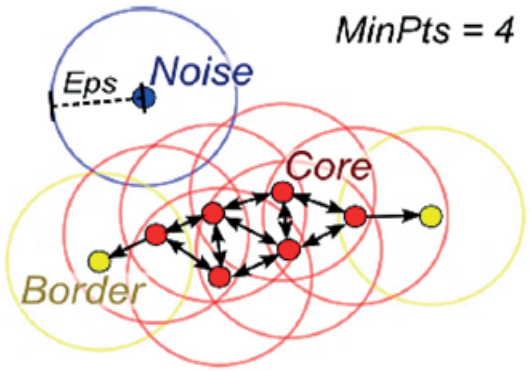

Fig. 2. DBSCAN algorithm concept relationship

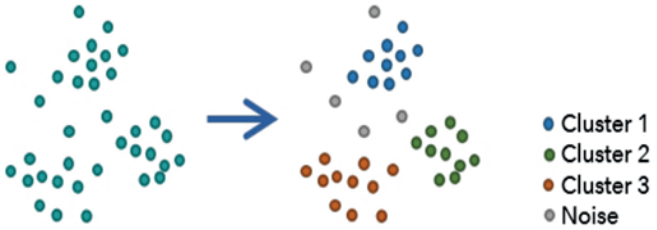

Fig. 3. DBSCAN algorithm running effect

The pseudocode of the DBSCAN algorithm is shown in Table 1.

Tab. 1. DBSCAN algorithm pseudocode

DBSCAN Algorithm
Input: data object D, object radius eps, minimum number of minPts Output: Clustering results Algorithm start: 1: Initialise a new class cluster C 2: for each object that is not marked in the data set P do 3: if Neps(p) Contained object≥minPts then 4: Add all the points in P to C 5: for all unprocessed objects in Neps(p) the q do 6; if Neps(q) contains at least minPts objects 7; Objects in Neps(q) that do not belong to the class cluster are added to C 8; end if 9: end for 10: else Label P for noise 11: end for END

MULTI-ATTRIBUTE DBSCAN OPTIMISATION ALGORITHM

Trace clustering follows the law of "birds of a feather flock together", which divides a bunch of unmarked data into piles after some similarity measurement method, which is called clustering. Each cluster comprises trajectory data points with high similarity. When the similarity between points is higher and the similarity between clusters is lower, the overall clustering effect will be better.

The original DBSCAN algorithm is a clustering algorithm based on density. The core of the algorithm is two parameters, namely the eps field and the minimum number of points, minPts. The eps field represents the field within the object radius eps, and Neps(p) is used to represent the set within the eps radius of point P. An object is called a core object if it contains at least minPts of other objects within the eps domain of the object. In this paper, the DBSCAN algorithm is extended on the basis of the AIS data set, making full use of the three attributes of ship position, speed and heading, and simplifying the originally complex trajectory segment to the similarity measure between vector points. In order to conform to the reality of the more complex marine traffic flow situation, two control variables SpdRange and DirRange are added to the input variables of the model, to control the ship's speed range and scope of course, according to the practical application scenarios. In the process of ship's trajectory clustering, constraints of the ship speed and the range of the course must be considered, because it may enter port at a low speed, and may also have left port at a high speed, so SpdRange and DirRange are considered. Ship trajectory points that are not only similar in spatial position but also have little difference in ship speed and basically the same ship direction are gathered into a cluster. The pseudocode of this algorithm is shown in Table 2 below.

Tab. 2 Multi-attribute DBSCAN extension algorithm pseudocode

Multi-attribute DBSCAN extension algorithm
Model inputs: data object D, object radius eps, minimum number of points minPts, speed threshold range SpdRange, ship direction threshold value DirRange Model output: clustering results start: 1:Definition of method 1: multi-attribute DBSCAN extension algorithm (dataSet D, P, eps, minPts, SpdRange, DirRange) 2:Initialise an empty clusterList 3: for each point P in dataset D do 4: queryNeighbourPoints(dataSet D, P, eps, minPts, SpdRange, DirRange) 5: if neighbourPts not NULL then 6: neighbourPts add to clusterList 中 7: for each cluster C in clusterList do 8: for each clusterC'' in clusterList do 9: if C and C && mergeCluster(C, C'') is TRUE 10: clusterList. remove(C'') 11: return clusterList 12: Define Method 2: queryNeighbourPoints(data, P, eps, minPts, SpdRange, DirRange) 13: Initialise an empty set cluster 14: for each point Q in data do 15: if distance(P,Q)<eps && |P.SOG-Q.SOG|<SpdRange && |P.COG-Q.COG|<DirRange 16: add Q to cluster 17: if cluster.size>minPts 18: Mark point P as the core object 19:: return cluster 20: or return NULL 21:Define Method 2: mergeClusters(cluster A,cluster B) 22: merge=FALSE 22: for each point Q in cluster B do 23: if point Q is the core object && cluster A contains the point Q 24: merge=TRUE 25: All objects in cluster B are added to cluster A END

The purpose of Method 2 (QueryNeighbourPoints) is to compare point P with all other ship track points to find track points similar to it. The core idea of this method is as follows: first, to judge whether there is a point-point P relationship in the ship trajectory data set that conforms to the designed similarity structure. The second step is to judge whether the number of track points contained in the cluster is greater than minPts. The third step is to judge whether the cluster contains at least minPts track points. If so, it means that the cluster is reasonable. Otherwise, the return object is null.

In the pseudocode, short (P, Q) < eps && | p. OG - Q.S OG | < SpdRange && | P.C OG - Q.C OG | < DirRange's role is to judge a vector points with another vector in line with the similarity measure scheme. With the formula expressed as follows:

$$D_{dist}(P,Q) = w1 * d1(P,Q) + w2 * d2(P,Q) + w3 * d3(P,Q)$$
$$w1 + w2 + w3 = 1$$

Given D_{dist} (P, Q), the structural distance between points P and Q, the structural distance between two points needs to meet conditions: short (P, Q) < eps && | p. OG - Q.S OG | < SpdRange && | P.C OG - Q.C OG | < DirRange, where short (P, Q) is the space distance between points P and Q;| p. OG - Q.S OG | is the speed deviation of the points P and Q;| P.C OG - Q.C OG | indicates points P and Q of the ship to the deviation range.

Method 3 (mergeClusters) consolidates the cluster of classes obtained through method 2 (queryNeighbourPoints). If the core point of class cluster A happens to be the boundary point of class cluster B, then according to the merge condition of the DBSCAN algorithm, class cluster A and class cluster B need to be fused.

SHIP SIMULATION

The improved multi-attribute DBSCAN extension algorithm is applied to ship trajectory extraction in waters near the main waterway of Jintang Bridge to carry out a simulation verification of the algorithm.

AIS DATA PRE-PROCESSING

By 1 June 2017 solstice JCP 4, AIS data after being decoded pre-treatment according to the ship MMSI number matching with AIS static information table are not accessible. Therefore, some static information is inaccurate and it required to adopt the method of manual entry of real AIS data.

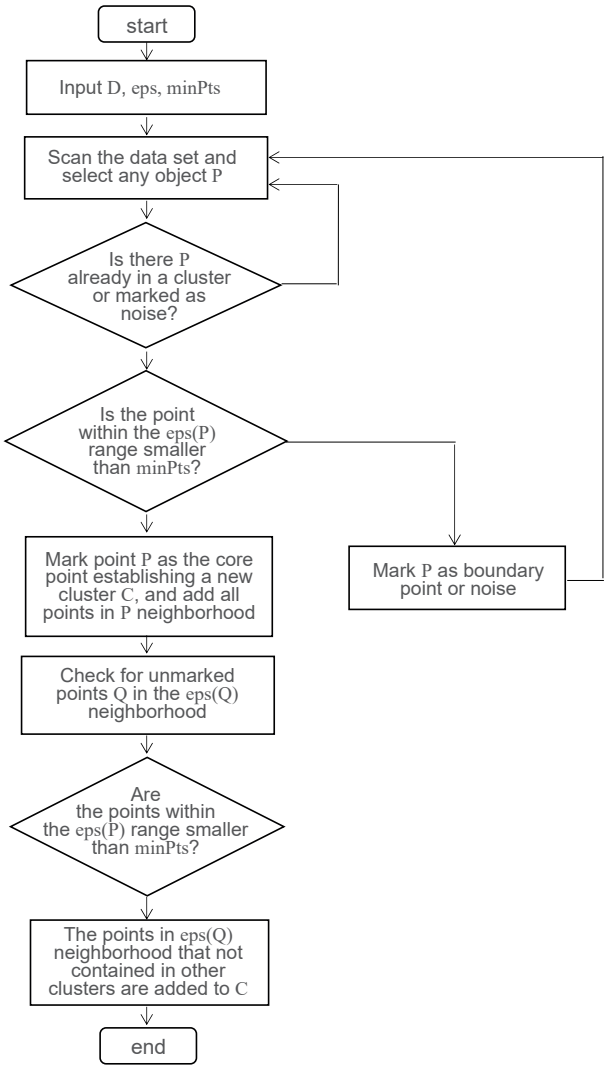

Fig. 4. Algorithm flow chart

After correction of AIS information that were incorrect for 221217 real AIS data of 638 ships, we removed redundant AIS data information used in the simulation. For the obtained AIS data samples, the proportions of each ship type were calculated statistically. The distribution of ships is shown in Fig. 5.

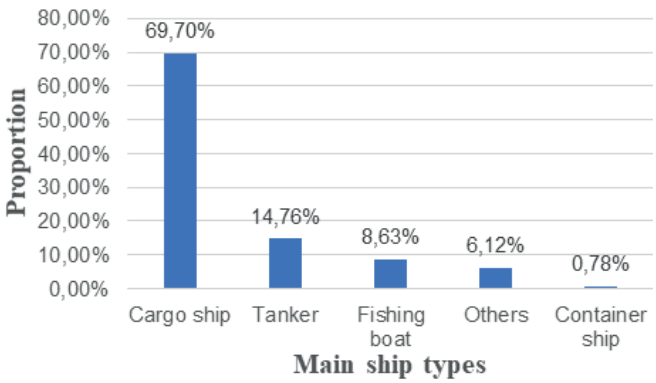

Fig. 5. Distribution of ships

As can be seen from Fig. 5, the main types of ships in the study area from large to small are: cargo ships (mainly liquid cargo ships and dry cargo ships), oil tankers, fishing ships and other types of ships (mainly pilot ships, dredging ships and tugboats). Among them, cargo ships, oil tankers and fishing vessels account for the majority, at 70.49%, 14.76% and 8.63% respectively, and amounting to 91.79% of the total number of ships.

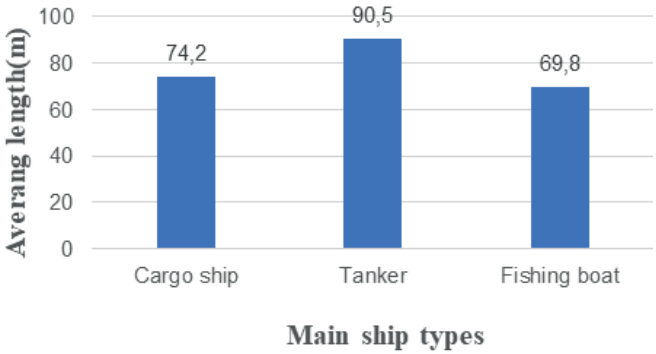

Fig. 6. Average ship length distribution

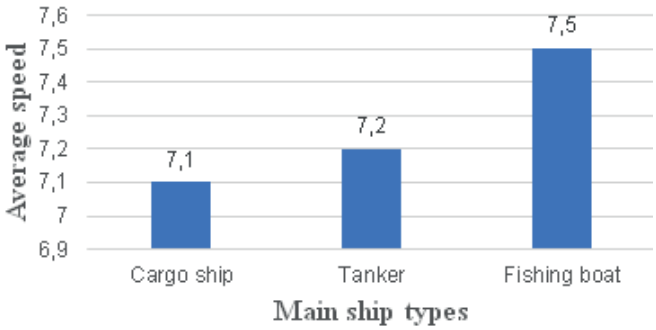

Fig. 7. Distribution of average ship speed

After screening, counting and calculating the lengths of the three ship types (cargo ships, oil tankers and trawlers), the average length distribution of all kinds of ships can be obtained, as shown in Fig. 6.

After screening, counting and calculating the speed of the top three types of ships (cargo ships, oil tankers and trawlers), the average speed distribution of all kinds of ships can be obtained, as shown in Fig. 7.

In order to analyse the main features of ship navigation in the water area, the multi-dimensional ship information in the water area is visualised and a four-dimensional ship information parallel coordinate chart is made. The relevant information of different types of ships is divided into four categories (cargo ship, oil tanker, fishing ship and others) in different colours. For each type of ship there are three corresponding attributes: speed, length and ship width, as shown in Fig. 8.

Ships of different tonnage have different requirements for speed control and water depth. In the process of entering and leaving port, the inertia and draft depth of large ships are larger, and the elimination of residual velocity is slower. Navigation must be in strict accordance with the regulations of the port authorities. The draft depth of small ships is relatively shallow and the manoeuvrability is good, which will cause small ships to travel too fast and be unable to enter or leave ports according to the requirements of ports. If the AIS data is not screened in advance, the clustering results will be very confusing, making it difficult to show the clustering effect, so the validity and correctness of the clustering results cannot be accurately judged.

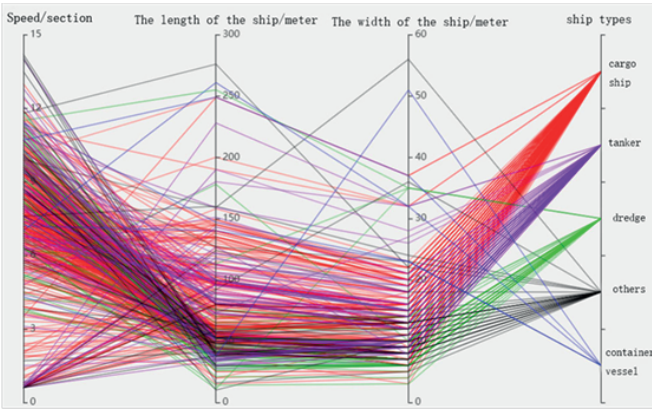

Fig. 8. Parallel coordinate chart of ship data

According to the distribution of ship types in the study area, their different ship track density and different ship lengths were displayed with the help of the BDP visual tool at the distance of 50 m and analysed according to the different ship lengths, so as to determine that the main objects of study were cargo ships and oil tankers. The process is as follows:

The track points of this water area are divided into two types: ship shelving point and ship movement point. In determining the speed threshold values of the shelving points and movement points of the ships in the study waters, it was found that when the sailing speed was within the range of 0–1 kn, the ship tracks presented the largest number of scatter points, and there were no continuous track points. These scatters were thus judged to be shelving points. If the filter speed value range is 0–1.1 kn, there will be continuous trajectory points, indicating that the current 1.1 kn is not the optimal speed threshold. Therefore, the optimal speed threshold is set as 1 kn to distinguish the shelving points and movement points. Track density visualisation is carried out for the shelving points of freighters and oil tankers with different lengths, showing the distribution of the ship track density when the speed is 0–1 kn. From the ship track density distribution, the hot spots of shelving points of different types of ships with different lengths can be intuitively found.

The following is a visual analysis of the trajectory density of the shelving points of cargo ships and oil tankers. First, the distribution of the trajectory density of the cargo ships' shelving points is shown in Fig. 9, Fig. 10 and Fig. 11. When the speed distribution is 0–1 kn, it is found from these three figures that the shelving point of each length presents a scattered distribution. Among them, 0–50 m and 50–100 m cargo ships are mainly anchored near Taepokou. The 100–150 m cargo ships are mainly anchored near the Yongjiang waterway.

Fig. 9. Track density distribution of lay down points of 0-50m long cargo ship

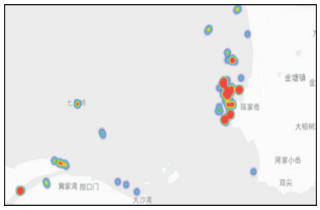

Fig. 10. Track density distribution of lay down points of 50-100m long cargo ship

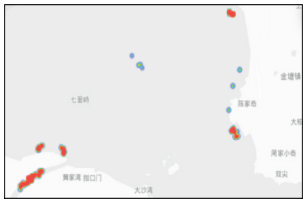

Fig. 11. Track density distribution of lay down points of 100-150m long cargo ship

Fig. 12 and Fig. 13 show the distribution of the trajectory density of the shelving points of oil tankers.

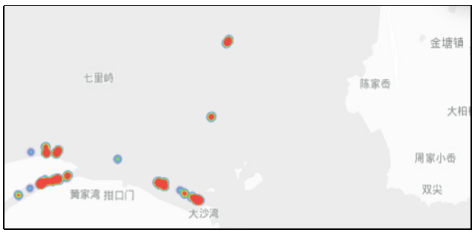

Fig. 12. Track density distribution of lay down points of 50-100m oil tankers

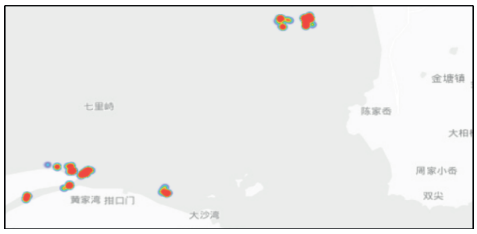

Fig. 13. Track density distribution of lay down points of 100-150m oil tankers

When the speed distribution is 0–1 kn, it can be seen from the above figures that the shelving points of each ship length also present a scattered distribution. As tankers with a length of 0–50 m do not have a shelving point, they are not shown. The 50–100 m and 100–150 m vessels are mainly anchored near Zhenhai, Beilun and the Yongjiang waterway. Visual analysis of the trajectory density of the movement points of cargo ships and oil tankers is carried out. The motion point trajectory density distribution of the cargo ships is shown in Fig. 14, 15 and 16.

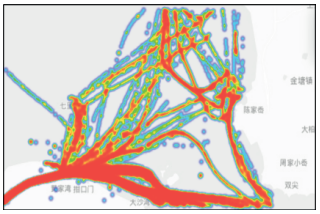

Fig. 14. Trajectory density distribution of 0-50m cargo ship

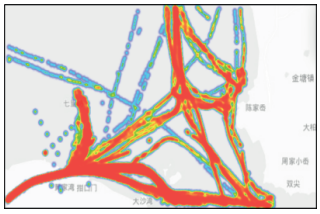

Fig. 15. Trajectory density distribution of 50-100m cargo ship

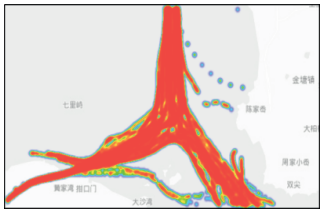

Fig. 16. Trajectory density distribution of 100-150m cargo ship

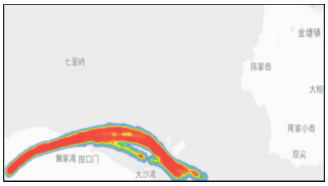

Fig.17. Trajectory density distribution of 0-50m oil tanker

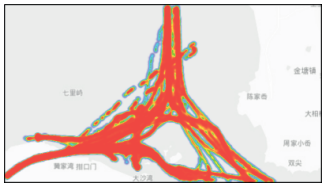

Fig. 18. Trajectory density distribution of 50-100m oil tanker

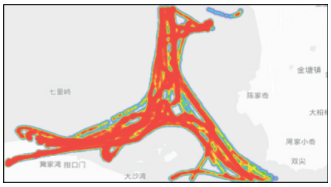

Fig. 19. Trajectory density distribution of 100-150m oil tanker

It can be seen from Figs. 14, 15 and 16 that the movement track of the 0–50 m ships is very similar to that of the 50–100 m ships, but the movement track distribution is chaotic. Comparatively speaking, ships with a length of 100–150 m have the most concentrated and regular movement tracks. Next, the trajectory density of the movement points of oil tankers was analysed visually as shown in Fig. 17, 18 and 19.

It can be seen from Figs. 17, 18 and 19 that the movement points of the tankers with a length of 0-50 m are distributed in a concentrated way, mainly in the channel near Beilun. The trajectory distribution of 50–100 m and 100–150 m length tankers is similar and relatively regular. However, from the comprehensive consideration of the track number, track density and track regularity, 100–150 m long cargo ships are more dominant and more representative in the study area. To sum up, in order to avoid the poor display of the ship trajectory clustering results, the cargo ships with a length of 100–150 m are selected as the research objects of this clustering experiment. These research objects are representative, and the clustering results can be used as a reference for future channel planning research of this study area.

CLUSTERING ANALYSIS

By 1 June 2017 solstice 4 JCP 4, a total of 221217 AIS data after decoding and data pre-processing, and removal of incorrect AIS data we selected the research object. It was ships of lengths of 100–150 m with the set speed threshold of 0–1 kn. After extracting the movement track points, a total of 40147 AIS data were used. The daily AIS data statistics are as shown in Table 3.

The ship shelving points and movement points are treated differently. According to the navigation speed threshold of 1 kn, the ship shelving point or the ship movement point in the water area of interest in this study is regarded as the critical value. Collision hazards are more likely to occur when the ship is moving, and the damage caused by an impact at speed is greater. Therefore, compared with the shelving point, analysis of the ship's movement point is more important. Therefore, trajectory clustering analysis is mainly carried

out for ship movement points. In this study, there are 511 AIS data of ship shelving points and 36,719 AIS data of ship movement points in total. The trajectory data sets of each ship trajectory cluster were obtained by inputting the optimal parameters obtained from multiple experiments through the ship trajectory cluster model (eps =0.015, minPts=45, DirRange=1.5, SpdRange=2), and the trajectory data sets of each ship trajectory cluster were visualised through the BDP visualisation platform, distinguished in different colours.

Tab. 3. AIS data statistics table

Time	June 1	June 2	June 3	June 4
AIS numbers	10420	9380	10693	9654

The traffic flow in this water area involves an intersection of three directions, which resembles the shape of a "man", and the traffic flow at the middle crossing is relatively complex. This paper simplifies the complex problem, divides the whole traffic flow into three data sets, and conducts clustering for each segment, each with its own route segment characteristics, as shown in Fig. 20.

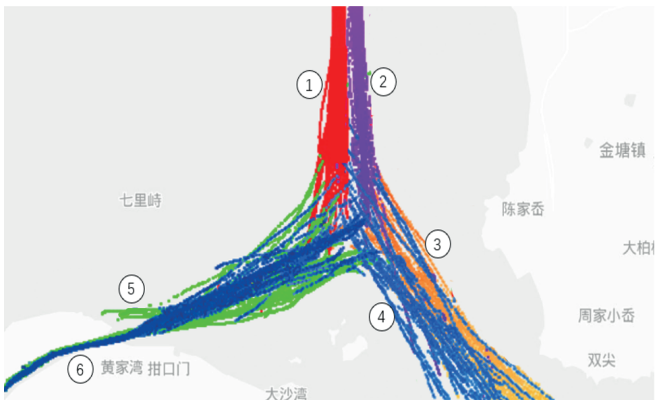

Fig. 20. Results of ship trajectory clustering

A total of 6 ship track class clusters were obtained, represented in different colours and numbered 1-6 respectively. The main channel of the bridge in and out of the Jintang Water area has been separated and is navigable, while the remaining sections have not been set up yet. Based on the analysis of the experimental results of ship trajectory clustering, ship trajectory clusters 1 and 2 of the main channel are found. Due to the implementation of the lane separation and navigation system in this water area, the ship trajectory cluster is relatively regular. All the ships passing through this area can navigate well within the established channel range, and the ship trajectory clustering results are consistent with the actual traffic flow.

Ship trajectory clusters 3 and 4 respectively represent the clusters entering and leaving the waters near Tapukou. Since the traffic separation system has not been implemented in this area, as shown in Fig. 18, some ship trajectories will deviate from the usual navigation tracks of most ships. It is found from the result of ship track clustering that the implementation of a traffic separation system is of positive significance to the regulation of ships entering and leaving ports.

Trajectory cluster 5 represents the cluster in the direction of Zhenhai. A small number of ships enter the Yongjiang Channel. Trajectory cluster 6 indicates that the river flows out of Yongjiang in other directions.

Based on the main channel clustering results (track clusters 1 and 2) in the water area of Jintang Bridge, the speed and course frequency of each track cluster are statistically analysed. The results are shown in Fig. 21 and Fig. 22.

It can be seen from Figs. 23 and 24 that, in track cluster 1, the average speed is 9.67 kn, and the interval with the most frequent occurrence of speed value is [5.0,7.7], [7.7,10.4], [10.4,13.1], which is similar to the southbound track cluster, with a large range of speed variation. Ships belonging to this track cluster also have significant speed reductions or increases while moving. Due to the particularity of the traffic flow direction in this water area, it is shaped like a "man". After a section of direct travel, ships often need to turn sharply, which has a great impact on their speed. The average course is 355.76°. The maximum range of course is from 154° to 189°, among which the ranges with the highest frequency of course values are [349,355], [355,0] and [0,5], which are in line with the actual direction of traffic flow.

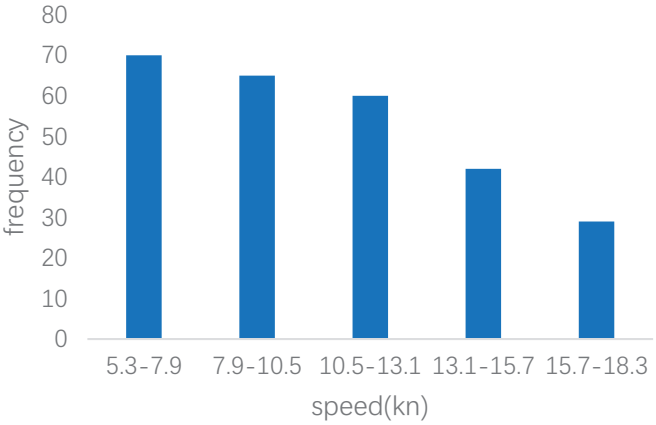

Fig. 21. Speed frequency distribution of trajectory class cluster 1

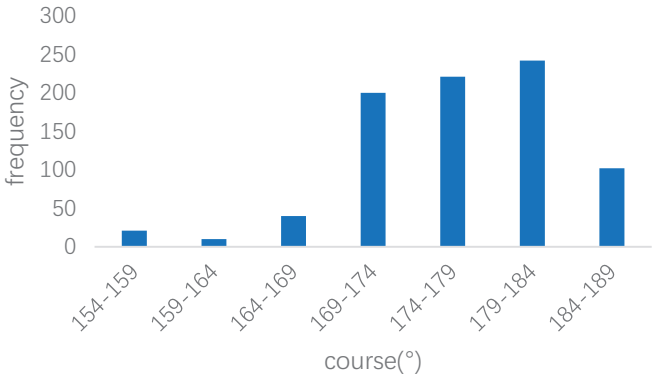

Fig. 22. Course frequency of trajectory class cluster 1

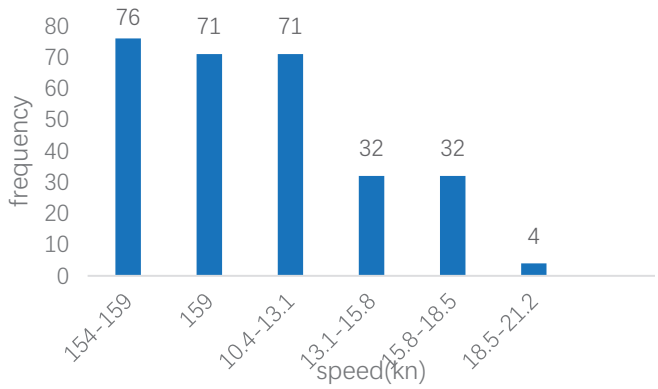

Fig. 23. Speed frequency distribution of trajectory class cluster 2

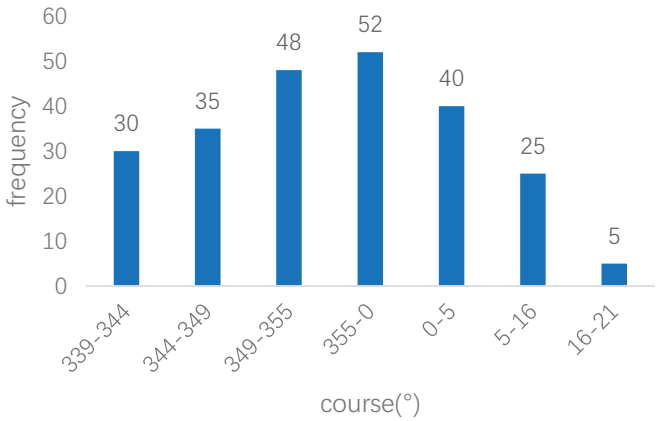

Fig. 24. Course frequency of trajectory class cluster 2

To sum up, the average speeds of ships heading south and north do not differ much, at 8.12 kn and 9.67 kn respectively and the average speed of ships heading north will be greater. The average heading of ships heading south and north is approximately opposite, which is 179.39° and 355.76° respectively. The analysis of the other four track clusters is the same as above. The ship navigation profile of each track cluster can be obtained by analysing the distribution of the speed and course frequency of each track cluster.

REPRESENTATIVE TRAJECTORY EXTRACTION

In order to further verify the reliability and accuracy of the multi-attribute DBSCAN algorithm, representative trajectories of class clusters are extracted for fitting. Each ship trajectory cluster has three attributes: position, speed and heading. Combined with the characteristics of vectors, a concept similar to representing vector points is proposed to compress the number of track points in the trajectory cluster. Existing studies have introduced the clustering method based on the trajectory line. In this paper, the method is the point-based clustering algorithm, so the output result of the algorithm is a collection of multiple trajectory points. The process is to calculate the average direction of all trajectory points of each trajectory cluster. In this direction, the trajectory cluster is divided into multiple parallel blocks, using the interval as the normal line. The width is determined based on domain knowledge or the results of multiple experiments. Finally, the parameter eps value used in the clustering algorithm is 0.15 as the block length. Each parallel block is calculated to represent the vector points, in which Avg(LAT) and Avg(LON), the central positions of all locus points, are taken as the starting points of the representative vector points. The average speed Avg(SOG) of all trajectory points is used as the magnitude of vector points. Avg(COG), the average course of all track points, is used as the direction to represent the vector points, and the three dimensions (centre position, average speed and average heading) together constitute the vector track points. After calculating the representative vector points of all parallel blocks, all the representative vector points are connected to form the representative trajectory, as shown in Figs. 23 to 34. In this paper, only the trajectory fitting to the represented trajectory cluster 1 and 2 is listed, and the remaining trajectories are similar to these.

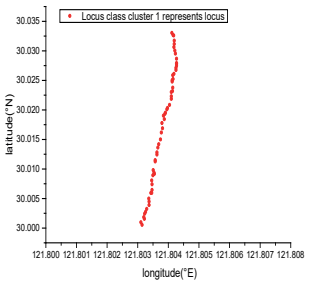

Fig. 23. Class cluster 1 represents the trajectory

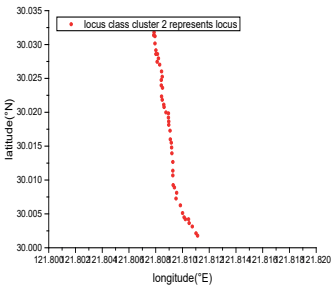

Fig. 24. Cluster 1 represents trajectory
Fig. 24. Cluster 2 represents trajectory

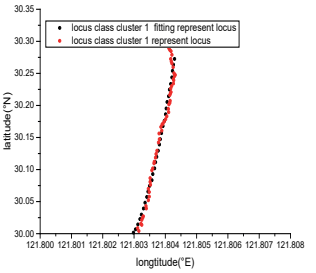

Fig. 25. Cluster 1 northbound of main channel represents trajectory fitting

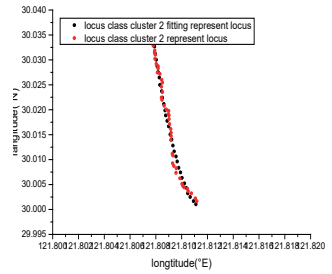

Fig. 26. Cluster 2 main channel southward represents trajectory fitting

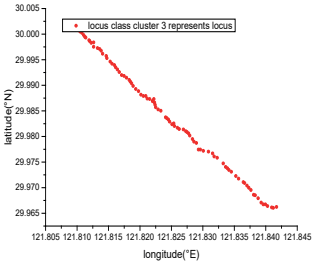

Fig. 27. Cluster 3 represents the trajectory

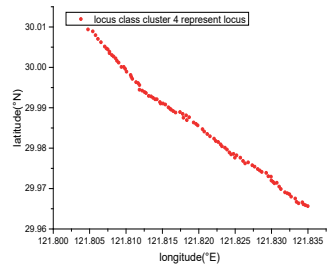

Fig. 28. Cluster 4 represents the trajectory

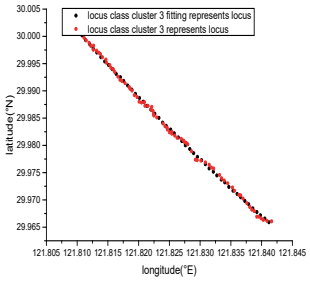

Fig. 29. Cluster 3 Dapu estuary represents the water area fitting

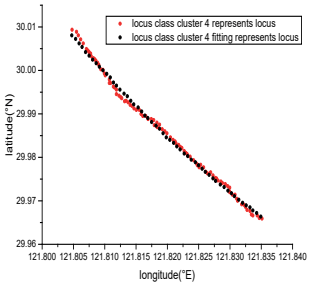

Fig. 30. Cluster 4 Dapokou represent the water area

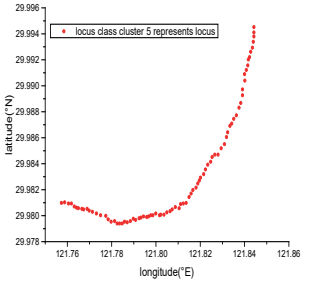

Fig. 31. Class cluster 5 represents the trajectory

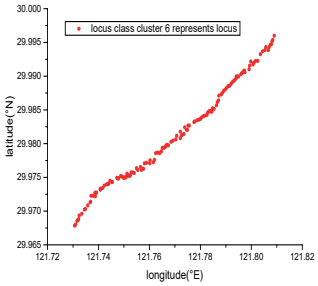

Fig. 32. Class cluster 6 represents the trajectory

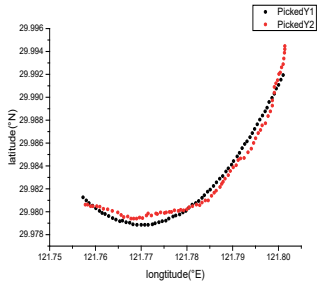

Fig. 33. Cluster 5 represents trajectory fitting

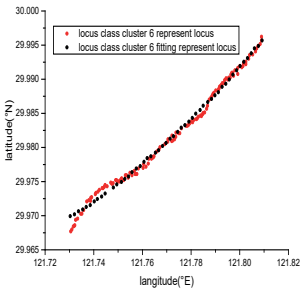

Fig. 34. Curve fitting of class cluster 6 out of Town harbor

RESULTS

From 40147 AIS data, 579 effective representative trajectory points were extracted to form 6 representative trajectory lines, accounting for about 1.44% of the total points.

The locus cluster 1 represents the starting position of the locus (30.033285°N, 121.804102°E) and the ending position (29.995572°N, 121.80227°E). At the same altitude, the distance between two points in geographical space is 2.266 nautical miles, and the azimuth angle formed between two points is 182.4°. By means of the regression model, the fitting degree of the original representative trajectory was calculated to be about 96.37%.

The locus class cluster 2 represents the starting position of the locus (30.00184°N, 121.811125°E) and the ending position is (30.032807°N, 121.8078°E). At the same altitude, the distance between two points in geographical space is 1.869 nautical miles, and the azimuth angle formed between two points is 349.1°. By means of the regression model, the original trajectory fitting degree is about 97.62%.

The locus class cluster 3 represents the starting position of the locus (30.001808°N, 121.809547°E) and the ending position (29.966182°N, 121.841095°E). At the same altitude, the distance between two points in geographical space is 2.696 nautical miles, and the azimuth angle formed between the two points is 142.5°. By means of the regression model, the original trajectory fitting degree is about 99.89%.

The locus class cluster 4 represents the starting position of the locus (29.965868°N, 121.835008°E) and the ending position (30.009402°N, 121.804738°E). At the same altitude, the distance between the two points in geographical space is 3.052 nautical miles, and the azimuth angle formed between the two points is 328.9°. By means of the regression model, the original trajectory fitting degree is about 99.55%.

The locus class cluster 5 represents the locus starting position of (29.994542°N, 121.801375°E), and the ending position of (29.980973°N, 121.75394°E). At the same altitude, the distance between two points in geographical space is 2.599 nautical miles, and the azimuth angle formed between the two points is 251.7°. By means of the regression model, the original trajectory fitting degree is about 97.12%.

The locus class cluster 6 represents the locus whose starting position is (29.967733°N, 121.730498°E) and ending position is (29.996267°N, 121.8094°E). At the same altitude, the distance between the two points in geographical space is 4.448 nautical miles, and the azimuth angle formed between the two points is 67.3°. By means of the regression model, the original trajectory fitting degree is about 99.44%.

To sum up, after calculation by the polynomial regression equation, the trajectory fitting degree of the 6 trajectory class clusters is respectively 96.37%, 97.62%, 99.89%, 99.55%, 97.12%, 99.44%, and the average fitting degree is 98.33%. The closer the value of R is to 1, the better the fitting degree of the regression equation for the observed value is, which indicates that the fitting degree of the ship representative tracks extracted is at a high level, and also reflects the feasibility and effectiveness of the ship representative track extraction algorithm. It is also proved that the multi-attribute DBSCAN extension algorithm has a high degree of fit. The representative trajectory results of the six ship trajectory clusters are shown in Figs. 26. The experimental results show that the clustering operation conducted by the multi-attribute DBSCAN extended algorithm has a very high degree of fitting.

CONCLUSION

The density-based clustering method is extended, and the multi-attribute DBSCAN extension algorithm suitable for ship trajectory clustering is introduced to establish an effective ship trajectory clustering model, after which the ship trajectories are analysed by using the ship automatic identification system AIS data. Combined with the idea of representing vector points, the representative trajectories of the ship trajectory class clusters are extracted, which proves that the multi-attribute DBSCAN algorithm has reliability and accuracy.

1. After decoding and pre-processing the AIS data, they are divided according to different ship types and lengths. Through

visual analysis of the ship trajectory density distribution, the type of ship object to be studied is determined.

2. A multi-attribute DBSCAN extension algorithm is proposed. The original DBSCAN algorithm is extended by combining the three attributes of ship position, speed and heading in the AIS data, and a multi-attribute DBSCAN extension algorithm is established. The ship trajectory clustering model is used to carry out experiments in the waters of the Jintang Bridge area. Ship trajectory clustering is carried out for ships with a length of 100–150 meters in this water area. The trajectory clustering results show that six ship trajectory clusters are obtained.

3. Corresponding to the actual ship traffic flow pattern in this water area, the feasibility of the multi-attribute DBSCAN extension algorithm is verified. The representative trajectory extraction algorithm was proposed, which combined the idea of representing vector points to extract the representative trajectories of six ship trajectory class clusters. The polynomial regression model was used to calculate the similarity fitting degree of each representative trajectory, verify the feasibility of the representative trajectory extraction algorithm, and prove the reliability and accuracy of the multi-attribute DBSCAN algorithm.

4. Using the ideas of representative vector points, the trajectory extraction algorithm is put forward, The average direction of all trajectory points of each trajectory class cluster is calculated. In this direction, the trajectory class cluster is divided into several parallel blocks with the width of λ, taking the interval λ as the normal. The width is determined by domain knowledge or the results of many experiments. Finally, the parameter EPS value of 0.15 used in the clustering algorithm is selected as the length of λ. Then, the representative vector points are calculated for each parallel block, and the center positions AVG (LAT) and AVG (lon) of all trajectory points in the parallel block are taken as the starting points of the representative vector points; the average speed AVG (SOG) of all trajectory points is taken as the size of the representative vector points; the average direction AVG (COG) of all trajectory points is taken as the direction of the representative vector points, with three dimensions (center position, average speed and average speed) The vector trajectory points are composed of two parts. After calculating the representative vector points of all parallel blocks, connect all the representative vector points to form the representative trajectory

5. Compared with the traditional DBSCAN algorithm, the multi-attribute DBSCAN extension algorithm proposed in this paper simplifies the originally complex trajectory segment to the similarity measure between vector points. In order to better conform to the complex marine traffic flow situation in reality, two control variables, SpdRange and DirRange, are added to the input variables of the model to control the range of ship speed and heading range, which is more in line with the environmental changes of sea navigation and makes the clustering accuracy higher.

6. Although the algorithm is not sensitive to noise, clustering of arbitrary shapes can be found, but the result of clustering has a good relationship with the parameters. DBSCAN uses fixed parameters to identify clustering, but when the sparsity of clustering is different, the same criteria may destroy the natural structure of the clustering; that is, sparse clustering will be divided into multiple clusters or dense and close clusters will be merged into one cluster.

REFERENCES

1. Yang, B. International standard of automatic ship identification system and its formulation. Standardization of Transportation, 2002 (01):42-44.
2. Zhao, L., Shi, G. A method for simplifying ship trajectory based on improved Douglas-Peucker algorithm. Ocean Engineering , Vol. 166, 2018: 37-46.
3. Zhao, L., Shi, G., Yang, J. Ship trajectories pre-processing based on AIS data. Journal of Navigation, 2018.
4. Wang, J., Zhu, C., Zhou, Y., Zhang, W. Vessel Spatio-temporal Knowledge Discovery with AIS Trajectories Using Co-clustering. Journal of Navigation, Nov 2017.
5. Zhang, Y., Shi, G., Li, S., Zhang, S. Vessel trajectory online multi-dimensional simplification algorithm. Journal of Navigation, 2020.
6. Zhang, L., Meng, Q., Xiao, Z., Fu, X. A novel ship trajectory reconstruction approach using AIS data. Ocean Engineering, 2018.
7. Yan, Z., Xiao, Y., Cheng, L., He, R., Ruan, X., Zhou, X., Li, M., Bin, R. Exploring AIS data for intelligent maritime routes extraction. Applied Ocean Research, 2020.
8. Wei, Z., Xie, X., Zhang, X. AIS trajectory simplification algorithm considering ship behaviours. Ocean Engineering, 2020.
9. Li, H., Liu, J., Wu, K., Yang, Z., Liu, R.W., Xiong, N. Spatio-temporal vessel trajectory clustering based on data mapping and density. IEEE Access, 2018, 6:58939-58954.
10. Sheng, P., Yin, J. Extracting shipping route patterns by trajectory clustering model based on AIS data. Sustainability, 2018, 10(2018):2327.
11. [11]Xiao, X., Shao, Z., Pan, J., Ji, X. Ship trajectory clustering model based on AIS information and its application. China Navigation, 2015, 38(02):82-86.
12. Zhou, H., Chen, Y., Chen, L. Clustering analysis and application of ship trajectory. Computer Simulation, 2020, 37(10):113-118+199.

13. Cui, K. Research on ship AIS trajectories prediction method based on machine learning. Zhengzhou University, 2020.

14. Liu, Y., Shi, B. Research on ship track clustering technology based on skeleton extraction. Information Technology, 2020, 44(03):50-53+58.

15. Jiang, Y., Xiong, Z., Tang, J. Ship trajectory clustering algorithm based on trajectory segment DBSCAN. China Navigation, 2019, 42(03):1-5.

16. Zhao, L., Shi, G., Yang, J. Adaptive hierarchical clustering of ship trajectory based on DBSCAN algorithm. China Navigation, 2018, 41(03):53-58.

17. Peng, X., Gao, S., Chu, X., He, Y., Lu, C. Clustering method of ship trajectory based on Spark. China Navigation, 2017, 40(03):49-53+68

18. Frey, B.J., Dueck, D. Clustering by passing message between data points. Science, 2007, 315(5814):972-976.

19. Jain, A.K., Dubes, R.C. Algorithms for clustering data. Technometrics, 2015, 32(32):227-229.

20. Yang, W., Long, H., Shao, Y., Du, Q. Research on density calculation based on Chebyshev distance and clustering method of K-means. Communications Technology, 2019, 52(04):833-838.

21. Chen, B. Research on spatio-temporal similarity of vehicle trajectories based on clustering algorithm. Fujian Normal University, 2015.

22. Design and implementation of ship trajectory clustering system based on bright [AIS]. Dalian University, 2016

CONTACT WITH THE AUTHORS

Deqaing Cui

e-mail: cuideqiang0213@163.com
Shanghai Maritime University,
Haigang streat, 201306 ShangHai,
CHINA

Xiaofeng Xu

e-mail: xfxu@shmtu.edu.cn
Shanghai Maritime University,
Haigang streat, 201306 ShangHai,
CHINA

Yun Li

e-mail: liyun@shmtu.edu.cn
Shanghai Maritime University,
Haigang streat, 201306 ShangHai,
CHINA

Yingjie Xiaoi

e-mail: xiaoyj@shmtu.edu.cn
Shanghai Maritime University,
Haigang streat, 201306 ShangHai,
CHINA