EURASIP Journal on Wireless Communications and Networking

RESEARCH

Open Access

A novel high-dimensional trajectories construction network based on multi-clustering algorithm

Feiyang Ren[1], Yi Han[2*], Shaohan Wang[2] and He Jiang[2]

*Correspondence: hy19970330@163.com
[2] COSCO SHIPPING Technology Co., Ltd., Shanghai, China
Full list of author information is available at the end of the article

Abstract

A multiple clustering algorithm based on high-dimensional automatic identification system (AIS) data is proposed to extract the important waypoints in the ship's navigation trajectory based on selected AIS attribute features and construct a route network using the waypoints. The algorithm improves the accuracy of route network planning by using the latitude and longitude of the historical voyage trajectory and the heading to the ground. Unlike the navigation clustering method that only uses ship latitude and longitude coordinates, the algorithm first calculates the major waypoints using Clustering in QUEst (CLIQUE) and Balance Iterative Reducing and Clustering Using Hierarchies (BIRCH) algorithms, and then builds the route network using network construction. Under the common PC specification (i5 processor), this algorithm forms 440 major waypoints from 220,133 AIS data and constructs a route network with directional features in 5 min, which is faster in computing speed and more suitable for complex ship trajectory differentiation and can extend the application boundary of ship route planning.

Keywords: Marine trajectories, High-dimensional data analysis, Multi-clustering algorithm, Machine learning, Data mining

1 Introduction

Maritime transportation plays an important role in the global economy, with more than 80% of the trading network occurring by sea. Thus, route planning is considered the main task faced by all crews and shipping companies. The quality of route planning is closely related to the safety and economic efficiency of ship navigation. Although researches have been conducted on ship navigation safety, maritime traffic accidents are still happening. In ship sailing, route planning is mainly done by the subjective judgment of experienced crew members. In other words, traditional route planning relies on experience and judgment, which has a high error rate. According to the Naus 2020 study [1], over 70% of ship collisions are related to the highly subjective judgmental operation of mariners. With the development of computational power and the growth of the internet of things (IOT), the use of data analysis along with practical path planning has become possible.

Data clustering is considered as the main method for dividing a huge amount of data into groups for more precise analysis. Cluster analysis is the grouping or clustering of data according to the inherent similarities and characteristics between the data [2–4]. The clustering result may show the target ship's trajectories and traffic volume distribution [5, 6]. As a standard data mining method, ship trajectory clustering integrates AIS data of different ships into different categories. It is beneficial for shipping companies and maritime authorities to understand marine traffic's operational status and characteristics. Density, graph, partition, and hierarchical-based clustering algorithms are often used for ship trajectory clustering. K-means clustering, a representative of partitioning-based clustering methods, has been widely used in related research for its simplicity and efficiency. Song [7] designed an improved K-means trajectory clustering method based on suburban curve fitting to get the traffic flow parameters of each direction and category at intersections. However, since this method only considers the case of smooth traffic, it cannot explain vehicle trajectory with fault discontinuity in the scenario of a complex traffic situation. Wang and Bai [8] used the Min–Max K-mean clustering error method to modify the global K-means algorithm to overcome the undesirable effects at initialization. Han [9] proposed an online learning model combining K-means clustering and gated recurrent unit (GRU) neural network for trajectory prediction. Tyagi and Trivedi [10] proposed a hybrid K-means algorithm to obtain clustering results for color images and refined the clustering results using the ant colony optimization (ACO) algorithm. Jiang [11] proposed an identification scheme for classifying and monitoring moving targets on sea based on structural database techniques and K-means. However, this method is sensitive to data noise and cluster center, which is less effective for noisy data. Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is a representative method of density clustering. Density clustering starts from the perspective of sample density and checks the connectivity between samples, and continuously extends the clusters based on the connectable samples to obtain the final clustering results. In 2017, Zhao [12] proposed a parameter determination DBSCAN algorithm based on statistical methods for trajectory clustering in waters with uneven distribution of ship trajectories. Yet, only the applicability in simple cases was considered. In 2019, Zhao [5] proposed a DP (Douglas-Peucker) compression and density-based trajectory clustering method for marine traffic pattern recognition on previous research and evaluated and compared a large number of ship navigation trajectories in Beilun-Zhoushan port, China. Wang [13, 14] proposed a ship trajectory clustering algorithm based on the hierarchical density of noise application space clustering based on Zhao's research. The research on trajectory clustering did not solve the problem of route planning, though it attempted to improve the clustering effect continuously from the perspective of optimization.

To better understand the ship navigation information, a maritime route network needs to be extracted from the ship's historical voyage trajectory, through which the network can help the relevant personnel to carry out route planning. In 2016, Dobrkovic [15] proposed for the first time the use of genetic algorithms to extract maritime traffic networks from AIS data. To enable long-term forecasting and planning of ship routes, in 2018, Dobrkovic combined quadratic trees and genetic algorithms to construct a maritime route network inclusive of incomplete and noisy AIS data [16]. Filipiak [17] pointed out the poor computational performance in Dobrkovic's study and proposed a parallel genetic algorithm combined with KD-B trees to extract the maritime route network from AIS data. Ni [18, 19] proposed an improved genetic algorithm for ship path planning that compensates for the inherent deficiencies of local optimization in order to achieve a balance between the local and global optimization capabilities of genetic algorithms in ship paths. Wang [20, 21] proposed a quadratic optimization genetic algorithm incorporating ship motion characteristics to aid automatic route planning in complex environments. Zhao [22] proposed a hybrid multi-iterative route planning method based on an improved particle swarm optimization-genetic algorithm, aiming to optimize the ship-related meteorological risks, fuel consumption, and navigation time, and to improve the diversity of route planning; However, only the effects of wind, waves, and anti-navigation on the ship were considered, while the effects of other maritime vessels on the ship were ignored. Chen [23] combined fuzzy control and genetic algorithm with building a route planning system for underwater vehicles, which can provide strong robustness. Route planning algorithm is an aspect of an unmanned ground vehicle obstacle avoidance system. Liu [24] proposed an improved A-Star algorithm for ship path planning that integrated route length, obstacle dynamics, navigation rules, and maneuverability constraints. In particular, the currents of the ocean were considered in the algorithm. Unfortunately, the precise maneuvering characteristics of the ship were not used. Sun [25] used fuzzy neural networks for scheduling the ship's path in complex navigation tasks. Also, fuzzy logic was used to process statistical data and neural networks optimize navigation routes. Proportion Integral Differential (PID) method was introduced in the decision system to ensure the stability of the decision system.

Though previous research has shown some solid results in the analysis of navigation history data and path predictions, those methods still lack connections with real world scenarios. In the first place, some studies find the major waypoints manually, which is highly subjective and error-prone, especially for the complex open water environment. In addition, those previous research calculated the ship's direction by only using the longitude and latitude information, which costs much computational power and sometimes can be mistaken. Most importantly, when performing trajectories clustering, AIS coordinate information is often used as input for better classification; however, with only longitude and latitude information, the output from those clustering processes can only generate results from a mathematical or statical perspective, while its practical performance can hardly be evaluated.

This paper proposes a new multi-level clustering algorithm based on high-dimensional ship AIS data to find the major waypoints on the ship trajectory and provides a basis for later ship navigation environment analysis.

2 Methodology

2.1 Methodological overview

The proposed method analyzes ship trajectories from high dimensions by automatically clustering paths with multi-clustering algorithms and shipping network reconstructions. Firstly, data pre-processing is performed by removing abnormal AIS data for noise cancellation purposes. Then AIS data are further processed in two steps: trajectory trimming and trajectory compression. Secondly, CLIQUE-BIRCH algorithm is used for trajectory clustering and waypoints discovery of AIS data, and clustering performance metrics are proposed to judge the method's performance. Finally, a newly proposed network construction method is used to connect the waypoints and construct a sea route network, and the constructed route network is evaluated with examples. Figure 1 gives the methodological overview of the research.

2.2 AIS data preprocessing

The presence of forwarding, loss, and data errors in AIS data can lead to many anomalies in ship trajectories built directly from AIS data. Therefore, AIS data needs to be pre-processed before building the model. Since the navigation situation of ships in and around ports is more complicated than that in open sea, and there are berthing, anchoring and wandering situations of ships, and the density of ships is higher, the navigation situation of ships in and around ports needs to be studied separately. On the other hand, since low-speed ships will affect the navigation efficiency, and the goal of the study is to build an efficient route network to assist in route planning, removing the data of low-speed ships will help improve the ship navigation efficiency, reduce the navigation cost and the navigation safety risk. In summary, three methods will be used to improve the quality of AIS data. First, according to the relevant ship management experience, the AIS data points will be matched with the ports and the data less than 185.2 km (100 nautical miles) from the ports will be eliminated; second, the data of ships sailing at less than 7 knots will be labeled and eliminated; finally, in order to reduce the total amount of AIS data studied, the Douglas-Peuker (DP) algorithm will be used for trajectory compression. The DP algorithm [5, 26, 27] has been widely used to remove redundant AIS data points from ship trajectories, while still preserving the original ship route shape characteristics, and the study does not lose generality.

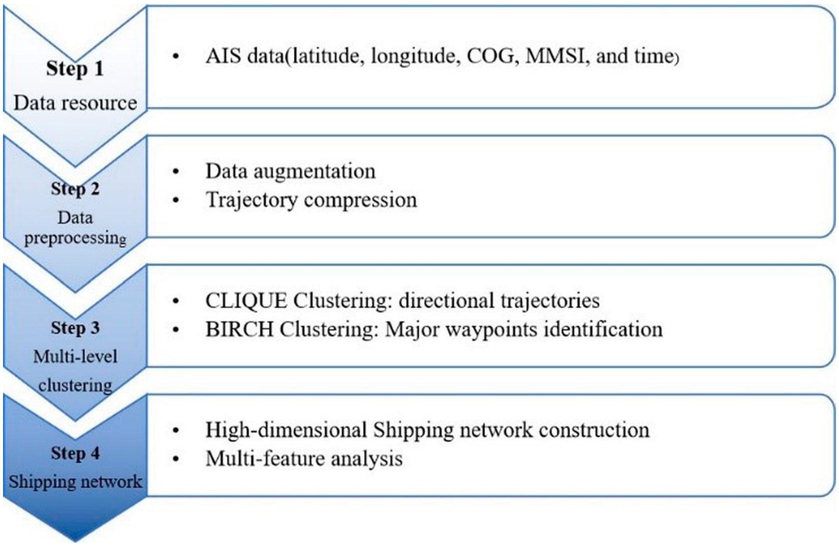

Step 1
Data resource
AIS data(latitude, longitude, COG, MMSI, and time)
Step 2
Data preprocessing
Data augmentation
Trajectory compression
Step 3
Multi-level clustering
CLIQUE Clustering: directional trajectories
BIRCH Clustering: Major waypoints identification
Step 4
Shipping network
High-dimensional Shipping network construction
Multi-feature analysis

Fig. 1 Methodological overview of the research

2.3 CLIQUE-BIRCH

Since the previous use of AIS data in trajectory clustering algorithms only contains latitude and longitude information, lacking consideration of other attributes, much information is often lost in the clustering results. At the same time, when introducing new attributes, it is necessary to calculate the values such as direction attributes from latitude and longitude information. For example, information such as draught, weather, and fuel consumption cannot be expressed by latitude and longitude, so the traditional algorithm fails to consider them, and the clustering results naturally cannot help relevant departments to make efficient route planning.

Therefore, a novel multi-level clustering algorithm network based on high-dimensional AIS data is proposed. First, the latitude, longitude, and Course over Ground (COG) from the AIS data points are input into the CLIQUE algorithm to cluster the navigation trajectories with directional features. The CLIQUE algorithm can efficiently handle high-dimensional data by automatically discovering the highest-dimensional subspaces in which high-density clustering exists. It is insensitive to the order of input tuples without assuming any canonical data distribution, and scales linearly with the size of the input data, and has good scalability when the dimensionality of the data increases [28]. Second, using BIRCH algorithm to find and generate waypoints on the identified navigation trajectory automatically. The algorithm can effectively identify the noise points and quickly cluster the clustered AIS data of the navigation track to efficiently identify the waypoints on the navigation tracks [29]. The identified waypoints add directional information compared to the waypoints obtained by the conventional method.

2.4 Network construction

After the waypoints are refined, a complete route network needs to be constructed from these waypoints. In previous studies, little attention was paid to the construction of route networks. Since waypoints are extracted from numerous AIS data, there must be some connection between waypoints and AIS data that can help to construct route networks. Therefore, a network construction method based on the connection between AIS data and waypoints is proposed. First, each waypoint will be the center of a circle, and a radius of size r will be set whenever necessary. The AIS data within each circle will be marked with correlation to that waypoint. Then, the route trajectory is extracted from the historical AIS data to traverse all the waypoints to build the circle, and the waypoints are connected in the order of connection to build the complete route network.

3 Model design

3.1 Definition of ship trajectory

Using MMSI to distinguish different ship trajectories, the ship's trajectory can be described by $\mathrm{Trajectory} = \{\mathrm{ship}_i | \mathrm{ship}_i, i = 1, 2, \ldots, m\}$, where ship_i is the trajectory of ship i and m is the number of ships, and ship_i is defined in (1):

$$\mathrm{Ship}_i = \left\{ p_i^k \middle| p_i^k = \left(\mathrm{MMSI}_i, \mathrm{lat}_i^k, \mathrm{lon}_i^k, T_i^k \right), k = 1, 2, \ldots, n \right\} \tag{1}$$

where k is the sequence number of AIS data points in each trajectory, n is the total number of AIS data points in each trajectory, p_i^k is the state vector of the kth AIS data point of the ith ship, and lat_i^k and lon_i^k are the coordinates of ship i at T_i^k.

3.2 AIS data preprocessing

The purpose of this step is to pre-process the AIS data. The first step is to prune the AIS data. The AIS data points are matched with the port information. The AIS data points near the port are removed, and the distance between the port warp points and the AIS data points is calculated using the Haversine formula:

$$\mathrm{hav}(\Theta) = \mathrm{hav}(\varphi_2 - \varphi_1) + \cos(\varphi_1)\cos(\varphi_2)\mathrm{hav}(\lambda_2 - \lambda_1) \tag{2}$$

with

$$\Theta = d/R \tag{3}$$

where φ_1 and φ_2 are dimensions, λ_1 and λ_2 are longitudes, d is the distance between the two places and R is the radius of the Earth.

The second step is to compress the trimmed AIS data by using the DP algorithm to improve the clustering efficiency without losing shape features. The steps of DP algorithm are as follows: for the trajectory composed of many AIS data points, the first step is to set the distance threshold D. The second step is to connect the first and last points of the trajectory into a straight line, find spot the vertical distance from all AIS data points on the trajectory to the straight line, and find the maximum distance $d_{\max}$; the second part uses $d_{\max}$ to compare with the pre-given threshold D. If $d_{\max} < D$, then all the middle points on this trajectory will be discarded, and take the straight line section as the approximation of the trajectory, and the processing of this section of the trajectory is finished; the third step, if $d_{\max} > D$, keep the AIS data point corresponding to $d_{\max}$, and use this store as the boundary to divide the trajectory into two parts, and repeat the method for these two parts, that is, repeat the second and third steps until all $d_{\max}$ is smaller than D, when the compression of the trajectory is finished. Figure 2 shows the process of compression of trajectories by the DP algorithm.

Obviously, the compression effect of DP algorithm is related to the threshold value, the higher the threshold value, the greater the compression degree, the more the AIS data points are reduced; Conversely, the lower the compression degree, the more the AIS data points are retained and the shape tends to be closer to the original trajectory.

3.3 CLIQUE clustering

AIS data are multidimensional data containing many different attributes, but traditional trajectory clustering methods often consider only latitude and longitude, or when introducing attributes such as direction, they need to be calculated with the help of latitude and longitude, which tends to cause a waste of computational resources; there may also be a situation in which the calculated direction is not consistent with the actual direction. Therefore, the CLIQUE algorithm was proposed to solve the above problems. CLIQUE clustering was proposed by Agrawal et al. [30] as a grid-based clustering method for discovering density-based clusters in a subspace. CLIQUE has the advantage

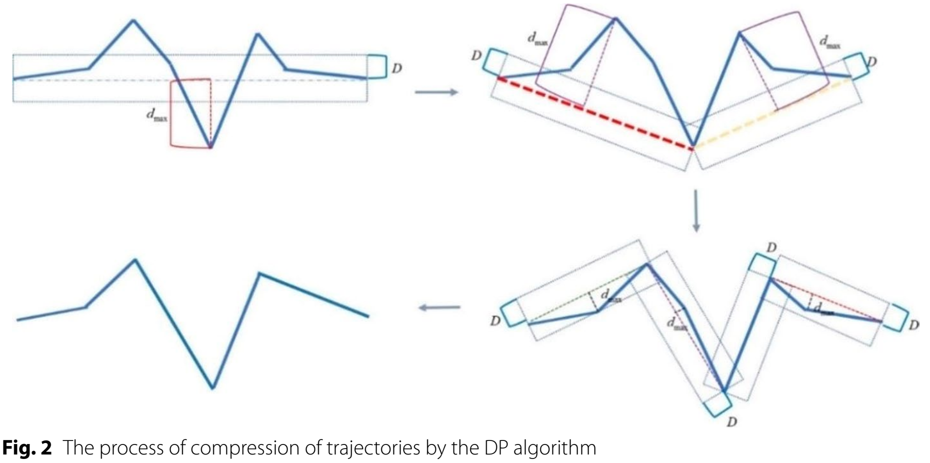
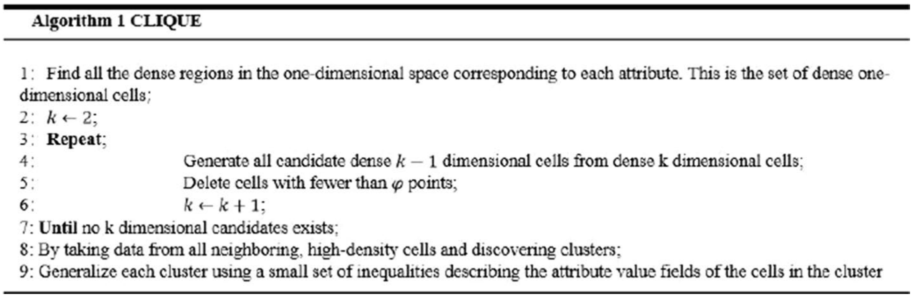


Fig. 2 The process of compression of trajectories by the DP algorithm

of efficient grid clustering, is insensitive to the input order of the data, and does not require the assumption of any canonical data distribution. It scales linearly with the size of the input data, has good scalability as the number of data dimensions increases, and is very effective for clustering high-dimensional data in large databases. CLIQUE works by dividing each dimension into non-overlapping intervals, thus dividing the entire embedding space of data objects into cells. It uses a density threshold to identify dense cells and sparse cells. A cell is dense if the number of objects mapped to it exceeds that density threshold. The main strategy of CLIQUE to identify candidate search spaces is to use the monotonicity of dense cells with respect to dimensionality. This is based on the a priori nature of frequent pattern and association rule mining usage. In the context of subspace clustering, the monotonicity is stated as follows: a k-dimensional (>1) cell c has at least 1 point only if each $(k-1)$-dimensional projection of c (which is a $(k-1)$-dimensional cell) has at least 1 point. CLIQUE performs clustering through two phases. In the first stage, CLIQUE divides the d-dimensional data space into a number of rectangular cells that do not overlap with each other and identifies dense cells from them. CLIQUE finds dense cells in all subspaces. To do this, CLIQUE divides each dimension into intervals and identifies intervals containing at least l points, where l is the density threshold. Then, CLIQUE iteratively connects the subspaces. CLIQUE checks whether the number of points in it satisfies the density threshold. The iteration terminates when no candidate is generated or none of the candidates are dense. In the second stage, CLIQUE uses the dense cells in each subspace to assemble clusters that may have arbitrary shapes. The idea is to use the minimum description length (MDL) principle to cover the connected dense cells using the maximal region, where the maximal region is a hyperrectangle into which each cell is dense and the region cannot be extended in any dimension of the subspace. It is difficult to find the best description of clusters in general. Therefore, CLIQUE uses a greedy algorithm. It starts with an arbitrary dense cell, finds the largest region covering that cell, and then continues the process on the remaining dense cells that have not yet been covered. The greedy algorithm terminates when all dense cells are covered. The steps of the CLIQUE algorithm are shown in Algorithm 1.

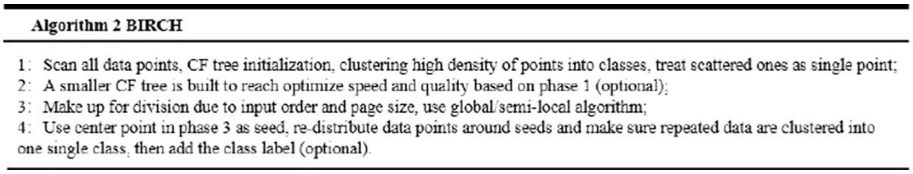


Algorithm 1 CLIQUE

```
1: Find all the dense regions in the one-dimensional space corresponding to each attribute. This is the set of dense one-dimensional cells;
2: $k \leftarrow 2$;
3: Repeat;
4:         Generate all candidate dense $k-1$ dimensional cells from dense k dimensional cells;
5:         Delete cells with fewer than $\varphi$ points;
6:         $k \leftarrow k+1$;
7: Until no k dimensional candidates exists;
8: By taking data from all neighboring, high-density cells and discovering clusters;
9: Generalize each cluster using a small set of inequalities describing the attribute value fields of the cells in the cluster
```

3.4 BIRCH clustering

The traditional method of finding waypoints requires a batch of manually identified waypoints, which are fed into the algorithm to help find waypoints. Such an approach depends on the quality of the manually identified waypoints, and if the quality is poor, the generated waypoints will not be referable. Therefore, the BIRCH algorithm is used to find the waypoints on the navigation trajectory automatically. The BIRCH algorithm [29] is a distance-based hierarchical clustering algorithm that takes memory space into account to obtain the best possible clustering results with limited memory (usually very small compared to the dataset) and to reduce the input and output of the dataset. The algorithm takes into account the time/space efficiency of the clustering process, the sensitivity of the data input and the accuracy of the final clustering results, particularly suitable for processing large data sets. The flow of the BIRCH algorithm is shown in Algorithm 2.



Algorithm 2 BIRCH

```
1: Scan all data points, CF tree initialization, clustering high density of points into classes, treat scattered ones as single point;
2: A smaller CF tree is built to reach optimize speed and quality based on phase 1 (optional);
3: Make up for division due to input order and page size, use global/semi-local algorithm;
4: Use center point in phase 3 as seed, re-distribute data points around seeds and make sure repeated data are clustered into one single class, then add the class label (optional).
```

The BIRCH algorithm aggregates information about clusters by clustering features (CF) description, and then clusters are clustered. Suppose a cluster contains N dimensional data objects $\{x_i\}$, then the clustering features of the cluster are defined as follows:

$$\mathrm{CF} = (N, \mathrm{LS}, \mathrm{SS}) \tag{4}$$

where N is the number of objects in the cluster, LN is the linear sum of N objects (i.e., $\sum_{i=1}^{N} x_i$), and SS is the sum of squares of objects (i.e., $\sum_{i=1}^{N} x_i^2$), which records the key metric for computing clustering and efficient use of storage. The measure of distance between clusters are derived by these clustering features:

The clustering feature tree in the BIRCH algorithm is a highly balanced tree that stores the features of clusters for hierarchical clustering. According to the definition of CF tree, the non-leaf nodes in the tree contain children, and they store the sum of the CF values of their children, i.e., the clustering features containing the children.

The CF tree contains two types of parameters: the non-leaf node branching factor B, the leaf node branching factor L and the threshold T. The branching factor B limits the maximum number of children per non-leaf node, i.e., each non-leaf node contains at most B children; the branching factor L limits the maximum number of children per leaf node; and T limits the maximum radius (or diameter) of the cluster in which a leaf node exists. Figure 3 is an example of a CF tree diagram.

In addition, the shape of the clustering feature tree can be changed by adjusting the size of the threshold and the branching factor, and then the clustering effect of different parameter combinations is evaluated using the Silhouette Score. Finally, based on the construction of the clustering feature tree, the clustering effect is evaluated based on the input class n_clusters (the optimal number of storage nodes). The nodes in the corresponding hierarchy are selected as clusters and are used as the clustering results and output.

3.5 Network construction

The waypoints were extracted from the AIS data through Sect. 3.4. In order to construct the route network, the following operations will be performed: First, each ship corresponds to a unique MMSI number, and each AIS data contains time information, so the route of each ship is extracted in chronological order according to the MMSI serial number. Second, each waypoint is taken as the center of a circle, and the radius r is set to match the AIS within this circle with the waypoints. Third, traverse all the circles formed by waypoints with the extracted sailing trajectory, and if there are AIS data points on the sailing trajectory fall into the circle with the waypoint as the center, mark this AIS data as having a connection with the waypoint, and then connect the waypoints sequentially according to the chronological order of AIS data. Finally, all the waypoints are traversed by different routes, and a route network with directions is constructed. The steps of the route network construction method are shown in Algorithm 3.

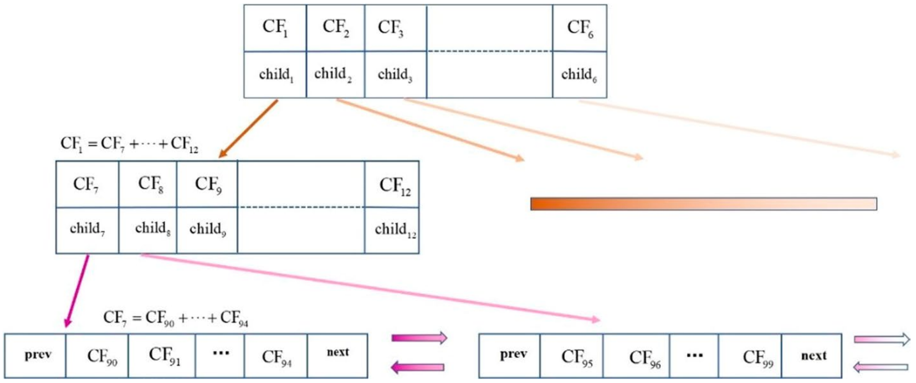
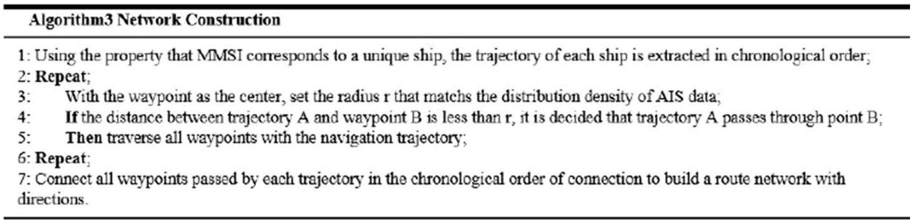


Fig. 3 CF tree ($B=6$, $L=5$)



```
Algorithm3 Network Construction
Using the property that MMSI corresponds to a unique ship, the trajectory of each ship is extracted in chronological order;
Repeat;
    With the waypoint as the center, set the radius r that matchs the distribution density of AIS data;
    If the distance between trajectory A and waypoint B is less than r, it is decided that trajectory A passes through point B;
    Then traverse all waypoints with the navigation trajectory;
Repeat;
Connect all waypoints passed by each trajectory in the chronological order of connection to build a route network with directions.
```

4 Result

This section presents a case of a proposed multilevel clustering algorithm network based on high-dimensional AIS data. For the proof of concept, the case study area was randomly selected and the regional geographic information was extracted as follows. Latitude: 37.105536°N to 40.940382°N; Longitude: 117.620811°E to 125.452704°E. In order to clearly demonstrate the effectiveness of the proposed algorithm and to avoid the influence of undesirable AIS data, in this case, the AIS data of container ships sailing at a speed no less than 7 knots, i.e., not in the vicinity of the port, were investigated. The configuration of this case study is shown in Table 1.

4.1 Data processing

The first step is to prune the AIS data. AIS data of 30 days from June 1 to June 30, 2021, were selected based on the geographic information and setting boundaries of the study area. First, a total of 220,133 AIS data were obtained by reading the initial AIS data, and 110,368 AIS data were obtained as the study data set by excluding the AIS data with speed not exceeding 7 knots and distance less than 185.2 km from the port (the distance of about 100 nautical miles from the port was considered as close to the port). The second step is to reduce the amount of AIS data by DP algorithm. While ensuring the shape characteristics of the route trajectory, 50 m was selected as the threshold value for each trajectory in order to reduce the AIS data points. Meanwhile, the value was determined based on the characteristics of the local AIS data and can be further improved by adaptive design to optimize the results. The purpose of this step is to improve the clustering speed and further obtain better clustering results. At the end of the compression

Table 1 The configuration of the case study

Item	Configuration
Boundary	Latitude: 37.105536°N to 40.940382°N Longitude: 117.620811°E to 125.452704°E
The number of research AIS data	220,133
AIS data sources	Bohai Bay, China on June 1 to June 30, 2021, provided by COSCO
The number of research ship trajectory	18,852
Ship trajectory sources	Bohai Bay, China on January 1 to June 30, 2021, provided by COSCO
Experimental environment	Processing unit: Intel(R) Core(TM) i5-1035O7 CPU @ 1.20 GHz 1.50 GHz Python versions: 3.7.3

process, the AIS data points are reduced from 110,368 to 25,420, with a compression ratio of 76.97%.

4.2 CLIQUE clustering: directional trajectories

Unlike road traffic, marine traffic is not restricted by roads, and ships have more freedom when sailing, so there exists this mixed area and situation of ships with different driving directions. In the route planning, if we can divide the navigation road in different directions according to the direction of the ship navigation, it will help to improve the ship navigation efficiency, reduce the navigation safety risk and the navigation cost. In order to divide the sailing trajectories with directional characteristics, after pre-processing the AIS data, the trajectories are clustered using CLIQUE, and the latitude, longitude and COG are input into the algorithm. Due to the large range of the Bohai Sea, in order to improve the accuracy of clustering, the Bohai Sea is divided into Laizhou Bay, Liaodong Bay, Bohai Bay, West Korea Bay and Bohai Strait according to the composition of the Bohai Sea. Since the amount of AIS data in Laizhou Bay, Liaodong Bay, Bohai Bay and West Korea Bay is similar to that of Bohai Strait, the same parameters are set for these four areas, i.e., the interval of the number of grid cells per dimension is set to 10 and the threshold of outlier points is set to 10; while for Bohai Strait, due to the relatively large amount of data, the parameters are adjusted accordingly, i.e., the interval of the number of grid cells per dimension is set to 10 and the threshold of outlier points is set to 10. The interval of the number of grid cells in each dimension is set to 45, and the outlier threshold is set to 35. The clustering results are shown in Fig. 4 (where a, b, c, d and e correspond to Laizhou Bay, Liaodong Bay, Bohai Bay, West Korea Bay and Bohai Strait, respectively). Each color represents a different direction of the channel, and the obtained trajectory AIS data points have latitude, longitude and COG.

4.3 BIRCH clustering: major waypoints identification

After obtaining the AIS data of the main channel, the BIRCH algorithm will be used to find the waypoints. The main three parameters of the BIRCH algorithm are set to a threshold value of 0.4, n clusters of 2, and a branching factor of 50. The node centers of the constructed clustered feature trees are used as clustering results and outputs. BIRCH provides a clustering method for very large datasets. By focusing on densely occupied regions, it makes sense of large clustering problems and creates a compact summary.

The effectiveness of clustering using the BIRCH algorithm is evaluated using the Silhouette_Score. The evaluation scores are shown in Table 2. A Silhouette_Score greater than 0.5 or better provides good evidence of the truthfulness of the clustering in the data. Therefore, the clustering effect of the selected parameters is ideal.

Figure 5 shows the results of BIRCH clustering, where A is Laizhou Bay, B is Liaodong Bay, C is Bohai Bay, D is West Korea Bay, and E is Bohai Strait. The blue AIS data points are the main shipping lanes, and the orange points are the clustering results, i.e., the waypoints found by BIRCH; a is the waypoints of Laizhou Bay, b Liaodong Bay, c Bohai Bay, d West Korea Bay, and e Bohai Strait. The generated waypoints contain information such as longitude, latitude, and COG.

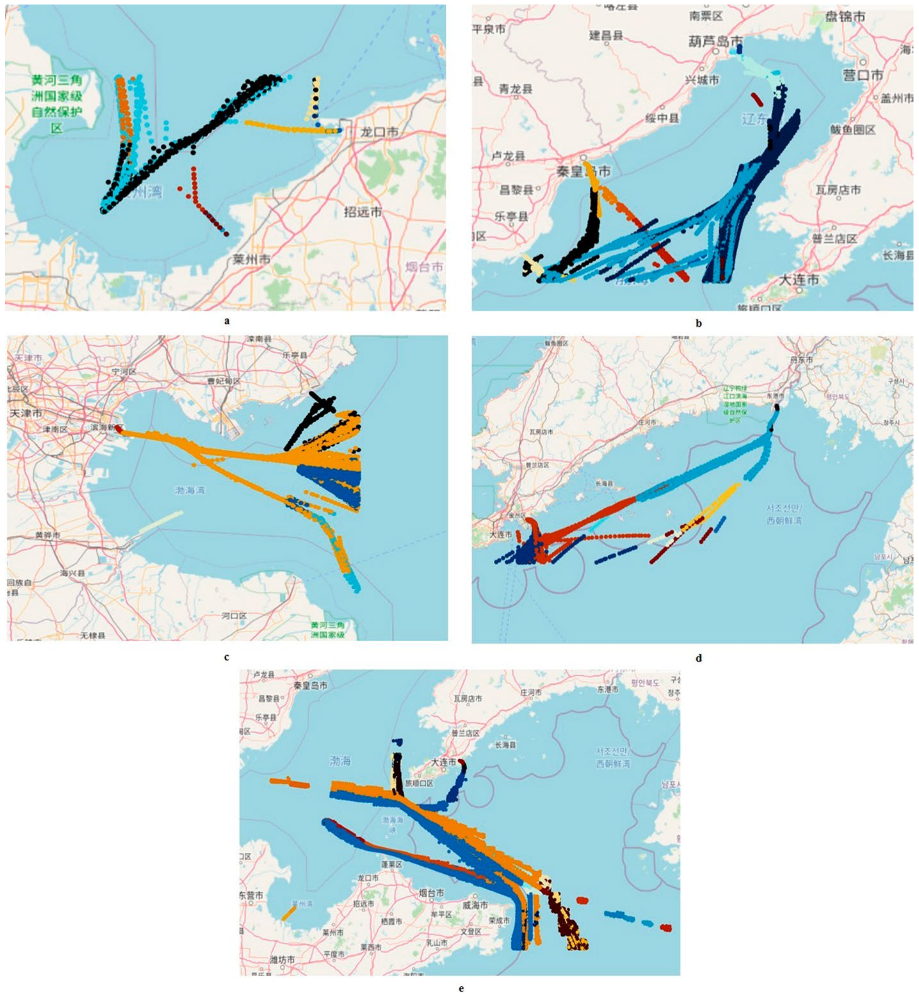


Fig. 4 The trajectories with directions for the five main areas in the Bohai Sea

Table 2 Silhouette_Score

n_cluster	The average silhouette_score
2	0.6425487921033189
3	0.5810285995548892
4	0.5765907817076483
5	0.5708288772927866

4.4 Network construction

After the waypoints are extracted, the waypoints will be connected according to the method proposed in Sect. 3.5. First, the ship trajectory data within the Bohai Sea from January 1 to June 30, 2021 were extracted, and 18,853 ship trajectories were extracted from the data within these six months according to the unique MMSI number and the time tag of AIS data corresponding to each ship. In order to obtain a more comprehensive route network, the waypoints extracted in Sect. 3.5 are

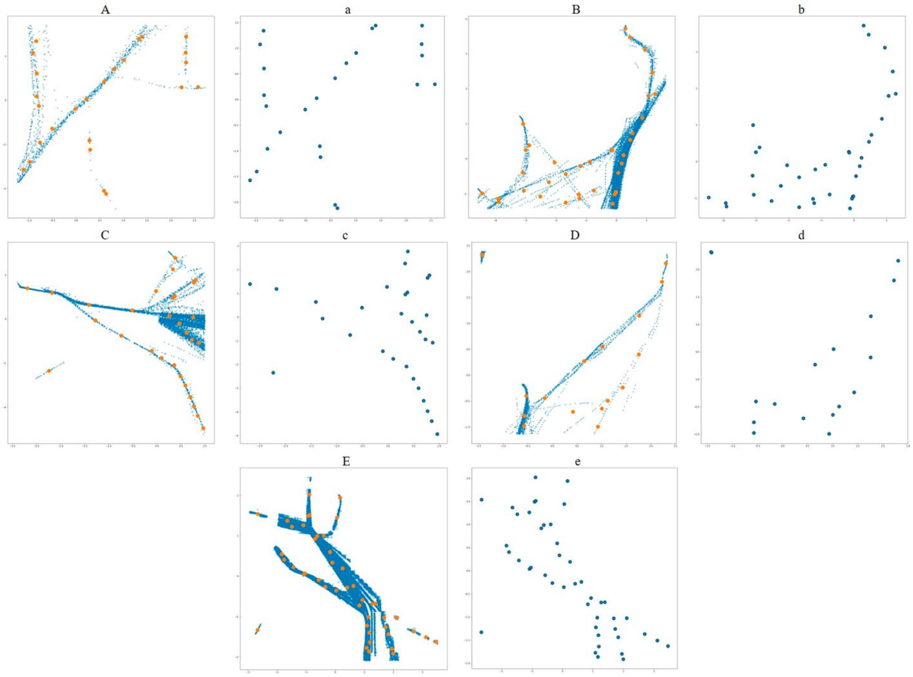


Fig. 5 The results of BIRCH clustering

optimized, and a more comprehensive search of the five main parts of the Bohai Sea was conducted, and finally a total of 440 waypoints were extracted. Since ports are also an important part of the route network when constructing the route network, a total of 29 ports in the Bohai Sea are also considered. According to the method in Sect. 3.5, the 440 waypoints and 29 ports are traversed by 18,853 ship trajectories to build the route network. Figure 6 shows the constructed route network.

As shown in Fig. 6, each waypoint is marked with a different color, and each color represents a different COG direction. The COG directions corresponding to different colors can be found by the gradient spectrum on the right in Fig. 6. As can be seen in Fig. 6, there are some port points that are not connected with other waypoints. This is because, only container ships are considered in the study, and no data of other ship types are used, while some ports are not container ports, so there are no records of container ship arrivals, so these ports are not connected with other waypoints when the route construction process is carried out. It also appears in Fig. 6, that there is no connection between two waypoints. Since there are errors and missing AIS data, the extracted route trajectory will inevitably have incomplete trajectories. In order to remedy this deficiency, 6 months of navigational trajectory data were used to minimize the inability to connect between waypoints caused by missing trajectories. As a whole, the constructed route network is complete and the routes in different directions can be clearly identified, which is beneficial for route planning afterward. The quality of the constructed route network will be discussed in the Discussion section.

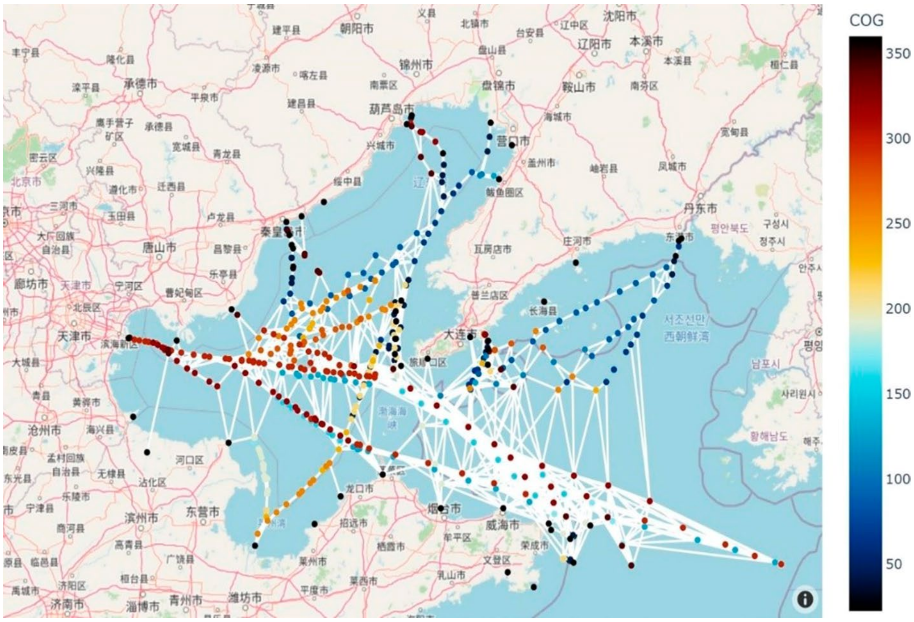


Fig. 6 Route network

5 Discussion

5.1 Comparison with clustering algorithm

In comparison with the methods in Sect. 4.2, the traditional DBSCAN algorithm, K-means algorithm, and CLIQUE algorithm without inputting directional information are used for AIS data trajectory clustering. Trajectory clustering from AIS data using DBSCAN, AIS data points identified as noise points are represented by black dots as shown in Fig. 7a. Too many noise points are generated using DBSCAN, and a large number of trajectories features as well as information are lost as shown in Fig. 7b. After removing the noise points, the retained trajectories are less, and many AIS data points are grouped in the same clusters, and no useful track information is obtained. Figure 7. a. b shows the problems encountered by DBSCAN when dealing with larger data volumes and uneven density datasets. As shown in Fig. 7.c, the K-means algorithm is used to cluster the trajectories of AIS data, and 10 classes of clusters are set in advance first. From the results, it can be seen that each cluster is interlaced together, and the AIS data points in the same cluster are scattered. Although the original trajectory characteristics can be retained, it is impossible to extract the routes according to the clustering results; As shown in Fig. 7d, the CLIQUE algorithm without inputting direction information is used, and for the obtained clustering results, each cluster is mixed together, and fewer trajectories are retained after removing a large number of AIS data points. Compared with several methods used in Fig. 7, the CLIQUE algorithm that inputs latitude, longitude, and COG can retain the original trajectories effectively, on the basis of which more information carried by COG is used to divide the flight paths with directional characteristics, leading to better clustering effect.

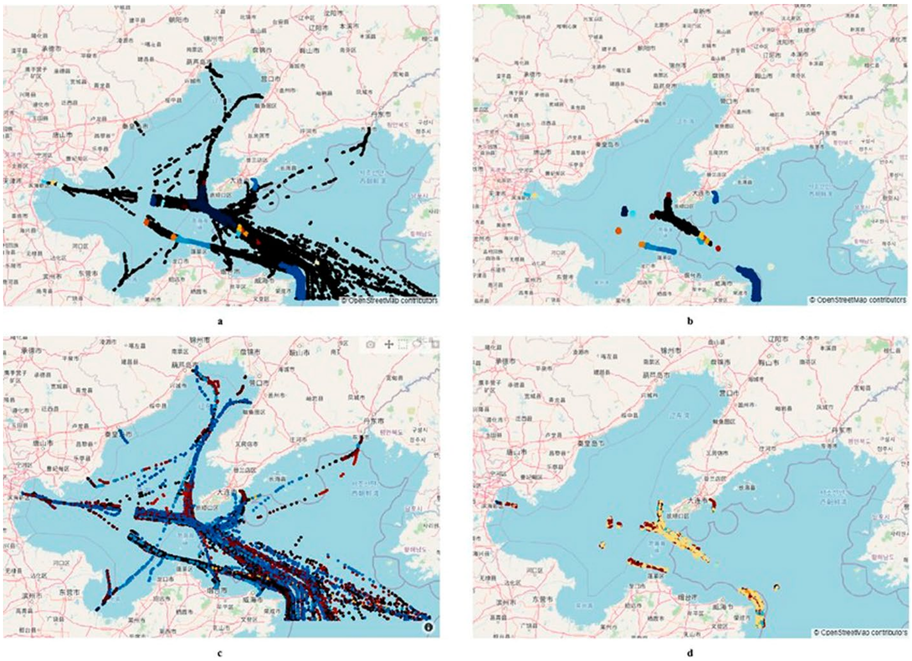


Fig. 7 Clustering results of Bohai route trajectories using other methods

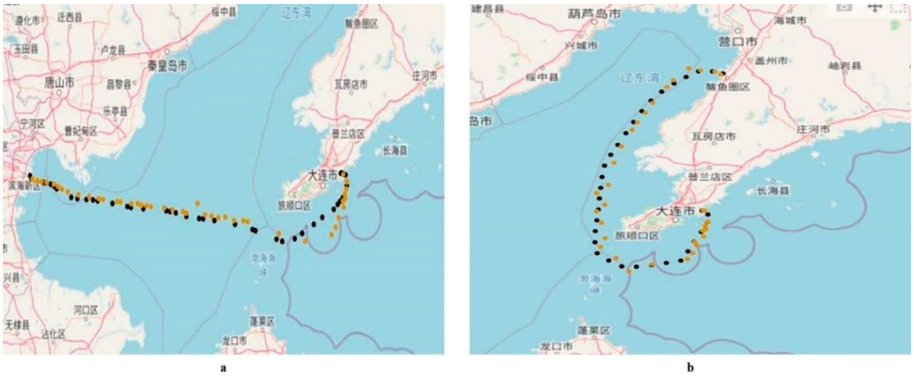


Fig. 8 Cases comparison

5.2 Case study

The evaluation method uses the output to construct a recommended route and then compares it with the actual proposed route. The vessel's route is compared with the route generated using the navigation software—Vessel Value Visualization. For testing purposes, two routes were selected within the Bohai Sea: Tianjin to Dalian and Baiyuquan to Dalian. As shown in Fig. 8, the black points are the trajectories planned by Vessel Value Visualization, and the orange points are the trajectories based on the constructed sea route network, where a is the comparison from Tianjin to Dalian and b is the comparison from Baiyuquan to Tianjin.

The paths provided by the constructed route network and the real route trajectory have different numbers of waypoints, so it is not easy to compare the two. When comparing trajectories, Hausdorff [13] distance is usually used for calculation, and this method consists of three main parts: vertical distance, parallel distance, and angular distance. This method calculates the distance between trajectory segments in three aspects: parallel distance, perpendicular distance, and angular distance.

For two way point based route trajectories $\text{traj}_A = \{a_1, a_2, \ldots, a_n\}$ and $\text{traj}_B = \left\{b_1, b_2, \ldots, b_n\right\}$, their Hausdorff distances are calculated by Eq. 5

$$\begin{aligned} H(\text{traj}_A, \text{traj}_B) &= \max\{h(\text{traj}_A, \text{traj}_B), h(\text{traj}_B, \text{traj}_A)\} \\ h(\text{traj}_A, \text{traj}_B) &= \max\{\min\{\left\|a_i - b_i\right\|\}\} \\ h(\text{traj}_B, \text{traj}_A) &= \max\{\min\{\left\|b_i - a_i\right\|\}\} \end{aligned} \tag{5}$$

where $\cdot$ denotes the Euclidean distance between the coordinate points in ship trajectory A and the coordinate points in ship trajectory B. $H(\text{traj}_A, \text{traj}_B)$ is the basic form of Hausdorff distance, i.e., it is the maximum value between $h\left(\text{traj}_A, \text{traj}_B\right)$ is and $h\left(\text{traj}_B, \text{traj}_A\right)$. In this design, the shape similarity between two trajectories can be obtained without considering their lengths. The similarity of the trajectories of the two routes in Fig. 8 is shown in Table 3, while Table 3 randomly selected six ports in Bohai. The trajectories of the routes generated by the network constructed in Sect. 4.4 are compared with the trajectories of the routes planned in Ship Vision, and the results are shown in Table 3.

As can be seen from Table 3, except between ports where container ships do not sail, the route planning made by using the way points found in Sect. 4.3 and the route network constructed in Sect. 4.4 is consistent with the real historical route and the recommended route by Vessel Value Visualization, and the results are better.

The proposed method has successfully verified that ocean trajectories can be clustered on the basis of higher dimensional data with a modified number of classes. When using CLIQUE-BIRCH for waypoint extraction, a total of one minute was used, and when traversing all waypoints with 18,853 navigational trajectories, the route network was constructed in less than 4 min, for a total time of less than 5 min. Compared with other methods such as genetic algorithm which takes more than 3 h on average to extract waypoints and more than 5 min on average to build route networks, the proposed method shows a significant reduction in computational time, saving computational resources and time costs. However, the current study focuses on only one additional dimension (direction) and the number of classes still needs to be modified manually, so that the self-clustering algorithm for classifying ocean trajectories remains a major problem.

Table 3 Trajectory similarity matrix between the 6 ports

	Tianjin	Dalian	Bayuquan	Weihai	Penglai	Yantai
Tianjin	1	0.983991	0.989840	0.978640	0.970282	0.973612
Dalian	0.983991	1	0.989725	0.968948	–	0.981115
Bayuquan	0.989840	0.989725	1	0.985780	0.954033	0.983947
Weihai	0.978640	0.968948	0.985780	1	–	0.981137
Penglai	0.970282	–	0.954033	–	1	–
Yantai	0.973612	0.981115	0.983947	0.981137	–	1

6 Conclusion

As a significant first step to realize a multi-featured clustering route network construction method based on real-world AIS data, a proof of concept is presented to exploit the numerous attributes contained in AIS data, useful information in AIS data is mined to cluster route trajectories with directions, and waypoints on the route are identified by one additional layer of the clustering algorithm, and a maritime route network is constructed by connecting waypoints according to the connections between AIS data points. Since the dataset used in the experiment is a real-world AIS dataset, the experimental results can be extended and can be used for waypoint finding and maritime route network construction in more sea areas.

The focus of this paper is on the search of sea waypoints and the construction of route networks. In the research process, only container ships with sailing speed not less than 7 knots are considered, and no other ship type factors are considered. Also, other factors affecting route selection, such as weather, fuel consumption and distance, are not studied. Therefore, in the subsequent studies, the selection of waypoints and the construction of route network by different factors will be considered more comprehensively.

7 Future work

Waypoint's detection and multi-featured network construction has potential to shape the future ship path planning field. In order to go beyond the proposed method, more features besides direction properties can be included in this work. Besides, based on the performance of current method, the clustering process can also be applied to real-time applications. In addition to that, the detection of waypoints can be further divided into classes to find the different layers of the network.

Abbreviations

IOT: The Internet of Things; AIS: Automatic Identification System; COG: Course over Ground; CF: Clustering Features; MMSI: Maritime Mobile Service Identify; GRU: Gated Recurrent Unit; ACO: The Ant Colony Optimization; PID: Proportion Integral Differential; DP: The Douglas-Peucker; BIRCH: Balance Iterative Reducing and Clustering Using Hierarchies; CLIQUE: Clustering in QUEst; MDL: Minimum Description Length; DBSCAN: Density-Based Spatial Clustering of Applications with Noise.

Acknowledgments

I would like to thank my advisor Mr. Yi Han for his guidance, advice, patience and encouragements throughout this research work. And all of my co-workers during this process, Mr. He Jiang Mrs. Yuanzhe Liu and Dr. Shaohan Wang for their help and advice with academic support and algorithm development.

Authors' contributions

FR contributes for the whole idea of the proposed method. The corresponding author YH supervised the whole progress and design the method. SW modified the algorithm and data analysis. HJ did part of the programing and data visualization work. All authors read and approved the final manuscript.

Funding

All funding supports are from Cosco Shipping Technology Co., Ltd.

Availability of data and materials

Data sharing is not applicable to this article as no datasets were generated or analyzed during the current study.

Declarations

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Shanghai Ship and Shipping Research Institute, Shanghai, China. [2]COSCO SHIPPING Technology Co., Ltd., Shanghai, China.

Received: 13 November 2021 Accepted: 10 March 2022
Published online: 21 March 2022

References

1. K. Naus, Drafting route plan templates for ships on the basis of AIS historical data. J. Navig. **73**(3), 726–745 (2020)
2. A.K. Jain, Data clustering: 50 years beyond K-means. Pattern Recogn. Lett. **31**(8), 651–666 (2010)
3. S. Wang, S.A. Zargar, F.-G. Yuan, Augmented reality for enhanced visual inspection through knowledge-based deep learning. Struct. Health Monit. **20**(1), 426–442 (2021)
4. S. Wang, R.-Y. Fong, F.-G. Yuan. Vibration-based damage imaging via high-speed cameras with 3D digital image correlation using wavelet transform. in *Sensors and Smart Structures Technologies for Civil, Mechanical, and Aerospace Systems 2021* (International Society for Optics and Photonics, 2021)
5. L. Zhao, G. Shi, A trajectory clustering method based on Douglas-Peucker compression and density for marine traffic pattern recognition. Ocean Eng. **172**, 456–467 (2019)
6. M. Grifoll, T. Karlis, M.I. Ortego, Characterizing the evolution of the container traffic share in the Mediterranean Sea using hierarchical clustering. J. Mar. Sci. Eng. **6**(4), 121 (2018)
7. J.-F. Song, S.-Y. Wang, H.-L. Zhao, Traffic flow detection at road intersections based on K-means and NURBS trajectory clustering. Math. Probl. Eng. **2020**, 1–6 (2020)
8. X. Wang, Y. Bai, The global Minmax k-means algorithm. Springerplus **5**(1), 1665 (2016)
9. P. Han et al., A combined online-learning model with K-means clustering and GRU neural networks for trajectory prediction. Ad Hoc Netw. **117**, 102476 (2021)
10. L. Tyagi, M.C. Trivedi, *Hybrid K-Mean and Refinement Based on Ant for Color Image Clustering* (Springer, Singapore, 2016)
11. Y. Jiang, et al. A novel classification scheme of moving targets at sea based on Ward's and K-means clustering (2018)
12. L. Zhao, G. Shi, J. Yang. An adaptive hierarchical clustering method for ship trajectory data based on DBSCAN algorithm. IEEE.
13. L. Wang et al., Ship AIS trajectory clustering: an HDBSCAN-based approach. J. Mar. Sci. Eng. **9**(6), 566 (2021)
14. F.-G. Yuan, et al. Machine learning for structural health monitoring: challenges and opportunities. in *Sensors and Smart Structures Technologies for Civil, Mechanical, and Aerospace Systems 2020* (International Society for Optics and Photonics, 2020)
15. A. Dobrkovic, M.-E. Iacob, J. Van Hillegersberg. *Maritime Pattern Extraction from AIS Data Using a Genetic Algorithm*. IEEE
16. A. Dobrkovic, M.E. Lacob, J.V. Hillegersberg, Maritime pattern extraction and route reconstruction from incomplete AIS data. Int. J. Data Sci. Anal. **5**(2–3), 111–136 (2018)
17. D. Filipiak et al., Extracting maritime traffic networks from AIS data using evolutionary algorithm. Bus Inf Syst Eng Int J Wirtschaftsinformatik **62**, 435–450 (2020)
18. S. Ni, Z. Liu, Y. Cai, Ship manoeuvrability-based simulation for ship navigation in collision situations. J. Mar. Sci. Eng. **7**(4), 90 (2019)
19. S. Ni et al., Modelling of ship's trajectory planning in collision situations by hybrid genetic algorithm. Pol. Marit. Res. **25**(3), 14–25 (2018)
20. L. Wang et al., Ship route planning based on double-cycling genetic algorithm considering ship maneuverability constraint. IEEE Access **8**, 190746–190759 (2020)
21. S. Wang et al., An efficient augmented reality (AR) system for enhanced visual inspection. Struct. Health Monit. (2019). https://doi.org/10.12783/shm2019/32278
22. W. Zhao et al., Multicriteria ship route planning method based on improved particle swarm optimization-genetic algorithm. J. Mar. Sci. Eng. **9**(4), 357 (2021)
23. J. Chen et al., Research on fuzzy control of path tracking for underwater vehicle based on genetic algorithm optimization. Ocean Eng. **156**, 217–223 (2018)
24. C. Liu et al., An improved A-star algorithm considering water current, traffic separation and berthing for vessel path planning. Appl. Sci. **9**, 1057 (2019)
25. K. Sun et al., Optimal path planning method of marine sailboat based on fuzzy neural network. J. Coast. Res. **93**(SI), 911–916 (2019)
26. C. Tang et al., A method for compressing AIS trajectory data based on the adaptive-threshold Douglas-Peucker algorithm. Ocean Eng. **232**, 109041 (2021)
27. L. Zhao, G. Shi, A method for simplifying ship trajectory based on improved Douglas-Peucker algorithm. Ocean Eng. **166**, 37–46 (2018)
28. R. Agrawal, C. Faloutsos, A. Swami, *Efficient similarity search in sequence databases* (Springer, Berlin, 1993)
29. T. Zhang, R. Ramakrishnan, M. Livny. BIRCH: an efficient data clustering method for very large databases. in *Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data* (Association for Computing Machinery, Montreal, 1996), p. 103–114
30. R. Agrawal et al., Automatic subspace clustering of high dimensional data for data mining applications. SIGMOD Rec. **27**(2), 94–105 (1998)

Publisher's Note