A Novel High-Dimensional Trajectories Construction Network based on Multi-Clustering Algorithm

Feiyang Ren
SSSRI: Shanghai Ship and Shipping Research Institute

Yi Han
Cosco Shipping Technology Co., Ltd

Shaohan Wang (wanghetongtt@hotmail.com)
Cosco Shipping Technology Co., Ltd https://orcid.org/0000-0002-1556-594X

He Jiang
Cosco Shipping Technology Co., Ltd

Research Article

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

Posted Date: November 18th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-1060086/v1

Version of Record: A version of this preprint was published at EURASIP Journal on Wireless Communications and Networking on March 21st, 2022. See the published version at https://doi.org/10.1186/s13638-022-02108-4.

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]

Abstract

A novel marine transportation network based on high-dimensional AIS data with a multi-level clustering algorithm is proposed to discover important waypoints in trajectories based on selected navigation features. This network contains two parts: the calculation of major nodes with CLIQUE and BIRCH clustering methods and navigation network construction with edge construction theory. Unlike the state-of-art work for navigation clustering with only ship coordinate, the proposed method contains more high-dimensional features such as drafting, weather, and fuel consumption. By comparing the historical AIS data, more than 220,133 lines of data in 30 days were used to extract 440 major nodal points in less than 4 minutes with ordinary PC specs (i5 processer). The proposed method can be performed on more dimensional data for better ship path planning or even national economic analysis. Current work has shown good performance on complex ship trajectories distinction and great potential for future shipping transportation market analytical predictions.

Key words: 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 researchers 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 brings a high rate of error. 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 increasing 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

1. Shanghai Ship and Shipping Research Institute;
2. COSCO SHIPPING Technology Co., Ltd.;
*Correspondence: hy19970330@163.com

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 and is less effective for noisy data. 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 on top of 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 is used to process statistical data and neural networks optimize navigation routes. PID

(Proportion Integral Differential) 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 work finds the major waypoints manually, which is highly subjective and error-pruned, 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. Thus, 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, the newly proposed edge's connection method is used to connect the waypoints to construct a sea route network, and the constructed route network is evaluated with examples. Figure 1 gives the methodological overview of the research:

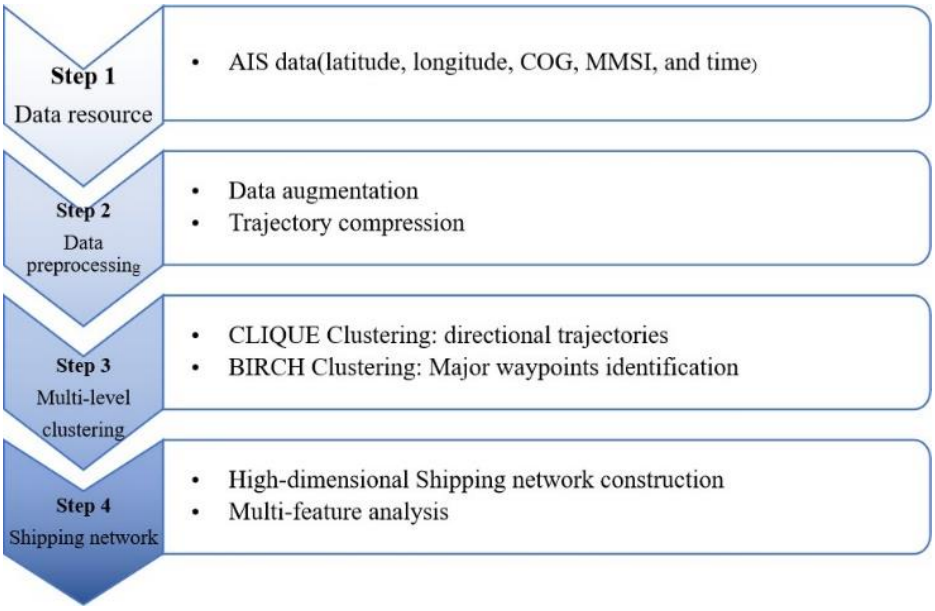

Step 1
Data resource
• AIS data(latitude, longitude, COG, MMSI, and time)
Step 2
Data preprocessing
• Data augmentation
• Trajectory compression
Step 3
Multi-level clustering
• CLIQUE Clustering: directional trajectories
• BIRCH Clustering: Major waypoints identification
Step 4
Shipping network
• High-dimensional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

Figure 1 Methodological overview of the research

2.2 AIS data preprocessing

Two methods are used to enhance the quality of AIS data. First, match AIS data points with ports and eliminate AIS displacements data of more than 185.2 nautical miles from the port. Second, since each vessel corresponds to a unique MMSI (Maritime Mobile Service Identify), the AIS data are further preprocessed using the time attributes and MMSI in the AIS data to avoid the connection between two sailing tracks. The Douglas-Peucker (DP) algorithm[5, 26, 27] has been widely used to remove redundant AIS data points from ship trajectories. It is also an effective way to simplify lines and other shapes in mapping research, such as visualization studies to facilitate further analysis and improve operational speed.

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 (Automatic Identification System) data is proposed. First, the latitude, longitude, and COG (Course over Ground) 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 (Balance Iterative Reducing and Clustering using Hierarchies) 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 Edge Construction

For the construction of shipping trajectories network on open waters, edge construction is employed on the connection with AIS coordinate and calculated major points. Edges are constructed as followed: First, each major point will act as the center of a circle on the graph with user defined radius. Historical AIS coordinate within the circle will be labeled as “related” with that specific point. Then based on the relationship with major points and the chronological AIS coordinate, each calculated major point will be connected to form the 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 $\mathrm{Trajectory}=\left\{ship_i \,\middle|\, ship_i ,\mathrm{i}=1,2,\mathrm{L}\ \ ,\mathrm{m}\right\}$, where ship_i is the trajectory of ship i and m is the number of ships, and ship_i is defined in (1):

$$Ship_i = \left\{ p_i^k \left| p_i^k = \left(MMSI_i, lat_i^k, lon_i^k, T_i^k \right), k = 1,2,\mathrm{L}\ ,\mathrm{n} \right. \right\} \tag{1}$$

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 k^{th} AIS data point of the i^{th} 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:

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

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.

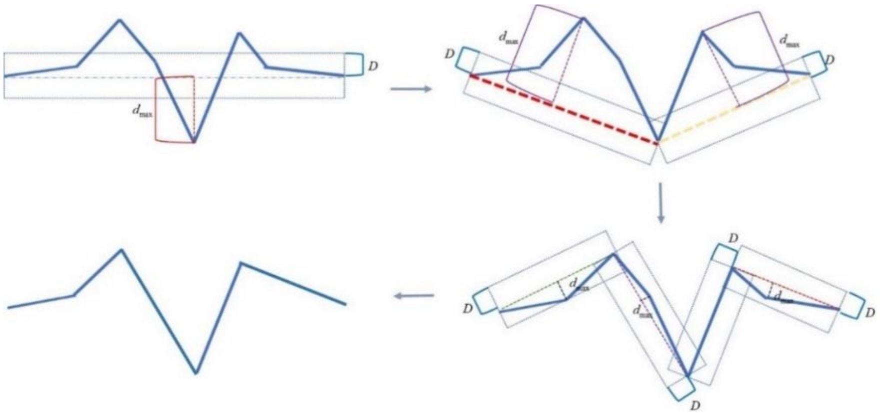


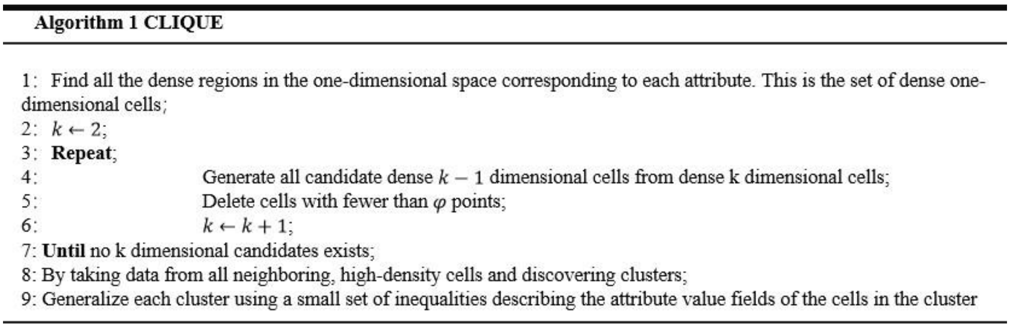
Figure 2 The process of compression of trajectories by the DP algorithm

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 is multidimensional that contain many different attributes, but the methods used in traditional trajectory clustering often only consider latitude and longitude, or when introducing directions, they need to be calculated with the help of latitude and longitude, which causes a waste of computational resources, and there may also be a situation that the calculated directions do not match with the actual directions. Therefore, the CLIQUE algorithm is used to solve the above problems. CLIQUE clustering was proposed by Agrawal et al. in 1998[30], a subspace clustering approach, and he applied a grid-based clustering method. The algorithm has the following advantages:(1) scalability; (2) the possibility of discovering clusters in subspaces in the high-dimensional space of the data set; (3) The clustering model has good explanatory power; (4) No order dependence on the input data; (5) No prior assumption that the dataset satisfies a certain probability distribution is required.

Unlike density-based clustering algorithms such as DBSCAN, the CLIQUE algorithm does not focus on the entire high-dimensional space. It can obtain clusters of subspaces of the original data space more efficiently. The steps of the CLIQUE algorithm are shown in Algorithm 1.

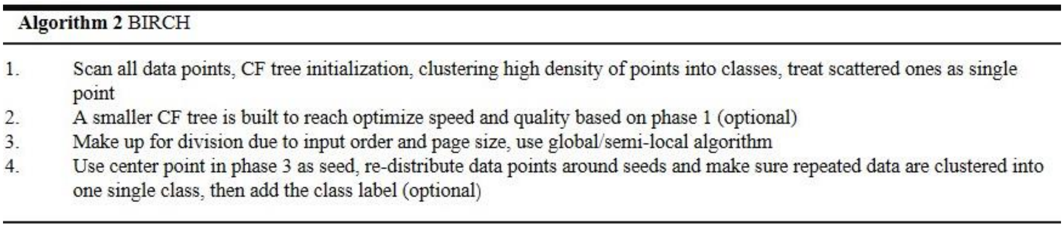


Algorithm 1 CLIQUE

```
Find all the dense regions in the one-dimensional space corresponding to each attribute. This is the set of dense one-dimensional cells;
$k \leftarrow 2$;
Repeat;
        Generate all candidate dense $k-1$ dimensional cells from dense k dimensional cells;
        Delete cells with fewer than $\varphi$ points;
        $k \leftarrow k+1$;
Until no k dimensional candidates exists;
By taking data from all neighboring, high-density cells and discovering clusters;
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:

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

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:

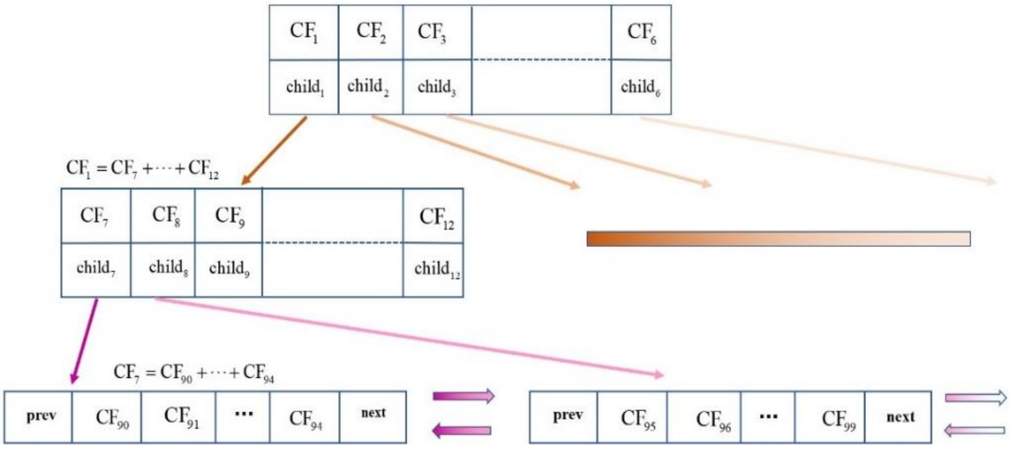


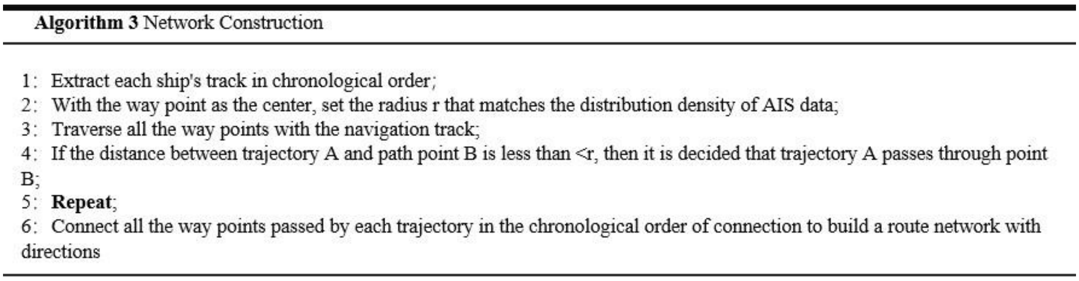
Figure 3 CF Tree (B=6, L=5)

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 BIRCH algorithm is described in the previous section. As is described in the previous section, the BIRCH algorithm….is a method which is needed to construct a directed route network, i.e., those waypoints that should be connected. Based on the historical AIS data, it is possible to find out which routes each track has passed through and up to the time of each AIS data point. Therefore, a directed sea route network can be constructed by simply connecting the waypoints in chronological order and the AIS order of the historical tracks. The route network is constructed as shown in Algorithm 3.



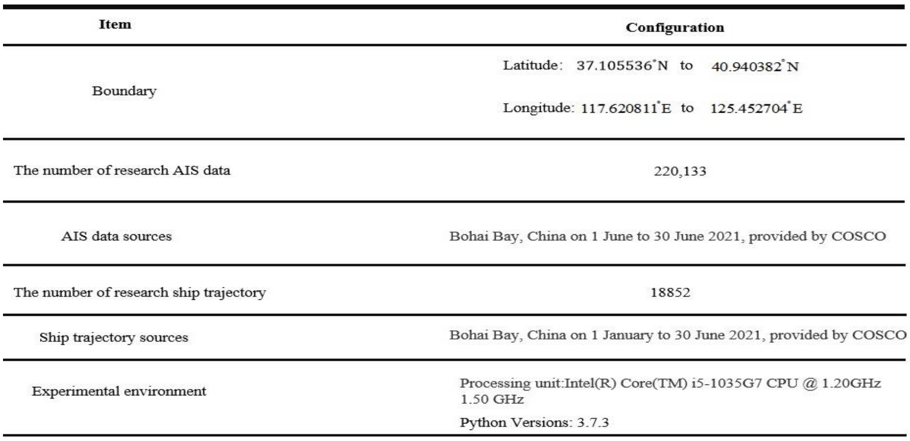
Algorithm 3 Network Construction

```
1: Extract each ship's track in chronological order;
2: With the way point as the center, set the radius r that matches the distribution density of AIS data;
3: Traverse all the way points with the navigation track;
4: If the distance between trajectory A and path point B is less than <r, then it is decided that trajectory A passes through point B;
5: **Repeat**;
6: Connect all the way points 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.

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 1 June to 30 June 2021, provided by COSCO
The number of research ship trajectory	18852
Ship trajectory sources	Bohai Bay, China on 1 January to 30 June 2021, provided by COSCO
Experimental environment	Processing unit:Intel(R) Core(TM) i5-1035G7 CPU @ 1.20GHz 1.50 GHz Python Versions: 3.7.3

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, 50m 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 process, the AIS data points are reduced from 110368 to 25420, with a compression ratio of 76.97%.

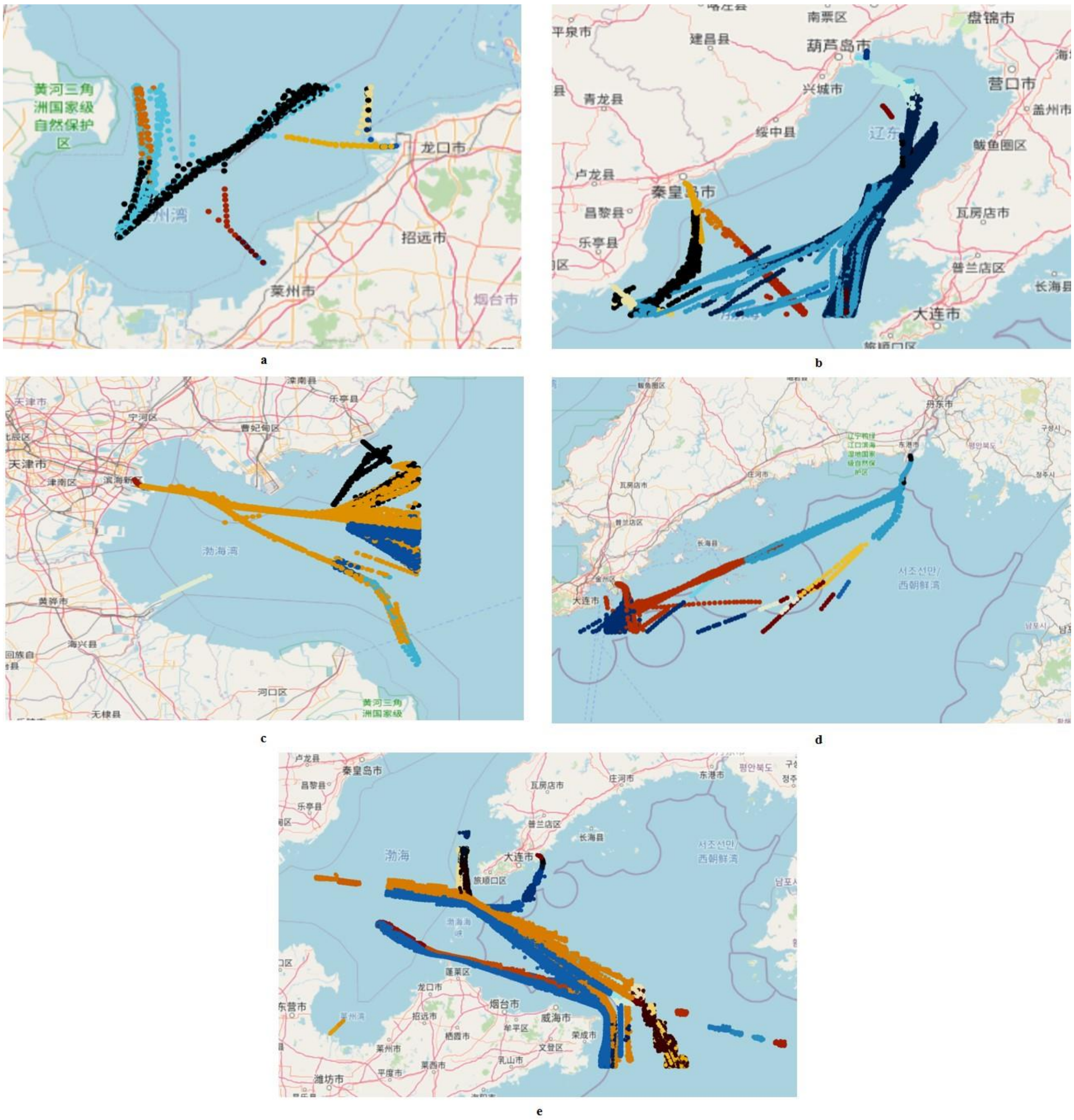


Figure 4 The main shipping lanes with directions in Bohai Bay

4.2 CLIQUE Clustering: Directional Trajectories

After processing the AIS data, the trajectory clustering is performed on the AIS data using CLIQUE. Without loss of generality, the two main parameters in CLIQUE clustering are set. Since the amount of AIS data in Laizhou Bay, Liaodong Bay, Bohai Bay, and West Korea Bay are similar and less compared with Bohai Strait's, the same parameters are set for these four regions, i.e., the interval of the number of grid cells in each dimension is set to 10, and the outlier threshold is set to 10; since Bohai Strait in Bohai

Bay, the clustering results are shown in Figure 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 with information such as 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.

Table 2 Silhouette_Score

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

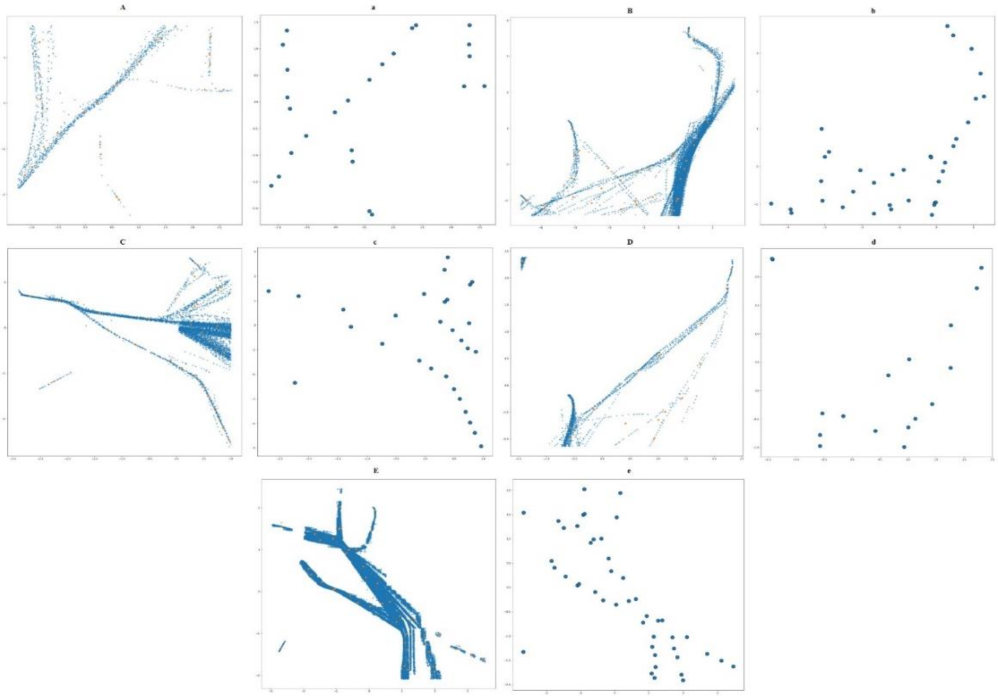


Figure 5 The Waypoints

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.

4.4 Network Construction

In order to demonstrate the connectivity between individual waypoints, the waypoints will be connected according to the method mentioned in Section 3.5. The network is constructed based on the ship's sailing trajectory from Bohai Bay for six months. Firstly, a total of 18,853 trajectories were extracted from Bohai Bay in half a year; Secondly, the ports and waypoints in Bohai Bay were connected according to the method proposed in subsection 3.5, i.e., 18,853 trajectories were used to traverse 29 ports and 440 waypoints obtained after optimization, and the trajectory network was constructed in less than 4 minutes on average.

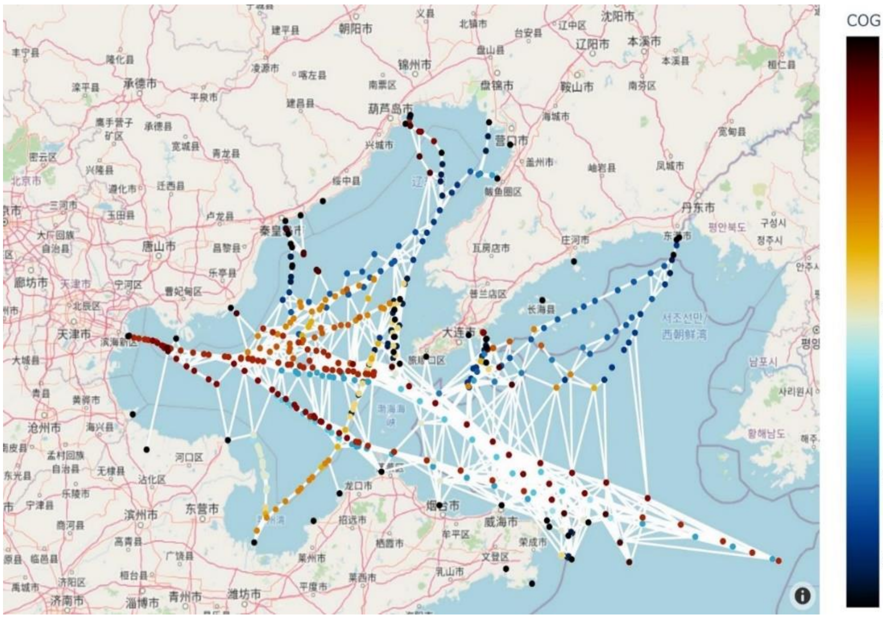


Figure 6 Route Network

Figure 6 uses color to visualize the orientation of the waypoints. The color of the waypoints indicates the average COG of the ship as it passes through the waypoints, based on the clustering results in the previous subsection.

5 Discussion

5.1 Comparison with Clustering Algorithm

For comparison with the methods in Section 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 Figure 7. a. Too many noise points are generated using DBSCAN, and a large number of trajectories features, as well as information, are lost as shown in Figure 7. b. 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 Figure 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 Figure 7.d, 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 Figure 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.

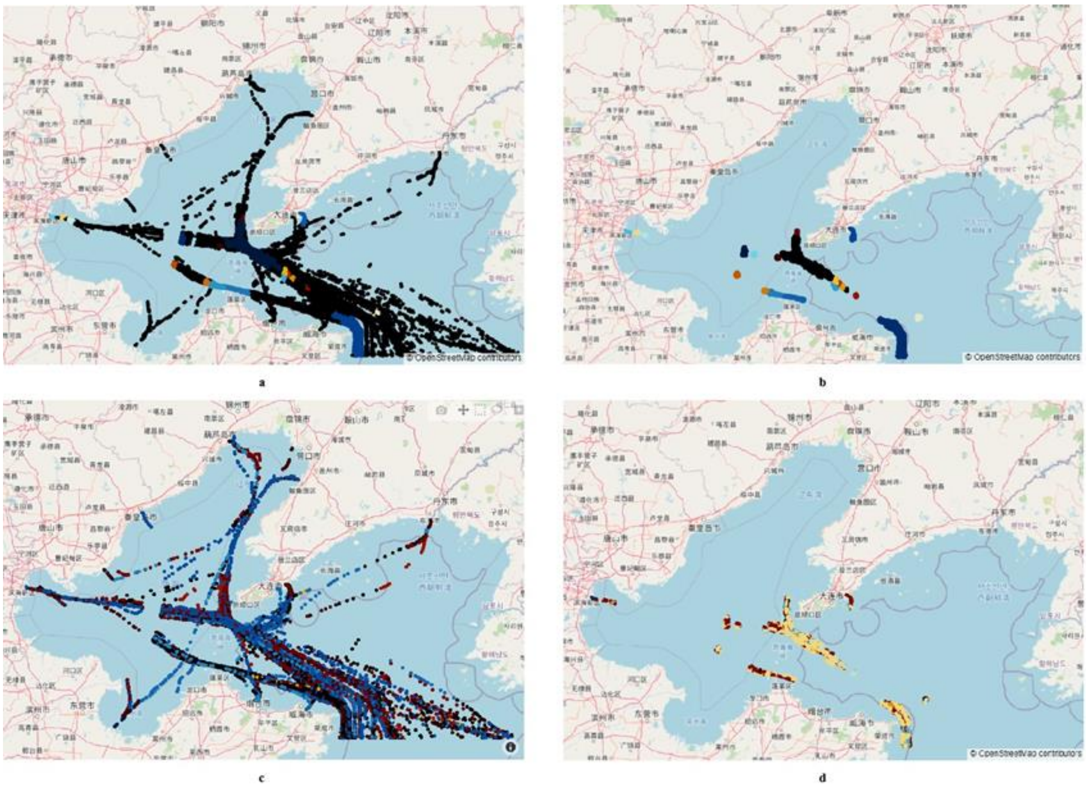


Figure 7 Other Methods

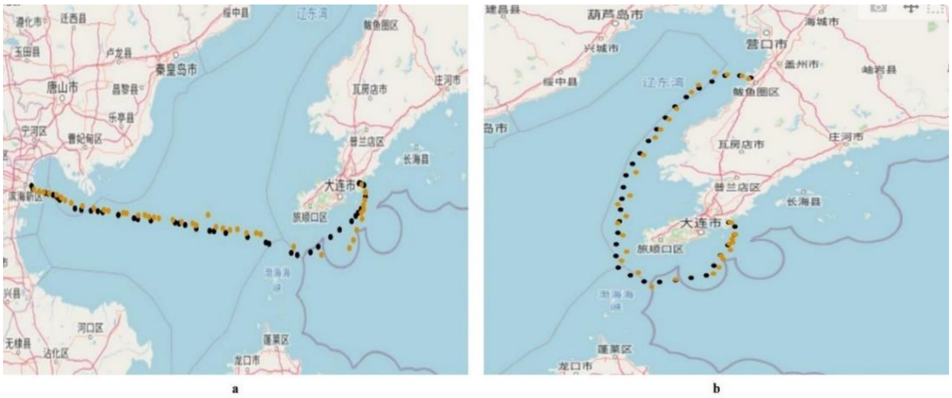


Figure 8 Cases Comparison

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 Bohai Bay. Tianjin to Dalian and Baiyuquan to Dalian. As shown in Figure 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 a comparison from Tianjin to Dalian and b is a 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.

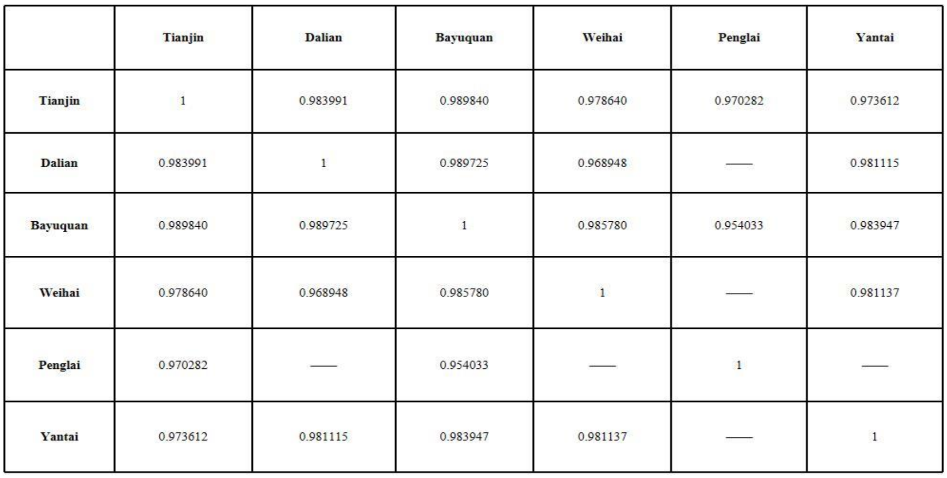
For two way point based route trajectories $traj_A = \{a_1, a_2, \cdots, a_n\}$ and$traj_B = \{b_1, b_2, \cdots, b_n\}$, their Hausdorff distances are calculated by Equation 5

$$\begin{aligned} H(traj_A, traj_B) &= \max\{h(traj_A, traj_B), h(traj_B, traj_A)\} \\ h(traj_A, traj_B) &= \max\{\min\{\|a_i - b_i\|\}\} \\ h(traj_B, traj_A) &= \max\{\min\{\|b_i - a_i\|\}\} \end{aligned} \tag{5}$$

where $\|\cdot\|$ denotes the Euclidean distance between the coordinate points in ship trajectory A and the coordinate points in ship trajectory B. $H(traj_A, traj_B)$is the basic form of Hausdorff distance, i.e., it is the maximum value between $h(traj_A, traj_B)$is and $h(traj_B, traj_A)$. 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 Figure 8 is shown in Table 3, while Table 3 randomly selected six ports in Bohai Bay, the trajectories of the routes generated by the network constructed in Section 4.4 will be 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 waypoints found in section 4.3 and the route network constructed in section 4.4 is consistent with the real historical route and the recommended route by Vessel Value Visualization, and the results are better.

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

The proposed method has successfully verified that marine trajectories can be clustered based on higher dimensional data with modified number of classes. Comparing to other methods like the genetic algorithm the proposed metho shows a reduction of computational time for more than 40%. However, current study only focuses on one addition dimension (the direction) and the number of classes still need to be modified manually therefore a self-clustering algorithm to classify the marine trajectories still remains as a major problem.

6 Conclusion

As a significant first step for enabling multi-featured clustering route network construction method based on real-world AIS data, a proof of concept is presented for using 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.

This paper focuses on the search of maritime waypoints and the construction of route networks. Only container ships with sailing speeds greater than or equal to 7 are considered, while fuel efficiency and weather conditions are not studied. Therefore, subsequent research in more depth will be needed.

7 Future Work

Route major points 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 on real-time applications. In addition to that, the detection of major points 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

Declarations

- **Availability of data and materials**

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

- **Competing interests**

  The author(s) declare(s) that they have no competing interests

- **Funding**

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

- **Author's contributions**

  F.R contributes for the whole idea of the proposed method. The corresponding author Y.H supervised the whole progress and design the method. S.W modified the algorithm and data analysis. H.J did most of the programing and data visualization work.

- **Acknowledgements**

  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.

Reference

1. Naus, K., *Drafting Route Plan Templates for Ships on the Basis of AIS Historical Data.* The Journal of Navigation, 2020. **73**(3): p. 726-745.
2. Jain, A.K., *Data clustering: 50 years beyond K-means.* Pattern Recogn. Lett., 2010. **31**(8): p. 651–666.
3. Wang, S., S.A. Zargar, and F.-G. Yuan, *Augmented reality for enhanced visual inspection through knowledge-based deep learning.* Structural Health Monitoring, 2021. **20**(1): p. 426-442.
4. Wang, S., R.-Y. Fong, and 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*. 2021. International Society for Optics and Photonics.
5. Zhao, L. and G. Shi, *A trajectory clustering method based on Douglas-Peucker compression and density for marine traffic pattern recognition.* Ocean Engineering, 2019. **172**: p. 456-467.
6. Grifoll, M., T. Karlis, and M.I. Ortego, *Characterizing the Evolution of the Container Traffic Share in the Mediterranean Sea Using Hierarchical Clustering.* Journal of Marine Science and Engineering, 2018. **6**(4): p. 121.
7. Song, J.-F., S.-Y. Wang, and H.-L. Zhao, *Traffic Flow Detection at Road Intersections Based on K-Means and NURBS Trajectory Clustering.* Mathematical Problems in Engineering, 2020. **2020**: p. 1-6.
8. Wang, X. and Y. Bai, *The global Minmax k-means algorithm.* Springerplus, 2016. **5**(1): p. 1665.
9. Han, P., et al., *A combined online-learning model with K-means clustering and GRU neural networks for trajectory prediction.* Ad Hoc Networks, 2021. **117**: p. 102476.
10. Tyagi, L. and M.C. Trivedi. *Hybrid K-Mean and Refinement Based on Ant for Color Image Clustering*. 2016. Singapore: Springer Singapore.
11. Jiang, Y., et al. *A Novel Classification Scheme of Moving Targets at Sea Based on Ward's and K-means Clustering*. 2018.
12. Zhao, L., G. Shi, and J. Yang. *An adaptive hierarchical clustering method for ship trajectory data based on DBSCAN algorithm*. IEEE.
13. Wang, L., et al., *Ship AIS Trajectory Clustering: An HDBSCAN-Based Approach.* Journal of Marine Science and Engineering, 2021. **9**(6): p. 566.
14. Yuan, F.-G., et al. *Machine learning for structural health monitoring: challenges and opportunities*. in *Sensors and Smart Structures Technologies for Civil, Mechanical, and Aerospace Systems 2020*. 2020. International Society for Optics and Photonics.
15. Dobrkovic, A., M.-E. Iacob, and J. Van Hillegersberg. *Maritime Pattern Extraction from AIS Data Using a Genetic Algorithm*. IEEE.
16. Dobrkovic, A., M.E. Lacob, and J.V. Hillegersberg, *Maritime pattern extraction and route reconstruction from incomplete AIS data.* International Journal of Data Science & Analytics, 2018. **5**(2-3): p. 111-136.
17. Filipiak, D., et al., *Extracting Maritime Traffic Networks from AIS Data Using Evolutionary Algorithm.* Business & Information Systems Engineering: The International Journal of WIRTSCHAFTSINFORMATIK, 2020. **62**.
18. Ni, S., Z. Liu, and Y. Cai, *Ship Manoeuvrability-Based Simulation for Ship Navigation in Collision Situations.* Journal of Marine Science and Engineering, 2019. **7**(4): p. 90.
19. Ni, S., et al., *Modelling of Ship's Trajectory Planning in Collision Situations by Hybrid Genetic*

Algorithm. Polish Maritime Research, 2018. **25**(3): p. 14-25.

20. Wang, L., et al., *Ship Route Planning Based on Double-Cycling Genetic Algorithm Considering Ship Maneuverability Constraint.* IEEE Access, 2020. **8**: p. 190746-190759.
21. WANG, S., et al., *An efficient augmented reality (AR) system for enhanced visual inspection.* Structural Health Monitoring 2019, 2019.
22. Zhao, W., et al., *Multicriteria Ship Route Planning Method Based on Improved Particle Swarm Optimization–Genetic Algorithm.* Journal of Marine Science and Engineering, 2021. **9**(4): p. 357.
23. Chen, J., et al., *Research on fuzzy control of path tracking for underwater vehicle based on genetic algorithm optimization.* Ocean Engineering, 2018. **156**: p. 217-223.
24. Liu, C., et al., *An Improved A-Star Algorithm Considering Water Current, Traffic Separation and Berthing for Vessel Path Planning.* Applied Sciences, 2019.
25. Sun, K., et al., *Optimal Path Planning Method of Marine Sailboat Based on Fuzzy Neural Network.* Journal of Coastal Research, 2019. **93**(SI): p. 911-916.
26. Tang, C., et al., *A method for compressing AIS trajectory data based on the adaptive-threshold Douglas-Peucker algorithm.* Ocean Engineering, 2021. **232**: p. 109041.
27. Zhao, L. and G. Shi, *A method for simplifying ship trajectory based on improved Douglas–Peucker algorithm.* Ocean Engineering, 2018. **166**: p. 37-46.
28. Agrawal, R., C. Faloutsos, and A. Swami. *Efficient similarity search in sequence databases*. 1993. Berlin, Heidelberg: Springer Berlin Heidelberg.
29. Zhang, T., R. Ramakrishnan, and 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*. 1996, Association for Computing Machinery: Montreal, Quebec, Canada. p. 103–114.
30. Agrawal, R., et al., *Automatic subspace clustering of high dimensional data for data mining applications.* SIGMOD Rec., 1998. **27**(2): p. 94–105.

List of Figures and Tables

Figure 7 shows the clustering results of Bohai Bay when the traditional DBSCAN algorithm, K-means algorithm and CLIQUE algorithm without inputting direction information are used for AIS data trajectory clustering.

As shown in Figure 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 a comparison from Tianjin to Dalian and b is a comparison from Baiyuquan to Tianjin.