Ship Classification Method for Massive AIS Trajectories Based on GNN

Tianyu Li, Huiqi Xu and Weigui Zeng
Coastal Defense Academy, Naval Aeronautical University, Yantai, Shandong, China

Abstract. Since criminals and maritime terrorism may tamper with AIS data and make the track suspicious, it is urgent to classify ships accurately and improve maritime navigation safety. Ship classification based on trajectory data can make up for the deficiency of traditional radar identification and optical identification which has important academic significance and practical value. The target recognition technology based on the traditional neural network can only process conventional Euclidean structure data, while the emerging graph neural network shows great advantages in processing non-Euclidean structure data. The ship trajectory data has the characteristics of the time and space domain and shows a non-Euclidean structure; therefore this paper proposes a classification and recognition method based on the graph neural network to process ship AIS data. First of all, the ship trajectory data is preprocessed and converted into graph data with vertices and edges. Then we use GNN to classify 4 types of ships including fishing vessels, passenger ships, oil tankers, and container ships. Finally, we compare the results with the SVM method. And it shows that this method is valid and proves that it is an effective method of ship classification.

Keywords. AIS; ship classification; neural network; GNN.

1. Introduction
Ship classification is widely used in both military and civilian fields, such as the detection of illegal ships, alertness to maritime terrorism, identify spy ships hidden as civilian ships, and combating smuggling by relevant departments [1]. At present, research methods from China and abroad for the classification of ship types are mainly based on traditional radar recognition and optical recognition, but these methods all have their limitations. For example, optical recognition relies on video surveillance equipment, which field of vision is limited and range of action is short. It is easily affected by meteorological factors such as rain and fog, especially under the meteorological conditions such as high humidity and low clouds at sea.; Although radar recognition is less affected by the environment, it has the problem of “visible but unclear”. It is easy to produce co-frequency interference clutter in a complex electromagnetic environment. In military applications, once the radar is turned on, it is easy to be detected and the safety is threatened. However, classification and identification of ships based on AIS data is less affected by the weather and it can automatically identify the status of the ship around the clock. More than that, this method has other advantages, such as it is not easy to be exposed by enemy reconnaissance., the data collection accuracy is high, and static data such as voyages and the attributes of the ships can be collected. In summary, AIS data is of great significance to the classification and identification of ships. AIS data has a large amount of data and a wide coverage area, and its classification and identification have certain challenges.

Traditional research methods mainly include clustering algorithms based on the distance between track points, machine learning algorithms after manually extracting features, and classification methods
based on neural networks. In recent years, compared with the method of combining manual feature extraction and machine learning, deep neural networks have gradually become a research hotspot. The neural networks currently used to classify ship trajectories mainly include traditional recurrent neural networks and convolutional neural networks such as CNN, MCDCNN, 1DCNN and RNN [2-13]. However, the data processed by the Convolutional Neural Network (CNN) is in the form of a matrix, which is based on the matrix arranged by the samples and it belongs to the Euclidean structure. If we regard the characteristics of the sample as nodes, then the traditional neural network nodes are independent from each other, and it does not make use of the connections between nodes. Otherwise, Recurrent Neural Network (RNN) is based on time series modeling, and its shortcomings are lack of sample characteristics and unable to use the connections between features recommended by different samples. This leads to incomplete feature learning and unsatisfactory classification results.

Graph Neural Networks (GNN) is a deep learning model for processing graph data. It can effectively use the feature connections between different samples, and it is popular in social networks, knowledge graphs, molecular chemistry and other fields. In the above application scenarios, the data has an obvious graph structure. Reference [14] used Graph Neural Network to classify and recognize signal modulation. This method effectively improves the accuracy of modulation type recognition when the signal-to-noise ratio is small.

This paper proposes a model based on GNN to realize ship classification. Ship trajectory information itself does not have a graph structure, but it contains temporal and spatial features, such as location features, distance features, speed features, etc. It can extract the connection relationship information between track points to establish a topological association network, which can effectively extract spatial features for machine learning. The technical route is as follows. First of all, we construct the mapping between the track point data with the vertices and edges in the graph structure. Then we extracted the key features of the vertices and assigned the weights of the edges. Therefore, the track point data is transformed into a graph data structure. Finally, we input GNN for training to realize the classification and recognition of ship types.

2. Data Preprocessing
The original data set is the AIS data of fishing vessels, passenger ships, oil tankers, and container ships in a certain area of the South China Sea for one year in 2020.

2.1. Build Ship Feature Table
After constructing the ship feature database, we extracted the 6 attributes of IMO, timestamp, heading, speed, latitude of the track point, and longitude of the track point from the AIS data as the value. And we used the IMO of the ship as the key value, which means that the trajectory characteristics of each ship are saved according to the IMO number. The track point data of each IMO is arranged in order of time stamp.

2.2. Data Cleaning
After data analysis, the dirty data that meets the data cleaning conditions are discarded. The dirty data mainly includes the abnormal position data and the redundant position data. Among which the abnormal position data means that the distance between two adjacent track points is too large when the time interval is short. And the redundant position data refers to the data features and attributes of two adjacent track points are exactly the same. The algorithm is as follows:

(1) For data with the same key, we calculated the time interval and the distance interval between the i+1th track point and the i-th track point. Considering the curvature of the earth, we use the Haversine formula as the distance calculation formula, as shown in equation (1). And the distance calculated by this formula is referred to as the Haversine distance, as shown in equation (2).

\[
haversin(\varphi) = \sin^2 \left(\frac{\varphi}{2}\right) = \frac{1 - \cos \varphi}{2}
\]  

(1)
\[ l = 2R \cdot \arcsin(\sqrt{\text{haversin}(x_{\text{lat}_1} - x_{\text{lat}_2}) + \cos(x_{\text{lat}_1})\cos(x_{\text{lat}_2})\text{haversin}(y_{\text{lon}_1} - y_{\text{lon}_2})}) \]  

In equation (2), \( l \) represents the distance between the two track points; \( R \) represents the radius of the earth, generally 6371Km; \( x_{\text{lat}_1} \) represents the latitude of point \( x_1 \), \( y_{\text{lon}_1} \) represents the longitude of point \( y_1 \), \( x_{\text{lat}_2} \) represents the latitude of point \( x_2 \), and \( y_{\text{lon}_2} \) represents the longitude of point \( y_2 \).

If the Haversine distance between two track points in a short time interval is too large, the \( i+1 \)th track point to the \( n \)th track point of the IMO ship is the abnormal position data and need to be discarded, where \( n \) represents the number of track points of the IMO ship under the same time window.

(2) If the \( i+1 \)th data point is exactly the same as the \( i \)-th data point, then the \( i+1 \)th track point is the redundant position data and needs to be discarded.

3. Ship Classification Model Based on GNN

3.1. The Graph

The graph is composed of edges and vertices, where edges are represented by \( e \) and vertices are represented by \( v \), as shown in figure 1. Each vertex in the graph contains its characteristics. The characteristics of the vertex can be represented by an \( X \times Y \)-dimensional matrix \( M \), where \( X \) represents the number of vertices, and \( Y \) represents the feature dimension of the vertex; The edges represent the relationship between the vertices, which is represented by \( X \times X \) dimensional matrix \( B \), called adjacency matrix. Matrix \( M \) and matrix \( B \) are the input of the graph neural network.

![Figure 1. Schematic diagram of the graph.](image)

3.2. Graph Data Construction

3.2.1. Build A Sample Database of Track Data. The time interval of AIS data collection is irregular. Some adjacent track points are separated by a few seconds, and some are separated by a few minutes or even more than ten minutes. The longer the time interval for collecting data, the worse the sample quality. Therefore, it is necessary to choose an appropriate and as small as possible time interval threshold \( TT \) (Time threshold) to ensure sample reliability.; The number \( N \) of track points contained in the sample segment determines the accuracy of classification and recognition. The more \( N \), the higher the accuracy of recognition. Therefore, it is necessary to choose a suitable and as many \( N \) as possible to ensure the validity of the sample. Generally speaking, the larger the time interval threshold, the more the number of track points in the sample segment, that is, the reliability and validity of the sample are a pair of contradictions. Therefore, it is necessary to find a balanced state to make the sample have more \( N \) under the premise of a smaller \( TT \). After testing, the experimental data used in this paper can meet the needs when the parameters \( TT=20 \) (which unit is seconds) and \( N=160 \) are selected.

Ships with the same key may have multiple trajectory segments. One trajectory segment is a sample, and the number of track points in a sample is \( N \). According to the sampling principle of \( TT=20 \), \( N=160 \), four types of ship trajectory samples are extracted, and the trajectory samples are saved as a three-dimensional matrix which called Sample_Matrix. The first dimension of the matrix is the number of
samples, the second dimension is the number of trajectory points per sample, and the third dimension is sample attributes, including (IMO, h, v, t-stamp, lat, lon). In which h is the heading, v is the speed, and t-stamp is the time Stamp, lat is the latitude value, and lon is the longitude value.

3.2.2. Build Topological Structure Graph Data. The total set of ship trajectory sample data is transformed into topological structure graph data, and the vertices and edges are described by appropriate features. Therefore, the adjacency matrix is constructed. Only by selecting appropriate feature data as the input of the neural network can the effectiveness of ship trajectory classification be improved. Through the research and analysis of traditional algorithms, heading and speed features are effective features for ship classification [15]. Therefore, this paper proposes to use heading features as vertical features and speed features as edge weights to construct an adjacency matrix. The specific process is as follows:

(1) Determine the receptive field. Calculate the mean Haversine distance \( \bar{l}_{ij} \) between all samples in the total sample set.

\[
\bar{l}_{ij} = \frac{\sum_{n=1,N}(l_{S_{in}}-l_{S_{jn}})}{N}
\]  

(3)

In equation (3), \( l_{S_{in}} - l_{S_{jn}} \) is the Haversine distance between the NTH track point in trajectory \( S_i \) and the NTH track point in trajectory \( S_j \).

Points with small Haversine distance indicates a strong distance feature connection relationship between vertices, and points with large Haversine distance indicates a weak distance feature connection relationship between the vertices. Sort the Haversine distance in order from small to large, and keep the data with the first 5% of Haversine distance value to ensure the strong connection between vertices, where the data exactly equal to 5% is the threshold of the relationship strength. The strong connection relationship is represented by 1, and the weak connection relationship is represented by 0. The relationship matrix \( R \) based on the spatial distance connection strength feature is constructed to determine the receptive field of the vertex. The dimension of the relationship matrix \( R \) is \( X \times X \), and \( X \) represents the number of samples which can also be called the number of vertices in the graph.

\[
R(i,j) = \begin{cases} 
0 & \text{Others;} \\
1 & \bar{l}_{ij} < Thr;
\end{cases}
\]  

(4)

In equation (4), \( Thr \) represents the threshold of relationship strength, and \( \bar{l}_{ij} \) is the average Haversine distance between samples.

(2) Define the weight of the edge. As shown in equation (5), The two-norm of the average speed difference between any two samples is calculated according to the speed characteristics, as the weight of the edge connecting the vertices, and the weight matrix \( E \) of the edge is constructed, whose dimension is \( X \times X \).

\[
E(i,j) = \text{norm}(\text{ave}_{v_{S_i}} - \text{ave}_{v_{S_j}})
\]  

(5)

In equation (5), \( \text{ave}_{v_{S_i}} \) is the average sailing speed of trajectory \( S_i \), and \( \text{ave}_{v_{S_j}} \) is the average sailing speed of trajectory \( S_j \).

(3) Construct an adjacency matrix. Multiply the weight matrix \( E \) of the edge by the relationship matrix \( R \) based on the characteristics of the spatial distance connection strength, and obtain the adjacency matrix \( B \). In equation (6), whose dimension is \( X \times X \).

\[
B = R \cdot E
\]  

(6)

After normalizing to \( B \), it shows in equation (7).

\[
B(i,j) = \frac{B(i,j) - \min (B)}{\max (B) - \min (B)} + \min (B)
\]  

(7)
In equation (7), min\((B)\) is the minimum value in matrix \(B\), and max\((B)\) is the maximum value in matrix \(B\).

(4) Determine the vertex characteristics. The heading feature of the track point is extracted as the vertex feature, and the vertex feature matrix \(M\) is constructed, the dimension of which is \(X \times 1\).

3.3. Network Structure
The GNN network structure proposed in this paper is shown in figure 2. The dots in the graph represent vertices, and different colors represent different labels. The graph neural network training process is to perform gradient dimensionality reduction training on the weight matrix \(W_t\). The input graph data features are passed through several layers of GNN, and the classification data with different labels is output. The relationship between the input and output of the graph neural network is as follows:

\[
h_i = \sigma \left( \sum_{n \in \text{Neigh}(n_i)} \overline{LAPRAS}(i,j)(W_t h_j) \right)
\]  

(8)

\[
\text{Figure 2. Schematic diagram of the neural network structure.}
\]

In equation (8), \(h_j\) is the feature value of vertex \(j\) in the input data, \(h_i\) is the feature value of output data vertex \(i\), \(\sigma\) is the activation function, and \(n_j \in \text{Neigh}(n_i)\) represents the value range of vertex \(j\) as the receptive field of vertex \(i\), \(W_t\) is the convolution kernel of graph convolution, \(\overline{LAPRAS}\) is the normalization of Laplacian matrix:

\[
\overline{LAPRAS}(i,j) = \overline{A}^{-1/2} \overline{B} \overline{A}^{1/2}
\]  

(9)

In equation (9), \(\overline{B} = \overline{B} + I, \overline{B}\) is the normalized adjacency matrix, and \(I\) is the identity matrix. \(\overline{A}\) is the degree matrix of \(\overline{B}\), which formula is \(A_{ii} = \sum_j B_{ij}\).

3.4. Model Training
The classification model GNN is based on the win7 operating system and is implemented using the python programming language, whose bottom layer is TensorFlow. The model training platform processor is 64 cores and the memory is 32G. Samples are divided into training sets and training sets. After all the sample data are scrambled, we take 80% as the training set to train the network model, and 20% as the test set to verify the effectiveness of the classification model. The model used a cross-entropy function and Adam optimization algorithm, the number of iterations is 2000, and the learning rate is 0.001.

As shown in figure 3, the ship trajectory recognition method based on graph neural networks is mainly divided into two parts: network training and network testing. The network training part inputs
the constructed training set graph data into the GNN network for machine learning. The output is the trained graph neural network; The network test part input the test set graph data into the trained graph neural network for vertical classification, and the classification results of each vertex to be tested, which is the ship category. This part was obtained to test the accuracy of the model.

4. Test Analysis

4.1. Evaluation Index

The classification of samples is measured by four indicators: Accuracy, Precision, Recall, and F1 value. Accuracy rate refers to the proportion of the number of samples that are correctly predicted to the total number of samples, as shown in equation (10), where $n_{\text{correct}}$ is the number of input sample categories that are the same as the input category after model identification, and N is the total number of samples:

$$\text{Acc} = \frac{n_{\text{correct}}}{N}$$  \hspace{1cm} (10)

Precision refers to the proportion of the number of samples that are correctly predicted to the total number of all predicted samples, as shown in equation (11), where $N_{\text{pre}}$ is the number of samples predicted by the model:

$$\text{Pre} = \frac{n_{\text{correct}}}{N_{\text{pre}}}$$  \hspace{1cm} (11)

Recall rate refers to the proportion of the number of samples that are correctly predicted to the total number of samples that should be predicted, as shown in equation (12), where $N_{\text{true}}$ is the total number of samples:

$$\text{Rec} = \frac{n_{\text{correct}}}{N_{\text{true}}}$$  \hspace{1cm} (12)

F1 value refers to the harmonic average of precision rate and recall rate, as shown in equation (13):

$$F_1 = \frac{2 \cdot \text{Pre} \cdot \text{Rec}}{\text{Pre} + \text{Rec}}$$  \hspace{1cm} (13)

4.2. Experimental Results

(1) As shown in tables 1 and 2, passenger ships have the highest recognition Precision and Recall rates, both reaching more than 90%. They are the easiest to be identified among the four ship types. This may be due to the fact that passenger ships are the most stable sailing and generally travel at a constant speed. The speed is high and the parking time is small during the driving, so the characteristics are the most obvious. (2) The fishing boat’s recognition Precision rate is 91.0%, and the Recall rate is 88.7%. It is easier to identify among the four types of boats. This may be because the fishing boat has different
movement modes according to different working conditions. For example, when it is in a fishing state, the trajectory of the ship is tortuous, the speed is low, and it is often at a berth; when it is in the sailing state, the trajectory is smooth and the speed is high, so the trajectory characteristics are the most obvious. (3) Tanker identification has the lowest recall rate, with 25.1% of oil tankers classified as container ships. The precision rate of container ships is the lowest. 15.3% of container ships are classified as oil tankers. This may be due to the fact that oil tankers and container ships are all merchant ships. The two ship types not only travel in close areas but also have similar speed and heading characteristics, which is prone to classification errors. (4) The precision rate and recall rate of oil tankers and container ships have opposite trends, so it is necessary to use another parameter F1 to measure. As shown in table 3, the F1 value of passenger ships is the highest, followed by fishing vessels. The F1 value of both tankers and container ships does not exceed 81%, indicating that tankers and container ships are easy to confuse. In addition, tankers have the lowest F1 value, indicating that the oil tanker is the most difficult to identify. According to equation (8), the accuracy of GNN ship trajectory classification method is 82.7%.

Table 1. Precision rate of GNN ship type recognition.

|       | Fish | Passenger | Tanker | Container |
|-------|------|-----------|--------|-----------|
| Precision | 0.910 | 0.941 | 0.757 | 0.715 |

Table 2. Recall rate confusion matrix.

| Predicted label | True label | Fish   | Passenger | Tanker | Container |
|-----------------|------------|--------|-----------|--------|-----------|
| Fish            | 0.887      | 0.023  | 0.043     | 0.047  |
| Passenger       | 0.032      | 0.916  | 0.028     | 0.024  |
| Tanker          | 0.024      | 0.027  | 0.698     | 0.251  |
| Container       | 0.032      | 0.007  | 0.153     | 0.808  |

Table 3. F1 value.

|       | Fish | Passenger | Tanker | Container |
|-------|------|-----------|--------|-----------|
| F1    | 0.898 | 0.928 | 0.726  | 0.759 |

4.3. Comparative Experiment

The method in this paper is compared with the classification of ships based on SVM. SVM classifies high-dimensional nonlinear sample features through hypersurfaces. The comparison results are shown in table 4. The accuracy of the GNN-based ship trajectory classification method is 17.3% higher than that of the SVM-based classification method. The GNN model classification effect is obvious, which proves that GNN can effectively identify ship types.

Table 4. Comparison of GNN and SVM classification results.

|     | Accuracy |
|-----|----------|
| GNN | 0.827    |
| SVM | 0.654    |

5. Conclusion

This paper proposes a classification and recognition method of ship trajectory types based on graph neural network GNN. The heading feature of the track point is used as the vertical feature of the graph data. The speed feature determines the weight of the edge connecting the vertices, and the distance feature between the track points is used as the connection strength of the edge to determine the receptive field. So far, the adjacency matrix based on vertex features and spatiotemporal features can be
constructed. This method uses machine learning to autonomously extract sample features and has strong generalization ability. It can be used to process non-Euclidean data structures, and can effectively use the feature connections between different vertices, which improves the accuracy of classification and recognition of ship trajectories. To be honest, the use of trajectory features to classify ships has limitations and one target recognition method cannot fully support the technical personnel’s judgment on targets. This research provides a type of data source for multi-source feature fusion target recognition of marine vessels.

However, this article does not consider ship size characteristics in the classification, which can affect the recognition rate of tankers and container ships. In addition, the method of constructing an adjacency matrix needs to be further optimized. The main follow-up work direction is to find ship characteristics that are more suitable for graph neural network input; to analyze the characteristics of cargo ships such as tankers and container ships, optimize algorithms, and improve the accuracy of classification and recognition of ship trajectory features based on GNN network.

References
[1] Xu P Y 2017 Research on the trajectory analysis of marine transportation ships China Water Transport (2nd Half) 17 (2) 16-17.
[2] Endo Y, Toda H, Nishida K and Kawanobe A 2016 Deep feature extraction from trajectories for transportation mode estimation Pacific-Asia Conference on Knowledge Discovery and Data Mining (Springer Cham) pp 4-66.
[3] Bi S G, Zhu X M, Zhou Y, Jiang Z W, Yu X H and Ling H 2019 Classification and identification of moving targets at sea 5th International Conference on Big Data and Information Analytics (BigDIA) pp 176-184.
[4] de Souza, E N, Boerder K, Matwin S and Worm B 2016 Improving fishing pattern detection from satellite ais using data mining and machine learning PloS One 11 (7).
[5] Laxhammar R and Falkman G 2011 Sequential conformal anomaly detection in trajectories based on hausdorff distance 14th International Conference on Information Fusion pp 1-8.
[6] Ljunggren H 2018 Using deep learning for classifying ship trajectories 21st International Conference on Information Fusion (FUSION) pp 2158-2164.
[7] Bomberger N A, Rhodes B J, Seibert M and Waxman A M 2006 Associative learning of vessel motion patterns for maritime situation awareness 9th International Conference on Information Fusion pp 1-8.
[8] Rhodes B J, Bomberger N A and Zandipour M 2007 Probabilistic associative learning of vessel motion patterns at multiple spatial scales for maritime situation awareness 10th International Conference on Information Fusion pp 1-8.
[9] Hebb D O 1949 The Organization of Behavior: A Neuropsychological Theory (New York: Wiley).
[10] Zheng Y, Liu Q, Chen E, Ge Y and Zhao J L 2014 Time series classification using multi-channels deep convolutional neural networks International Conference on Web-Age Information Management pp 298-310.
[11] Górecki T and Łuczak M 2015 Multivariate time series classification with parametric derivative dynamic time warping Expert Systems with Applications 42 (5) 2305-2312.
[12] Cui Z, Chen W and Chen Y 2016 Multi-scale convolutional neural networks for time series classification arXiv:1603.06995.
[13] Wang Z, Yan W and Oates T 2017 Time series classification from scratch with deep neural networks: A strong baseline 2017 International Joint Conference in Neural Networks (IJCNN) pp 1578-1585.
[14] Liu Y, Liu Y and Yang C 2020 Modulation recognition with graph convolutional network IEEE Wireless Communications Letters 9 (5) 624-627.
[15] Cui T T, Wang G L and Gao J 2020 Ship trajectory classification method based on 1DCNN-LSTM Computer Science 47 (9) 175-184.