ISSN: 2710-0588

Discussion on the Following Distance of Ships in the Deep Water Channel of the Yangtze Estuary based on AIS data

Bo Tu

Merchant Marine College, Shanghai Maritime University, Shanghai, China

Abstract

Ship intelligent navigation has been a research hotspot in the field of port and navigation, and ship following pitch is an important aspect of ship intelligent navigation. When the ship is sailing in the ocean, the ship can keep a large enough spacing. However, in the Yangtze River Estuary Channel, the number of ships passing through the same channel at the same time is large, and the ships cannot maintain an arbitrarily large tracking distance. When using the AIS data to make statistics on the heeling distance of the deep water channel in the north channel of the Yangtze River estuary, it was found that the heeling distance of ships did not have a positive linear relationship with the ship length. In this paper, we analyze several key factors that affect the vessel tracking distance, and through the random forest algorithm, we can predict the spacing of different vessels entering the deep water channel of the Yangtze River estuary.

Keywords

Big Data; AIS; Ship Following Gap; Random Forest Algorithm.

1. Introduction

Vessel tracking spacing is an important part of intelligent navigation. When a ship navigates in narrow waterways and divided navigation system waters, it becomes more important for the ship to maintain a safe following spacing. The study of ship spacing first appeared in the field of ships, and Fujii [1] believed that the ship field was related to traffic density, ship speed, and sea area, and the longitudinal spacing of ships should be kept more than 7 nautical miles. British scholar Good-win [2] made further research on ship domain based on international maritime collision avoidance rules and proposed a ship domain model consisting of three sectors, the ship domain of 300-400m in length, 1.2 nautical miles for the bow part and 0.6 nautical miles for the stern part. The British scholar DAVIS [3] proposed the concept of dynamic sector based on the study of ship domain, and believed that there exists a super domain around the ship as a circle with a diameter of 5.4 nautical miles.Hense [4] et al. used AIS data to analyze the ship domain, and obtained a comfortable ship domain model for the open water of southern Denmark through 4 years of observation and statistics, and believed that the ship longitudinal direction should be kept at 8 knots, which will make the driver feel most comfortable. Domestic He Liangde [5], Hou Haiqiang [6], Ming Li and others [7] have established the ship following spacing model based on ship characteristics combined with road following theory [8], but marine ships are less maneuverable than road vehicles, which is not enough to meet the actual situation. In this paper, a large amount of AIS data is used to find out the following spacing of different ships at different moments. The statistical analysis was done to find out the relationship between the following pitch and the ship length, ship speed, ship type, relative ship speed with the previous ship, and find out the importance between them.

ISSN: 2710-0588

2. Vessel Tracking Distance and its Influencing Factors

2.1. Measure of the Sample of Ship Following Distance

The ship following distance defined in this paper is the actual distance between the stern of the previous ship and the bow of the ship at the same time. In the deep water channel of the Yangtze River estuary, the channel is a divided navigation system, and the ship is not allowed to chase over or cross the channel. In this paper, firstly, the ship's latitude and longitude are converted into coordinates under X-Y coordinate system, and the positive half-axis of X-axis represents the longitudinal distance of the ship moving from the origin, and the positive half-axis of Y-axis represents the lateral distance of the ship moving from the origin. Based on the change of time, the ship spatio-temporal trajectory map of the ship in X-axis direction is drawn, and then the interval of 20s time is used to make truncation line on the spatio-temporal trajectory, and the distance between the two ships in X-axis direction before and after the same moment is the spacing.

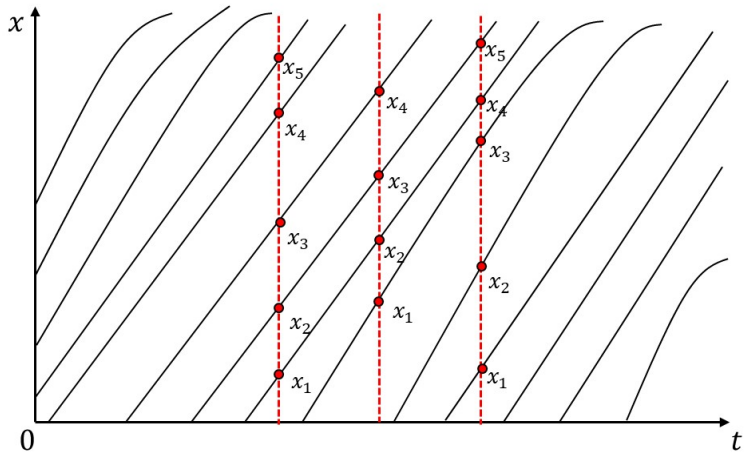

Figure 1. Ship following gap definition

The ship following distance satisfies the following relation:

$$Gap = X_1 - X_2 - P_1 - P_2 \tag{1}$$

Where Gap denotes the heeling distance of the front and rear ships, X_1 denotes the distance of the front ship in the longitudinal direction, X_2 denotes the distance of the rear ship in the longitudinal direction, P_1 denotes the distance of the AIS antenna of the front ship from the transom of the front ship [9], and P_2 denotes the distance of the AIS antenna of the rear ship from the bow of the rear ship.

2.2. Influencing Factors of Ship Following Distance

The influencing factors for the ship to keep a certain distance between the ships are mainly the length of the ship, the speed of the ship to the ground, the relative speed between the ship and the ship in front of the ship and the type of the ship. These factors, to a certain extent, determine the spacing between the ship and the vessel in front of the tracking. In the study of the spacing between ships in the deep water channel of Yangtze River Estuary, the influencing factors are selected as shown in Table 1.

ISSN: 2710-0588

Table 1. The factor of ship following gap

Influencing factor	Symbol	Unit
Length of boat	L	m
Speed over ground	V	kn/h
Relative speed to previous ship	RV	Kn/h
Type	T	dimensionless

3. Data Processing and Experimental Methods

In this paper, we adopt the following approach: firstly, we select the AIS data of the desired area, then we preprocess the data, then we normalize the data features and discretize some attributes [10], then we divide the data into training and prediction sets, train the random forest model, and finally we analyze the results.

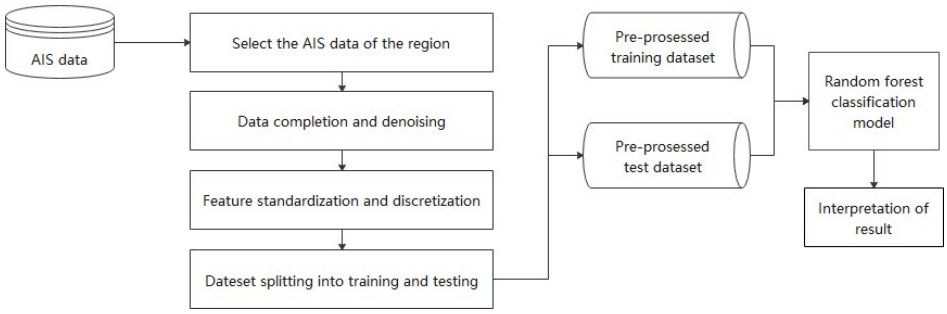

Figure 2. Data processing

3.1. Data Source and Selection

The data source of this paper is from the AIS data of the East China Sea region of Shanghai Maritime Bureau for one year in 2018, and selected by using the algorithm, a rectangular area with a vertical distance of 4 nautical miles and a horizontal distance of its specific location as shown in Figure 3. The navigable water depth of this water area is more than 12.5, and the total AIS data for the whole year of 2018 is more than 1.6 million, and the annual volume of passing ships is 23,852, among which cargo ships and container ships account for more than 70%.

The data source of this paper is from the AIS data of the East China Sea region of Shanghai Maritime Bureau for one year in 2018, and the algorithm is used to select the deep water channel of the North Channel $31°6'.897-31°9'.24N$, $122°13'.232-122°17'.136E$, a rectangular area with a vertical distance of 4 nautical miles and a horizontal distance of its specific location as shown in Figure 3. The navigable water depth of this water is more than 12.5, and the total AIS data for the whole year of 2018 is more than 1.6 million, and the annual volume of passing ships is 23,852, among which cargo ships and container ships account for more than 70%.

Figure 3. Data position

ISSN: 2710-0588

3.2. Basic Data Analysis

In this area, there are seventeen types of passing ships in total, among which the proportion of bulk carriers is the highest, accounting for about 35% of the total data; the length of passing ships is concentrated between 100-300m, the width of ships is concentrated above 20m, and the speed of ships is concentrated between 6-14kn.

3.2.1. Distribution of Ship Types

Figure 4 shows the distribution of ship types in the study area. From the results of statistical analysis, it can be seen that: the highest proportion of cargo ships is about 50%, followed by container ships accounting for about 25%, followed by petrochemical ships accounting for about 9%. Due to the different types of ships, the maneuverability of ships varies greatly. In the actual navigation, the following distance with the previous ship will also be different to some extent. Therefore, it is also very important to analyze the influence of ship type on ship following distance. In this paper, according to the proportional distribution of each type of ships, the ships are divided into cargo ships, container ships, petrochemical ships, tugboats, engineering service ships, ro-ro ships, passenger ships and other classes, in total 8 types.

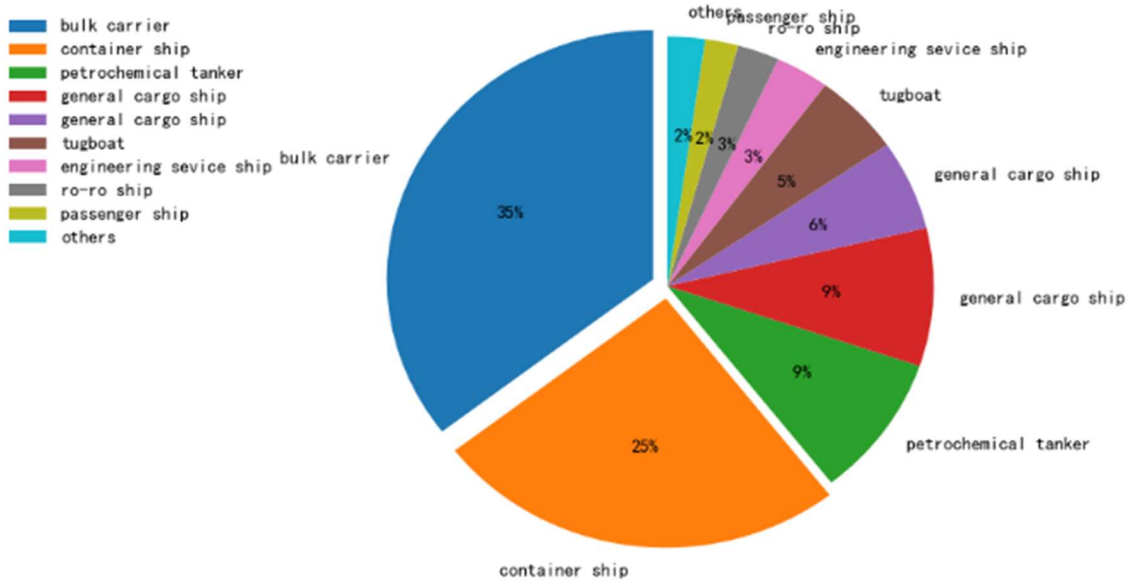

Figure 4. Ship type

3.2.2. Distribution of Ship Length and Width

Figure 5 and Figure 6 show the distribution of ship length and width in the selected waters of the northern trough of the Yangtze River. The ship length and width reflect the size of the ship tonnage to a certain extent, and also determine the maneuvering performance of the ship to a certain extent. When the ship length is long, it tends to keep a larger distance with the previous ship, but in the actual navigation, how big is their actual association is also a more important issue. Most of the ships in the selected area of this paper are concentrated in the length of 100-300m, and the width of the ships is concentrated in more than 20m. Therefore, when classifying the ships in this paper, the ship lengths are divided into six categories of ships: 0-100m, 100-150m, 150-200m, 200-250m, 250-300m and more than 300m.

ISSN: 2710-0588

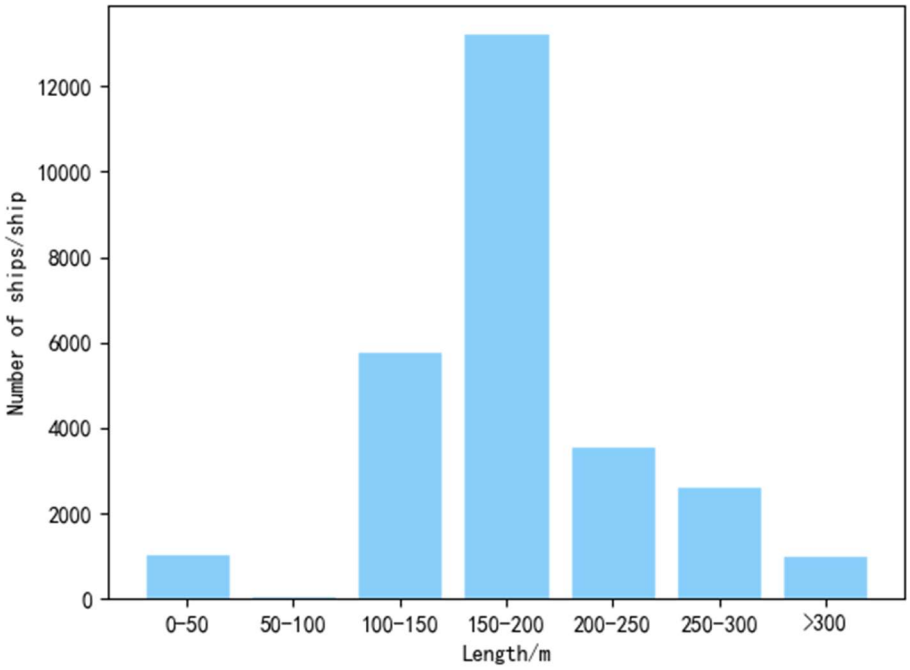

Figure 5. Ship length

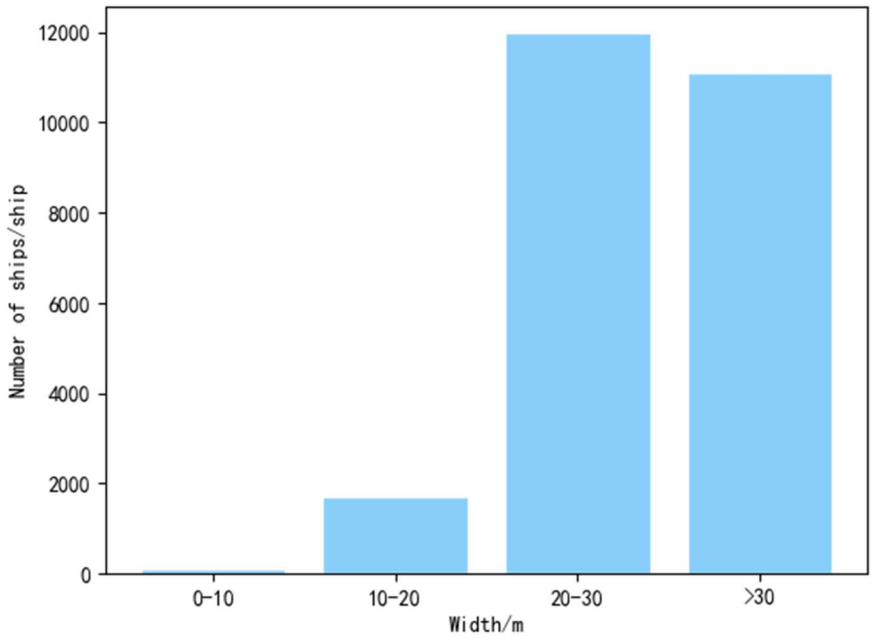

Figure 6. Ship width

3.2.3. Vessel Speed Distribution

Figure 7 shows the speed distribution of ships passing through the selected area in this paper, and it can be seen from the figure that there are especially few ships with speed less than 4kn and speed more than 16kn, and most of the ships' speed is concentrated in 6-14kn. types, and filter the corresponding data for later analysis.

ISSN: 2710-0588

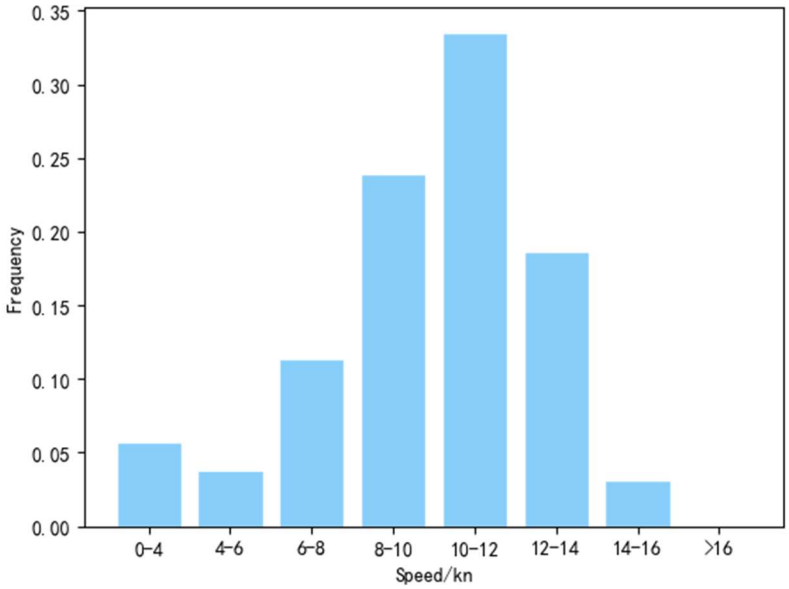

Figure 7. Ship speed

3.2.4. Data Attributes Standardization and Discretization

The main attributes of this paper are shown in Table 3, where the ship length, ship speed to ground, relative speed and pitch are quantitative data, and the ship type and pitch to ship ratio are qualitative data. For quantitative data 2.3.2 Data attributes standardization and discretization.

The main attributes of this paper are shown in Table 3, where the ship length, ship speed to ground, relative speed and pitch are quantitative data, and the ship type and pitch to ship ratio are qualitative data. For quantitative data x, the standardized attributes x' are determined by Equation 2.

$$x' = \frac{x - \mu}{\delta} \tag{2}$$

Where μ is the overall mean of each attribute and δ is the overall standard deviation of each attribute. Normalizing the data allows the attributes to contribute equally to the objective function used for model training.

Discrete the qualitative data into 10 categories of ship types and 13 categories of spacing based on the distribution of ship types in the figure above. The data discretization was divided as follows.

1) Ship type: according to the specific types of ships: 1={dry cargo ship}, 2={bulk carrier}, 3={petrochemical ship}, 4={container ship}, 5={general cargo ship}, 6={liquid ship}, 7={tug}, 8={engineering service ship}, 9={ro-ro-ro ship}, 10={other}.

2) Divided by ship spacing:

1=[0,500),2=[500,1000),3=[1000,3001500),4=[1500,2000),5=[2000,2500),6=[2500,3000),7=[3000,3500),8=[3500,4000),9=[4000,4500),10=[4500,5000).

ISSN: 2710-0588

Table 2. The symbol factor of ship following gap

Influencing factor	Symbol	Unit
Length	L	m
Speed over ground	V	kn/h
Relative speed to previous ship	RV	kn/h
Type	T	dimensionless
Gap	Gap	m
Gap/length	$Ratio$	dimensionless

4. Experimental Methods

4.1. Principle of Random Forest Classification Algorithm

The random forest algorithm belongs to one of the machine learning methods. It is a supervised learning algorithm that integrates multiple decision trees. Based on the type of data processing results, random forest can complete both classification and regression applications. The core idea of random forest algorithm is based on the optimal classification and regression of the results of many random decision trees to explain the effect of input variables on explanatory variables. Decision tree is the core of random forest, and the optimal generation of decision tree is guaranteed according to conditional entropy and information gain.

The basic principle is: First, the Bagging method is used to randomly select parameter combinations with different numbers of variables and different numbers of samples from the dataset D with the number of samples N to form numerous random decision trees and determine the classification node of each tree.

Then calculate the information entropy of the data set D, and the conditional entropy between each variable Xi and D. Information entropy can be called the average uncertainty of the data set. Conditional entropy H(D Xi) is the change of entropy value brought by Xi variable under the premise of the occurrence of data set D.

The information entropy of dataset D is H(D):

$$H(D) = E[-\log D_i] = -\sum_{i=1}^{n} D_i \log_2 D_i \tag{3}$$

In the formula: Di is the probability of the occurrence of the i-th data in the data set D.

The joint entropy H(D, Xi) between the data set D and the variables is:

$$H(D, X_i) = -\sum_{i=1}^{n}\sum_{j=1}^{m} p(D, X_i)\log(D, X_i) \tag{4}$$

The conditional entropy between the data set D and the variable is:

$$H(D \mid X_i) = H(D, X_i) - H(X_i) \tag{5}$$

Then, the information gain g(D, Xi) of each variable Xi relative to the data set D is calculated to measure the degree of reduction of the uncertainty of the data set D by the feature Xi. The decrease in uncertainty is the increase in the probability of D occurring. The formula for calculating information gain is as follows:

ISSN: 2710-0588

$$g(D, X_i) = H(D) - H(D \mid X_i) \tag{6}$$

A large number of random decision trees that can be processed in parallel can be generated based on the combination of randomly selected sample variables and parameters, and these random decision trees together form a random decision forest. Each decision tree will give a classification result based on the classification tree according to the data to be classified, and finally summarize the result of the decision tree with the highest repetition in all the forests as the classification result of the random decision forest.

4.2. Experimental Design

Taking the four influencing factors of the ship following distance in the deep water channel of the Yangtze Estuary as the input, the distance is classified by the random forest classification algorithm, and thirteen types of distance are given according to the statistical analysis. The ratio of training samples to prediction samples is 7:3, and the samples are randomly sampled to ensure the extensiveness of the model. The sample training set and test set are shown in Table 3.

Table 3. Sample set and train set

Type	sample set	train set
gap_label_1	90701	30452
gap_label_2	72511	24098
gap_label_3	55134	18556
gap_label_4	43158	14345
gap_label_5	34790	11608
gap_label_6	28741	9487
gap_label_7	24970	8328
gap_label_8	19378	6388
gap_label_9	13807	4576
gap_label_10	12629	4102

4.3. Analysis of Results this Experiment Uses Python Language.

When the number of decision trees is 136 and the number of decision tree nodes is 2, the classification effect of the random forest algorithm is the best, and the verification accuracy of the out-of-bag error is 85.66%. Classification and identification, the classification accuracy of random forest algorithm is 86.12%.

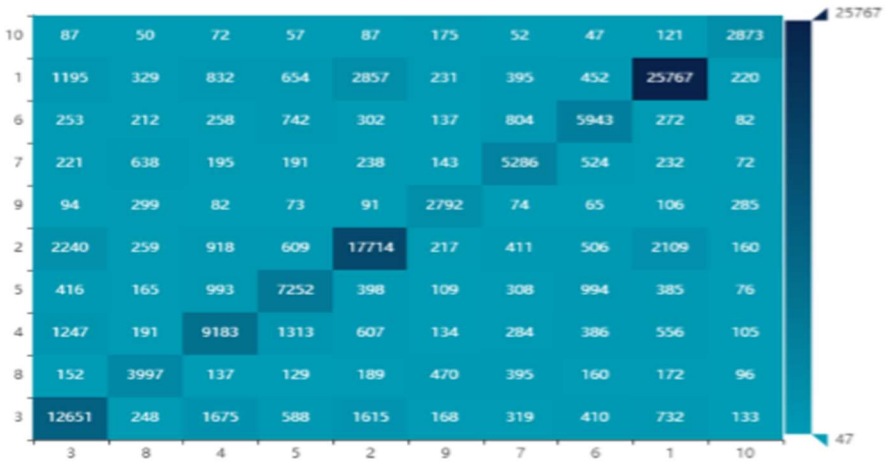

Figure 8. Random forest algorithm training sample prediction results

ISSN: 2710-0588

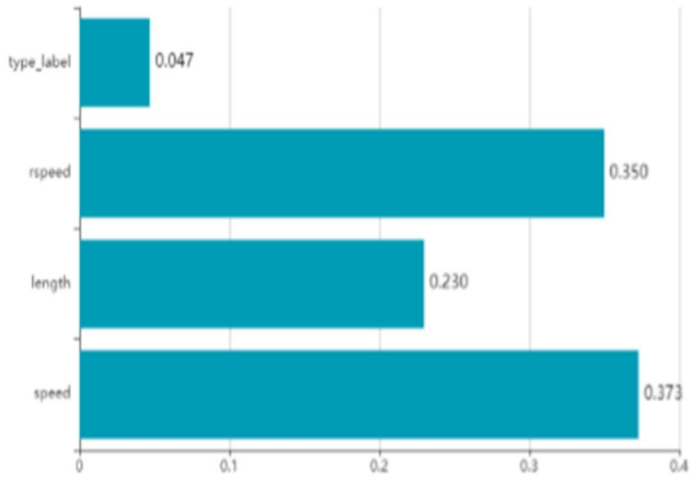

Figure 9. Feature weights of each influencing factor

It can be seen from Fig. 9 that the ship speed has the greatest influence on the following distance, followed by the relative ship speed, and the ship type has the smallest influence on the following distance, and its weight only accounts for 4.7%.

5. Conclusion

When a ship is sailing in the deep water channel of the Yangtze Estuary, the following distance it maintains is related to the length, speed, relative speed and type of the ship, and the speed of the ship has the greatest impact on it, and the ship type has the least impact on it. This paper also uses the random forest classification algorithm to train a random classification model using the one-year ship-following distance of the deep-water channel of the Yangtze Estuary as a training set. It is 86.12%, which is a good performance.

References

[1] Fujii Y . Traffic Capacity[J]. Journal of Navigation, 1971, 24(4):543-552.

[2] Goodwin, E. M . A Statistical Study of Ship Domains[J]. J. of Navigation, 1973, 28(01):130.

[3] Davis P V , Dove M J , Stockel C T . A Computer Simulation of Marine Traffic Using Domains and Arenas[J]. Journal of Navigation, 1980, 33(02):215.

[4] Hansen M G , Jensen T K , Lehn-Schi Ler T , et al. Empirical Ship Domain based on AIS Data[J]. Journal of Navigation, 2013, 66(06):931-940.

[5] He Liangde, Jiang Ye, Yin Zhaojin, et al. Model of the following distance between inland ships[J]. Chinese Journal of Transportation Engineering, 2012(01):55-62.

[6] Hou Haiqiang, Li Yicheng, Chu Xiumin. Model of the distance between ships in the busy waters of the Yangtze River [J]. Journal of Dalian Maritime University (No. 4): 21-24.

[7] Ming Li, Liu Jingxian, Wang Xianfeng. Calculation model for the safe longitudinal spacing of super-large ships [J]. China Navigation, 2014(4):40-43.

[8] Wang Dianhai, Jin Sheng. Review and prospect of vehicle following behavior modeling [J]. Journal of China Highway and Transportation, 2012, 25(001):115-127.

[9] Pan Jin, Wang Yong, Huang Yifei, et al. Research on the evaluation method of ship-bridge collision probability based on AIS data [J]. Journal of Huazhong University of Science and Technology (Natural Science Edition), 2019(11).

[10] Gkerekos C , Lazakis I , Theotokatos G . Machine learning models for predicting ship main engine Fuel Oil Consumption: A comparative study[J]. Ocean Engineering, 2019, 188(Sep.15):106282.1-106282.14.