Ship classification based on random forest using static information from AIS data

Yitao Wang1*, Lei Yang1, Xin Song1 and Xuan Li1

1College of Aerospace Science and Engineering, National University of Defense Technology, Changsha, Hunan, 410073, China
*Corresponding author’s e-mail: ytwang_15@nudt.edu.cn

Abstract. With the wide use of automatic identification system (AIS), a large amount of ship-related data has been provided for marine transportation analysis. Generally, AIS reports the type information of ships, but there are still many ships with type unknown in AIS data. It is necessary to develop algorithms which can identify ship type from AIS data. In this paper, we employ random forest to classify ships according to the static information from AIS messages. Moreover, the importance of static features is discussed, which explains the reason why some classes of ships are misclassified. The method of this paper is proved to be effective in ship classification using static information.

1. Introduction

Automatic identification system (AIS) is a navigation aid system which broadcasts the static, dynamic, voyage and safety information of ships [1]. In order to strengthen maritime safety, the International Maritime Organization (IMO) requires ships to install AIS equipment since 2004 [2]. The wide application of AIS promotes the development of intelligent shipping, and also plays an important role in route extraction [3], pollution prevention [4], maritime traffic feature discovery [5] and ship abnormal behavior detection [6-7].

AIS has 27 kinds of messages, in which message 5 contains the static and type information of ships. The ships type may be missing or mislabelled for various reasons [8], which probably implies the occurrence of illegal activities such as smuggling and illegal fishing [9]. It is necessary to develop the ship classification methods based on AIS data. Some researches have been done to solve this problem. Random forest [10] and KNN [11] is used to recognize ships type from static characteristics (such as ship size and draft). However, the static features considered in these methods are inadequate [10-11] and the AIS data are collected from a small region [11]. Kraus [12] combines static, geographical and dynamic features from AIS data in German Bight to identify six types of ships, while this method has the problem of data leakage.

In this paper, random forest is applied to classify ships by learning the potential pattern of static features from AIS data (message 5) received by the ocean satellite HY-1C/D [13] and HY-2B/C. The rest of this paper is organized as follows. In section 2, we firstly preprocess the static information and extract the static features, then the principle of random forest algorithm is briefly introduced. In section 3, the performance of random forest is evaluated by experiments on five types of ships. In section 4, conclusions and future works are presented.
2. Methodology

2.1. Data preprocessing and feature extraction

Figure 1 shows the quantities of the top 20 kinds of ships in message 1 received by HY-2C. The abscissa in figure 1 is the codes of ship type, and the codes of passenger ships, cargo ships, tankers, fishing boats and tugs are 60–69, 70–79, 80–89, 30 and 52 respectively. The red curve is the accumulated proportion of ship quantity. According to the statistical results in figure 1, the five categories of ships mentioned above account for 93.44% of the total number of ships, and we select them to train a random forest.

![Figure 1. Quantities of different types of ships.](image)

The original AIS static messages contain five fields named A, B, C, D and draught which can reflect the size information of ships. A, B, C and D are the distances from the reference point O used for reporting position to bow, stern, port and starboard respectively. The length and width of a ship can be calculated by equation (1).

\[
\begin{align*}
\text{length} & = A + B \\
\text{width} & = C + D
\end{align*}
\] (1)

Except for the above 7 ship size features, other geometric features are extracted and they are defined as equation (2):

\[
\begin{align*}
\text{ldivw} & = \text{length} / \text{width} \\
\text{area} & = \text{length} \times \text{width} \\
\text{grith} & = \text{length} + \text{width}
\end{align*}
\] (2)

Considering that ships static information can be incorrect, the obviously abnormal data should be removed in pretreatment process. Since the distributions of static data are various in different types of ships, the data of each type of ships need to be processed separately. Taking tankers as an example, its standardized features distribution is shown in figure 2(a). We use box plot to identify and remove outliers, and this method has no requirement on data distribution and is more conservative when filtering outliers. The specific methods are as follows. Firstly, calculate the upper quartile \( Q_u \), lower quartile \( Q_l \) and the interquartile range \( IQR \) of each feature for one type of ships. Secondly, delete the data whose features are outside \([Q_l - 3IQR, Q_u + 3IQR]\). Figure 2(b) shows the data distribution of tankers after removing outliers. Figure 2(c) shows the distribution of original data of five types of ships, and figure 2(d) shows the data distribution of all ships after removing outliers. Table 1 shows the number of static data used in this paper.
2.2. Random Forest
Random forest is an integrated learning algorithm based on decision tree, which has the advantages of low deviation, low variance and strong generalization ability. The method for creating random forest is as follows:

- Randomly select n samples from the training samples, and use these samples to create a decision tree. The decision tree is trained by CART algorithm: m features are randomly selected at each node to be split, and then the feature k and threshold $t_k$ that can minimize the cost function (equation (3)) are selected from the features, using which the sample set at the node is divided. $G_{\text{left/right}}$ is the impure of the left and right subsets (equation (3)), $m_{\text{left/right}}$ is the number of samples divided into left and right subsets, and $P_{i,n}$ is the proportion of class n samples in the i-th node;
- When the decision tree reaches the maximum depth or the cost function cannot be reduced, stop node splitting and terminate the creation of the decision tree;
- Repeat the first step to create a large number of decision trees to form a random forest. The prediction category of the sample is determined by the voting of the decision tree in random forest.

Table 1. Quantities of static messages.

| Type          | Passenger ships | Cargo ships | Tankers | Fishing boats | Tugs |
|---------------|-----------------|-------------|---------|---------------|------|
| Number        | 1,285           | 61,477      | 21,094  | 5,968         | 2,046|

Figure 2. Distribution of original static data and static data after removing outliers.
\[
J(k, t_k) = \left( \frac{m_{\text{left}}}{m} \right) G_{\text{left}} + \left( \frac{m_{\text{right}}}{m} \right) G_{\text{right}} \\
G_j = 1 - \sum_{n=1}^{q} \rho_{i,n}^2
\]

3. Results and discussion

In order to optimize the parameters of random forest and objectively evaluate the performance of 1D-CNN, we use stratified sampling to divide the data set into training set, validation set and testing set according to the ratio of 6:2:2.

The best parameters of the classifier are obtained by grid search. Considering the imbalance of the number of samples, the sample loss is weighted in training process. The weight of class I samples is the ratio of the total number of samples to the number of class I samples. Table 2 shows the best parameters for random forest. The total accuracy and F1 score of the classifier on testing set 0.8614 and 0.8693. Figure 3 shows the confusion matrixes of random forest on the validation set and testing set, where class 0 to class 5 represent passenger ships, tugs, tankers, fishing ships and cargo ships respectively. The model tends to confuse tankers with cargo ships, as well as misclassify passenger ships, tugs and fishing boats.

![Figure 3. (a). confusion matrix on validation set, (b). confusion matrix on testing set.](image)

To explain the confusion matrixes, we use t-SNE [14] to visualize the static features. t-SNE is a method of data visualization, it can map data from high-dimensional space to low-dimensional space. If two samples are similar in high-dimensional space, the distance between their maps in low-
dimensional space will be close. Figure 4 shows the visual results of static features. The data distributions of passenger ships, tugs and fishing boats after dimensionality reduction overlaps partially, which also occurs on cargo ships and oil tankers. This shows that the static information of the misclassified ships is similar. Figure 5 shows the importance of static features. It can be seen that the ship's dimensional characteristics such as A, length and length-width ratio can better describe the ships features than draught. In addition, although the features C and D contribute little to the classifier, these two parameters have been reflected in the length-width ratio, girth, area and width, which proved the static features constructed in this paper are effective in the ship classification task.

4. Conclusion
In this paper, we use random forest to classify ships according to the static information in AIS data. Firstly, the static features are defined and extracted from AIS data, then experiments are conducted to get the best parameters of random forest, finally we attempt to explain the confusion matrix by data visualization. Moreover, the importance of features is analysed, which proved the static features proposed in this paper are effective in ship type identification.

Experiments show that the static features may be similar between some kinds of ships. Using only static characteristics is not enough to distinguish five types of ships. In future work, we plan to combine static and dynamic information from AIS data to improve the performance of ship type classification.

Acknowledgments
HY-1C/D and HY-2B/C data were obtained from https://osdds.nsoas.org.cn. The authors would like to thank NSOAS for providing the data free of charge.

References
[1] Technical characteristics for an automatic identification system using time-division multiple access in the VHF maritime mobile band, Recommendation ITU-R M.1371-4. 2010. 3-4, 101-107. Available online: Template BR_Rec_2005.dot (itu.int).
[2] Tichavská M., Cabrera F., Tovar B., et al. (2015) Use of the Automatic Identification System in Academic Research. In: International Conference on Computer Aided Systems Theory. In: Computer Aided Systems Theory – EUROCAST 2015. Switzerland. pp. 33-40.
[3] Sheng, P., Yin, J. (2018) Extracting Shipping Route Patterns by Trajectory Clustering Model Based on Automatic Identification System Data. Sustainability, 10 (7): 1-13
[4] Ambjörn C. (2008) Seatrack Web forecasts and backtracking of oil spills - an efficient tool to find illegal spills using AIS. In: 2008 IEEE/OES US/EU-Baltic International Symposium. Tallinn, Estonia. pp. 1-9.
[5] Fournier, Mélanie, Casey H.R., Rezaee S., et al. (2018) Past, present, and future of the satellite-based automatic identification system: areas of applications (2004–2016). WMU Journal of Maritime Affairs, 17(3): 311-345.
[6] Rong H., Teixeira A.P., Soares C.G. (2020) Data mining approach to shipping route characterization and anomaly detection based on AIS data. Ocean Engineering, 198:106936.
[7] Castaldo F., Bastani V., Marcenaro L., et al. (2014) Abnormal vessel behavior detection in port areas based on Dynamic Bayesian Networks. In: International Conference on Information Fusion. Salamanca, Spain. pp. 1-7.
[8] Iphar, C., Napoli, A., Ray, C. (2015) Detection of false AIS messages for the improvement of maritime situational awareness. In: OCEANS 2015 - MTS/IEEE, Washington, DC. pp. 1-7.
[9] Longépé, N., Hajduch, G., Ardianto, R., Joux, R. et al. (2018) Completing fishing monitoring with spaceborne Vessel Detection System (VDS) and Automatic Identification System (AIS) to assess illegal fishing in Indonesia. Mar. Pollut. Bull, 131: 33-39.
[10] Zhong H., Song X., Yang L. (2019) Vessel Classification from Space-based AIS Data Using Random Forest. In: 2019 5th International Conference on Big Data and Information Analytics (BigDIA). Kunming. pp. 9-12.

[11] Damastuti, N., Siti Aisjah, A. (2019) Classification of Ship-Based Automatic Identification Systems Using K-Nearest Neighbors. In: 2019 International Seminar on Application for Technology of Information and Communication (iSemantic). Semarang. pp. 331-335.

[12] Kraus, P., Mohrdieck, C., Schwenker, F. (2018) Ship classification based on trajectory data with machine-learning methods. In: 2018 19th International Radar Symposium (IRS). Bonn. pp. 1-10.

[13] Zhang K., Ding Z., Gao J., et al. (2019) Design and Application of AIS on HY-1C/D Satellite. Spacecraft Engineering, 28(05): 26-31.

[14] Maaten, L.V.D., Geoffrey, E.H. (2008) Visualizing Data using t-SNE. Journal of Machine Learning Research, 2579-2605.