Characteristics analysis of ship traffic flow based on Data Mining

Pengfei Qin

School of Merchant Marine, Shanghai maritime University, Shanghai 201306, China.

aqpf1997115@163.com

Abstract

Firstly, the AIS data is cleaned and processed to obtain reasonable and valuable traffic flow data and select the characteristic data to be analyzed. Then it introduces the principle of the intelligent algorithm that needs to be used, expounds the specific process of the analysis, finds out the relationship between the speed and course of different types of ships, which is used to judge the type of unknown ships; finally, it makes an example analysis to demonstrate the scientificity and rationality of the method. Finally, the ship type of unknown ship can be predicted well, which shows the feasibility of the method in practical application.

Keywords

Hip traffic flow; data mining; fitting experiment.

1. Introduction

As the research object of marine traffic engineering or the management object of ship traffic management, marine traffic or ship traffic has its definite meaning, the combination of ship motion and the overall behavior of ships in the designated area [1]. With the increase of port throughput, the number of ships entering and leaving the port is also increasing, and the composition of ships is also changing. In addition, with the continuous growth of international and domestic waterway cargo transportation, the transportation network between ports is becoming more and more complex, and the ship traffic flow in port waters will become more and more complex. It is of great practical significance to study the popularity of ship traffic and put forward reasonable measures and suggestions for it, so as to ensure the navigation safety and improve the navigation efficiency in water area [2]. In reference [3], using data mining theory and technology, the spatial and temporal characteristics of navigation environment and the probability distribution characteristics of heading and speed change rate of main channel in Xiamen Bay are obtained. In reference [4], a clustering algorithm based on density is improved. In reference [5], the comprehensive distance is regarded as the distance between tracks in KNN classification, and finally the ship track classification is realized.

Characteristics of ship traffic flow.The popularity of ship traffic is characterized by macro and micro aspects. The ship traffic flow model includes five basic elements, namely, the location, direction, width, density and speed of ship traffic flow[3]. Ship traffic flow has specific characteristics in the formation and movement stage, which will change with time and place. According to the differences between the whole and individual in the research process, it can be divided into macro characteristics and micro characteristics. Macro characteristics refer to the overall behavior of all ships in a certain water area or time period, including ship flow, composition of traffic flow, space-time distribution characteristics, time distance distribution and waiting time distribution between ships; micro characteristics refer to the individual of a single ship under certain conditions under the influence of other ships and various factors. Behavior characteristics, including ship position, direction, speed, width, ship spacing and ship to ship time.

2. Method introduction

Ship traffic flow data mining is based on the obtained characteristic data of ship traffic flow. The main steps include: data processing, fitting inspection and determination of fitting function. The main process is shown in Figure 1.

2.1 Data processing

The obtained AIS data is preprocessed, and the speed of the ship traffic flow is selected as the research object. The total number of samples is n, the frequency of each speed interval is f_i^* The theoretical probability of normal distribution is pi , The calculation formula of mean μ and variance σ 2 of samples is:

$$\mu = \frac{1}{n}\sum_{i=1}^{k} xif_i^* \tag{1}$$

$$\sigma^2 = \frac{1}{n-1}\left[\sum(xi^2 f_i) - \frac{1}{n}(\sum xif_i^*)^2\right] \tag{2}$$

2.2 Fitting test

In this paper, Chi- square test is used to test whether the velocity distribution of ship traffic flow obeys the normal distribution n (μ, σ2). This method is mainly based on the deviation degree between the actual value and the theoretical value to see whether the set test conditions are met. The samples are divided into group K. because there are two constraints (μ, σ) in the normal distribution, the constraint number γ = 2, so the degree of freedom is:

$$\boldsymbol{DF = k - \gamma - 1} \tag{3}$$

significant level α = 0.05,

$$\chi^2 = \sum \frac{(f_i^* - np_i)}{np_i} \tag{4}$$

if $\chi^2 < \chi^2_{0.05}$, It shows that the difference between the actual value and the theoretical value is not large, and the ship speed of the traffic flow follows the normal distribution. Otherwise, it shows that the difference between the two is large, and other methods need to be selected for fitting test.

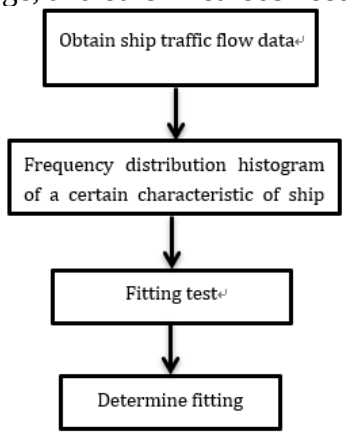

Figure 1 main process

2.3 Data aggregation and classification

Through the research, it is found that the speed and course of different kinds of ships are quite different, so there may be some inherent relationship. In order to find out their relationship, we first use k-means clustering algorithm to have an intuitive feeling, and then use KNN algorithm to predict the speed and course of unknown ships to identify the ship type of unknown ships.

K-means algorithm is a hard clustering algorithm, which is the representative of a typical prototype based objective function clustering method. It is a certain distance from the data point to the prototype as the optimized objective function, using the method of function seeking the extreme value to get the adjustment rules of iterative operation. K-means algorithm takes Euclidean distance as similarity measure, which is to find the optimal classification corresponding to an initial clustering center vector V, making the evaluation index minimum. The algorithm uses the sum of square error criterion function as the clustering criterion function.

2.4 KNN algorithm

KNN (k-nearest neighbor) classification algorithm is one of the simplest machine learning algorithms. Its main idea is: first calculate the distance between the samples to be classified and the training

samples of the known classes, and then find the nearest k neighboring points from them; then judge the data categories of the samples to be classified according to the categories of the k neighbor samples. Most of the K samples belong to the categories of the samples to be classified.

3. Case test

3.1 Speed distribution law

In order to further analyze the distribution law of traffic flow speed, the ship inlet speed and frequency are counted, as shown in Table 1.

Table 1 observation value and frequency calculation of ship inlet speed

Speed group (kn)	Group median (xi)	Number (fi*)	Frequency (fi)
[3.5,4.5]	4	80	0.05
[4.5,5.5]	5	64	0.04
[5.5,6.5]	6	80	0.05
[6.5,7.5]	7	191	0.12
[7.5,8.5]	8	382	0.24
[8.5,9.5]	9	413	0.26
[9.5,10.5]	10	239	0.15
[10.5,11.5]	11	80	0.05
[11.5,12.5]	12	32	0.02
[12.5,13.5]	13	16	0.01
[13.5,14.5]	14	16	0.01
total		1593	1

According to the data in Table 1 and the above formula, first calculate the mean value and variance of the sample

$$\mu = 8.36(\text{kn})$$

$$\sigma^2 = 3.61$$

The theoretical frequency pi of normal distribution is obtained by looking up the normal distribution table. For example, the process of calculating the probability of ship speed in the range of 3.5-4.5kn is as follows:

$$p(3.5 \le X \le 4.5) = \varphi\left(\frac{4.5-\mu}{\sigma}\right) - \varphi\left(\frac{3.5-\mu}{\sigma}\right) = 0.02$$

Table 2 calculation table of theoretical frequency Pi and $\chi 2$

Group number	Speed (kn)	Number (fi*)	Theoretical frequency (pi)	$\frac{(f_i - np_i)2}{np_i}$
1	[3.5,4.5]	80	0.02	7.2739
2	[4.5,5.5]	64	0.04	0.0012
3	[5.5,6.5]	80	0.09	2.8011
4	[6.5,7.5]	191	0.16	1.6012
5	[7.5,8.5]	382	0.18	3.1647
6	[8.5,9.5]	413	0.21	1.8407
7	[9.5,10.5]	239	0.14	1.1451
8	[10.5,11.5]	80	0.08	1.7661
9	[11.5,12.5] [12.5,13.5] [13.5,14.5]	32 16 16	0.03 0.01 0.003	7.83 0.0003 2.6344
total		1593		19.594

The combination of frequencies less than 64 is divided into 9 groups, so k = 9, DF = 6, $\chi 2$ 0.05 = 12.592 and χ 2 = 19.594 > 12.592 are obtained by looking up the table. Therefore, it is considered that the observed ship speed does not obey the normal distribution, which may be due to the complex navigation mode of the channel and the influence of human factors

3.2 The relationship between speed and course

The AIS data of a certain section in a year are selected to find out two types of ships with the largest number. The names of ship types are 80 and 70 respectively. First, a scatter diagram is made to get an intuitive understanding.

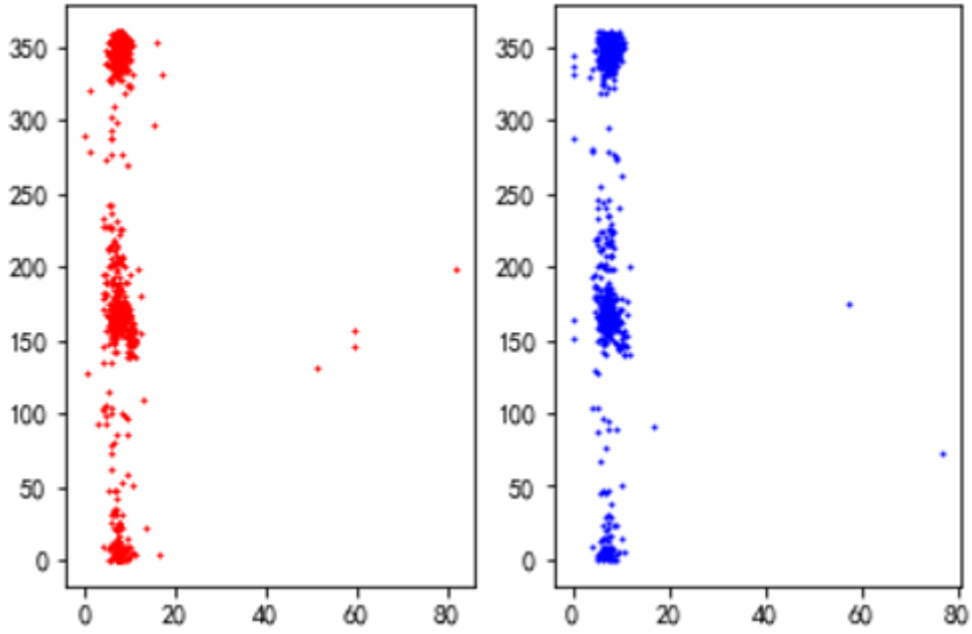

Figure 2 Scatter diagram of speed and course of two ships

It can be seen from the figure 3 that the speed and course distribution of the two ship types are similar, so it is not reliable to judge the ship type only by experience. Therefore, K-means algorithm is used for clustering analysis. The process is as follows:

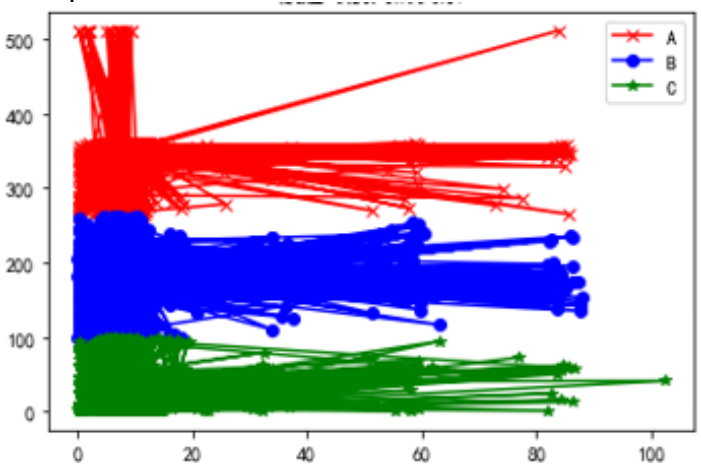

Figure 3 Clustering results

Although obvious clustering is obtained, the ship type can not be estimated quantitatively, so KNN algorithm is used.

The specific process is as follows:

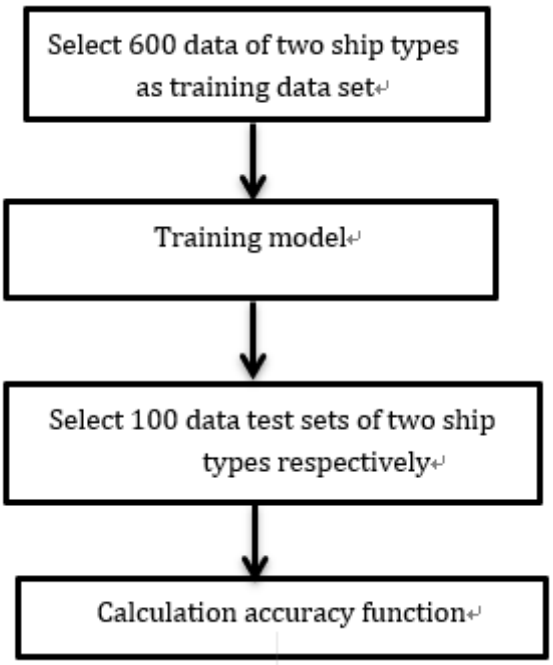

Figure 5 flow of KNN algorithm

Table 3 sample data

Speed	Course
7.9	345.5
8.3	356.8
6	188.3
7.4	343.6
7.2	340.8
7.4	349.2
6.7	351.3
6.8	353.3
7	341.8

The final results are shown in the table 4 .

Table 4

Type of ship	70	80
Accuracy	75%	70%

The accuracy is not very high, maybe because the number of training data sets is still small, or the selected data is not representative.

4. Conclusion

Due to the increasing amount of AIS data, if we do not use data mining technology to analyze its internal laws, it will be difficult to get the characteristics of maritime traffic flow. In this paper, we use the knowledge of mathematical statistics to study the speed law of marine traffic flow. Although we can predict the unknown ship type, the accuracy is not very high. We can consider to improve the distance calculation method of KNN algorithm to improve the accuracy. In the future, we can do further research on other characteristics of marine traffic flow, such as position, draft, Captain, etc. Research.

References

[1] Wu Zhaolin, Zhu Jun. Maritime Traffic Engineering (Second Edition) [M]. Dalian: Dalian Maritime University Press, 2004.
[2] Zheng bin, Chen Jinbiao, Xia Shaosheng, Jin Yongxing. Characteristics analysis of marine traffic flow data based on data mining [J]. China navigation, 2009,32 (01): 60-63.
[3] pan Jiacai, Shao Zheping, Jiang Qingshan. Research on the application of data mining in the analysis of maritime traffic characteristics [J]. China navigation, 2010,33 (02): 60-62.
[4] Li yongpan, Liu Zhengjiang, Cai Yao, Zheng Zhongyi. Analysis of maritime traffic characteristics based on AIS data constraint clustering [J]. Shipbuilding Engineering, 2018,47 (01): 176-179. Liu Lei, Chu Xiumin, Jiang Zhonglian, Zhong Cheng, Zhang Daiyong.
[5] algorithm of ship trajectory classification based on KNN [J]. Journal of Dalian Maritime University, 2018,44 (03): 15-21.