ATLANTIS PRESS

Advances in Economics, Business and Management Research, volume 85

2nd International Conference on Economy, Management and Entrepreneurship (ICOEME 2019)

Important Variables Identification and Proactive Evaluation of Real-time Ship Traffic Sailing Risk in Waterway

Moyang Zhao
Navigational College
Jiangsu Maritime Institute
Nanjing, China 211170

Shukui Zhang
Navigational College
Jiangsu Maritime Institute
Nanjing, China 211170

***Abstract*—In order to further improve accident prediction accuracy of real-time ship traffic in waterway, based on ship detector data and traffic accident data collected on two downstream waterways of Yangtze River, important variables were sifted with random forest (RF) model from the initial data of waterway status within 20-40min before the traffic accident, then new Bayesian network(BN)model was established with 4 most important variables combined with Gaussian mixture model (GMM) and maximum expectation (EM) algorithm. Compared with BN model previous studied built with direct initial data, the new models complexity is not only reduced and its prediction effect is improved, with the accident prediction correct rate of 81.29%.**

***Keywords*—waterway; sailing risk; proactive evaluation; random forest; Bayesian network**

I. INTRODUCTION

The ship traffic accidents in the waterway not only cause casualties and heavy property losses, but also bring economic losses to environmental governance that are difficult to estimate due to the occurrence of major malignant accidents such as oil spill and oil leakage. At home and abroad research on vessel traffic accident based on historical data using support vector machine (SVM) [1], the grey system theory [2], Markov [3], fractal theory [4] or combination model[5]to forecast the future traffic accident.

This article choose two of the Yangtze River channel morphology and environment parameters in close waterway, first using the random forest model screening was carried out on the important parameters affecting the safety of navigation, using variable after the screening to build model, and then using the improved Bayesian network model within the channel on real-time risk evaluation of ship traffic flow, and test the model's transferability, so used to the other channel.

II. RESEARCH DATA

This paper selects the Caoyaoxia waterway and Dantu waterway of the Yangtze River as the research channel, the data collected from Caoxiexia waterway are used for model construction and evaluation, the data collected from Dantu waterway are used for the transferability test of the model. The data of 72 ship traffic accidents occurring in Caoxiexia waterway and Dantu waterway from March to October 2017 are currently adopted. The specific accident type and the number of accidents are shown in “Table I”.

TABLE I. SUMMARY OF ACCIDENT DATA

Name of the channel	Type and number of accidents						Amount
Caoxiexia waterway	collision /30	sink /11	Contact loss /4	stranding /2	on the rock /1	Other/10	58
Dantu waterway	collision /8	sink /3	Contact loss /1	stranding /0	on the rock /0	Other/2	14

According to the previous research results, the corresponding channel traffic flow state data within 20-40min before the accident is selected to build the model. Each accident information extraction in the accident only four ships near detector or VTS monitoring system, in the order from upstream to downstream, respectively named four ship name of detector for N_1, N_2 and N_3 and N_4 interchange. The input of the model includes three variables: Difference and average of the initial data, all kinds of variables, all variables are listed in “Table II”. Among them, the number 1 ~ 4 and N_1 ~ ship detector N_4 interchange one-to-one correspondence, Q (1), V (1) and O (1) respectively the ship detector N1 within 20 to 40 minutes before the accident, ship the average velocity of traffic flow density and average channel share, N - Q (12) ship detector N1 and N_2 of traffic flow density difference, A - Q (13) ship detector N_1 and N_3 average density of traffic flow, other variable naming is similar to the above, For example, n-v (13) represents the speed difference between vessel detector N1 and vessel detector N_3.

TABLE II. CANDIDATE DIFFERENT VARIABLES FOR MODELING

Optional variable	Initial variable	Variable difference	Variable mean
variable name	Q(1),V(1),O(1), Q(2),V(2),O(2), Q(3),V(3),O(3), Q(4),V(4),O(4)	N-Q(13),N-V(13),N-O(13), N-Q(14),N-V(14),N-O(14), N-Q(23),N-V(23),N-O(23), N-Q(24),N-V(24),N-O(24)	A-Q(12),A-V(12),A-O(12), A-Q(34),A-V(34),A-O(34)

III. MODEL BUILDING

A. Stochastic Forest Model

Because many used in model building optional variables, considering the complexity of the model calculation and to avoid the problem of model in fitting, so before modeling, screening for optional variables, in order to identify a greater influence on the safety of ship traffic flow within the channel variables, and with these variables as the input variables of the model. RF is a statistical learning algorithm proposed by Breiman in 2001, which can effectively measure the importance of input variables. The RF calculation steps of variable significance are as follows:

- K samples were extracted from the initial training sample set $H = \{(x_i, y_i)\}(i = 1,2,\cdots,n)$ by bootstrap to form a new sample set, written H_s, build a CART with H_s.
- At each node of the CART, M variables (m≤M) are randomly selected from all variables M for CART node segmentation.
- Repeat the above 2 steps s times to build a RF model.
- Calculate the out-of-bag CART data generated each time (out-of-bag, OOB) H_{OOB}, $H_{OOB} = H - H_s$.
- H_{OOB} was predicted and classified by CART, and the correct classification times were summed up.
- For each independent variable i, change the value of x_i in H_{OOB}, Then x_i and CART were used to predict and classify H_{OOB}, and the correct classification times were calculated.
- According to the two steps (5) and (6), the number of correct classifications is calculated, and the descending value of classification accuracy (error) after data H_{OOB} changes x_i is calculated.
- The declining value (error) of classification accuracy after changing x_i in all s CART data H_{OOB} is calculated, and the importance of x_i is obtained. The larger the error is, the more important the variable is.

This paper uses Matlab computing platform to achieve variable importance of RF computing steps. 30 variables were input into the prepared RF calculation program, and the calculation results of the importance of each variable were shown in “Fig. 1”. Thereinto, the ordinate represents the declining value of classification accuracy after the change of a variable, and this value represents the importance of the variable.

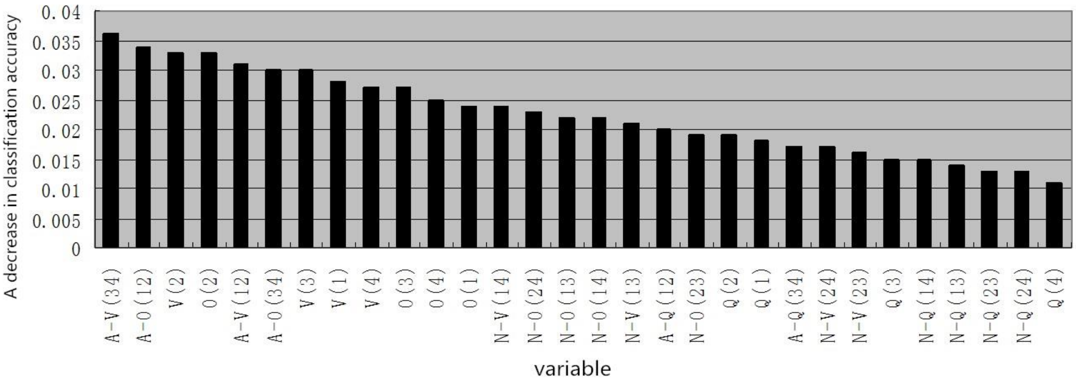

Fig. 1. Distribution of importance of different variables.

The preliminary research results show that the Bayesian network model constructed by using the data of the two ship detectors nearest to the traffic accident location within 20-40min before the traffic accident takes place has the best prediction effect, and this model selects 6 variables in total. In this article, therefore, the selection of variables should be less than 6, in order to prove the role of key parameters identification, through comparison and analysis, and considering the model prediction accuracy requirement, finally select A - V (34), A - O (12), V (2) and O (2) the most important four variables as the input variables, four variables respectively traffic story so before 20-40 min downstream of the ship within the average velocity, the upstream channel average share, N2 detector measured ship speed and channel share.

B. BN Model Building

Because of the uncertainty of navigation safety of ships in the waterway, the Bayesian network has been widely used in the engineering field because of its obvious advantages in the expression and reasoning of uncertain knowledge. Again at the same time because of the vessel traffic situation in the waterway information acquisition may be missing, and the advantages of the maximum expected algorithm based on gaussian mixture model is able to effectively deal with the lack of research data, this article USES the algorithm to construct the Bayesian networks, and real-time navigation of ship traffic flow within the channel risk carries on the forecast.

The construction is completed in three steps:

The first step is data input. The ship traffic flow data, traffic accident data and non-accident data were input into the Bayesian network, and the training samples and test samples were randomly determined. The ratio of the two data was 1:1 to ensure the reliability of model test.

The second step is classifier training. Using training sample data and EM algorithm, the classifier is trained and constructed.

The third is the application of the third step classifier. The inference engine and test data of the constructed classifier are input into the classifier and the output is the posterior probability that the test samples belong to each category. In order to improve the accuracy of classification results, an appropriate safety valve value can be determined according to the actual observation in the specific classification. If the posterior probability is greater than the safety valve value, it is considered that the corresponding test sample is traffic accident data; otherwise, it is considered as non-traffic accident data.

IV. THE MODEL RESULTS

A. The BN Model Results

Based on the relevant data of 58 traffic accidents and 600 non-traffic accidents in caoyaoxia waterway, the state data of ship traffic flow in the waterway within 20-40 min before the accident were screened by important variables, and the classification results were compared with the previous research results, as shown in “Table III”.

TABLE III. COMPARISON OF PREDICTION CORRECT RATE FOR TWO BN MODELS

Systematic name	The correct rate of unfiltered important variables	Correct rate of screening important variables
Accident classification	72.96%	81.29%
Nonaccident classification	74.43%	76.53%
The overall classification	73.91%	76.59%

The “Table III” shows that although preliminary study adopts two ship detector before and after the traffic accident scene within 20 to 40 minutes before the accident of the BN model built by testing data in the accident and overall accuracy prediction accuracy is higher, at the same time also has good classification accuracy in the accident, but the application of after RF filter selected four important variables to construct the BN model under the same conditions of prediction effect is better, in not only the accuracy and improve the overall accuracy of small amplitude, 76.53% and 76.59%, respectively, Moreover, the accident classification accuracy rate is as high as 81.29%. By comparison, it shows that the BN model established after RF screening important variables not only reduces the model complexity, but also has better prediction effect.

B. Model Portability Testing

In order to compare the universality of the model established after screening important variables with the model established only with initial data in the previous study, this paper conducted transferability test for the two models. The test results of transferability of the two models are shown in “Table IV”.

TABLE IV. TRANSFER-ABILITY OF TWO BN MODELS

Systematic name	The correct rate of unfiltered important variables	Correct rate of screening important variables
The accident classification	51.36%	68.41%
Nonaccident classification	67.28%	94.63%
The overall classification	67.32%	91.86%

Results Display, although the model established based on the screening of important variables has a decrease in the classification accuracy of traffic accidents, which is 68.41%, the classification accuracy of non-traffic accidents is as high as 94.63%, and the overall classification accuracy is also as high as 91.86%. The test results of the model based on the direct use of initial data are not ideal in three aspects. The test accuracy is 51.36%, 67.28% and 67.32%, respectively.

V. CONCLUSION

After RF screening important variables, the BN model established not only reduces the model complexity, but also has better prediction effect, with an accuracy rate of 81.29% in accident classification.

The BN prediction model based on the data in one channel can be used to predict traffic accidents in other channels. The transferability of BN model based on screening important variables is better than that of BN model based on initial data.

REFERENCES

[1] Li Jun. Research on ship traffic accident prediction based on support vector machine [D]. Wuhan: Wuhan University of Technology, 2008.

[2] Wang Baokuo. Research on the application of grey forecasting of ship traffic accidents [J]. China sailing, 2011, 34(1): 59-62.

[3] Wang Qi, Wang Zhipeng. Markov grey model for maritime traffic accident prediction [J]. China sailing, 2013, 36(4): 119-124.

[4] Chen Zhiyu, Hu Shenping, Hao Yanbin. Water traffic accident prediction based on fractal theory [J]. Shanghai Maritime University, 2009, 30(3): 18-21.

[5] Chen Weijiong, Hao Yuguo. Combined evaluation model of port navigation environment safety [J]. Journal of transportation, 2007,15(1): 75-77.