ISSN 2616-5775 Vol. 7, Issue 7: 123-130, DOI: 10.25236/AJCIS.2024.070716

Prediction of the main dimensions of yachts based on random forest

Liu Yangbing[1,a,*], Sun Chengmeng[1,b], Lin Haihua[1,c]

[1] Naval Architecture and Port Engineering College, Shandong JiaoTong University, Weihai, China
[a] 530177209@qq.com, [b] scmeng717@163.com, [c] 7216219@qq.com
**Corresponding author*

Abstract: *The main dimensions data is an important technical parameter that affects the performance of yachts, and it is an important task in the preliminary design stage to determine those values by statistically revealing the relations between the main scales of yachts. Establishing formulas between the main dimensions of yachts through regression analysis is a traditional solution. As an ensemble learning algorithm, the Random Forest algorithm is more suitable for learning the features in the sample data of the main dimensions of yachts and provides better prediction results. On this study, we analysed the distribution pattern of the collected main dimension data of yachts, and use the* Spearman's rank correlation *to calculate the correlation between different dimensions of the yacht data, generating correlation coefficient matrix, and employ multiple linear regression and box plot methods to identify outliers in the data, thereby enhancing data usability. The Random Forest algorithm (RF) is used to predict the draft and weight of yachts based on length and width. At the same time, the BP neural network algorithm is used to compare the performance of RF. The results show that the Random Forest algorithm can effectively improve the accuracy of main dimension prediction.*

Keywords: *Main Dimensions, Random Forest, BP Neural Network, Regression Analysis, Spearman's Rank Correlation*

1. Introduction

Yacht design is a quantitative process that consists of iterations to satisfy specified requirements [1]. The main dimensions of a yacht, which include length, beam, draft, and displacement, are the fundamental measurements that indicate the external size of the yacht. They are closely related to the design of the hull line of the yacht and significantly affect its performance. Determining value of the main dimensions is one of the crucial tasks during the preliminary design stage.

The regression formula of design parameters can be used in combination to determine the main dimensions of the hull during the preliminary design stage [2-3]. This data typically originates from ship-shaped database, which are collections of data such as the principal dimensions of a specific series of ship types, and designers will supplement the database through methods such as spline interpolation, and then obtain the regression formula for a specific design parameter through regression analysis [4]. Regression analysis is not the only method for determining the principal dimensions of a ship's hull. Artificial Neural Networks are also frequently used [5-6], and when there is sufficient data, neural networks can predict the characteristics of ships such as resistance and seakeeping [7]. However, some similar tasks have the appearance of Random Forest algorithm.

Some studies have attempted to use multiple regression analysis[8] and neural networks[9] to establish predictive models for the principal dimensions of specific types of ships, such as asphalt carriers[10], marine law enforcement vessels[11], and yachts[12]. Indeed, there is a paucity of research utilizing Random Forest for predicting the main dimensions of ships, especially yachts.

Random Forest is an ensemble learning algorithm based on decision trees, proposed by Leo Breiman in 2001[13]. Regression Forest (RF) can perform both classification tasks and regression tasks, further introduces random attribute selection in the training process of decision trees, in regression tasks, RF based on the construction of Bagging ensemble with decision trees as the base learner. Random Forest algorithm has been applied in AIS-based classification of abnormal ship behaviors[14], roll motion response of damaged ships in beam seas[15], ship fuel consumption[16], ship detection[17], selection of prototype ship[18],and other aspects. These cases successfully demonstrate that RF exhibits excellent performance in achieving classification and fitting tasks for nonlinear data. Compared to traditional regression analysis methods for determining the main dimensions of ships, RF should be able to achieve smaller prediction errors and higher prediction accuracy.

Through this study, we utilized the collected yacht data to establish a prediction model for the main dimensions of yachts using the Random Forest algorithm. Additionally, we analyzed and processed this data to enhance the reliability of the dataset. We analyzed the correlation between the main dimension data through the correlation coefficient matrix and eliminated outliers through box plots and multiple linear regression analysis. Furthermore, we compared the Random Forest prediction model with the neural network prediction model on this dataset and found that the Random Forest algorithm achieved the same or even better prediction results compared to the neural network algorithm.

2. Data Preprocessing

We collect the main dimension data of monohull and catamaran yachts ranging from 5 meters to 40 meters, including small and medium-sized modern sharp-chined yachts as well as luxury yachts, from yacht sales websites utilizing web crawling technology. In order to eliminate the influence of any abnormal data, we preprocess the dataset first.

2.1. Correlation analysis of data

Correlation analysis can measure the degree of correlation between two variables. Generally, if the data in each dimension satisfies a normal distribution, we would use Pearson correlation analysis to calculate the correlation coefficient matrix. The normal P-P plot for the main dimensions of the yacht is shown in Figure 1. In the P-P plot, if the sample points are closer to the diagonal line, it indicates that the data distribution in this dimension satisfies a normal distribution. However, the main dimensions of yachts do not all satisfy a normal distribution, so we use Spearman correlation analysis to calculate the correlation coefficient matrix.

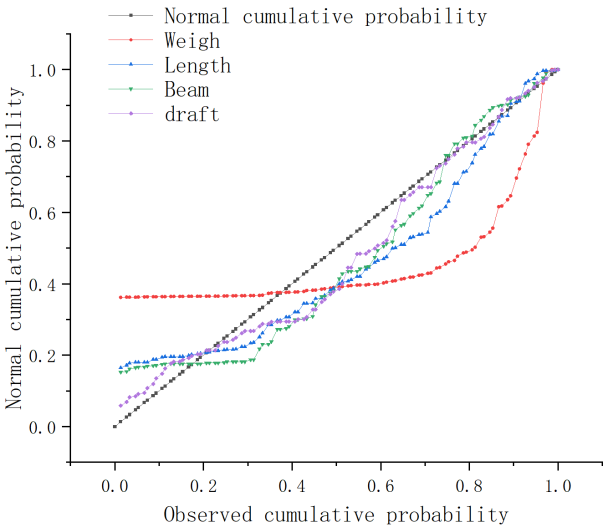

Figure 1: Normal P-P plot for yacht weight, length, width, and draft.

The calculation formula for the Spearman's rank correlation is [19-20]:

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \tag{1}$$

Where d_i is the difference in ranks of the two main dimensions for which correlation is being calculated, n is the number of samples, and r_s is the rank correlation about the two dimensions.

The value range of the Spearman's rank correlation is between -1 and 1. The Spearman's rank correlation is closer to 1, the correlation between dimensions is higher. The correlation coefficients of the main dimensions of the yacht are plotted as a heatmap, as shown in Figure 2. The color of the heatmap is more vivid when the correlation between main dimensions is higher. From the Heatmap, it can be seen that there is a high correlation between the displacement, draft, length, and width of the yacht.

When designing yachts, there is generally a specific proportional relationship between the length and width of the yacht. Disregarding the influence of other objective factors, we use the length and width to predict the weight and draught of the yacht, because these main dimension data can roughly calculate the performance of the yacht.

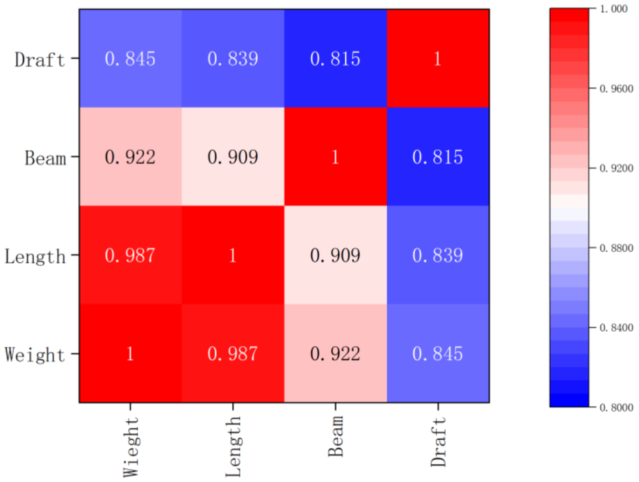

Figure 2 Heatmap of the Spearman correlation matrix.

2.2. Outlier filtering

Firstly, a box plot can be used to visually identify outliers in the data. A box plot is constructed by plotting 50% of the data within the "box," where the edges of the box represent the two quartiles. Specifically, 25% of the data is less than the lower quartile (Q1), and another 25% of the data is greater than the upper quartile (Q3). The box extends outwards with two whiskers (or "outer fences"), and the length of the whiskers is 1.5 times the distance between the upper and lower quartiles. Data points beyond the whiskers are considered outliers. Additionally, a box plot can provide information about the distribution of the data by observing the position of the mean and median. We plotted the collected data on yacht weight and draught as a box plot, which is shown in Figure 3.

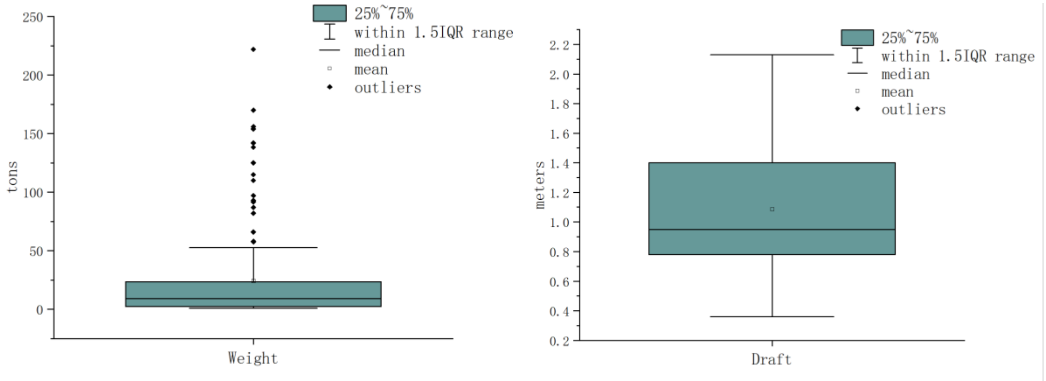

Figure 3: Box plot for yacht draft and displacement (or weight)

As shown in Figure 3, the draught data is relatively evenly distributed. The box plot of yacht weight shows that the yacht weight data exhibits a clear negative skew and has a large variance. This is because, objectively, as the length and width of the yacht increase, the weight of the yacht increases at a cubic rate. The box plot of yacht weight has obvious outliers. Through analysis, we have excluded sample data with yacht weights exceeding 150 tons. To further analyze these outliers, we introduce the method of multiple linear regression analysis.

Use linear regression to establish a predictive model for draft based on the length and width of the yacht, and apply the same approach to predict the displacement. Calculate the studentized residual, Cook's distance, and leverage value for each sample data point. When the absolute value of the studentized residual for this sample point is greater than 2, we consider this outlier to be a potential anomalous value, and specific analysis of the sample data is required to make a final judgment. Similarly, when the Cook's distance is greater than 0.027 or the leverage value is greater than 0.060, the same procedure should be carried out.

As shown in the Table 1, the R2 of the linear prediction model has been significantly improved after handling outliers. An R^2 value closer to 1 indicates that the model has a higher degree of explanation for the predicted items. The formula for calculating R2 is shown below:

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_{i=1}^{n} \left(y_i - \widehat{y_i}\right)}{\sum_{i=1}^{n} \left(y_i - \overline{y_i}\right)} \tag{2}$$

SSE represents the Sum of Squared Errors, which is the sum of the differences between the predicted values and the sample values; SST represents the Total Sum of Squares, which is the sum of the differences between the dependent variable and its mean value.

Table 1: Comparison of determination coefficients of the linear prediction model before and after handling outliers.

Predicted item	R2 before processing	R2 after processing
Weight (displacement)	0.749	0.818
Draft	0.637	0.857

As shown in Table 1, after handling outliers, the R^2 of the linear model for predicting draft is 0.818, indicating that the model can express 81.8% of the information about yacht draft in the dataset. Similarly, the model for predicting yacht weight can explain 85.7% of the information. This is not entirely in line with our expectations, which is one of the reasons why we use the Random Forest algorithm to improve prediction accuracy.

3. Random Forest model training

To retain as much information present in the dataset as possible, the method of cross-validation is employed to select 80% of the data as the training set, while the remaining 20% is used as the test set.

Random Forest uses Bootstrap Sampling to select T randomly distributed subsets from the original dataset, where T is the number of base learners. For regression tasks, the base learners of Random Forest are decision trees. Each subset trains a decision tree, and the Bootstrap Sampling method ensures that each subset contains approximately 63.2% of the original data. Random Forest ultimately outputs the simple average of the results from each decision tree as the final prediction. Figure 4 shows the training process of the Random Forest algorithm.

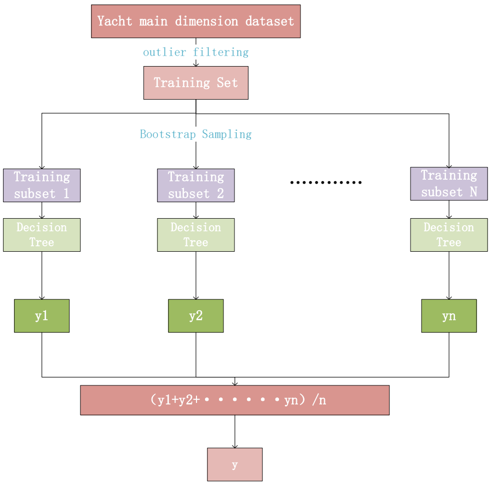

Figure 4: Random Forest Training Process

The training parameters of Random Forest include the criteria for node splitting, the number of decision trees, the minimum number of samples required for node splitting, the minimum number of samples at leaf nodes, and the maximum depth of trees. Generally, if the depth of decision trees in Random Forest is too deep, the number of decision trees is excessive, or node splitting occurs too frequently, it can lead to overfitting in Random Forest, manifested by a significant difference in R^2 between the training set and the test set. However, increasing the number of trees in the forest, designing a smaller number of samples for splitting, and setting a reasonable depth for decision trees can help enhance Random Forest's ability to fit complex data. After adjusting the training parameters of Random Forest, it was found that the algorithm performed optimally when the number of decision trees was around 100, the minimum number of samples required for node splitting was 3, the minimum number of samples at leaf nodes was 2, and the maximum depth of trees was 5.

The R^2 values for the prediction of yacht draft and weight using the Random Forest algorithm are shown in Table 2.

Table 2: The degree of fit of the Random Forest algorithm for predicting yacht draft and weight.

Predicted item	R^2 of the training set	R^2 of the test set
Weight (displacement)	0.996	0.983
Draft	0.997	0.994

It can be observed that the small difference between the R^2 data of the training set and the test set indicates that the Random Forest prediction model has good generalization ability and exhibits no signs of overfitting or underfitting. The Random Forest algorithm achieves a 98.3% degree of explanation for yacht draft information and an even higher 99.4% degree of explanation for yacht weight information. Compared to multiple linear regression, the performance of Random Forest is more satisfactory. However, how does it compare to other types of artificial intelligence algorithms? We established a neural network prediction model using the same dataset and compared its prediction results with those of the Random Forest. The neural network is trained with one input layer, one output layer, and three hidden layers. The hidden layers use the ReLU activation function to enhance computational efficiency and introduce sparsity, which reduces computational complexity and improves the generalization ability of the model. The Adam optimizer is used to optimize the weights of the neural network. Afterwards, we will evaluate the prediction models based on their fitting degree to the two main scale data of displacement and draught, as well as the prediction error.

4. Evaluate the models

Compare the prediction effects of three prediction methods, namely multiple linear regression, random forest algorithm, and neural network, on the same training set, and evaluate the performance of the models based on the R2 and mean squared error (MSE) of the prediction models. The prediction results of each sample in the test set by the three models are shown in Figure 3.

4.1. Assessment standards

The coefficient of determination, R2, has been introduced earlier. It is widely used in regression analysis, curve fitting, and model evaluation. It helps researchers understand the fitting effect of the model and thereby determine whether the model is suitable for a specific dataset. MSE is the sum of the squares of the residuals between the predicted values and the true values. A smaller MSE indicates better fitting performance of the model. Its Equation is Formula 2.In this formula,'i' is the number of sample, and $\widehat{y_i}$ is the predicted value.

$$MSE = \frac{\sum\left(y_i - \widehat{y_i}\right)}{n} \tag{3}$$

4.2. Result Comparison and Analysis

The evaluation indicators of the three predictive models are shown in Table 3. The Random Forest algorithm achieves the best fitting effect and the smallest prediction error when applied to the data of water draft and displacement. The BP neural network also achieves a good fitting effect on the data of water draft and displacement, with a reduced prediction error compared to multiple linear regression. The multiple linear regression model performs well in predicting water draft of yachts, but its performance in predicting displacement is not as good, resulting in a relatively large Mean Squared Error (MSE).

Table 3: Evaluation of the three predictive models.

predictive model	R2 related to draft	MSE related to draft	R2 related to Weight	MSE related to Weight
multiple linear regression	0.818	0.032	0.857	213.48
random forest	0.983	0.005	0.994	5.701
BP neural network	0.955	0.021	0.969	27.262

On the test set, the performance of each model also confirms the evaluation results in Table 3. As shown in Figure 5 and Figure 6, the predictions of the main dimensions of yachts by the Random Forest algorithm are closer to the sample data. The comparison of draft value prediction results is shown in Figure 5. The BP neural network algorithm and the multiple linear regression algorithm have similar prediction results for the water draft of small and medium-sized yachts with shallow water draft, while for the prediction of draft of large and medium-sized yachts with larger water draft values, the multiple linear regression algorithm has a larger error. Figure 6 shows the prediction results of each algorithm for displacement. The prediction results of the multiple linear regression method have a large MSE, which would affect the observation of the prediction results of other models if plotted, so the results are not plotted. Both the BP neural network algorithm and the Random Forest algorithm can predict displacement data well, while the BP neural network has a large prediction error on the 25th test set sample data, and the Random Forest avoids the occurrence of errors through bootstrap sampling and Bagging ensemble.

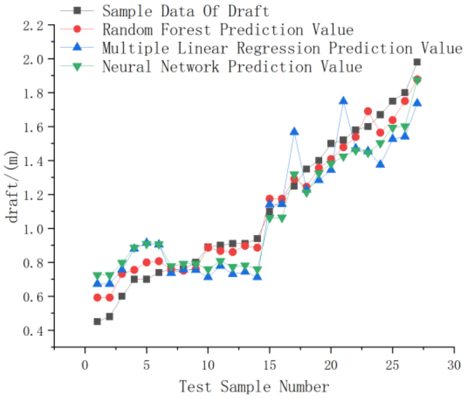

Figure 5: Comparison of draft data of samples in the test set with the prediction values from the Random Forest algorithm, Neural Network algorithm, and Multiple Regression algorithm.

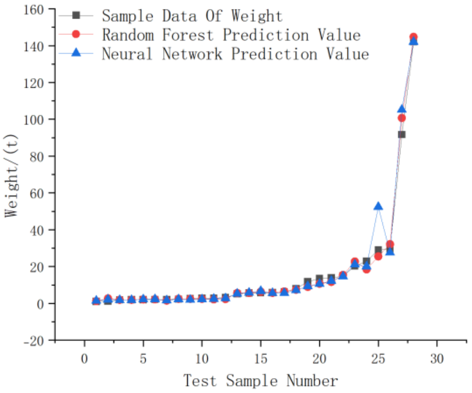

Figure 6: Comparison of weight data of samples in the test set with the prediction values from the Random Forest algorithm, Neural Network algorithm, and Multiple Regression algorithm.

5. Conclusion

Through this study, we discuss a method that applies the Random Forest algorithm to the initial design stage for determining the main dimensions of yachts. We hope to quickly determine the draft and displacement (weight) of yachts based on their length and width during the design phase. We discuss the preprocessing of yacht`s sample data and how to obtain the optimal prediction model by Spearman's rank correlation and Regression Analysis. From this, we can see the great potential of combining yacht design, or ship design, with artificial intelligence. Based on the comparison of performance of those three models, the following conclusions can be drawn:

(1) There is a high correlation between the length, width, and weight data of yacht, and the draft data also has a relatively high correlation with other main dimension data. However, compared to the prediction results for draft, the prediction effect for displacement (or yacht weight) is better. Higher correlation leads to higher prediction accuracy when using machine learning methods for the prediction model.

(2) The multiple linear regression model has poor prediction performance for severely skewed and discrete yacht weight data, but it performs better for draft data that is close to a normal distribution. When predicting the main dimensions of yachts, multiple linear regression can be used as a simple prediction method.

(3) The BP neural network exhibits excellent performance in predicting the main dimensions of yachts. However, there may be some deviations in the data predictions for specific samples, which could be related to the particularity of the collected dataset.

(4) The prediction values obtained by the Random Forest algorithm for the main dimensions of yachts are closer to the actual values. Random Forest algorithm can be an excellent tool for predicting the main dimensions of yachts during the initial design stage.

References

[1] Bülent İbrahim Turan and Mehmet Akman. Practical Design Framework for Lobster-Type Motor Yachts[J]. Journal of ETA Maritime Science, 2023, 11(4) : 282-289.

[2]Tomasz Cepowski. Determination of regression formulas for main tanker dimensions at the preliminary design stage[J]. Ships and Offshore Structures, 2019, 14(3) : 320-330.

[3] Tomasz Abramowski and Tomasz Cepowski and Peter Zvolenský. Determination of Regression Formulas for Key Design Characteristics of Container Ships at Preliminary Design Stage[J]. New Trends in Production Engineering, 2018, 1(1) : 247-257.

[4] Radojcic D V, Zgradic A B, Kalajdzic M D, et al. Resistance and Trim Modeling of a Systematic Planing Hull Series 62 (with 12.5°, 25°, and 30° Deadrise Angles) Using Artificial Neural Networks, Part 1: The Database[J]. Journal of Ship Production and Design, The Society of Naval Architects and Marine Engineers, 2017, 33(03): 179–191.

[5] Radojcic D V, Kalajdzic M D, Zgradic A B, et al. Resistance and Trim Modeling of a Systematic Planing Hull Series 62 (with 12.5°, 25°, and 30° Deadrise Angles) Using Artificial Neural Networks, Part 2: Mathematical Models[J]. Journal of Ship Production and Design, The Society of Naval Architects and Marine Engineers, 2017, 33(04): 257–275.

[6] Bertram and Mesbahi. Simple Design Formulae for Fast Monohulls[J]. Ship Technology Research, 2004, 51(3) : 146-148.

[7] Romero-Tello P. and Guti..rrez-Romero J.E. and Serv..n-Camas B.. Prediction of seakeeping in the early stage of conventional monohull vessels design using artificial neural network[J]. Journal of Ocean Engineering and Science, 2023, 8(4) : 344-366.

[8] Liu Yang, Zhang Mingxia, Zhao Yuhao. Regression Model for Principal Dimensions of Bulk Carriers Based on New IMO Regulations [J]. Ship & Boat, 2013, 24(04): 78-81.

[9] Liu Fei, Lin Yan, Li Na, et al. Research on Principal Dimension Models of Marine Surveillance Ships and Fishery Administration Ships [J]. Ship Science and Technology, 2012, 34(07): 49-54.

[10] Li Hongwei, Guan Guan. Analysis of Principal Dimensions of Marine Law Enforcement Vessels Based on Multiple Regression Methods [J]. Ship Standardization Engineer, 2020, 53(03): 55-58.

[11] Zhang Qiuping, Jin Tingyu, Yao Wen, et al. Statistical Analysis of Principal Dimension Elements of Asphalt Carriers [J]. China Ship Repair, 2022, 35(05): 59-62, 72.

[12] Guan Guan, Zheng Mengtian, Ji Zhuoshang. Research on the Application of Big Data Technology in Determining the Principal Dimensions of Ships [J]. Applied Science and Technology, 2018, 45(02): 1-5.

[13] Breiman L. Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author)[J]. Statistical Science, Institute of Mathematical Statistics, 2001, 16(3).

[14] H. Rong and A.P. Teixeira and C. Guedes Soares. A framework for ship abnormal behaviour detection and classification using AIS data[J]. Reliability Engineering and System Safety, 2024, 247 : 110-105.

[15] Ran X L ,Qiu T L ,Ping Z W .Rapid prediction of damaged ship roll motion responses in beam waves based on stacking algorithm[J].Journal of Hydrodynamics,2024,36(2):394-405.

[16] Zhou Yi et al. Predicting ship fuel consumption using a combination of metocean and on-board data[J]. Ocean Engineering, 2023, 285(P2)

[17] N. Li et al. SHIP DETECTION BASED ON MULTIPLE FEATURES IN RANDOM FOREST MODEL FOR HYPERSPECTRAL IMAGES[J]. ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2018, XLII-3 : 891-895.

[18] Zhang Mingxia, Zhao Tongming, Wang Siyi. Application Research of Random Forest Method in Selection of Prototype Ship Types [J]. Applied Science and Technology, 2023, 50(05): 126-132, 174.

[19] Jiefang Jiang and Xianyong Zhang and Zhong Yuan. Feature selection for classification with Spearman's rank correlation coefficient-based self-information in divergence-based fuzzy rough sets[J]. Expert Systems With Applications, 2024, 249(PB) : 123633-

[20] Guanqiao Wang and Heng Zhou and Bohang Lin. Evaluation of Carbon Emissions Redcution Performance Based on TOPSIS and K-Means Clustering Algorithm[J]. Academic Journal of Computing & Information Science, 2023, 6(7)