Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 5
AIDML 2024

Research and Case Study on Insurance Underwriting Decision Making Based on LSTM, SVM and Random Forest Modeling

Yunzhi Zhou [*, #], Xiubi Huang [#]

School of Automation, Guangdong University of Technology, Guangzhou, China, 510006

* Corresponding author: laurazhou1@protonmail.com

[#]These authors contributed equally.

Abstract. In recent years, the world has suffered from a variety of extreme weather events, which are gradually becoming a crisis for insurance companies and homeowners. To address the challenges of insurance companies in their underwriting decisions for a specific region, this paper proposes a binary classification model for determining whether they should be underwritten, with an output of 1 (underwritten) or 0 (not underwritten). Firstly, extreme weather events are identified by screening the number of casualties, property and crop losses and quantifying the impact of these events. Then, a system of metrics was developed, including the use of LSTM models to predict risk indicators. Correlation analyses were conducted to verify the validity of the indicator system, and then the quantitative data were fed into SVM models and random forest models to solve the classification problem. Finally, recommendations for insurance companies and homeowners were made based on the results, and two regions on different continents were selected to demonstrate the effectiveness of the model.

Keywords: Insurance Underwriting Decisions; Binary Classification Model.

1. Introduction

In recent years, the losses caused by natural disasters are increasing year by year [1]. The world has suffered multiple extreme weather events that have progressively become crises for insurers. This has caused insurers to change how and where they are willing to underwrite, and weather-related accidents have led to higher property insurance premiums. This paper develops a model for how to best position property insurance so that the model is resilient to future claims costs while ensuring the sustainable health of insurers.

2. The establishment of indicator system

2.1. The analysis of some indicators

In Google Scholar and other periodical search sites, using property insurance as the keywords, we reviewed related literature, marked the keywords involved in these 34 articles with frequency, and finally selected the indicators with higher frequency as the judgment indicators of whether the study area is insured or not, which include three aspects: risk, business interests and market demand. The following indicator system as figure 1 is finally established:

Content from this work may be used under the terms of CC BY-NC 4.0 licence (https://creativecommons.org/licenses/by-nc/4.0/).
Published by Warwick Evans Publishing.

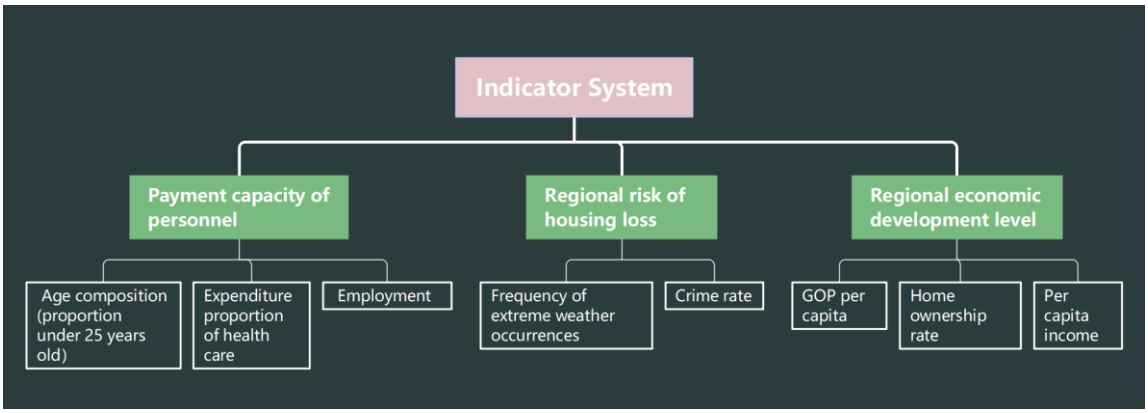

Figure 1. Property insurance evaluation index system

The following data was obtained by looking at data sites such as U.S. Bureau of Economic Analysis (BEA), worldpopulationreview, etc. Figure 2 and figure 3 show home ownership rates and personal income by state, respectively. The specific data are shown in table 1.

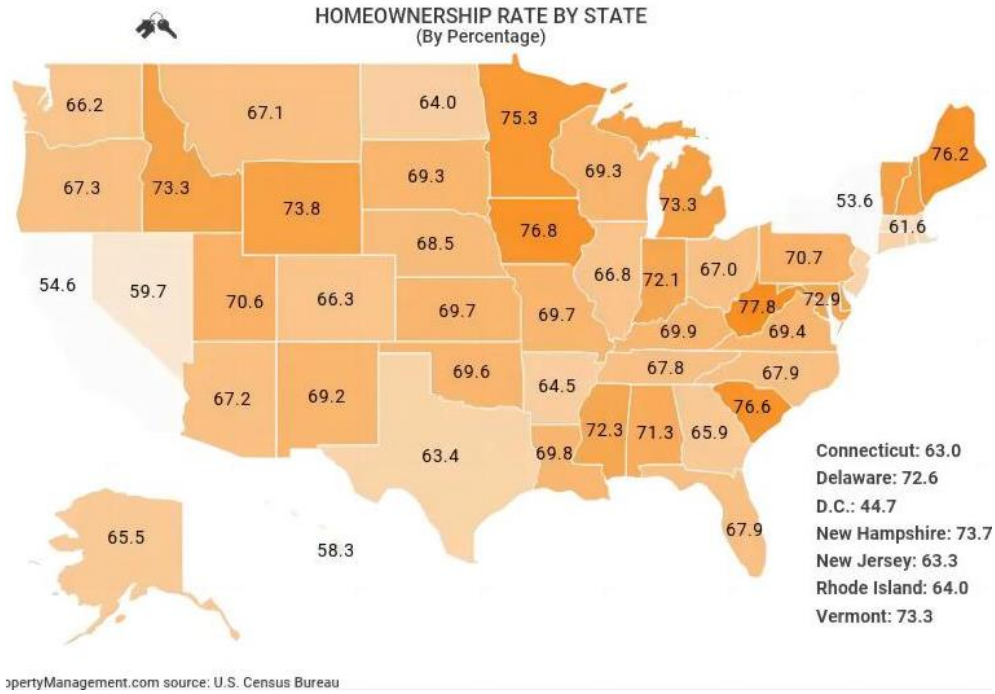

Figure 2. homeownership rate by state

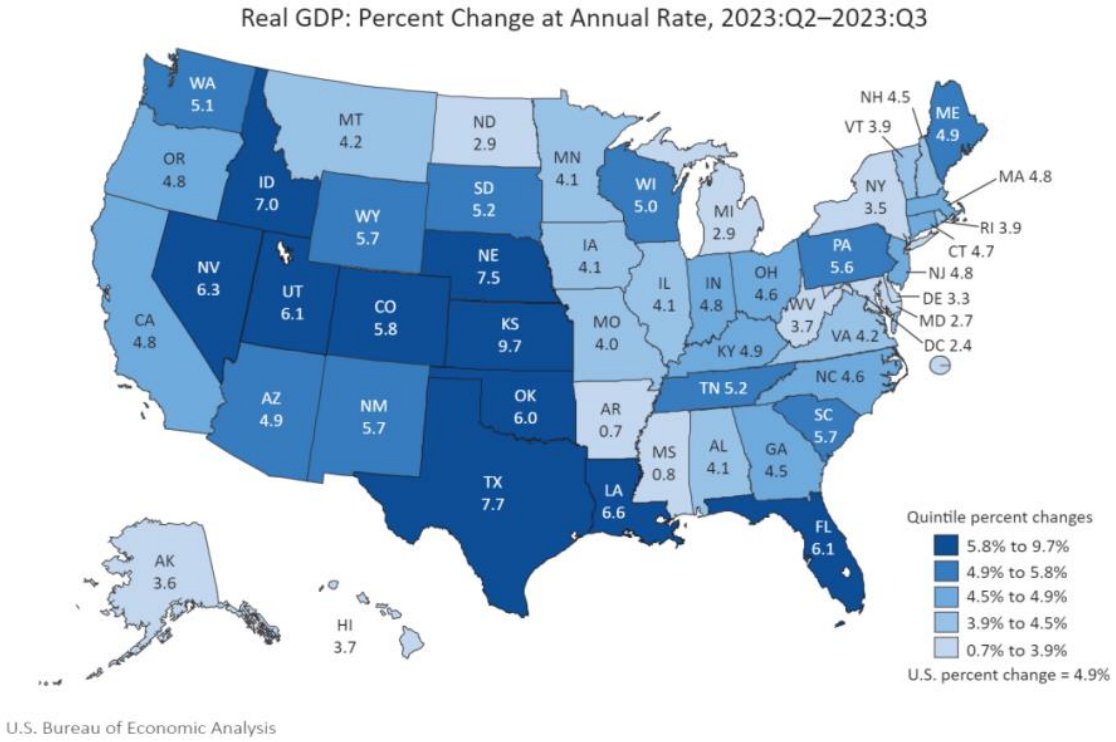

Figure 3. Personal income: percent change at annual rate

Table 1. Evaluation indicators

region	the number of extreme weather events in the past decade		per capita income	home ownership rates
ALABAMA	167		33777	71.30%
ALASKA	93		43054	65.50%
…				
VIRGINIA	403		50764	69.40%
WASHINGTON	308		31922	66.20%
WEST VIRGINIA	417		40188	77.80%
WISCONSIN	305		38114	769.30%
WYOMING	71		16676	73.80%

Spearman's rank correlation coefficient is used to solve the correlation problem if the sample data of a random variable or its transformed values do not obey a normal distribution [2]. In the following, the correlation analysis of the indicators is carried out to produce a heat map to verify the correctness of the indicator system.

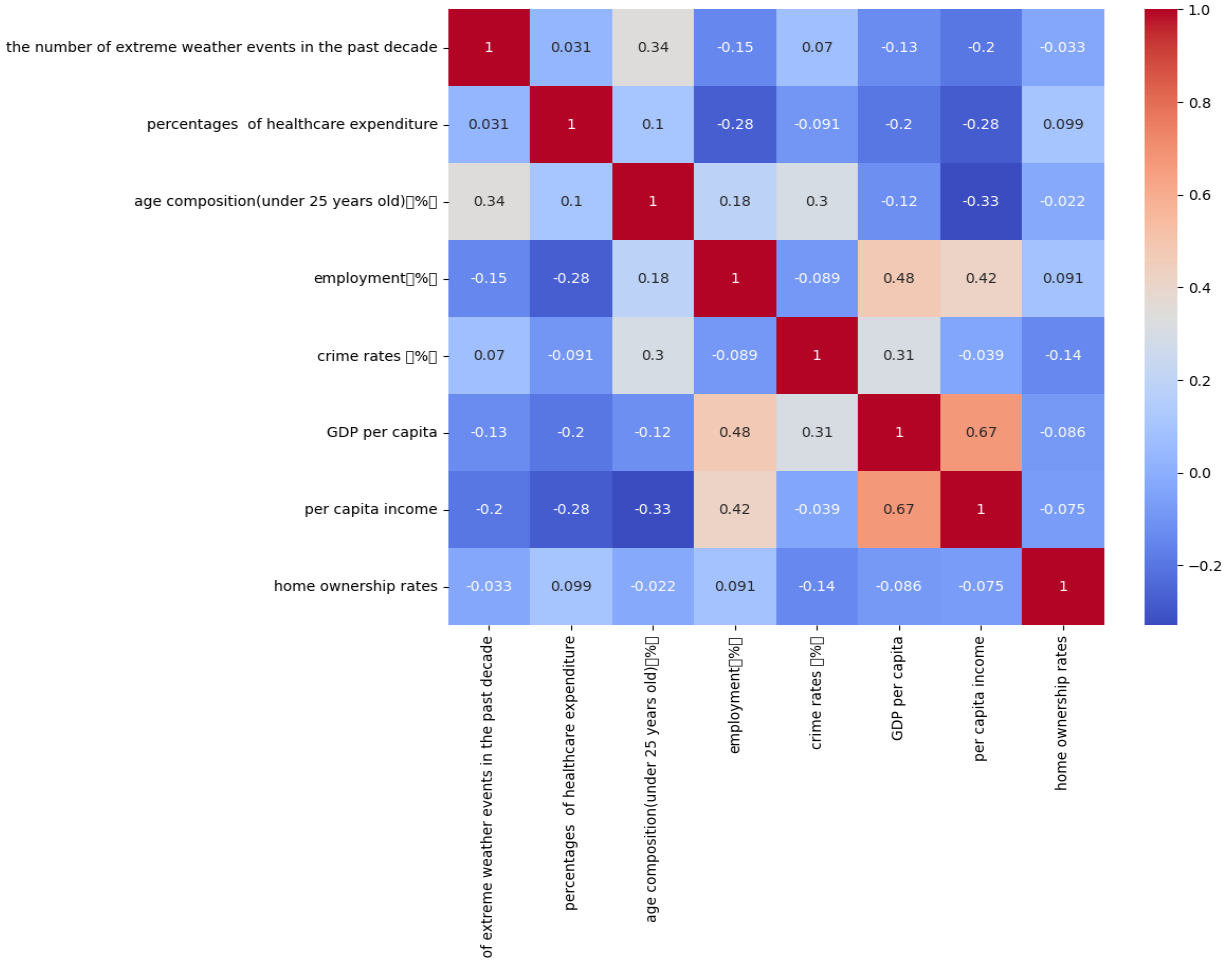

Figure 4. Correlation analysis

As can be seen in the figure 4, there is a weak correlation between the indicators, which is in line with the reality that one indicator cannot fully measure all the factors affecting property risk. The evaluation system established in this paper is more rigorous.

2.2. LSTM model for prediction

We constructed an LSTM recurrent neural network model to predict the next 10 years based on extreme weather data from the past 50 years. The frequency of extreme weather events each year is used to assess the risks faced by insurance companies underwriting the policy.

An LSTM network (Long Short-Term Memory) is a special kind of Recurrent Neural Network (RNN) that processes input data by traversing time steps and updating the RNN state. It is designed to solve long-term dependency problems [3] and uses previous time steps as inputs to predict the time series or subsequent values of the series. Train the LSTM neural network for time series prediction by first training the regression neural network with sequence output, where the response (goal) is to shift the values by a time step of the training sequence [4]. That is, at each time step of the input sequence, the LSTM neural network learns to predict the value at the next time step.

The steps are as follows:

1. Data preprocessing: given time, area and number of extreme weather events are organized into time series data, and data from the past 50 years are used to predict the number of extreme weather events in the next decade).

XTrain time as input

YTrain Extreme weather counts as output

Input data used by XTest for prediction (last 10 years not involved in training)

Output data used by YTest for prediction (last 10 years not involved in training)

2. Create LSTM network: Based on the data format from the previous step, create an LSTM network structure.

3. Training an LSTM (Long Short-Term Memory) network with data from the past 50 years

4. Forecast: Use trained networks to predict the number of extreme weather events in each region over the next decade.

We used the MATLAB build model to input the first 50 years of extreme weather data of the states into the model, with the previous 70% data as the training set and the last 30% data as the test set to evaluate the prediction effect of the model. The predicted results are shown in the figure 5:

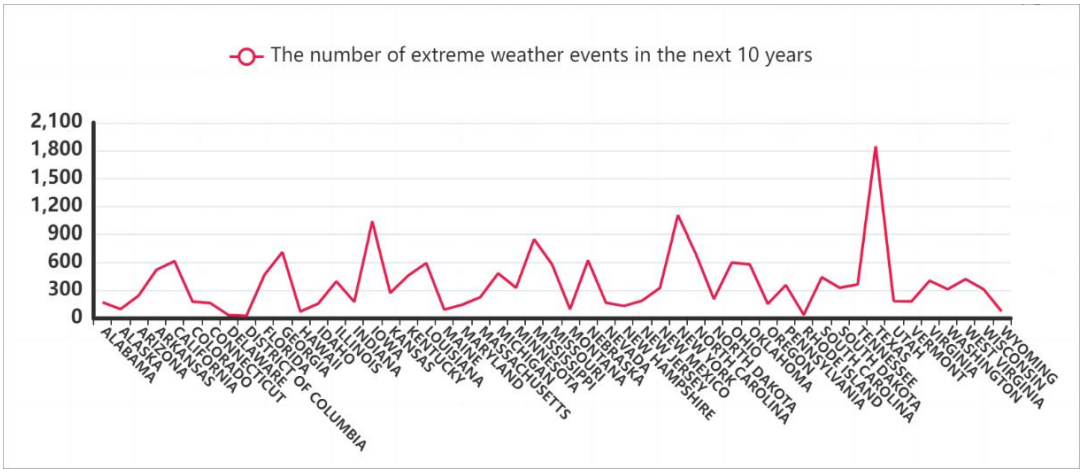

Figure 5. Prediction of the number of extreme weather events in the next decade

2.3. Insurance underwriting decision making based on SVM

Support vector machine (SVM) is a binary classification model with a linear classifier defined in the feature space. This method applies structural risk minimization theory and can better solve practical problems such as small samples and nonlinearity [5], and it can achieve the global solution by transforming the classification problem into a quadratic programming (QP) problem [6]. The learning strategy of SVM is interval maximization, which can be formalized as a problem of solving convex quadratic programming, and is also equivalent to the minimization problem of the regularized hinge loss function [7].

In this paper, different regions are used as objects to be classified and solved using SVM model.

For the classification problem of nonlinear data, the core of the SVM algorithm is the data mapping using kernel functions to find the optimal hyperplane. The use of kernel functions can effectively improve the ability of SVM to handle linear not separable problems, making the classification model simpler [8] [9]. In choosing the kernel function, we must weigh the properties of the kernel function against the intrinsic features of the data. For example, the Sigmoid kernel functions and the polynomial kernel functions are more complicated in the parameter setting, which increases the model uncertainty and instability to some extent. In contrast, the RBF kernel function requires only one parameter to build the model, thus improving the efficiency and accuracy. Both the Sigmoid and RBF kernel functions showed good performance when processing large-scale datasets. However, the performance of the Sigmoid kernel functions is relatively poor when faced with high-dimensional data.

Considering that the regional property risk assessment problems are non-linear, small data volume and complex characteristics, the Gaussian kernel function (RBF) is the most suitable choice of kernel

function. By using a Gaussian kernel, we can construct an efficient and accurate classification model to meet the needs of practical applications.

The results obtained are as follows figure 6 and table 2:

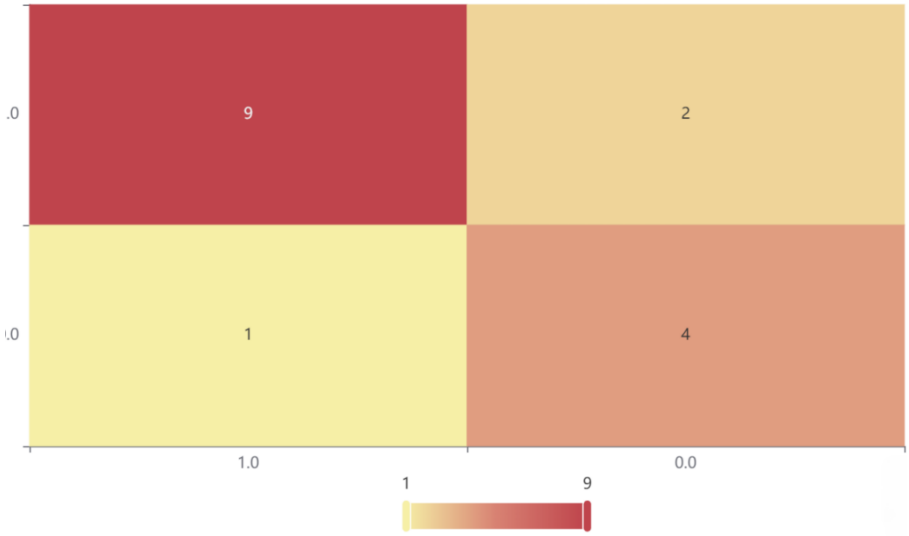

Figure 6. Gauss radial basis kernel function confusion matrix

Table 2. Model summary table

Name	Parameter name	Parameter values
Model parameter setting	Data preprocessing	Yes
	Training set ratio	0.7
	Error-term penalty coefficient	1.0
	Kernel	Rbf
	Numerical values of the kernel function system	0.01
	Multi-classification of the decision functions	Ovr
	Model convergence parameters	0.001
	Maximum iteration times	500
Model evaluation effect	accuracy	81.2%
	Accuracy (composite)	82.7%
	Recall rate (composite)	81.2%
	f1-score	0.817

In summary, various indicators are taken as the independent variable, while whether the policy is insured is taken as the dependent variable, and the training set proportion is set to 0.7 to conduct SVM modeling. The RBF kernel function was selected and took the error term penalty coefficient C=1.0.

It can be seen from the above table: the accuracy of the final model in the test set is 81.2%, accuracy (comprehensive) is 82.7%, recall (comprehensive) is 81.2%, and f1-score (comprehensive) is 0.817. This model works better.

The classification of the SVM model is as follow table 3: (only some results are shown due to space limitations)

Table 3. Whether or not to underwrite

Region	Yes (1) or No (0)
ALABAMA	1
ALASKA	0
……	
WEST VIRGINIA	1
WISCONSIN	1
WYOMING	0

2.4. Insurance underwriting decision making Based on random forest

Random forest is a class of combination methods, designed specifically for decision tree classifiers [10]. We use it to make decisions about whether to write coverage.

The results are as follow table 4:

Table 4. Model summary

Parameter Name	Parameter Values
Training Time	0.714s
Train sets	0.7
Data Shuffle	no
Cross Validation	3
Evaluation criteria for node splitting	gini
Number of decision trees	100
Sampling was put back	true
Extra-bag data testing	false
Maximum proportion of features considered when partitioning	auto
Minimum number of samples for the splitting of the internal nodes	2
Minimum sample number of the leaf nodes	1
Minimum weight of the samples in the leaf node	0
Maximum depth of the tree	10
The maximum number of leaf nodes	50
The threshold of node division impurity	0
Accuracy	96.1%
Recall	96.1%
Accuracy rate	96.2%
F1	0.96

In summary, all indicators are taken as the independent variable, and whether the policy is insured as the dependent variable, and the training set proportion is set to 0.7 to conduct random forest modeling. It can be seen from the table above that the accuracy of the final model is 96.1%, accuracy (comprehensive) is 96.1%, recall (comprehensive) is 96.2%, and f1-score (comprehensive) is 0.96. The model works better and is better than the SVM model.

Classification of different areas of the US using random forest modeling, the results are shown in table 5:

Table 5. Determine whether to underwrite

Region	Yes (1) or No (0)
ALABAMA	1
ALASKA	1
……	
VIRGINIA	1
WASHINGTON	1
WEST VIRGINIA	1
WISCONSIN	1
WYOMING	1

We looked up the relevant index data for London in Europe and Sydney in Oceania (where two extreme weather events occur on different continents) to demonstrate our model, as shown in the table 6 below:

Table 6. Demonstrate the model

Region	London	Sydney
the number of extreme weather events in the past decade	120	100
the number of extreme weather events in the next decade	125	101
percentages of healthcare expenditure	0.095	0.056
age composition (under 25 years old) (%)	0.18	0.22
employment (%)	0.75	0.82
crime rates (%)	0.01	0.008
GDP per capita	45000	30000
per capita income	38000	25000
home ownership rates	0.6	0.55
Yes (1) or No (0)	1	1

After the prediction using random forest, the results are as follow table 7:

Table 7. whether to underwrite or not

Region	Yes (1) or No (0)
London	1
Sydney	1

As a result, insurance companies can cover insurance in areas where extreme weather events are increasing.

3. Conclusion

In view of the challenges posed by insurance companies and owners in recent years, this study proposed a forecasting model based on neural network and applied to the underwriting decisions of the insurance industry. By screening data on casualties, property and crop losses, we successfully identified extreme weather events and constructed a set of risk index prediction system including LSTM model. Moreover, we conducted a correlation analysis to verify the effectiveness of the index system and solved the classification problem using the SVM model and the random forest model.

Through an in-depth analysis of the model results, we provide specific suggestions for insurance companies and property owners. The application of the model not only helps insurance companies to assess underwriting risks more accurately and develop more reasonable premium strategies, but also

can provide risk early warning and preventive measures for owners to reduce potential losses. At the same time, we selected two regions with different continents as cases to show the effectiveness of the model in practical application.

Overall, the neural network-based prediction model proposed in this study provides a new solution for the insurance industry to cope with extreme weather events. The model has strong practicability and promotion value, and is expected to provide strong support for the sustainable development and risk management of the insurance industry. In the future, we will continue to optimize the model algorithm and expand the application field to better serve the insurance industry and social and economic development.

References

[1] Zhang Yue. Research on the size measurement of catastrophe insurance fund in China [D]. Sichuan: Southwest University of Finance and Economics, 2023.

[2] Zhang Litian, Bu Qingjie, Yang Guihua, et al. Problems of the correct use of commonly used mathematical and statistical methods in academic papers in the field of environmental sciences [J]. Journal of Environmental Science, 2007 (1): 171 - 173.

[3] A SVM based ship collision risk assessment algorithm[J]. Ocean engineering,2020, 202 (Apr.15): 107062. 1 - 107062. 11.

[4] Yao Tianlei, Chen Xiliang, Yu Peiyi. Review of generative reinforcement learning studies based on sequence modeling [J/OL]. computer science.

[5] L. Zhou and Y. Cui, Parameter Optimization and Its Application of Support Vector Machines Based on Improved Particle Swarm Optimization Algorithm.2022 4th International Conference on Intelligent Information Processing (IIP), Guangzhou, China, 2022, pp. 213 - 216.

[6] Zou Xiang, Liu Ton, Huang Pu, et al. An intelligent prediction method of natural gas consumption based on LSTM [J]. Automation Expo, 2023, 40 (6): 60 - 64.

[7] Chen Yamin, Wang Ke, Wang Zhan, Wang Huiqin, Li Yuan, Zhen Gang. Pigment classification method for mural multispectral images based on multiscale superpixel segmentation [J/OL]. Progress in laser light and optoelectronics.

[8] R. Zhang, W. Wang. Facilitating the applications of support vector machine by using a new kernel [J]. Expert Systems with Applications, 2011 38 (11): 14225 - 14230.

[9] G. F. Smits, E. M. Jor dan. Improved SVM regression using mixtures of kernels [C]. International Joint Conference on Neural Networks. IEEE Xplore, 2002.

[10] Huang Mingfeng, Du Hai, Wang Qing et al. Typhoon extreme wind speed prediction for cities and offshore waters along the southeast coast of China [J]. Journal of Building Structures, 2024, 45 (05): 104 - 114.