Port container throughput prediction method based on SSA-SVM

Weiyuan Wu [1, #], Long Ma [2, #, *], Shangzhi Gao [3, #]

[1] School of Management of Tianjin University of Technology, Xiqing, Tianjin, China, 300384

[2] School of Business of Central South University, Hunan, Changsha, China, 410083

[3] School of Science and Preparatory of Dalian Minzu University, Dalian, Liaoning, China, 116650

* Corresponding author: melon770301@163.com

[#]These authors contributed equally.

Abstract. To improve the prediction accuracy of the port cargo throughput and the applicability of the prediction model, and then provide data support for the port construction to meet the needs of port decision-making, take the monthly cargo throughput data of Shanghai Port from January 2009 to December 2022 as an example, use Pearson correlation analysis to screen 12 import and export impact factors. This paper improves the traditional SVM model, uses SSA (Sparrow Search Algorithm) to optimize the parameters c and g in SVM, and uses the model to predict. Compared with the model that uses the grid search algorithm to optimize the parameters of SVM, the model has a significant improvement in fitting and robustness, its predicted value is closer to the actual value, the prediction performance is better, and it can better reflect the actual state of the port.

Keywords: Port throughput prediction, Correlation analysis, SSA, SVM.

1. Introduction

With the implementation of the national strategy of "Belt and Road", ports are playing an increasingly important role, and port throughput has become an important indicator for measuring the development of port construction. Accurate forecasting of port throughput is important for national and regional development, and accurate forecasting results can provide support for major government decisions and help enterprises plan layout, transformation and upgrading.

At the moment, many domestic and foreign scholars are studying the port throughput forecasting problem, and the methods used can be basically classified into statistical method [1], machine learning algorithms[2], and error correction models[3]. Statistical measures include time series, gray forecasting, and Markov models. To address the problem of seasonal errors in port throughput, Hui Li[4]used the holt-winter algorithm to forecast the cargo throughput of the Yangtze River mainline, but the forecast results only gave the nonlinear growth dynamics of the throughput of major cargo types without providing specific forecast values. Javed. F et al.[5]. explored the SARIMA model for the major international ports' container throughput forecasting problem. However, the time span of seasonal cycle forecasting is too large, and to meet short-term throughput forecasting needs, Yu et al. [6]. collected historical data of Fujian coastal ports from 2001-2016, the gray model GM(1,1) combined with Markov chain, and used the Markov chain to account for the stochastic nature of variable fluctuations ,and Corrections are made to the forecast values. Considering that container handling volume is influenced by many factors and is a completely gray information system, which is not effective with a single forecasting model, He.C et al. [7]. constructed a combined forecasting model that combines fractional order GM(1,1) and BP neural net to forecast the container throughput of Jinji Port Group for 2021-2025.

The classical methods in machine learning algorithms are mainly neural net and SVM. Baochai D. et al.[8]. combined pca and BP neural net , to achieve data dimensionality reduction by pca, and then used BPNN model for prediction. Fu et al. [9]. presented an ABC-BPNN forecasting model based on manual bee swarm optimization, which optimized the weight and threshold value of BP and effectively improved the prediction accuracy compared with the traditional BP. Changan Li et al. [10].

then used the ant colony algorithm to optimize the initial weight and threshold value of BP to achieve throughput prediction. Gao et al. [11]. used LSTM recurrent neural network (RNN) to achieve the everyday volume of upcoming containers into the yard containers. Jin et al. [12]. used PCA with dimensionality reduction and then used SVR to predict the Tianjin port throughput in the new period. Combining advantages of direct and iterative forecasting models, Changli Song et al. [13]. constructed a multi-step hybrid forecasting method based on support vector machine model to achieve throughput forecasting for major cargoes at Dalian port. And considering the problem of possible outliers in time series, Zhankun Guo et al. [14]. proposed a LOF-SSA-LSSVM forecasting model based on local outliers.

2. Methodology

2.1. Pearson Correlation Analysis

In the prediction model with multiple features, the correlation between features will alter the accuracy of the model and the significance of forecast coefficient. In order to explore whether there is a strong correlation between features and filter the features, we choose to use Pearson Correlation Coefficient to solve the correlation between features, the arithmetic formula is as follows:

$$R(X,Y)=\frac{E\left[\left(X-\overline{X}\right)\left(Y-\overline{Y}\right)\right]}{\sqrt{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)}\sqrt{\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)}} \tag{1}$$

E is the mathematical expectation, $\overline{X}$ is the average of X. $\overline{Y}$ is the average of Y. When the Pearson Coefficient is greater than 0.3, we believe that elimination is required.

2.2. Principle of Sparrow Search Algorithm

The Sparrow Search Algorithm[15] is a swarm intelligence optimization algorithm based on a series of behaviors of the sparrow group in the foraging process. It consists of three parts: discoverer, follower and alerter. The discoverer is in the nearest position to the food in the sparrow group, leading the group foraging; followers update their location based on the finder to increase the probability of obtaining food; the alerter is a part of the discoverer and follower, generally accounting for 10 % - 20 % of the group. When the alerter detects the predator 's attack, it alerts the sparrow population. The sparrows in the population will approach each other to reduce the probability of being preyed on and make anti-predation behavior.

The location update formula of the discoverer is as follows:

$$X_{i,j}^{t+1}=\begin{cases} X_{i,j}^{t}\cdot\exp\left(-\dfrac{i}{\alpha\cdot iter_{\max}}\right) & R_2<ST \\ X_{i,j}^{t}+Q\cdot L & R_2>ST \end{cases} \tag{2}$$

where $X_{i,j}^{t+1}$ means the situation of the i th individual in the j th dimension at the time of the $t+1$ iteration, α is a odd number between $[0,1]$, $iter_{\max}$ represents the maximum number of iterations, and R_2 represents an early warning value between[0,1]. ST indicates a safety value between [0.5,1],When the warning value is greater than the security value, it implys that a sparrow spotted the predator and promptly sent an attention signal. The population adjusts the search strategy and makes

anti-predation behavior close to the safe area. Q is a random number that obeys a normal circulation, and L represents a matrix of 1 by d, where each component in the matrix is 1.

The follower 's situation renew formula is as follows:

$$X_{i,j}^{t+1}=\begin{cases} Q\cdot\exp\left(\dfrac{X_{worst}-X_{i,j}^{t}}{i^{2}}\right) & i>\dfrac{n}{2} \\ X_{p}^{t+1}+\left|X_{i,j}-X_{p}^{t+1}\right|\cdot A^{+}\cdot L & otherwise \end{cases} \tag{3}$$

Among them, X_p and X_{worst} represent the optimal position and the least position in the group respectively. A performs a matrix of 1 by d, in which the elements are randomly assigned 1 or - 1, and $A^{+}=A^{T}\left(AA^{T}\right)^{-1}$. When $i>0.5n$, it indicates that the less adaptable i th adder does not have access to food and needs to fly elsewhere for.

Each iteration has SD sparrows as alerters, and their position update formulas are as follows:

$$X_{i,j}^{t+1}=\begin{cases} X_{best}^{t}+\beta\cdot\left|X_{i,j}^{t}-X_{best}^{t}\right| & f_i>f_g \\ X_{i,j}^{t}+K\cdot\left(\dfrac{\left|X_{i,j}^{t}-X_{worst}^{t}\right|}{f_i-f_w+\varepsilon}\right) & f_i=f_g \end{cases} \tag{4}$$

β is an odd number that obeys the typical ordiary distribution. K is a random number between - 1 and 1. f_i is the fitness amount of the current sparrow individual while f_g and f_w is the current global best and worst fitning values, respectively. ε is the smallest constants to escape the denominator appearing zero.

2.3. Support Vector Machine--SVM

SVM can become a nonlinear classifier by introducing the kernel function technique[16]. The basic idea is to compare the input space to a feature space through a nonlinear conversion, so that the decision hyperspace model in the input space corresponds to the decision hyperplane model in the feature space.

ζ_i represents a slack variable that measures the distance between the misclassified sample and the hyperplane, and C is a penalty factor.

$$\begin{aligned} &\min\phi(w,\zeta)=\frac{1}{2}\left\|w^{2}\right\|+C\sum_{i=1}^{n}\zeta_i \\ &s.t\quad y_i\left[\left(w\cdot x_i+b\right)\right]-1+\zeta_i\geq 0\quad i=1,...n,\zeta_i\geq 0 \end{aligned} \tag{5}$$

$$\begin{aligned} &\min_{\alpha}\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j x_i^{T}x_j-\sum_{i=1}^{m}\alpha_i \\ &s.t.\quad\sum_{i=1}^{m}\alpha_i y_i=0 \\ &0\leq\alpha_i\leq C,\quad i=1,2,3...,m \end{aligned} \tag{6}$$

We can transform the problem into a dual problem by constructing a langrange function. In order to curtail the space complication and realize the nonlinear problem, we introduce the kernel function

to solve the hyperplane equation according to the quadratic programming technique and the kernel function.

$$f(x)=\sum_{i=1}^{m}\alpha_i y_i k(x,x_i)+b\text{, while } k(x,x_i)=\varphi(x_i)\cdot\varphi(x)=\exp\left(\frac{-|x_i-x|^2}{g^2}\right) \text{ is kernel function.}$$

2.4. Port Throughput Prediction Model Based on SSA-SVM

The influencing elements of port cargo throughput are complex. SVM is an effective method to predict port cargo throughput, Moreover, SVM can be combined with optimization algorithms to optimize and improve the parameters of the model to further improve the accuracy of model prediction. The accuracy of SVM prediction model is closely related to parameters c and g. The value of c will affect the classification accuracy of the classifier, and the value of g will affect the division of feature space. If the value of g is too large, it will lead to excessive fitting, and if the value is too small, it will lead to insufficient fitting[17]. The parameters determined by artificial random setting or classical grid search method largely depend on people ' s subjective ideas, and the accuracy is not reliable. The intelligent optimization of SVM model parameters by SSA algorithm is that it can automatically search the SVM model parameters in the global range to achieve the global optimal accuracy[16].

The support vector machine (SVM) model has a good ability to deal with small sample data, and can effectively solve two very important parameters in the SVM regression model: penalty coefficient c and RBF kernel function g. In this paper, the SVM regression model is used to construct the port cargo inbound and port cargo outbound models, and the most important parameters c and g of the SVM regression model are optimized using the grid search method (GS) and the sparrow optimization algorithm (SSA) in the process of model construction and analysis. The prediction accuracy of the GS-SVM regression model and the SSA-SVM regression model will be compared respectively. The construction flow chart of the combined model and algorithm is shown in Figure 1.

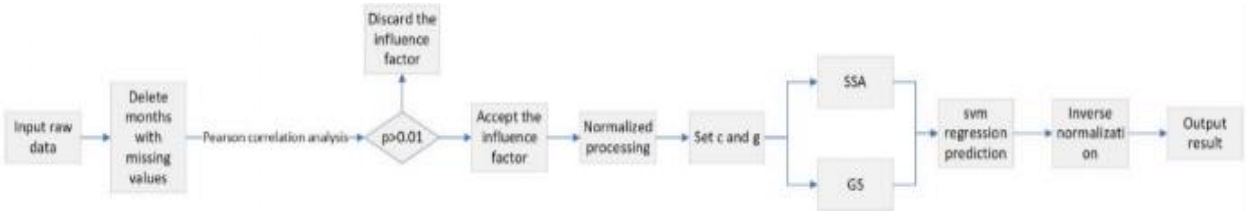

Figure 1. Flow Chart of Port Throughput Forecast Based on SSA-SVM

3. Example analysis

3.1. Data source and pre-processing

To contrast the prediction effectiveness of GS-SVM and SSA-SVM models, this paper takes Shanghai port as an example to collect monthly cargo throughput from Shanghai Municipal Bureau of Statistics from January 2009 to December 2022. Existing research shows that the factors affecting port cargo throughput are complex, but it can be divided into the port's own factors such as container throughput, port freight equipment, port employment and economic factors such as GDP, trade import and export, per capita consumption of urban residents. This paper chooses the container throughput and the import and export trade volume as the main factors, which can be accurate to the month, considering that the port cargo throughput is composed of the port cargo inbound and outbound, this paper divides the container throughput into the port incoming and the port outgoing and the import and export value of trade is divided into import value and export value according to the classification standard of product type and market type.

The prediction models require the data dimensions of independent and dependent variables to be consistent, but the lack of data will impact the construction of the prediction model, hence, this article

adopts the method of data preprocessing by deleting the missing values. That leaves 107 data sets from January 2011 to December 2019. To avert the effect of the singular data on the convergence of the model training, the MAPMINMAX method is used to normalize the data:

$$X_i = \frac{X_i - X_{\min}}{X_{\max} - X_{\min}} \tag{7}$$

3.2. Screening of critical influencing factors based on Pearson correlation analysis

This paper takes Shanghai port as an example to build a forecast model of cargo inflow. There are numerous factors that impact the cargo arrival at Shanghai port. But not every factor has a significant correlation with the volume of goods entering Shanghai Port. If all these factors are involved in the training model without screening, the convergence rate of the model will be reduced, and the robustness and generalization of the model will be affected. Therefore, Pearson correlation analysis was used to screen the possible related factors, and the factors with significance level greater than 0.01 were excluded. The factors after the final screening were shown in Table 1.

Table 1. Factors affecting cargo throughput of Shanghai port

Influencing factors of cargo volume of Shanghai Port	Influencing factors of Shanghai port cargo output
Total imports: value of the month (US $billion)	Total exports: value of the month (US $billion)
Total imports: state-owned enterprises: value of the month (US $100 million)	Standard container throughput: Out of Port: monthly value (10,000 TEU)
Total imports: foreign-invested enterprises: the value of the month (billion U. S. dollars)	Exports by market: Hong Kong, China: value of the month (US $billion)
Total imports: private sector: value of the month (US $100 million)	Total exports: state-owned enterprises: value of the month (US $100 million)
Total imports: general trade: value of the month (billion U. S. dollars)	Total exports: foreign-invested enterprises: value of the month (US $100 million)
Total imports: processing trade: value of the month (billion U. S. dollars)	Total exports: private sector: value of the month (US $100 million)
Total imports: mechanical and electrical products: value of the month (US $100 million)	Total exports: General Trade: value of the month (US $100 million)
Total imports: high-tech products: value of the month (US $100 million)	Total exports: high-tech products: value of the month (US $100 million)
Total import and Export: customs area: import: value of the month (billion U. S. dollars)	Total exports: mechanical and electrical products: value of the month (US $100 million)
Foreign direct investment: contract amount: tertiary industry: value of the month ($billion)	Exports by market: Japan: value of the month (US $100 million)
Foreign direct investment: contract amount: wholly foreign-owned: value of the month (US $100 million)	Exports by market: United States: value of the month (US $100 million)
Foreign direct investment: contract amount: current month value (US $billion)	Exports by market: EU: Monthly Value (US $100 million)

3.3. Port throughput forecast based on SSA-SVM

3.3.1. Critical parameter setting

Taking Shanghai port as an example to build a forecast model of cargo inflow. In this paper, GS-SVM prediction model and SSA-SVM prediction model are used to set the key parameters C and G. In this paper, cross-validation method is used in GS-SVM prediction model whose C value is 0.5 and G value is 2.8284. In the SSA-SVM model, we assume the number of sparrow population to be 20, the proportion of finders to be 0.7, and assume the value range of C and G to be 0 to 100. After

optimizing the parameters of sparrow optimization algorithm, the value of C is 30.4, the G value is 99.9.

3.3.2. Model training and validation

Taking Shanghai port as an example to build a forecast model of cargo inflow. In this paper, all the normalized data are used as the training data, the forecast value of port inflow can be obtained by the de-normalization of the data predicted by the trained model. The predicted value of port inflow of SSA-SVM regression model is fitted to the original value as shown in Figure 2, while the predicted value of port inflow of GS-SVM regression model is fitted to the original value as shown in Figure 3.

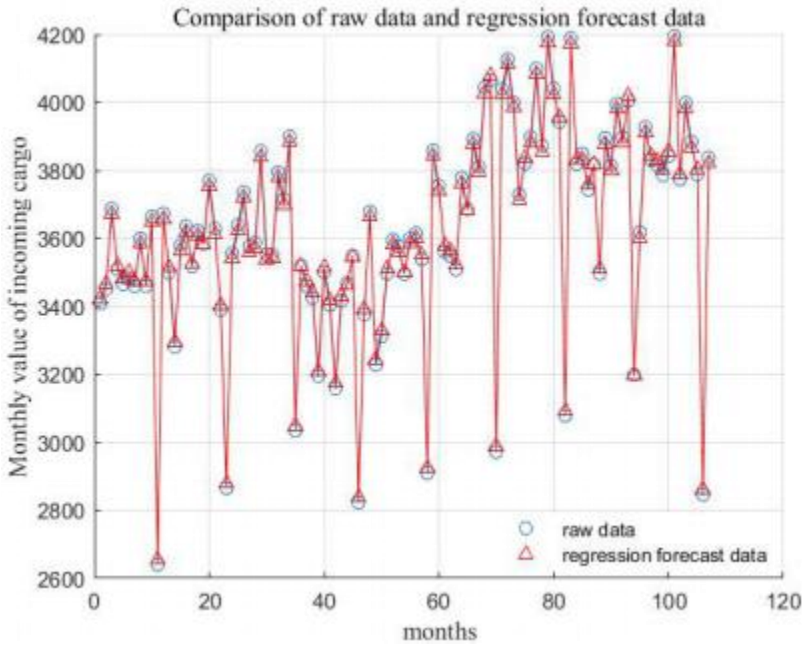

Figure 2. Prediction results of cargo inflow at Shanghai Port based on SSA-SVM model

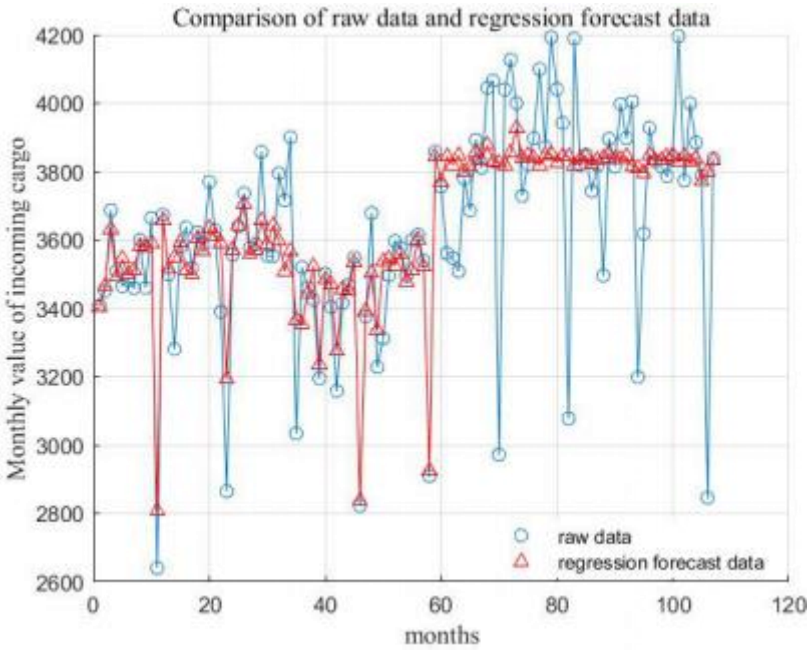

Figure 3. Prediction results of cargo inflow at Shanghai Port based on GS-SVM model

3.4. Prediction results and accuracy test

In this paper, the regression evaluation indexes MSE and R2 were used to test the predictive validity of the model. The following table 2 shows the comparison of the prediction accuracy between the GS-SVM model and the SSA-SVM model.

Table 2. comparison of prediction accuracy of each model

	Prediction model of cargo outbound volume		Prediction model of cargo inbound volume	
	MSE	R^2	MSE	R^2
GS-SVM	0.00315456	89.7938%	0.0183649	53.6583%
SSA-SVM	$0.17741e^{-}05$	99.7817%	$9.16163e^{-}05$	99.8299%

Since SSA algorithm can search parameters C and G of SVM model, the global optimal solution can be obtained. As the results shown in Table 2, Compared with GS-SVM, SSA-SVM can significantly improve the fit degree of predicted and original values, and significantly reduced the MSE of predicted and original values.

4. Conclusion

According to the conclusion of this paper, the SSA-SVM prediction model, which is built in this paper, is verified for the prediction of Shanghai Port Cargo Volume and Shanghai Port Cargo Volume, the two indexes R2 and MSE, which can be used to evaluate the prediction model, are better than GS-SVM. In other words, the parameters c and g obtained by sparrow optimization algorithm are more accurate than those obtained by grid search algorithm. Compared with grid search method, Sparrow optimization algorithm can better optimize the parameters of SVM regression prediction model. Based on the existing research, this improved SVM regression forecasting model proposes a new forecasting method of port cargo throughput, which can be used in the practice of port cargo throughput forecasting.

References

[1] R Lyman Ott, Micheal T Longnecker. An introduction to statistical methods and data analysis [M]. Cengage Learning, 2015: 1305465520.

[2] Zhou, Zhi-Hua. Machine learning [M]. Springer Nature, 2021: 978 - 981 - 15 - 1967 - 3.

[3] Wang R, Tan Q. Dynamic Model of Port Throughput's Influence on Regional Economy[J]. Journal of Coastal Research, 2019, 93 (SI): 811 - 816.

[4] Li Hui. Analysis and forecast of cargo throughput of the Yangtze River Trunk line based on Holt Winters Algorithm [J]. China Water Transport, 2021 (4): 29 - 32.

[5] Javed Farhan, Ghim Ping Ong. Forecasting seasonal container throughput at international ports using SARIMA models [J]. Maritime Economics & Logistics, 2018, 20 (1): 131 - 148.

[6] Wang Yu, Wang Zhiming. Combined throughput prediction of Fujian coastal ports based on grey model and Markov Chain[C]// Proceedings of 2018 International Symposium on Social Science and Management Innovation (SSMI 2018): 112 - 119.

[7] He Chen, Wang Huipo. Container throughput forecasting of Tianjin-Hebei port group based on grey combination model [J]. Journal of Mathematics, 2021, 2021: 8877865.

[8] Baochai D. Research on Prediction of Port Cargo Throughput based on PCA-BP Neural Network Combination Model[C]//2020 5th International Conference on Information Science, Computer Technology and Transportation (ISCTT). IEEE, 2020: 518 - 523.

[9] Huang Fucheng, Liu Dexin,An Tiansheng,Cao Jie. Port Container Throughput Forecast Based on ABC Optimized BP Neural Network [J]. IOP Conference Series: Earth and Environmental Science,2020, 571 (1): 012068.

[10] Li Changan, Lu Xueqin, Wu Zhong qiang. Throughput prediction of port based on Back Propagation Neural Network Optimized by Ant Colony Algorithm[J]. Acta Metrologica Sinica, 2020, 41 (11): 1398 - 1403.

[11] Gao Yinping, Chang Daofang, Fang Ting, et al. The Daily Container Volumes Prediction of Storage Yard in Port with Long Short-Term Memory Recurrent Neural Network [J]. Journal of Advanced Transportation, 2019, 2019: 5764602.

[12] Jinyu Wei, Yuqiao Tang, Yang Yu, Xueshan Sun. Research on Port Throughput Prediction of Tianjin Port Based on PCA-SVR in the New Era [C]. Deng, Z. (eds) Proceedings of 2019 Chinese Intelligent Automation Conference. Lecture Notes in Electrical Engineering, 2019, 586: 57 - 64.

[13] Song Changli, Ji Lianjie, Guan Feng, et al. Research on throughput prediction of dalian port's main cargo based on support vector machine[J]. Journal of Dalian Ocean University, 2019, 34 (5): 752 - 756.

[14] Guo Zhankun, Jin Yongwei, Liang Xiaozhen, et al. Prediction model of port container throughput based on outlier detection[J]. Mathematics in Practice and Theory, 2019, 49 (17): 26 - 34.

[15] Xue J, Shen B. A novel swarm intelligence optimization approach: sparrow search algorithm [J]. Systems Science & Control Engineering, 2020, 8 (1): 22 - 34.

[16] Jin aibing, Zhang Jinghui, Sun Hao, Wang Benxin. Intelligent Prediction and early warning model of slope instability based on SSA-SVM [J]. Journal of Huazhong University of Science and Technology Science, 2022, 50 (11): 142 - 148.

[17] Li Shufeng, Li Jia, Zhang Yufeng, Wang Dapeng, Yuen Pei-sen. Study on Particle swarm optimization Support vector machine outage prediction [J]. Journal of Nanjing University of Science and Technology, 2022, 46 (4): 460 - 466.