Scientific and Social Research
2022, Volume 4, Issue 7

Forecast of Logistics Demand in the Pearl River Delta Region Based on PCA-GA-SVM Model

Yiyong Ye*

School of Economics and Management, Wuyi University, Jiangmen 529020, Guangdong Province, China

***Corresponding author:** Yiyong Ye, yyyong2022@163.com

Abstract: Regional economic development is highly correlated with the change of regional logistics. This paper selects the freight volume as the representative index to reflect the development of regional logistics, and constructs the prediction index system of regional logistics demand. Accordingly, the principal component analysis method is used to reduce the data dimension of the prediction index, and the complexity of the prediction model. Further, the support vector regression model is optimized by genetic algorithm which is constructed by using the advantages of support vector machine algorithm in dealing with nonlinear and small sample size problems. The empirical analysis shows that the prediction model based on PCA-GA-SVM has very good prediction accuracy, it can provide valuable reference for regional logistics development and management.

Keywords: Regional logistics; Principal component analysis; Support vector regression; Genetic algorithm

Online publication: July 14, 2022

1. Preface

Regional logistics and economic development are inseparable. On the one hand, the large-scale and modern development of regional logistics not only provides sufficient power for regional economic development, but also helps to enhance industrial advantages, and promote the high-quality development of regional economy. Additionally, the sustainable development of regional economy also provides a solid foundation for the healthy development of the logistics industry. Further, the optimization and upgrading of the economic and industrial structure respectively, are conducive for the rapid development of the emerging logistics industry, leading to gradual developed of the logistics industry into a new economic growth point. At present, the research on the development of logistics industry has attracted much attention, especially in the fields of logistics theory research, logistics model, and logistics technology innovation, which has made considerable achievements, providing an effective basis for the development of the logistics industry, and the decision-making of logistics management departments. Based on the high correlation between regional logistics and economic development, this paper forecasts the development status of regional logistics through regional economic indicators, realizes the scientific and accurate prediction of regional logistics demand, and provides a reliable quantitative basis for regional logistics management and development planning.

2. Literature review

At present, the research on regional logistics demand forecast is mainly reflected in two aspects; (1) To use logistics data itself to predict the logistics demand, including logistics demand forecast in specific regions

and fields [1-6]; (2) To use regional economic data to predict the logistics demand [7-13], including the logistics demand forecast as a regression problem, and drawing lessons from the prediction method in the field of economics of logistics demand forecast research, by building a prediction model, study the relationship of regional economic development and the quantity of logistics demand, analyze the influence of regional economic development on the development of logistics, and subsequently obtained certain research results.

Due to the short development time of China's logistics industry, there is relatively lack of logistics statistical data, however, there are many economic indicators that may affect the logistics demand, therefore, in the actual research it is necessary to ensure the accuracy of prediction model. Secondly, the index system of the model should not be too complex, lastly less amount of indicator data should be sufficient for the model development. Combining with the existing research foundation, this paper firstly uses the principal component analysis method to reduce the data dimensionality of the predictive index system to reduce the complexity of the model, and then uses the support of vector machine model, which is optimized by the genetic algorithm to solve the problems of small sample size and nonlinearity, thereby can effectively meet the above forecast requirements.

3. Prediction model construction

3.1. Construction of predictive index system

According to the close relationship between regional economy and the regional logistics, through correlation analysis, and integration with the availability of index data, the economic indicators which are selected in this paper includes; Gross domestic product, industrial output value, agricultural output value, tertiary industry output value, fixed asset investment, regional total retail sales, regional total foreign trade, per capita disposable income, and the forecast index is freight volume. In contrast, macro policy, logistics service level, transportation network, and other factors are not considered, due to the difficulty to quantify these statistical indicators, additionally these index data are difficult to obtain.

3.2. Principal component analysis (PCA)

Principal component analysis (PCA) is a statistical method which is used to reduce dimension. With the help of orthogonal transformation, the original variables are recombined into a new group of several linear unrelated comprehensive variables. At the same time, it can capture few comprehensive variables and reflect the information of the original variables as much as possible. Therefore, PCA can effectively reduce the dimension of the data space studied, subsequently, reduce the complexity of the predictive index system.

3.3. Support-vector regression model (SVR)

The kernel method originates from statistical learning theory, mainly applied in the field of pattern recognition, and it's largely used is to find and learn interrelations in a set of data, which is an effective way to solve the problem of nonlinear pattern analysis. Support vector regression is a classical model of nuclear methods, which performs well in nonlinear small sample size and nondeterministic data, and has a wide application prospect in the complex nonlinear prediction and comprehensive evaluation [6]. More importantly, SVR aims at the minimization of structural risk, and achieves a good balance between the degree of model learning, and the generalization ability of the model, which effectively improves the shortcomings of other nonlinear intelligent algorithms.

3.4. Genetic algorithm (GA)

Genetic Algorithm originated in computer simulation studies on biological systems. It is a stochastic search algorithm that draws lessons from natural selection and natural genetic mechanisms in the biological community. Unlike conventional algorithms, genetic algorithms do not rely on the gradient information,

but search for optimal solutions by simulating natural evolutionary processes, which use some coding technology, and act on digital strings called chromosomes to simulate the evolutionary process of groups composed of these strings. Genetic algorithm has many advantages, such as the universality of feasible solutions, do not need auxiliary information, internal heuristic random search characteristics, not easy to trap local optimal, using the natural evolution mechanism to show complex phenomenon, can quickly and reliably solve the very difficult problem, has inherent parallelism and well computing ability, have scalability, easy to mix with other techniques, and others [6].

3.5. The PCA-GA-SVM prediction model

The support vector regression model achieves good statistical laws and better generalization ability with small statistical sample size, however, there are also some shortcomings; (1) Since training the SVM needs to solve the quadratic programming problems, when the number of training samples and the indicator dimensions is large, the operation time of the model will become longer; (2) The selection parameter of the support vector machine. Since the selection of the penalty coefficient and kernel functions in the model has a great influence on the predicted results, the current common practice is to combine the personal experience, and use the method of trial and error to determine the parameters of the model. Due to the lack of systematic theoretical guidance, this processing method has a certain degree of subjectivity, which leads to deviations in the predicted results. In order to solve this problem, a genetic algorithm is introduced in this paper. With the help of the powerful global search ability of the genetic algorithm, the parameters of the support vector machine regression model are genetically encoded and searched globally by using the real number coding method. As the final parameters of the support vector regression model, the specific algorithm flow is as follows:

(1) Use principal component analysis to reduce the dimension of sample data, and extract principal component information according to the cumulative contribution rate of eigenvalues.

(2) Initialize the value range of support vector regression model parameters (penalty coefficients and kernel parameters).

(3) Initialize the individual parameters of the genetic algorithm, and randomly generate M chromosomes according to the real number coding method to generate the initial population P(t) of the genetic algorithm.

(4) According to the gene sequence of the chromosome bit string, the selected factor combination set is obtained according to the selection strategy.

(5) For each individual in the initial population, the support vector regression program is used to calculate the predicted output value corresponding to the target value of the training sample, the misclassification rate of the training sample can be calculated, and finally the individual fitness value of the chromosome can be obtained.

(6) Execute M cycles to complete the calculation of the fitness value of each individual in the initial population.

(7) Perform selection, crossover, and mutation operations to form the next generation of subpopulations.

(8) For the optimal individual in the new population, the grid search method is used to search its nearby area, and the optimal individual is replaced by the searched parameter combination.

(9) Execute the iteration termination criteria, and stop it if the iteration termination conditions are met; otherwise, change the child to the new parent, and go to step 5 until the iteration termination conditions are met. At this time, the optimal individual in the population is the solution of parameter inversion.

(10) Using the obtained combination of optimal parameters (penalty coefficient and kernel parameter), substitute it into the support vector regression program in step (2) to predict and analyze the data.

4. Empirical research

4.1. Data collection and preprocessing

The data used in this paper obtained from the Statistical Yearbook of 9 cities in the Pearl River Delta of Guangdong Province (1990-2020), of which the data from 1990-2015 is used as training data to determine the parameters of the model, and the data from 2016-2020 is used as test data for Test the predictions of the model.

Due to the different dimensions and large differences in the values of each prediction index, before applying the data, the original data was initially standardized, then the sample data is converted into the interval [0,1], to eliminate the impact of dimensional differences on the comparability of sample data. The data conversion formula is as follows:

$$T = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

In this formula, X represents the original data; X_{min} represents the minimum value of the original data; X_{max} represents the maximum value of the original data; and T is the transformed data.

4.2. Indicator correlation analysis and data dimension reduction

Firstly, SPSS software was used to analyze the correlation between independent and dependent variable in the prediction index. It was found that the correlation coefficients of independent variable and dependent variable were all over 0.93, indicating a significant positive correlation. In addition, there is also a high degree of autocorrelation among the 8 independent variables, and the correlation coefficients are all higher than 0.90, indicating these variables may exist multicollinearity as independent variables in the regression model. Therefore, in order to ensure the prediction effect of the model, dimension reduction should be conducted on the sample data initially to reduce the mutual interference and influence among indicators. In this paper, PCA was used to reduce the data dimension. With 90% cumulative variance contribution rate as the selection standard, the first and second principal component indexes were selected to replace the original 8 input indexes, which not only reduced the complexity of the index system, but also improved the prediction accuracy of the model.

4.3. Model training and testing

Firstly, to initialize the model parameters; 3-e-SVR is used as the type of support vector machine; The radial basis (RBF) kernel function is used as the kernel function; The value range of the penalty parameter C is [0, 100]; and The value range of the radial basis kernel function parameter is [0,100]. The maximum evolutionary generation of the genetic algorithm is 300, the maximum population is 40, the crossover probability is 0.3, and the mutation probability is 0.01. Then use the data from 1990 to 2015 for training, and the model calculates a stable iterative value of the best fitness after 300 iterations. The result is shown in **Figure 1**.

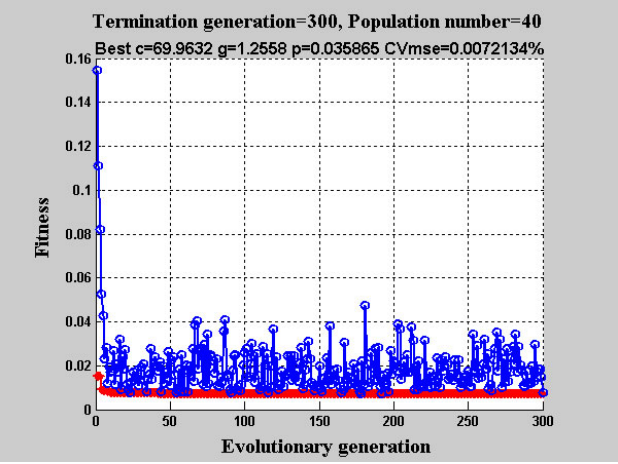

Figure 1. Fitness curve

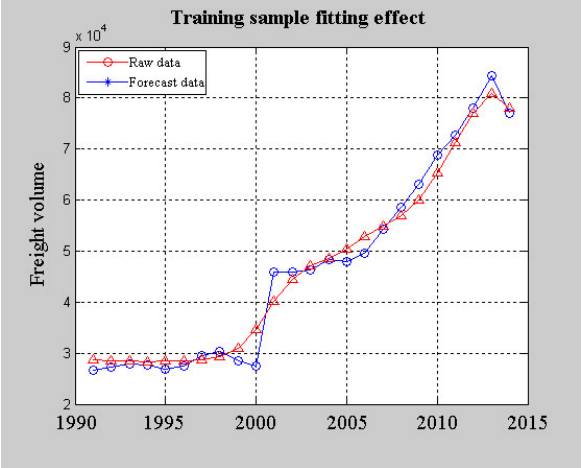

Figure 2. Training sample fitting effect

Figure 1 showed that, after the optimization and selection of the genetic algorithm, the penalty parameter C of the support vector machine model is 69.9632, and the value of g is 1.2558. At this time, the minimum mean square error of the training set is 0.0072134%. Then substitute these two parameters into the SVM model for calculation, and the training results of the training data set are shown in **Figure 2**. It can be seen that the SVM model with optimized parameters has an ideal fitting effect on the sample data, except for individual samples, the forecast data of other years are basically the same.

In order to verify the superiority of the PCA-GA-SVM model with the unchanged experimental data, the standard SVM and PSO-SVM models were used for prediction, and compared with the predicted results of the PCA-GA-SVM model, the results are shown in **Figure 3** and **Figure 4** for prediction effects and prediction error, respectively.

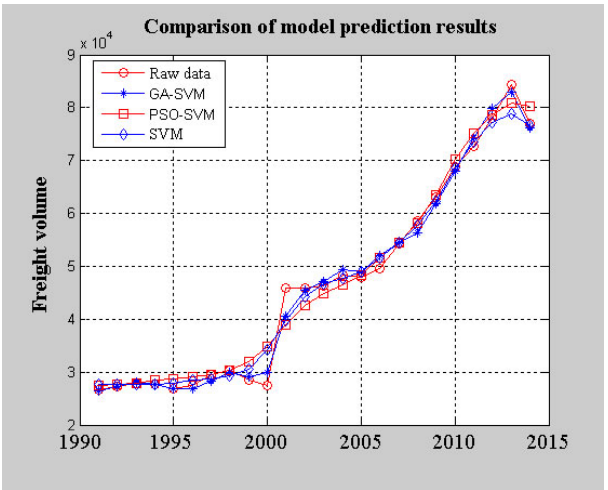

Figure 3. Comparison of the prediction effects

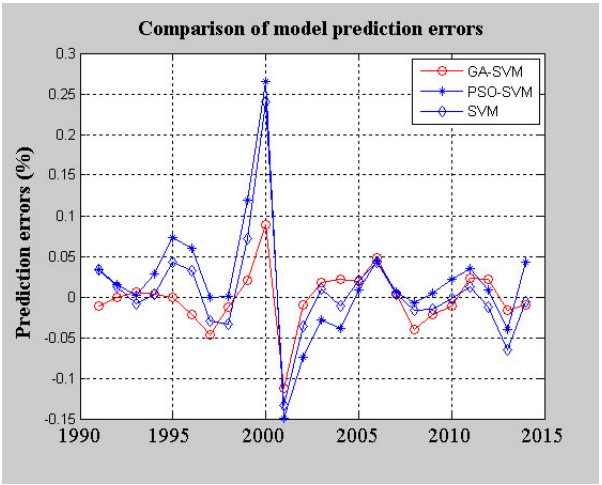

Figure 4. Comparison of the prediction error

Combining with **Figure 3** and **Figure 4**, it can be seen that the PCA-GA-SVM model has more advantages in prediction accuracy, better fitting effect for nonlinear data, and better compatibility with outlier data compared to the standard Svm and Pso-Svm models. In order to accurately calculate the sample training error values of the three models, this paper uses the average relative percentage error to evaluate the prediction performance, as shown in **Table 1**.

Table 1. Prediction error

Method	PCA-GA-SVM	PSO-SVM	standard SVM
Average error	2.45%	5.61%	4.74%

It can be seen that the error of the PCA-GA-SVM model is the lowest, and its prediction error fluctuation is also the smallest. The main reason is that the model uses the global optimization ability of the genetic algorithm to search for the optimal penalty parameter C and kernel function, in contrast these parameters cannot be achieved using the traditional Svm and Pso-Svm model.

The trained model is used to predict the logistics demand from 2016 to 2020, and compare it with the raw freight volume to test the actual effect of the model, and the results are shown in **Table 2**.

Table 2. Model prediction error

Raw data	81023	93318	94376	91535	90341
Forecast data	80023	91347	95854	90064	87972
Prediction error	1.25%	2.16%	-1.54%	1.63%	2.69%

As shown in **Table 2**, the average prediction error of the model is below 2%, and the accuracy is very high, indicating that the model not only has a good fitting ability to the training sample data, but also has a very strong generalization ability to the test sample data, therefore, the model can be used to forecast the freight volume in the Pearl River Delta in the short term.

5. Conclusion

Based on the close relationship between regional economy and logistics, this paper constructs a PCA-GA-SVM regional logistics demand prediction model according to the characteristics of logistics data. The empirical analysis shows that the model has higher prediction accuracy, and its prediction results are better, compared to the standard SVM and PSO-SVM model. In additional, the PCA-GA-SVM model can realize the accurate prediction of logistics freight volume in the short term, thus providing an effective reference for the management decision-making of the logistics management department.

Acknowledgments

This research was supported by the Specialty and Innovation Projects for Colleges and Universities of Guangdong Province (Project No. 2020WTSCX097), and the Guangdong Philosophy and Social Science Planning Project (Project No. GD20XGL14).

Disclosure statement

The author declares no conflict of interest.

References

[1] Hu P, 2019, Research on Urban Logistics Demand Forediction and Development Countermeasures Based on Arima-BP, Tianjin University of Technology.

[2] Wu M, Li B, 2018, Research on Logistics Industry Development Based on GM (1,1) Grey Prediction Model: Takes the Economic Growth Background in Henan Province as an Example. Henan Science, 36(08): 1305-1312.
[3] Ji Z, 2019, Guizhou Province in Guizhou Guizhou Based on Gray-Markov. Logistics Technology, 42(02): 145-149.
[4] Li L, Yue Y, Tian W, 2019, Evaluation and Prediction of Logistics Capacity in Beijing, Tianjin, and Hebei Based on the Gray Model of PCA and Markov Residues. Journal of Beijing Jiaotong University (Social Science Edition), 18(02): 129-142.
[5] Ma H, Liao Y, 2018, SVR Based on Genetic Algorithm. Logistics Technology, 37(03): 61-64+149.
[6] Cao Z, Yang Z, Liu F, 2018, Regional Logistics Demand Prediction of Support Vector Regression Machine. Journal of Systems Science, 26(04): 79-82+90.
[7] Chen G, 2019, The Cold Chain Logistics Demand Forecast of Henan Fresh Agricultural Products Based on the Grey Model. Modern Trade Industry, 40(11): 56-58.
[8] Liang Y, Yang H, Su H, 2018, Prediction and Analysis of Cold Chain Logistics Demand for Agricultural Products in Tianjin Based on Multiple Linear Regression. Southern Agricultural Machinery, 49(18): 230-231.
[9] Guo M, Li H, 2018, Demand Prediction of Fruits and Vegetables Based on PCA-RBF Neural Network Model. Jiangxi Agricultural Journal, 30(10): 137-141.
[10] Yin Y, Wang D, 2018, Grain Logistics Demand Prediction Analysis in Quanzhou Port Based on BP Neural Network Model. Technology and Industry, 18(11): 82-85.
[11] Tsai W, Huang H, 2019, Combined Model Analysis of Port Logistics Demand Based on BP-RBF Neural Network. Journal of Zhengzhou University (Engineering edition), 40(05): 85-91.
[12] Lu Y, 2018, Forecast of Dalian Port Based on Combined Prediction Model. China Market, 2018(27): 21-24.
[13] Zhao X, Zhang Jun, 2018, Tmall Double 11 Logistics Demand Forecast Based on the Interval Gray Number Prediction Model. Journal of Chongqing Technology and Business University (Natural Science Edition), 35(06): 40-45.