A Probe into the Influence of Major Infectious Diseases on the Grain Yield of Each Province based on ε-SVR Method

Xiaoxing Tong[1], Liang Meng[2] and Guo Yu[1]
[1]*Department of information technology, Jiaxing Technician Institute, Zhejiang, China*
[2]*School of Environmental and Geographical Sciences, Shanghai Normal University, Shanghai, China*

Keywords: ε-SVR, Major Infectious Diseases, The Grain Yield of The Province That Year, Prediction.

Abstract: The influence of major infectious diseases to the grain yield of the province was investigated by establishing a new prediction method based on ε-support vector regression(ε-SVR). the train model was built from historical data, including the grain yield of Beijing, Tianjing etc affected by SARS-CoV in 2003, Guangzhou in 1961 affected by cholera, Xinjiang in 1986 affected by Hepatitis E. It is proved that γ in radial basis kernel function is 0.01, penalty coefficient C is 1.0e + 7, loss function P is 10, the average relative error of model fitting is 1.96%, and the decisive coefficient is 0.99. We predict the production data of Gansu, Shanxi and Guangdong affected by SARS in 2003 and that of Guangdong affected by break-bone fever in 1978.The average relative error was 3.27%. However, after removing the two factors of the proportion of the infected population and the proportion of dead population, the model was built again. The average relative error of model fitting was 1.97%, and the average relative error of prediction was 3.31%. It shows that the major infectious diseases only have a small impact on grain yield. This model provides a new method for regional grain yield prediction and national macro-control in the short term.

1 INTRODUCTION

At the beginning of 2020, the global outbreak of COVID-19 caused the spread of the virus and the number of infected people. Due to its robust infection, the global multi national declaration issued in early April to stop grain exports to ensure its food supply. Under the epidemic situation, accurate prediction of the current year's grain output can help countries and regions better grasp the development trend of agricultural ecology, preserve the basic requirements of people's life, and even stabilize the people's hearts. Therefore, under the current situation, the analysis of the relationship between the epidemic situation and the current year's grain output has become an urgent questioned relating to people's livelihood. Although there are many grain prediction methods in the market, such as Nerlove model, system dynamics model, IPSO-BP model, time series model, there are still few studies on regional grain yield prediction under large-scale infectious diseases, and lack of theoretical basis for the government to provide grain macro-control.

In recent years, with the development of artificial intelligence and machine learning technology, SVM, an algorithm with strong generalization ability and wide applicability, has been widely used in the fields of agricultural production prediction, classification and image recognition, such as drip irrigation emitter flow prediction, identification of corn, soybean and rice, modern Agrometeorological analysis, research on irrigated cultivated land, research on topographic data of tea garden, and the final solution of SVM is convex quadratic programming problem, which is superior to neural network in dealing with local extreme value. In this research project, due to the differences of epidemic viruses, there are some characteristics, such as more serious outbreaks in individual regions and mild outbreaks in individual regions. SVM algorithm is easier to get the global optimal solution. The traditional SVM algorithm is only limited to binary classification, ε- SVR is an algorithm that can expand the regression problem on the basis of traditional binary classification.

Tong, X., Meng, L. and Yu, G.
A Probe into the Influence of Major Infectious Diseases on the Grain Yield of Each Province based on E-SVR Method.
DOI: 10.5220/0011356500003440
In *Proceedings of the International Conference on Big Data Economy and Digital Management (BDEDM 2022)*, pages 895-899
ISBN: 978-989-758-593-7

2 THEORETICAL BASIS

2.1 SVR and Its Kernel Function

As a binary classification model (non-zero is 1), SVM is based on the linear classifier with the largest interval in the feature space. In order to make SVM use of continuous values as regression prediction, the SVR model is proposed after optimization by multiple classification iterations. The existing SVR models can be divided into the following categories:

Linear kernel function

$$K(x, x_i) = x^t \cdot x_i \tag{1}$$

Multiform kernel function

$$K(x, x_i) = [(x \cdot x_i) + 1]^q \tag{2}$$

Where q is the order of polynomials, the resulting classifier is a polynomial of order Q.

(3) Radial basis function (RBF)

$$K(x, x_i) = \exp\{-\gamma \mid x - x_i \mid^2\} \tag{3}$$

The feature of radial basis function (RBF) classifier is that the center of each basis function corresponds to a support vector, and the output weights are automatically determined by the algorithm. The inner product function is similar to the neural center characteristics of human brain, and different S-parameter values have different classification surfaces.

(4) S-shaped kernel function

$$K(x, x_i) = \tanh[v(x \cdot x_i) + c] \tag{4}$$

The kernel function consists of a multilayer perceptron network with a hidden layer. The weights of the network and the number of nodes in the hidden layer are automatically determined by the algorithm, and there is no problem of local minima bothering the neural network.

Christopher J.C. Burges has experimented and compared linear kernel function, polynomial kernel function and radial basis function, and different kernel functions have their own advantages and disadvantages for different databases. There are also studies based on UCI benchmark database data analysis, which show that the performance of radial basis function is slightly better.

In this paper, the radial basis function (RBF) is determined to be the best kernel function of the model.

2.2 Important Parameters in Kernel Function γ, C, P

γ: Set up kernel function γ Value of, with γ The results show that the test set has a bad effect on classification and good training classification effect. It is easy to generalize the error of fit, generally 0.01.

C: Penalty factor C represents how much you value outliers, the greater C values, the less you want to lose them. When the value of C is large, the punishment for error classification increases, while the punishment for error classification decreases when the value of C is high. When C is larger and approaches infinity, it means that the classification error is not allowed and it is easy to over fit; when C tends to 0, we are no longer concerned about whether the classification is correct and is easy to be undefeated. In this study, the effect of large-scale infectious diseases can not be ignored because of the small sample, so the value of C is larger, when the value is 1.0e+7, the fitting degree of the model prediction value and original value is the highest.

P represents the parameter B in the loss function of SVM. The loss function in SVM is defined as the sum of hinge loss function and a regularization term.

3 GRAIN YIELD PREDICTION MODEL

3.1 Training Sample

In this paper, the data of grain output of 21 provinces including Beijing, Tianjin, Hebei, Shanxi and Inner Mongolia during the SARS epidemic in 2003, Guangdong Province during the cholera epidemic in 1961 and Xinjiang Province during the hepatitis E epidemic in 1986 were selected as training samples. Among them, the number of SARS virus infected in Beijing was 2434 in 2003, the number of cholera infected in Guangdong Province was 4319 in 1961, and the number of hepatitis E infected in Xinjiang was 119280 in 1986. The absolute number of infectious diseases was large, which can enhance the wide applicability of the model.

3.2 Forecast Sample

In this paper, the data of grain output in Guangdong, Shaanxi and Gansu during the SARS epidemic in 2003 and the data of grain output in Guangdong during the dengue epidemic in 1978 were selected as the prediction samples.

4 DATA PREPROCESSING AND METHOD ANALYSIS

4.1 Data Preprocessing

According to the existing research, for the machine learning method to study the influencing factors of grain yield, the main factors are the sowing area of grain crops, the amount of chemical fertilizer, the effective irrigation area of grain crops and so on. Considering that the main purpose of this paper is to study the prediction of grain yield under the epidemic situation, the factors of epidemic degree, the number of local farmers (number of rural employees) and the change trend are added. In recent years, the development trend of grain yield can cover other secondary factors such as fertilizer application.

To sum up, the main factors are finally classified into the following four categories: 1) epidemic impact: because of the differences in the population of each province, it can not accurately explain the severity of the epidemic simply by the two dimensions of infected population and dead population, so the proportion of infected population and dead population in the total population at the end of the year is selected as the index of the severity of the epidemic; 2) In recent years, the grain sown area has changed according to the policy, and the impact on grain yield is also very intuitive and obvious; 3) Agricultural population: the "rural employed population" is used to replace the "rural employed population". The rural employed population in recent five years can better represent the change trend of agricultural population; 4) The local grain output of the previous year can be used as one of the most direct basis for the prediction of the grain output of that year. In order to better reflect the trend of grain output change, the local grain output data in recent four years are selected.

Table 1: Training sample data.

SN	Province	Year	Infectious disease	Proportion of infected persons	Proportion of deaths	Grain yield in the n-4 year (10000 tons)	Grain yield of the year (10000 tons)
1	Beijing	2003	SARS	1.67E-04	1.01E-05	201.0	58.0
2	Tianjin	2003	SARS	1.74E-05	1.19E-06	174.9	119.3
3	Hebei	2003	SARS	3.10E-06	1.48E-07	2746.3	2387.8
4	Shanxi	2003	SARS	1.34E-05	6.04E-07	821.7	958.9
5	Neimenggu	2003	SARS	1.21E-05	1.05E-06	1428.5	1360.7
…	……	…	…	…	…	…	…
21	Ningxia	2003	SARS	1.03E-06	1.72E-07	293.3	270.2
22	Guangdong	1961	cholera	1.07E-04	1.06E-05	1230.0	990.5
23	Xinjiang	1986	Hepatitis E	1.68E-02	9.93E-05	407.5	547.7

4.2 Method Analysis

After experiments, the kernel function is analyzed γ Make adjustments when γ= 01, the average relative error of training sample is about 2% γ> The average relative error of the training samples is still about 2%, but the average relative error of the prediction samples increases obviously γ< 01 and continued to decrease, the average relative error of training samples gradually increased, indicating that the fitting degree decreased.

When C < 1.0E+5, the average relative error of training sample is about 5%, and the smaller C is, the lower fitting degree is. When C > 1.0E+5 and gradually increases, the average relative error of the training sample is gradually reduced. When C is 1.0E+7, the average relative error of the training sample is 1.96%, and the average relative error of prediction sample is 3.27%, The average relative error of training samples decreases slightly, but the average relative error of prediction samples increases greatly, which indicates that the prediction effect decreases.

P represents to adjust the parameter B in the loss function. When p > 10 and gradually increases, the average relative error of training samples gradually increases and the fitting degree decreases. When p < 10 and gradually decreases, the average relative error of training samples gradually decreases, but the average relative error of prediction samples increases greatly, which indicates that the prediction effect decreases.

5 PREDICTION RESULTS OF GRAIN YIELD OF PREDICTION SAMPLES IN THE CURRENT YEAR

It turns out that when γ= 0.01, C = 1.0E+7, P = 10, the average relative error of training samples is 1.96%, the coefficient of determination is 99.0%, and the average relative error of prediction samples is 3.27%, which can meet the demand of regional grain yield prediction in the year of infectious diseases, while the average relative error of prediction samples is 3.27% ε- SVR has strong generalization ability due to its modeling of a small number of cases and parameter optimization, so it is based on SVR ε- SVR grain yield model can provide accurate reference data for regional short-term grain yield prediction.

Table 2: Training sample results.

SN	Province	Infectious disease	Actual grain yield of that year(10000 tons)	Fitted Grain yield that year(10000 tons)
1	Beijing	SARS	58.0	56.1
2	Tianjin	SARS	119.3	121.0
3	Hebei	SARS	2387.8	2384.7
4	Shanxi	SARS	958.9	950.9
5	Neimenggu	SARS	1360.7	1345.9
…	…	…	…	…
21	Ningxia	SARS	270.2	267.1
22	Guangdong	cholera	990.5	1003.2
23	Xinjiang	Hepatitis E	547.7	544.0

Table 3: Forecast results.

SN	Province	Infectious disease	Actual grain yield of that year(10000 tons)	Fitted Grain yield that year(10000 tons)
1	Gansu	SARS	789.3	767.4
2	Shanxi	SARS	968.4	964.1
3	Guangdong	SARS	1430.4	1319.0
4	Guangdong	break-bone fever	1632.0	1665.9

In order to study whether the large-scale infectious diseases have a significant impact on grain production in that year, the original samples were modeled again by removing the two parameters of the epidemic degree (the proportion of infected population and the proportion of deaths). The results show that the average relative error of training samples is 1.97%. The average relative error of prediction samples is 3.31%, and the coefficient of determination also reaches 0.99. The model also has a good reference value for short-term grain yield prediction.

Table 4: Forecast results Including and Excluding epidemic data.

SN	Province	Infectious disease	Including epidemic data		Excluding epidemic data	
			Actual grain yield of that year(10000 tons)	Fitted Grain yield that year(10000 tons)	Actual grain yield of that year(10000 tons)	Fitted Grain yield that year(10000 tons)
1	Gansu	SARS	789.3	767.4	789.3	765.9
2	Shanxi	SARS	968.4	964.1	968.4	964.0
3	Guangdong	SARS	1430.4	1319.0	1430.4	1317.8
4	Guangdong	break-bone fever	1632.0	1665.9	1632.0	1664.1

6 CONCLUSION

6.1 Prediction Model Reliability

Due to the limited samples of large-scale infectious diseases in China, it belongs to small sample data analysis, ε- The SVR method has excellent ability of fitting and generalization for a small number of samples.

Stay γ= 0.01, C = 1.0E+7, P = 10, the model can better fit the sample data, and the prediction effect is also better ε- The grain yield model of SVR is accurate and reliable.

6.2 The Impact of Infectious Diseases

After removing the two parameters of the proportion of the infected population and the proportion of dead population representing the epidemic degree, the new model can still better fit the modeling sample data and forecast the target sample. Although the average relative error is slightly larger than that of the model with epidemic parameters, it is based on the existing domestic large-scale infectious disease epidemic data modeling. The epidemic situation had limited influence on the grain yield in the area of that year.

This method can provide a theoretical reference for the national macro-control of food production, and it is a new research direction.

REFERENCES

Burges C J C.A tutorial on support vector machines for pattern rec-ognition [J]. Data Mining and Knowledge Discovery, 1998 (2) :121-167.

Chen Xiaolu, Wang Yanfang, Zhang Hongmei, Liu Fenggui, Shen Yanjun Extraction method of irrigated arable land in the Chahannur Basin based on the ESTARFM NDVI [J] Chinese Journal of ecological agriculture (Chinese and English), 2021,29 (06): 1105-1116 DOI: 10.13930/j.cnki. cjea. 200880.

CHENG Peng, WANG Xi-li. Influence of SVR Parameter on Non-linear Function Approximation[J]. Computer Engineering, 2011,37(03):189+191+194.

Gao Xinyi, Han Fei Grain yield prediction of support Vector Machine Based on hybrid intelligent algorithm [J] Journal of Jiangsu University (NATURAL SCIENCE EDITION) ,2020,41(03):301-306.

Guo Lin, Bai Dan, Wang Xinduan, et al. Establishment and validation of flow rate prediction model for drip irrigation emitter based on support vector machine [J]. Transactions of the Chinese Society of Agricultural Engineering, 2018,34(02):74-82.

Hu Chenglei, Liu Yonghua, Gao Juling Research on prediction method of grain yield based on IPSO-BP mode [J] China Journal of agricultural chemistry, 2021,42 (03): 136-141.

Li Donglin, Zuo Qiting, Zhang Wei, Ma Junxia Agricultural water resources allocation model in Tarim River Basin based on Nerlove approach [J] water resources protection,2021,37(02):75-80.

Li Tong, Dong Weihong, Zhang Qichen, Wen chuanlei Analysis and prediction of grain water footprint in Heilongjiang province based on time series model [J] Journal of drainage and irrigation mechanical engineering, 2020,38 (11): 1152-1159.

Li Ying, Chen huailiang Review of Machine Learning Approaches for Modern Agrometeorology [J] Journal of Applied Meteorology, 2020,31 (03): 257-266.

Liang Ji, Zheng Zhenwei, Xia shiting, Zhang Xiaotong, Tang Yuanyuan Crop recognition and evaluationusing red edge features of GF-6 satellite [J] Journal of remote sensing, 2020,24 (10): 1168-1179.

LIN Sheng-liang, LIU Zhi. Parameter selection in SVM with RBF kernel function[J]. Journal of Zhejiang University of Technology, 2007(02):163-167.

Vapnik V N.The nature of statistical learning theory[M].New York:Springer, 1999.

Wang Qian, Huang Kai Simulation of Agricultural Water Footprint and Analysis of Influencing Factors in Beijing Based on System Dynamics [J] systems engineering,2021,39(03):13-24.

Wu Danhua, Zhou Limei Grain yield prediction based on BP neural network [J] Agricultural Engineering Technology,2020,40(27):51-53. DOI: 10.16815/j.cnki.11-5436/s.2020.27.008.

XIAN Guang- ming, ZENG Bi- qing. ε- SVR algorithm and its application[J]. Computer Engineer ing and Applications, 2008(17):40-42.

Xiong H L, Zhou X C, Wang X Q, et al. Mapping the spatial distribution of tea plantations with 10 m resolution in Fujian province using Google Earth Engine [J] Journal of Earth Information Science, 2021,23 (07): 1325-1337.