RESEARCH ON PM2.5 CONCENTRATION COMBINATION FORECASTING MODEL BASED ON COR-SVM

Xiaoyu Feng, Peng Tian, Yujie Shi, Mei Zhang*

College of Science, Nanjing Agricultural University, Nanjing, China -
956043527@qq.com, 1241642647@qq.com, 1506957832@qq.com,zhangmei2006@njau.edu.cn

Commission III, WG III/8

KEY WORDS: Concentration of PM2.5, Correlation Coefficient, Multivariate Linear Fitting, Support Vector Machine, Combination Forecasting

ABSTRACT:

PM2.5 is a pollutant that can enter the lungs, threatening human health and affecting people's living and traveling. In this paper, we use multivariate linear regression, support vector machine and their combined prediction method to predict the concentration of PM2.5. It is significant for the convenience of healthy life. This paper is based on a series of meteorological data such as O_3 concentration, CO concentration, SO_2 concentration, PM2.5 concentration and PM10 concentration from 2014 to 2018 in Beijing. By calculating the correlation coefficient between the concentration of PM2.5 and the concentration of the other four components, the multivariate linear regression equation was fitted by using the correlation coefficient with high correlation as the factor of multiple linear regression. Then we use support vector machine regression prediction method to predict the concentration of PM2.5. The combined prediction method is obtained by weighing the two prediction results. It is found that the prediction method of support vector machine is better in dealing with large-scale and small sample data prediction, and the multi-linear fitting method is better in processing short-term prediction. The combined prediction results based on correlation coefficients combine the advantages of the two prediction methods, and the prediction results are more reasonable.

* Corresponding author

1. BACKGROUND

PM2.5 is an inhalable pollutant that threatens human health and affects human life, which is increasingly important for the reasonable prediction of PM2.5 concentration. The predictions for PM2.5 include gray prediction, regression prediction, BP neural network prediction, support vector machine prediction (Pacelli et al., 2011; Li, 2008). Compared with Regression Prediction, Gray Prediction requires a small amount of sample data and flexible prediction. Regression Prediction has a large demand for samples. support vector machine(SVM) regression prediction has obvious advantages in linear and nonlinear small sample prediction through machine learning. Different prediction methods have different prediction tendency, and the accuracy of predictions is different in different environments. By weighing combination prediction of different methods(Wang and Liu, 2018), there are often more reasonable predictions. The use of combined forecasting models to reasonably predict the concentration of PM2.5 provides convenience for human habitation, travel, etc., and it has a strong practical significance.

2. DATA PROCESSING

The research data in this paper contains Beijing meteorological monitoring data from 2014 to 2018. The data includes a series of meteorological data such as PM2.5 concentration, PM10 concentration, SO_2 concentration, CO concentration, and O_3 concentration. After deleting the outliers and interpolating the missing data, the data is normalized to eliminate the influence of unnecessary noise such as dimension on the research (Quan et al.,2019).

For each given PM2.5 concentration data column $\{y_n\}$, where

4.1 Support Vector Machine (SVM)

The support vector machine is a statistical learning method that seeks the optimal hyperplane and rationally classifies the samples according to the supervised learning method **(**Fan and Wang,2019**)**. It

$$y^{(1)}_i = \frac{y^{(0)}_i}{\max\{y^{(0)}_j\}(j=1,2...)}(i=1,2.....) \qquad (1)$$

Where $y^{(0)}_i$ is the i-th value of the initial PM2.5 concentration sequence $y^{(1)}_i$ is the i-th value of the normalized PM2.5 concentration sequence. For the SO_2 concentration, the CO concentration, and the O_3 concentration series $\{x1^0_n\}$ $\{x2^0_n\}$ $\{x3^0_n\}$, the same processing is performed to add a new sequence $\{x1^1_n\}$ $\{x2^1_n\}$ $\{x3^1_n\}$. Where $\{x1^0_n\}$ represents the SO_2 concentration series, $\{x2^0_n\}$ represents the CO concentration series, and $\{x3^0_n\}$ represents the O_3 concentration series, which respectively correspond to the PM2.5 concentration value series.

3. THE CORRELATION COEFFICIENT METHOD AND MULTIPLE LINEAR REGRESSION

3.1 Correlation Coefficient Method

For the two samples x, y, through the formula **(**Hu and Yao, 2019**):**

$$\rho(x,y) = \frac{\mathrm{cov}(x,y)}{\sqrt{Var(x)Var(y)}} \qquad (2)$$

Where is the correlation coefficient, cov(x, y) is the covariance of sample x and sample y, and Var(x) is the variance of sample x. The correlation coefficient of the two-sample factor is obtained, and the Pearson correlation coefficient test is used to determine the factor with high correlation with PM2.5. The correlation coefficient is introduced into the prediction model, used as a factor of multiple linear regression, and the prediction function based on the microscopic factor is fitted to predict.

4. SUPPORT VECTOR MACHINE AND ITS REGRESSION PREDICTIONS

is a two-class classification model, which is the maximum interval classifier of feature space. It can convert linear and nonlinear classification optimization problems into convex quadratic linear programming problems by citing Lagrange multipliers to simplify the problem.

4.2 Support Vector Machine Application Principle

4.2.1 Linear Classification Divider

(1) Basic mathematical logic (Lin et al.,2018)

① The first problem solved by the linear classification divider (Cheng, 2018) is to find the optimal linear hyperplane. The mathematical representation is as follows:

$$\max \frac{1}{\|W\|} \qquad s.t. y_i(W^T X_i + b) \geq 1, (i = 1,2...n) \tag{4}$$

Where $\|W\|$ is the two norm of the coefficient matrix and X_i is the sample vector element. When $\|W\|$ takes the minimum value, it means that the optimal linear hyperplane is found.

③ Following the introduction of the Lagrange multiplier, the problem is transformed into a linear programming problem, which is expressed as follows:

$$L(W,b,a) = \frac{1}{2}\|W\|^2 - \sum a_i(y_i(W^T X_i + b) - 1), \quad a_i \geq 0 \tag{5}$$

$$\theta(W) = \max_{a \geq 0} \quad L(W,b,a) \tag{6}$$

4.2.2 Nonlinear Classification Divider

By citing the kernel function (Jiang and Wang ,2017), the nonlinear classification problem (Chen,2010) is implicitly mapped to the high-dimensional vector space, therefore, the nonlinear programming problem is transformed into a linear solution problem.

(1) For example:

$$\begin{cases} \max \ \sum_{i=1}^{n} a_i - \frac{1}{2}\sum_{i,j=1}^{n} a_i a_j y_i y_j \langle \phi(x_i), \phi(x_j) \rangle \\ s.t., ai \geq 0, i = 1,2,...,n \\ \sum_{i=1}^{n} a_i y_i \geq 0 \end{cases} \tag{7}$$

Where $\phi(x)$ is a mapping, and a low-dimensional space is mapped to a high-dimensional space. a_i is a Lagrange multiplier. Using the kernel function $K(x_i, x_j)$, then converting to the following expression:

$$\begin{cases} \max \ \sum_{i=1}^{n} a_i - \frac{1}{2}\sum_{i,j=1}^{n} a_i a_j y_i y_j K(x_i, x_j) \\ s.t., a_i \geq 0, i = 1,2,...,n \\ \sum_{i=1}^{n} a_i y_i \geq 0 \end{cases} \tag{8}$$

(2) The kernel function

A function that implicitly maps low-dimensional vector elements to high-dimensional vector spaces, can avoid the complex calculation of SVM in high dimensional space. Transform nonlinear classification problems into high-dimensional linear programming problems.

The general kernel function (Tan, 2019) has a polynomial form and a Gaussian form of the kernel function. You can also plan the kernel function according to your own needs. The more applicable is the Gaussian form, for example:

$$K(x_i, x_j) = \exp(-\frac{\|X_1 - X_2\|^2}{2\sigma^2}) \tag{9}$$

By adjusting different parameters σ, the kernel function meets the requirements and the planning classification problem is solved reasonably.

5. COMBINATION FORECAST

Different forecasting methods have different emphasis, and the accuracy of prediction is different. It is possible that the combined prediction of multiple prediction methods will be better.

Two different forecasting methods adopted in this pap er are weighed and relaxed to obtain new forecasting results. Mathematically expressed as:

$$\begin{cases} F_{new} = \omega_1 F_{svm} + \omega_2 F_{reg} \\ \omega_1 + \omega_2 = 1 \end{cases} \quad (10)$$

Where F_{new} is the new predicted value, F_{svm} and F_{reg} are the values of the support vector machine and the regression prediction, and ω_1 and ω_2 are the relaxation coefficients.

Equivalent to turning the problem into:

$$\min_{\omega_1+\omega_2=1} \sum \left\| F_{new} - F_{true} \right\| \quad (11)$$

Different sample spaces have different ω_1 and ω_2. The more data of the training samples, the more accurate is the determination of ω_1 and ω_2. The size of ω_1 and ω_2 can also indicate the applicability of the two methods.

6. RESULT ANALYSIS

6.1 Calculation Results of Correlation Coefficient

Beijing was selected as the research object, and the real-time meteorological monitoring data of Beijing in the past five years was collected. After pre-processing the data, the correlation coefficients of PM2.5 and NO_2, O_3, SO_2, and PM10 were calculated as follows:

Factor	Correlation coefficient
NO_2	0.76
O_3	-0.51
SO_2	0.50
PM10	0.98

Table 1. Correlation coefficient between PM2.5 and each factor

6.2 Linear regression Fitting Equation Prediction

The three factors with higher correlation with PM2.5 determined above are the concentrations of NO_2, O_3 and PM10, then multiple linear fitting equations of PM2.5 are obtained. The data from January 1st to 15th, 2015 were selected for multivariate linear equation fitting to predict the data of No. 16. The prediction equation is:

$$C_{PM2.5} = 116.7 - 1.14C_{NO_2} - 2.84C_{O_3} + 0.96C_{PM10} \quad (12)$$

$C_{PM2.5}$ is the concentration of PM2.5, C_{NO_2} is the concentration of NO_2, C_{O_3} is the concentration of O_3, and C_{PM10} is the concentration of PM10.

The residual analysis chart is shown in Figure 1. The fitted equation passes the residual test.

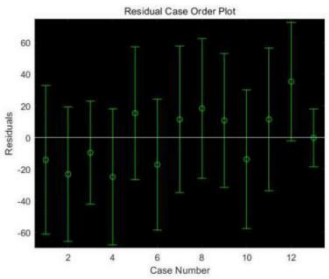

Figure 1.Residual analysis chart

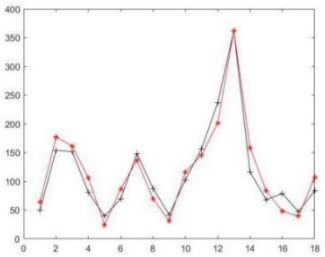

Figure 2. Forecast map

Due to the lack of data, the data of No. 1 was removed, and the regression equation fitted by the data from No. 2 to No. 15 found that there was a mismatch in the residual test on the No. 9 data, thereby eliminating the data of No. 9 to ten. The three-day data fits the regression equation, the equation above passed the

residual test. From this equation, the PM2.5 concentration for the five days from the 16th to the 20th was calculated, and the regression results were: 158.88, 83.32, 48.04, 39.45, 107.02, and the 18-day prediction chart are shown in Figure 2. The relative errors are: 0.36, 0.22, 0.39, 0.16, 0.28.

PM2.5 (ug/m³)	2014	2015	2016	2017	2018	2018(prediction)	Relative
Jan.	108.48	106.87	75.47	109.06	34.31	58.43	0.70
Feb.	163.58	108.09	48.95	68.68	49.42	47.50	0.04
Mar.	98.35	88.26	94.90	62.68	90.07	81.94	0.09
Apr.	86.60	73.58	66.99	52.30	63.08	63.85	0.01
May	62.28	52.74	55.69	60.35	56.31	63.13	0.12
Jun.	59.46	59.81	64.19	42.47	47.46	49.08	0.03
Jul.	86.33	62.23	73.59	53.02	47.78	59.34	0.24
Aug.	64.42	46.13	50.75	34.35	29.26	33.93	0.01
Sep.	69.76	51.67	58.72	55.09	26.48	56.23	0.54
Oct.	110.80	77.33	83.74	49.68	38.52	50.63	0.24
Nov.	104.30	126.80	106.83	41.35	64.50	-169.03	5.65

Table 2. Forecast of the monthly average of PM2.5 by month

6.3 Support Vector Machine Fitting Results

6.3.1 Forecast by month average (baseline forecast)

Based on the meteorological data of Beijing from 2014 to 2017, as a training sample, the average estimate is made according to the month. The relative error test was performed with the data of 2018. The concentration factors of CO, NO_2 and O_3 were used as training samples. The estimated results are shown in Table 2. The results of this group can be used as reference data, and some factors can be changed later for comparative study.

Using the support vector machine regression prediction method, we can conclude the data prediction accuracy in the second, third, fourth, sixth and eighth months is good, and the relative error is less than 10% from the table 2. However, in other months, especially the average forecast results of January and November are very large, but the results of relative multiple linear regression have been significantly improved. For the frequent cold air activities in Beijing in January, the number of winds increased, making air pollutants such as PM2.5 not easy to accumulate, which may result in the average concentration of PM2.5 in January 2018 compared to previous years. There has been a sharp drop. However, from the method, because the predicted time sample size is too small, the factors associated with the introduction of the support vector machine are not high, which may cause the prediction to be inaccurate.

Therefore, the set of data is used as the reference data for prediction. The control variable method is used to find the key factors affecting the prediction.

6.3.2 Horizontal and Vertical Prediction

Previous predictions used a fixed month of the year as a training sample to predict a fixed month of the forecast year. For example, the average concentration of PM2.5 in January 2018 is predicted, and the average concentration of PM2.5 in January of 2014~2017 is

selected for prediction. Actually, due to climate change, the actual corresponding effect of the corresponding month is not good. The overall variation of the four seasons is not much different. However, when the local refinement reaches the month, there will be a big difference, which also brings difficulties for the prediction. Therefore, instead of making a horizontal month forecast, it is changed to a vertical forecast. The 12 consecutive months of 2017 are used to predict the average concentration of PM2.5 in January 2018, and the analogy is to predict the first six months of 2018.

The predicted average PM2.5 concentration for January and February of 2018 is: 49.8574, 51.6255, and the relative error of is 0.4690 and 0.046. It can be found that the prediction of the average PM2.5 concentration in the longitudinal direction is more stable. One is the next month, the data is more correlated, and the overall trend is more stable. From the data, it can be seen that there is a certain period in the data processing for the month, in January 2017. The starting point ends in March of 2018. The cycle is not a fixed 12 months, and the length of the cycle varies according to the interannual variation.

Jan	Feb	Mar	Apr	May	June
109.06	68.68	62.68	52.30	60.35	42.47
July	Aug	Sep	Oct	Nov	Dec
53.02	34.35	55.09	49.68	41.35	41.35

Table 3. Beijing PM2.5 average concentration (ug/m^3) from January to December 2017

6.3.3 Sample Size and Fixed Sample Size of Training

Although the selected data is from 2014 to 2018, the amount of data is large, but the effective data volume is still small. Only the average PM2.5 concentration data of each month is used as a sample. The influence of the sample data volume on the results is sometimes huge. Properly increasing the density of the sample for this purpose may lead to better predictions. For example, we can select the data from January 1st to 16th of 2014 as the training sample, then predict the PM2.5 concentration from 17th to 20th, and calculate the error with the true concentration. The calculation result is shown in Table 6.

When making predictions, use the data from No.1 to No.16 of 2014 to directly predict the PM2.5 concentration from No.16 to No.20, but in fact there will be an accumulation of errors. For example, the prediction of No.17 will add up to the prediction of the error of No.16,and so on. You can take a head-to-tail prediction which fixing the size of predicting sample. For example, in the prediction of the 17th, the real data of No. 2 to No. 16 is used for prediction, and so on.

Concentration （ug/m^3）	Actual value	Predictive value	Relative Error
17th	178.60	142.59	0.20
18th	129.87	139.83	0.19
19th	117.71	153.19	0.18
20th	7.43	139.86	17.82

Table 4.Forecasts for the 17th to the 20th of January 2014 (16 days forecast 4 days)

Concentration ug/m^3	Actual value	Predictive value	Relative Error
11th	257.80	270.6068	0.05
12th	33.16	597.8492	17.03
13th	99.43	172.4170	0.7341
14th	92.17	127.0624	0.3786
15th	117.59	138.6316	0.1789

Table 5. Forecasts for the 11th to the 15th of January 2014 (10 days forecast one day)

Comparing Table 4 with Table 5, combined with Table 2, can be seen. The SVM regression prediction method for PM2.5 concentration is better than the prediction of a more detailed range in a wider range of predictions. It

is also in line with our actual experience, from large-scale to inter-annual PM2.5 expected concentration predictions, actually avoiding internal contingency in a small range. For example, the difference in PM2.5 concentration between January 11 and January 12, 2014 is large, and it is a small accidental change. However, in the large range, the monthly expectation of PM2.5's annual expectation tends to be flat. It shows that the change trend of PM2.5 in a small period of time is not stable, but it can be seen from Table 2 that PM2.5 changes over a large period of time and tends to be flat.

6.3.4 Combined Forecasting

Using the data from the first six months of 2018, two sets of data are obtained from the first two different prediction methods. According to formula (11), plan a better ω1, ω2, and using a new method to predict the average PM2.5 concentration for the three months of July, August and September. Then the relative error is calculated.

Month 2018	True value	Multiple linear fit	SVM prediction results	Combined forecasting result
Jan.	34.31	44.77	58.43	45.48
Feb.	49.42	45.76	47.50	45.85
Mar.	90.07	96.44	81.94	95.68
Apr.	63.08	59.20	63.85	59.44
May	56.31	54.97	63.13	55.40
Jun.	47.46	41.62	49.08	42.01
Jul.	47.78	46.17	59.34	46.86
Aug.	29.26	33.47	33.93	33.50
Sept.	26.48	43.05	56.23	43.74

Table 6.Combined prediction results (PM2.5 concentration (ug/m^3))

The data of the corresponding month in the previous four years is the training sample, and the prediction formula of the corresponding month is linearly fitted, and the average concentration of PM2.5 corresponding to the month of 2018 is predicted accordingly. From the prediction data of the first six months, it is determined that the coefficient of ω_2, that is, the linear regression prediction value is 0.95, thereby combining the predicted PM2.5 concentrations of July, August, and September. As can be seen from the results, the combined prediction improves the accuracy of the prediction and makes the prediction more stable.

8. THE RESULT ANALYSIS

The prediction of PM2.5 concentration is affected by a variety of environmental factors. Wind, cold storm and other factors will affect the concentration of PM2.5, and unpredictable natural environmental factors increase the difficulty of prediction.

In this paper, the correlation coefficient method and support vector machine regression prediction method are combined. After the correlation coefficient method is used to extract other microscopic factors related to PM2.5 concentration, these related microscopic factors are added as reference factors to the support vector machine. In the regression model, the predicted results are obtained.

After using the correlation coefficient method to obtain the microscopic factors related to PM2.5 concentration, the multiple linear equations of PM2.5 for NO_2, O_3 and PM10 were fitted by multiple linear regression method. The multivariate linear fit has a large demand for the sample, and the short-term fitting equation is not completely applicable for a long
time, and the error is large. For example, when using formula (12) to process the prediction in table 6, it is impossible for the predicted value to be negative. The reason is that the sample size selected by formula (12) is too small, which is suitable for the short-term prediction corresponding to the sample. But for long-term monthly forecasts, there is no fixed linear fit formula. Using the data from January of the previous four years for total linear regression, the results were better when predicting January 2018 data.

Comparing linear regression prediction with support vector machine regression prediction, it can be seen that the support vector machine regression model has a strong advantage in predicting the monthly expected concentration of PM2.5, and it can accurately predict the expected concentration. However, the support vector machine loses its advantage in the face of a smaller range of daily expected concentration predictions, but the multiple linear regression equation prediction is more stable. In this paper, the combined forecasting model combines the prediction advantages and disadvantages of the two methods to a certain extent, and improves the stability of the forecast to some extent.

In general, the support vector machine prediction model based on the correlation coefficient method has a certain effect in predicting the concentration of PM2.5, and can predict the expected concentration of PM2.5 on the macroscopically. However, there are still many challenges in the prediction of PM2.5 concentration. The current prediction model can only achieve a smooth prediction within a certain range, and the accuracy is still relatively low. The relationship between PM2.5 and other factors can be further studied, and the nonlinearity fitting can be used to improve the accuracy of prediction. The selection of a time series for prediction may also be another breakthrough for the problem.

REFERENCES

Jinghua Chen, 2010. Lagrangian multiplier method and its extension. *Journal of Hubei Normal University (Natural Science Edition)*, 30(04), 108-111.

Fengwei Cheng, 2018. A linear classification algorithm based on SVM. *Journal of Taiyuan University(Natural Science Edition)*, 36(04), 52-54.

Wenting Fan, Xiao Wang, 2019. PM2.5 prediction based on improved firefly optimization support vector machine. *Computer Systems*, 28(01), 134-139.

Liang Hu, Zeqing Yao, 2019. Total entropy weight method based on correlation coefficient and its application. *Journal of PLA University of Science and Technology (Natural Science Edition)*, 1-4.

Yingying Jiang, Huiwen Wang, 2017. Multi-attribute fuzzy decision classification algorithm based on nonlinear utility function. *Journal of Huaibei Normal University (Natural Science Edition)*, 38(02), 28-35.

Xiaodong Li, 2008. Prediction model of China's grain yield based on grey relational support vector machine . *Journal of Hebei University of Technology (Natural Science Edition)*, (04), 76-80+103.

Xiangliang Lin, Rui Yuan, Yuqiu Sun, Chao Wang, Changsheng Chen, 2018. Basic theory and research progress of support vector machine. *Journal of Yangtze University* , 15(17), 48-53+6.

Pacelli, Vincenzo,Bevilacqua, Vitoantonio,Azzollini, Michele. 2011. An artificial neural network model to forecast exchange rates. *Journal of Intelligent Learning Systems and Applications*, 3(2), 57-69.

Yanming Quan, Zhenbo Qi, Weishi Li, Rui Zhang, 2019. Multi-camera calibration of one-dimensional calibration using normalization algorithm. *Optics Journa*l, 1-14.

Shuaixin Tan, 2019. SVM intrusion detection method based on piecewise kernel function. *Software Guide*, 1-5.

Wei Wang, Hong Liu, 2018. Application of PM2.5 concentration prediction model. *Journal of Liaoning University of Science and Technology*, 41(01), 75-80.