Prediction of SO_2 Concentration Based on Fuzzy Time Series and Support Vector Machine

Linliang Zhang[1, a *], ZhaoXia Li[1, b], Yuejun Ma[1, b] and Pengcheng Yue[1, b]

[1]Shanxi Transportation Research Institute, XueFu Street 79th, Taiyuan City, Shanxi Province, China

[a]zhanglinliang@163.com [b]lzx87117@126.com

Keywords: SO_2 concentrations; Support vector machine; Fuzzy time series; Prediction model

Abstract. The existing prediction methods for SO_2 concentration mainly have the disadvantages such as no unified sources and influences, sensitive to small sample, easy to fall into local optimum etc. In order to solve these problems, a method for SO_2 concentration prediction is proposed, based on fuzzy time series and support vector machine (SVM). The method takes the seasonal variation of SO_2 concentration as the basis, takes the four seasons as the time series, and takes the 24 hours as the graining window width, then extract the characteristic values of the original sample data through the Gaussian kernel function for SVM training, and optimize model parameters by k-fold cross validation method combined with the grid division. Finally, a SO_2 concentrations prediction model is established by using one-hour average SO_2 concentrations as sample data, and the calculation process is realized by using LIBSVM tool. The results show that the prediction method of SO_2 concentration based on fuzzy time series and SVM is not restricted by the machine rational theory, can solve small-sample learning problems, and has good nonlinear fitting effect.

Introduction

Human bodies will be potentially affected when the concentration of SO_2 in the atmosphere is above 0.5 parts per million(ppm); Most people start to feel pungent when the SO_2 concentration is between 1 to 3 ppm; People will appear ulcer and pulmonary edema until stifle even die when the SO_2 concentration is between 400 to 500 ppm. Furthermore, SO_2 has synergistic effect with fume in the air. When the concentration of SO_2 in the atmosphere is 0.21 ppm and the concentration of fume is greater than 0.3 mg/L, it can make the increased incidence of respiratory diseases and the rapid-deterioration of chronic diseases. Great Smog of 1952, Meuse River Valley smog of 1930 and Donora Smog of 1948 are all the results of the synergistic effect. Therefore, SO_2 in the atmosphere threaten our human health directly. It will help us to seek effective measures to control and improve the atmospheric environment if the concentration variation of SO_2 can be predicted effectively. Many international and domestic academics have made many attempts in this field and proposed some prediction models. Multi-factor regression analysis method is the most common air pollutant concentration prediction method, which assumes that the pollutant concentration is associated with some specific factors and makes the influencing factors as independent variables, makes pollutants concentration as the dependent variable, carries out correlation analysis on the pollutants concentration and influence factors. But, for the prediction of SO_2 concentrations, this method has many uncertainties. First of all, Studies have found that the influence factors of SO_2 concentration include emissions of pollutants, layout and types of the pollution sources, weather conditions, fuel structure, etc., but there are still many factors which are not found or difficult to determine; Secondly, even if all the factors have been found, still there is a causal, fuzzy and coupling relationship between the influencing factors, causing the multiple factors regression analysis method difficult to accurately analysis the correlation between factors and factors and the correlation between factors and SO_2 concentration. For this problem, we can make full use of machine learning algorithm to solve the problem of the complex nonlinear model.

Artificial neural network, which developed rapidly in recent years, is a typical modeling method based on machine learning algorithms. The studies using neural network model to predict the concentration of SO_2 have achieved good effect, but it is difficult to avoid problems such as network training over-fitting or easy falling into local optimum [1, 2, 3]. Support Vector Machine (SVM) is a machine learning technique based on statistical learning theory, showing unique advantages in solving the small sample, nonlinear problems. It follows the structural risk minimization principle, which can effectively prevent recurrent neural networks over-fitting and the local minimum problems. Therefore, SVM has become a research hotspot in the field of pollution prediction [4, 5]. In this paper, an improved method combining SVM and the time series is proposed. This method extract the characteristic values of the sample data using fuzzy time series and use them as the input of SVM so as to improve the accuracy of the model, then solve the nonlinear fitting problem in high dimensional space using Gaussian kernel function. This method can effectively solve the instability problem of using multi-factor regression model to predict, and affords a new prediction method for SO_2 and other atmospheric pollutants.

Support Vector Machine

Support vector machine (SVM) can establish a hyper plane as sample classification decision surface, use the kernel function to do fitting analysis or modeling for samples, and use the convex quadratic optimization to get the global optimal solution.

In recent years, in the field of environmental pollution prediction and air quality prediction, many researches on regression analysis or fitting prediction using SVM have begun.

This paper is mainly to predict the concentration change of atmospheric pollutant SO_2, complete the concentration sample data mapping from low dimensional space to high dimensional space based on the SVM using Gauss kernel function, achieve the transition from linear to nonlinear relationship, and solve the nonlinear intrinsic dependence problems of the SO_2 concentration prediction, that is：

$$\Phi\text{：} X \rightarrow H \qquad x \rightarrow \phi(x) \tag{1}$$

Where H represents the high dimensional space, and represents the mapping function. If

$$K(x, x') = (\phi(x),\ \phi(x')) \tag{2}$$

Where $K(x, x')$ represents the kernel function, according to the Cover theorem, data sets in low dimensional space are usually linearly inseparable, but will be linearly separable when be mapped to a higher dimensional space. Through the kernel function $K(x, x')$ it can achieve $x \rightarrow \phi(x)$ mapping indirectly, so that the low dimensional data space is mapped into high dimension space to achieve linear transformation.

Based on the above method, use Kernel function to do the linear calculation after the high dimensional map, and the decision function is obtained:

$$\begin{aligned} f(x) &= \mathrm{sign}((\boldsymbol{\omega}, x) + b) \\ &= \mathrm{sign}(\textstyle\sum_j y_j \alpha_j K(x, x_j) + b) \end{aligned} \tag{3}$$

$$\boldsymbol{\omega} = \textstyle\sum j\ \alpha_j\ K(x, x_j) \tag{4}$$

Where *w* represents the weight vector, *b* represents the threshold, α_j represents the Lagrange multiplier. After a high dimensional map, the main problem of linear SVM will be changed to:

$$\begin{cases} \min J(\boldsymbol{\omega},\ b) = \dfrac{1}{2}\boldsymbol{\omega}^T\boldsymbol{\omega} \\ \text{s.t.} \quad y_j\left(\boldsymbol{\omega}^T\phi(x_j) + b\right) \geq 1 \end{cases} \tag{5}$$

The corresponding dual form is

$$\begin{cases} \max W(\boldsymbol{\alpha}) = \boldsymbol{e}^{\mathrm{T}}\boldsymbol{\alpha} - \dfrac{1}{2}\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{\alpha} \\ \text{s.t. } \boldsymbol{y}^{\mathrm{T}}\boldsymbol{\alpha} = 0 \qquad \alpha_j \geq 0, \forall j \end{cases} \tag{6}$$

Where *e* represents the unit vector, *Q* represents $l \times l$ matrix, and $Q_{ij} = y_i y_j k(x_i, x_j)$.

Fuzzy Granulation of Time Series

Establishment of Fuzzy Granulation Gaussian Model. Information granulation was proposed by Professor L.AZadeh in 1979. It is a research method to take the similar objects (in the form of group) as a whole or segment the whole object effectively. The granulation methods include the nom fuzzy granulation method (*c*-granulation) and fuzzy granulation (*f*-granulation).

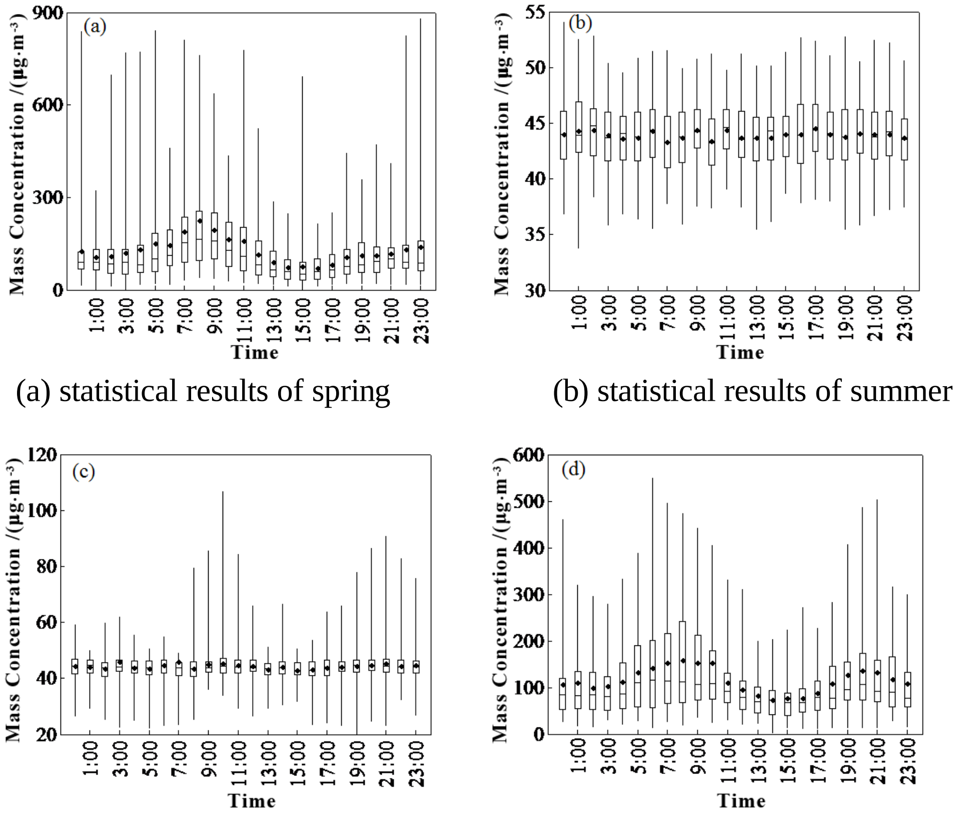
In this paper, we use the fuzzy granulation method, which divide the sample data into several subsequences (also known as the operating windows) according to the time sequence, then generate a number of fuzzy sets (i.e., fuzzy particles). In the process of fuzzy information granulation, the quality of the fuzzy degree affects the final granulation results. Commonly used fuzzy particles mainly have ladder type, triangle type, parabolic type, Gaussian type and so on. Here we use asymmetric Gaussian fuzzy model. The membership functions and images are as Eq.7 :

$$A(x,m,\sigma,\mu) = \begin{cases} \exp\left[-(x-m)/\sigma^2\right], x \leq m \\ \exp\left[-(x-m)/\mu^2\right], x > m \end{cases} \tag{7}$$

Time Series Analysis of SO_2 Concentration Change. Time series analysis refers to that using a set of digital sequence in chronological order (time series) to analyze the statistical laws reflected by a random data sequence used in order to solve practical problems. Usually a time series will have fore elements: trend, seasonal fluctuations, cycle fluctuations and irregular fluctuations. Studies have found that the SO_2 concentration change has two time series elements: the seasonal fluctuation and the diurnal cycle fluctuation.

According to the traditional astronomical seasons division method, we divide twelve months into 4 seasons: January to March as the spring, April to June as the summer, July to September as the autumn, October to December as the winter, Now discuss the statistical regularity of the hourly monitoring data of SO_2 concentration in a certain monitoring point from April 2014 to March 2015. Statistical results are shown in Fig. 2. In Fig. 2 the straight lines represent the variation range of SO_2 concentration monitoring data, the upper endpoint represents the maximum value, and the lower endpoint represents the minimum value. Rectangles represent dispersion degree of SO_2 concentration variation, and the upper edge of the rectangular represents 75% quantile values, the edge represents 25% quantile values, rectangular square represents the average value, the rectangular horizontal lines represents the median.

From Fig. 2 we can see that the seasonal differences of SO_2 concentration are obvious. It can be seen that the maximum value of SO_2 concentration in spring is about 20 times of the maximum value in summer, and the minimum values in the four seasons are basically the same, so in spring the concentration change is the largest, and in summer it is the smallest. Focus on the dispersion degree, we can see that in summer and autumn, the variation of SO_2 concentration is small, and the difference between the 25% and 75% quantile values is about 3-6μg·m^{-3}, while the concentration variation in spring and winter is much larger than that in summer and autumn, and the difference between the 25% and 75% quantile values is about 50-180μg·m^{-3}. Focus on the median and the average value, we can see that in summer and autumn, the median and average are basically the same, and the data are evenly distributed, while in spring and winter, the median value is lower than the average, and the monitoring data is distributed in the shape of a gourd, and the two ends are distributed unevenly. The average values appear peak in the morning (7:00-9:00) and reach a trough in the afternoon (14:00-16:00), and the whole average values present cycle fluctuation like a sine curve.



(a) statistical results of spring

(b) statistical results of summer

(c) statistical results of autumn

(d) statistical results of winter

Figure 2. The hour statistical results of SO_2 concentration from April 2014 to March 2015

Establishment of SO2 Concentration Time Series Prediction Model

Fuzzy Granulation of Sample Data. Divide the sample data in April 2014 to 2015 March by time series into spring part, summer part, autumn part and winter part, and predict the concentration change trend and scope of the last one day of the four seasons, in order to verify the accuracy of the forecast results. Use the monitoring data from January 2015 to March 2015 as the study object of the SO_2 concentration variation regularity in spring, and predict the 24 hour concentration change trend in March 31, 2015. We have 1848 effective sample data and the range is $y_1 \in (8.6,883)\mu g \cdot m^{-3}$.

Use the monitoring data from April 2014 to June 2014 as the study object of the SO_2 concentration variation regularity in summer, and predict the 24 hour concentration change trend in June 30, 2014. We have 2150 effective sample data and the range is $y_2 \in (33.7,54.1)\mu g \cdot m^{-3}$.Use the monitoring data from July 2014 to September 2014 as the study object of the SO_2 concentration variation regularity in autumn, and predict the 24 hour concentration change trend in September 30, 2014. We have 2208 effective sample data and the range is $y_3 \in (21.3,107.1)\mu g \cdot m^{-3}$.Use the monitoring data from October 2014 to December 2014 as the study object of the SO_2 concentration variation regularity in winter, and predict the 24 hour concentration change trend in December 31, 2014. We have 2059 effective sample data and the range is $y_4 \in (1.1,550.2)\mu g \cdot m^{-3}$.

Divide the original SO_2 concentration sample data of four seasons into many sub windows using the selected 24 hour granulation window width, and filter out a fuzzy particle by Gaussian function for every sub windows. The fuzzy particles of all sub windows form the input samples of the training model. Taking spring as an example, the variation of the initial concentration with time is shown in Fig. 3. There are 1848 valid data, so the number of granulation sub windows is equal to 1848 divided by granulation window width 24, then rounded to 77. The sample data after fuzzy granulation is shown in Fig. 4, and three kinds of concentration curves are plotted, where U indicates the maximum value of SO_2 concentration change, R indicates the mean value of SO_2 concentration change, L indicates the minimum value of SO_2 concentration change.

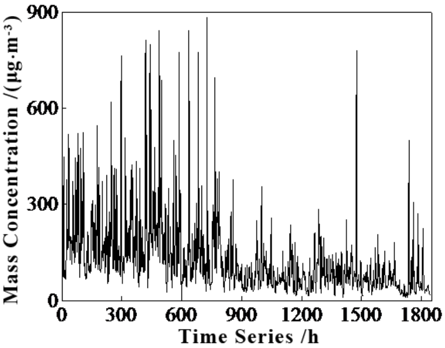


Figure 3. SO_2 concentration in the spring

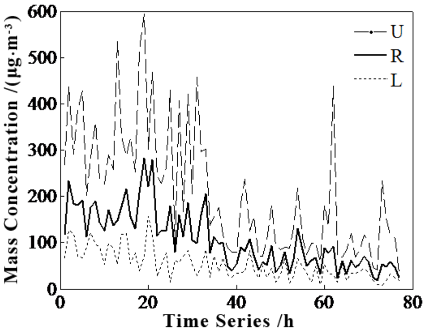


Figure 4. SO_2 concentration after fuzzy granulation

Implementation of Regression Prediction based on LIBSVM. Use LIBSVM toolbox to do the regression prediction for the processed sample data set. Steps are: Normalize the sample data set. Try to search for the optimal regression parameters of the penalty parameter C and kernel function parameter G by using the K-fold cross validation method in the range of $[2^{-10}, 2^{10}]$. After getting the best regression parameters, do the regression prediction for the minimum, average and maximum values of the sample set. The prediction results are shown in Fig. 5, and it can be seen that the three fitting results are consistent with the original data, and the model has good predictive ability. Similarly, we can get the prediction model summer, autumn and winter.

Table 1 lists the comparison results of the optimal values of the penalty parameter C and the kernel function parameter G of three fuzzy granulation objects (minimum L, average R and maximum U) for models of spring, summer, autumn and winter. The comparison results of the actual and predictive values of the three fuzzy granulation objects for the four season models are shown in Fig. 6. As can be seen from Table 1, the optimal value of the model parameters can ensure that the standard error δ_{MRE} is small enough to effectively reduce the degree of dispersion of the input sample set and improve the generalization performance of the model.

Monitoring point is located in the provincial nature reserve, and follows the first level standards in accordance with the national standard GB3095-2012, which stipulates that the upper bound of the mean value of 24 hours is 50μg·m^{-3}.As shown in Fig. 6(a), the spring predicted value is 37.136μg·m^{-3}, and the actual average value is 29.35μg·m^{-3}, so both of the actual values and predicted values are less than 50μg·m^{-3}, meeting the first level standard. As shown in Fig. 6(b), the summer predicted value is 42.596μg·m^{-3}, and the actual average value is 44.125μg·m^{-3}, so both of the actual values and predicted values are less than 50μg·m^{-3}, meeting the first level standard. As shown in Fig. 6(c), the autumn predicted value is 39.474μg·m^{-3}, and the actual average value is 44.404μg·m^{-3}, so both of the actual values and predicted values are less than 50μg·m^{-3}, meeting the first level standard. As shown in Fig. 6(d), the winter predicted value is 63.351μg·m^{-3}, and the actual average value is 65.579μg·m^{-3}, so both of the actual values and predicted values are less than 150μg·m^{-3} and more than 50μg·m^{-3}, meeting the second level standard, because in winter, the emission of SO_2 is influenced by the heating supply. Obviously, the model prediction results can represent the property of the actual gas concentration.

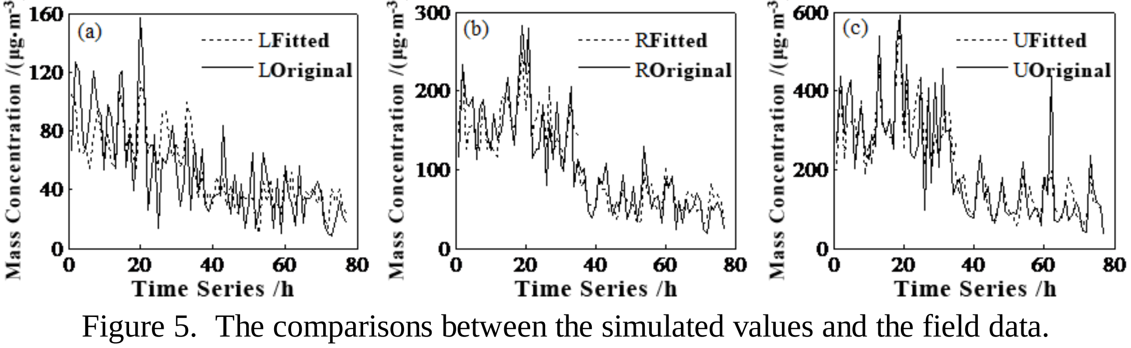


Figure 5. The comparisons between the simulated values and the field data.

Table 1 The performance parameters and error analysis of the model

Time series model	Object	Optimum parameters		δ_{MRE}	r^2	Predictive value /(μg·m^{-3})	Actual value /(μg·m^{-3})	Absolute error /(μg·m^{-3})
		G	*C*					
spring	L	1	0.5	0.014	0.580	23.029	15	8.029
	R	1	1	0.018	0.736	37.136	29.35	7.786
	U	1	0.25	0.053	0.639	66.287	48.2	18.087
summer	L	0.5	0.25	0.002	0.494	39.432	39	0.432
	R	0.5	0.25	0.001	0.511	42.596	44.125	1.529
	U	0.036	0.25	0.002	0.478	45.797	51.6	5.803
autumn	L	0.016	0.25	0.005	0.417	33.402	22	11.402
	R	1	1	0.007	0.435	39.474	44.404	4.93
	U	0.25	0.25	0.007	0.556	54.607	79.8	25.193
winter	L	1	0.25	0.019	0.703	29.154	12.9	16.254
	R	1	0.35	0.033	0.542	63.351	65.579	2.228
	U	1	0.25	0.057	0.562	134.737	178	43.263

Summary

Using time series model to predict the concentration of SO_2 can avoid considering all the pollution sources and influencing factors, and can describe the future trends and variation range of SO_2 concentration precisely. The time series model which is established by the combination of SVM and fuzzy granulation has high prediction accuracy and satisfactory generalization performance. Using Gaussian function to extract the characteristic value of the three kinds of fuzzy granulation objects can improve the accuracy of the model and improve the fitting effect. Due to that there is still no perfect theoretical basis for the factors that affect the performance of SVM, we can focus on it and further improve the accuracy of the SO_2 concentration model.

References

[1] C.B. Liu, X.F. Wang and F. Pan: Predicting Air Pollutant Emissions from a Medical Incinerator Using Grey Model and Neural Network, Applied Mathematical Modelling. Vol. 39 (2014) No.5-6, p.1513..

[2] Nunnari. G, S. Dorling and U. Schlink: Modelling SO2 Concentration at a Point with Statistical Approaches, Environmental Modelling & Software. Vol. 19 (2004) No.10, p.887..

[3] Birant D: Comparison of Decision Tree Algorithms for Predicting Potential Air Pollutant Emissions with Data Mining Models, Journal of Environmental Informatics. Vol. 17 (2011) No.1, p.46.

[4] L. Chen, D.M. Wu and Q. Chen: Prediction of Air Pollution based on Wavelet Analysis and Support Vector Machine, Journal of Xian University of Science & Technology. Vol. 30 (2010) No.6, p.726..

[5] Yeganeh, B., Motlagh, M. S. P., Rashidi, Y., and Kamalan: Prediction of CO Concentrations based on a Hybrid Partial Least Square and Support Vector Machine Model. Atmospheric Environment. Vol. 55 (2012) No.3, p.357.