Prediction of PM$_{10}$ Concentrations in the Ningdong Base

Fengjun Li[1, a], Xiaole Guo[1, b]

[1]School of Mathematics and Statistics, Ningxia University, Yinchuan 750021, China

[a]email: muzi152148@163.com, [b]email: 501875738@qq.com

Keywords: PM$_{10}$; SVR; Prediction; Ningdong Base

Abstract. Utilization of PM$_{10}$ concentrations and meteorological data of Ningdong Base, the PM$_{10}$ concentration is predicted based on LS-SVR, BP-ANN and traditional MLR models. The results show that the LS-SVR model can better describe the nonlinear dependence between PM$_{10}$ concentration and meteorological factors, and predict the PM$_{10}$ concentration more accurately.

Introduction

Predicting air environmental quality has great significance for people to understand the status of air quality in the future [1]. The existing prediction methods [1-4] need to construct function relation between influencing factors and prediction results, and have to determine the weight value reasonably. There is very complex nonlinear relation between factors and results in practice.

Support vector machine (SVM) is based on the small samples statistical learning theory (SLT). It has a simple mathematical form, intuitive geometric interpretation and good generalization ability. Meanwhile it can effectively avoid the local extremum, overcome the "dimension disaster", and has fewer artificial setting parameters. Least squares optimization of support vector regression machine (LS-SVR) has been widely used in image noise processing, structure detection and so on [4]. But rarely make the new application in the field of environment. In this paper, the LS-SVR, back propagation (BP) ANN and MLR models are used to predict PM$_{10}$ concentration in Ningdong base. And finally, comparative analysis is made among these methods.

Theory

Using SVM to solve nonlinear mapping is similar to solve the classification problems. The low-dimensional nonlinear regression problem is converted into the higher dimensional linear regression problem by kernel function. The principle is based on the given training set:

$$T = \left\{ (x_1, y_1), (x_2, y_2), \mathbf{L}, (x_l, y_l) \right\} \in (X \times Y)^l, \tag{1}$$

where $x_i \in X = R^n$ is input, $y_i \in Y = R$ is output, and the aim is to find a real function $f(x)$, then use any input x to infer the corresponding output y [5]. The hyper plane which the desires of classification problems is actually the solution of regression problem. This SVM is called support vector regression machine (SVR). The SVM topology structure is similar to a three-layer feed forward neural network. The hidden layer node is corresponding to the input sample and a support vector inner product of the kernel function, and the output node is corresponding to a linear combination of the hidden layer and output.

The least squares can be used to solve the regression problem in order to accelerate the solving speed, and the model is called the LS-SVR [6]. Data are from four hours PM$_{10}$ concentration values of ambient air point of Ningdong environmental monitoring station. Respectively,

monitoring during the winter (January 1 to January 31, 2012), spring (March 15 to April 15, 2012), summer (June 10to July 15, 2012) and autumn (August 20to September 22, 2012), daily monitoring for 24 hours. At the same time meteorological data are collected including wind speed, temperature, relative humidity, sunlight, rainfall, etc. The 1000 groups of data are chosen from raw data for training, and 800 groups of data are selected to verify.

Models

PM_{10} concentration is affected by a variety of weather conditions. By reference [3-5] and on-the-spot investigation, the initial concentration (PM_{10}^0, ug/m³), wind speed (v, m/s), temperature (T, □) and relative humidity (H, %), pressure (p, kPa) , rainfall (R, mm), sunlight (s, H) are selected as input parameters.

The LS-SVR model is established in Matlab7.8. The kernel function of SVR model should be constructed, and must satisfy the Mercer theorem. Considering the suitability and convenience, the Gauss function of radial basis function (RBF) is selected:

$$k(x, x_i) = \exp(-\frac{\|x - x_i\|^2}{s^2}) , \tag{2}$$

where x_i is training data points, s is the width of function parameter.

The BP-ANN model is established by the ANN toolbox of Matlab7.8. The input layer has seven nodes, and one output node. The empirical formula is used to select the hidden layer nodes:

$$l = \sqrt{mn} + \frac{m}{2} , \text{ for } l = 3 , \tag{3}$$

where n is the number of input layer neurons, m is the number of output layer neurons[11]. We select the most commonly used type S function as the transfer function:

$$f(x) = \frac{1}{1 + e^{-x}} . \tag{4}$$

The MLR model is established by the statistics toolbox of Matlab7.8. The model is as below:

$$y = b_0 + \sum_{i=1}^{n} b_i x_i + e_i , \tag{5}$$

where x_i is the value of the i-th input variable, y is the measured PM_{10} concentration, constant term b_0 and regression coefficient b_i are obtained by least square method, e_i is regression error. The solving process of regression in fact is the process of minimizing the average error e .

Results and analysis

Model accuracy comparison

The table 1 shows the statistical results of the three prediction model. Under the same variance, LS-SVR model has the smallest relative error and the highest percentage.

Table 1 The statistical results of prediction model

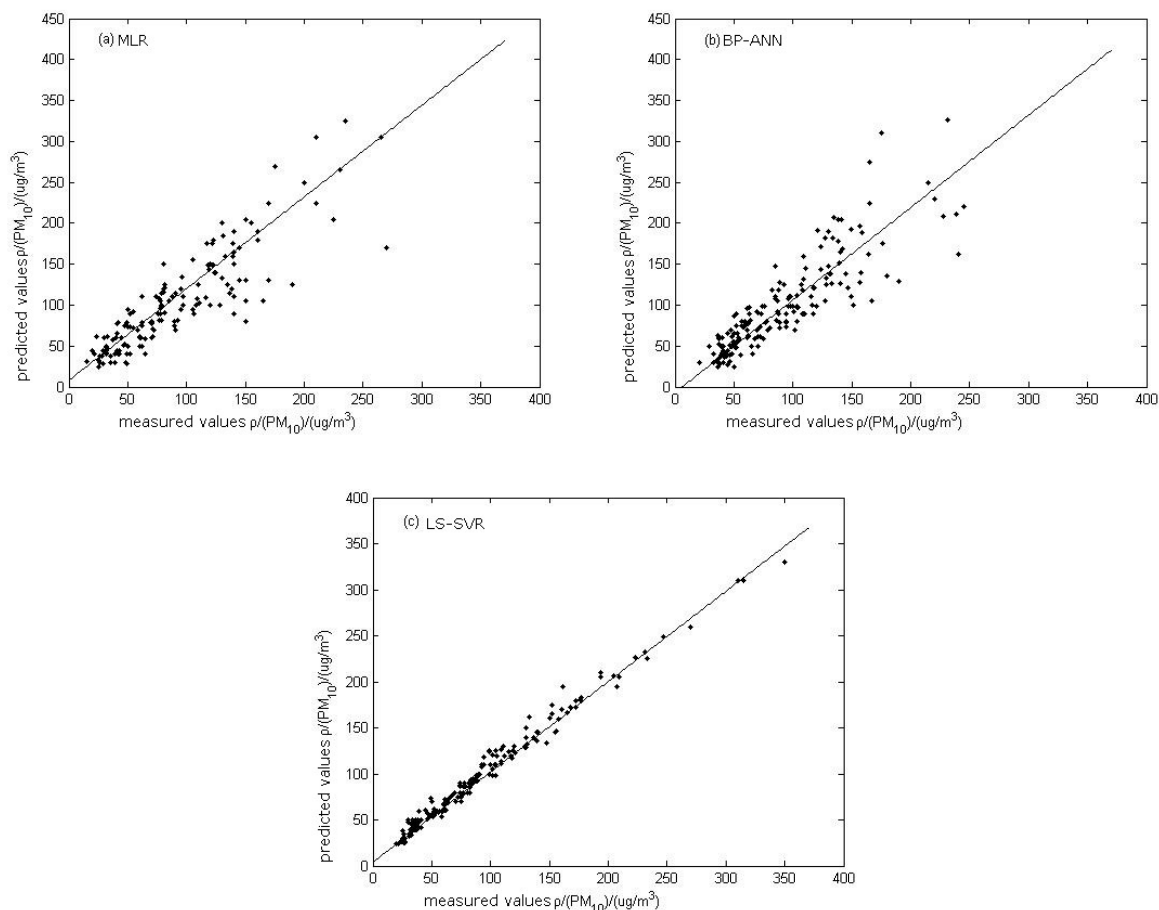|	Variance	Relative Prediction Error	Percentage
LS-SVR	0.0089	0.0039	98%
BP	0.0089	0.0272	89%
MLR	0.0089	0.0396	67%

Fig.1 Scatter plots of predicted and observed of three models

The (a), (b), (c) in Figure 1 show the results of MLR, BP-ANN and LS-SVR, respectively. The related coefficients $|R|$ with significance level $a = 0.05$ follows below, MLR model is only 0.5. The BP-ANN model is 0.83. The LS-SVR model is up to 0.97, and the slope is more close to 1. It means that the predicted value by LS-SVR and the measured values have significant correlation. The (a) in Figure 2 shows the fitting degree under various time series model, its error is shown in table 2. Visibly, the LS-SVR model is more suitable for predicting the PM_{10} concentration.

Table 2 Different error of the prediction model

Test indicators [2]	LS-SVR	BP-ANN	MLR
E_{MA}	18.67	29.60	37.00
E_{RMS}	34.78	42.45	53.60

Influence factors analysis

The LS-SVR, BP-ANN model can't directly output the weights of influence parameters. The MLR model which is established by the statistics toolbox of Matlab7.8 is set up as follows:

$$r(PM_{10}) = 504.993 + 0.383 \times r(PM_{10}^0) - 7.712 \times v - 28.69 \times R$$
$$-1.143 \times T - 0.281 \times H - 0.363 \times p + 0.051 \times S \tag{6}$$

where $r(PM_{10})$ is predicted concentration, $r(PM_{10}^0)$ is the initial concentration. Visibly, initial concentration, rainfall, wind speed, temperature and pressure can influence the PM_{10}.

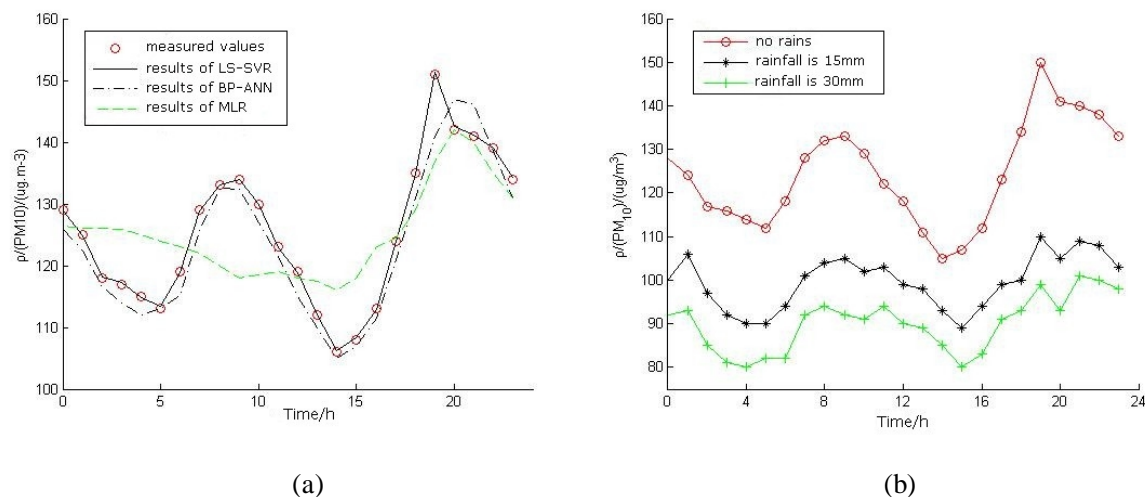PM_{10} concentration under different rainfall of 24 hours by LS-SVR is shown in figure 2(b).

(a)   (b)

Fig.2 The result comparison of 24 hours variation

The increased rainfall can obviously reduce the PM_{10} concentration, and the average reduction is 17. 28%. The reduction of PM_{10} concentration will decline with the increase of rainfall, but not rendering arithmetic proportion. We can project that the influence factors of precipitation after saturation other influence factors will be dominant.

Conclusion

Compared with the BP-ANN and traditional MLR model, the LS-SVR prediction model can well predict PM_{10} concentration. By controlling the emissions of the pollution sources or increasing rainfall precipitation can effectively reduce the concentration of PM_{10}.

Acknowledgements

References

[1] L. Z. Shi. Source Apportionment, concentration prediction, and meteorological effect of Particulate Matter (PM_{10}) in Urban Air (Central South University, Changsha 2011).

[2] L. Z. Shi, Q. H. Deng and C. Lu. Prediction of PM_{10} mass concentration based on BP Artificial neural network: Journal of Central South University (Science and Technology). Forum Vol. 1969-1974(2012), p. 43.

[3] S. Y. Gong and J. L. Feng. Relationships among relative humidity and PM_{10} concentration and atmospheric visibility in Shanghai: Research of Environmental Sciences. Forum Vol. 628-632 (2012), p.25.

[4] B. Zhen, Y. G. Feng and J. Li. Characterization and source apportionment of $PM_{2.5}$ and PM_{10} in Hangzhou: Environmental Monitoring in China. Forum Vol. 44-48(2010), p.26.

[5] A. Russo, R. C. Frank and G. L. Pedro. Air quality prediction using neural networks with stochastic variables: Atmospheric Environment. Forum Vol. 822-830(2013), p. 79.

[6] Jian Ling: Applications of SVM to predict Silicon Content in Hot Metal(Zhejiang University, Zhejiang 2006).