SVR Smoggy Forecast Model Based on Data mining

Zihan Wu

Hebei University-Rockwell Automation Laboratory2, Baoding 071000, China

Abstract

In recent years, the haze in the north of China is more serious. Effective haze forecast is more important. A SVR(Support Vector Regression) model is proposed based on Grid(Grid optimization) and PSO (Particle Swarm Ptimization)

Keywords

PM2.5 Forecasting; SVR; Factor Analysis; Grid Traversal Method; PSO.

1. Introduction

Haze occurs in the near-surface layer of a weather disaster. When visibility is reduced due to atmospheric fog and haze happen, so would social economy and people's lives have an impact. At the same time, the fog and haze occurs in the atmosphere near the formation, making the air pollution, air quality is decreased, causing serious harm to human health[1].

Nans-han Zhong, academician of Chinese Academy of Engineering pointed out: Haze not only have a serious impact on the respiratory system, but also on the cardiovascular, cerebrovascular, nervous system, etc[2]. Effective haze forecast not only can prompt children, old people, stay indoors, but also to avoid physical exertion.Xue-kuan Ma[3], the chief of the central meteorological station point out: Although the haze can be predicted, but to improve the accuracy of the Prediction is a world problem.

SVR has an unique advantage in solving small sample and nonlinear problem. In this paper, we discuss the feasibility of applying support vector regression method to forecast PM2.5[4]. GA method is used to optimize parameters of SVR, so as to improve the accuracy of the prediction model

2. Basic Principles

2.1 Basic principles of support vector regression machine

SVR is a new generation of machine learning technique developed by Vapnik[5]based on statistical learning theory. It can solve the small sample, nonlinear, high dimension and local minimum problems. It has become one of the hot research topics in machine learning field, and has been applied to the classification, regression, time series forecasting and so on[6-8].

The basic idea of SVR is shown in Figure 1-1. The hollow circle and fork is two kinds of samples. $H:\omega^T x+b=0$ is a hyperplane between them. $H1:\omega^T x+b=-1$, $H2:\omega^T x+b=1$ is the shortest distance from the H which through various types of samples. The interval is Δ.

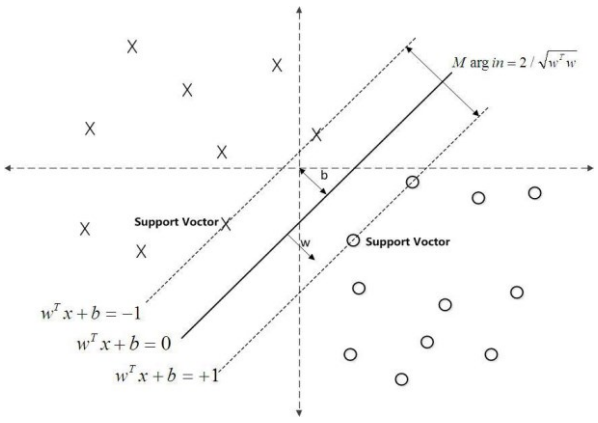

Fig.1 The basic principles of SVM

A training set $\{X_i, Y_i\}, i = 1,2,...,l, Y_i \in \{-1,1\}, X_i \in R^d$ .In this space, there is a hyperplane $H : \omega^T x + b = 0$ can accurately classify the samples of attribute space. At the same time, there is also a hyperplane parallel to the hyperplane A and B. The sample of the two kinds is right to fall on $H1$ and $H2$ respectively which distance H recently, While other training samples will be located outside the B and A, meet the constraints: $M \arg in = 2/\sqrt{\omega^T \omega}$ .To find such a hyper plane, only need to maximize the interval $M \arg in$. Build the following formula:

$$\begin{cases} \min \left\| \omega^2 \right\| / 2 \\ s.t. Y_i(\omega \cdot X_i + b) - 1 \geq 0 \end{cases}$$

Seeking the optimal formula, you can get the optimal hyperplane. SVR algorithm and classification algorithm are essentially the same, The optimal hyperplane is not only the desires of maximum interval, but also meet the classification hyperplane of minimum deviation.

2.2 Particle Swarm Ptimization

Particle swarm optimization (PSO) [9]is a stochastic algorithm for simulating swarm intelligence. The particle swarm optimization is simplified to a single particle in a space, and the movement of its time and continuous time is analyzed[10].

The basic idea is to initialize a group of random particles and to achieve the optimal solution by iteration. After each iteration, the particles are sorted by the best fit of the history. The optimal value of particle history is selected as the first particle, the other in order of priority. The behind each particle update their own by tracking the two values: The first one is the optimal solution Pid particle i found itself. Another is the optimal solution for the entire particle group $P(i-1)d$.

Particle i in N-dimensional space is expressed as: $x_i = (x_{i1}, x_{i2}, ..., x_{iN})^T$

Speed is expressed as: $v_i = (v_{i1}, v_{i2}, ..., v_{iN})^T$

Individual extreme value is expressed as: $p_i = (p_{i1}, p_{i2} ...., p_{iN})^T$

PSO evolution equation can be described as:

$$v_{id}^{t+1} = \omega \cdot v_{id}^t + c_1 r_1 (p_{id}^t - x_{id}^t) + c_2 r_2 (p_{(i-1)d}^t - x_{id}^t) \quad \ldots \quad (1)$$

$$x_{id}^{t+1} = x_{id}^t + v_{id}^{t+1} \quad \ldots \quad (2)$$

$i = 1,2,3,\cdots,\ m, d = 1,2,\cdots, D; \omega$, which is a non negative constant, is called an inertia factor. c_1, c_2 is called learning factor. r_1, r_2 is a random number between [0,1]. t is the number of iterations.

3. Air pollution forecast models

3.1 The selection of input factors

In this paper, the establishment of a training model temperature, relative humidity, wind speed, air pressure, water vapor condensation point (dew point) and precipitation sample.

Factor analysis is a method of multivariate analysis. In a multivariate (indicator) system, factor analysis is often used in comprehensive evaluation and monitoring. Factor analysis is mainly to study the relationship between correlation matrix and covariance matrix. Multiple variables are transformed to a factor, thus to achieve the relationship between the original data and the factors.

Table 1 Correlation table of factor analysis form sample

Correlation coefficient					
	Temperature	Dew point	Humidity	Air pressure	Wind speed
pm2.5	0.074	0.747	0.697	-0.482	-.561

In this paper, we use the statistical software SPSS to analyze the training sample. The results show that: the correlation PM2.5 and precipitation is small, and humidity, dew point, wind speed, temperature, air pressure, and the correlation is big. Therefore, larger correlation meteorological factors is chosen as the training samples.

Table 2 for sphericity test and KMO test of sample

KMO and Bartlett		
Kaiser-Meyer-Olkin		0.759
Bartlett test	chi-square	2483.375
	Df	105
	Sig.	0.000

Test results show that: KMO (test statistic) is 0.759. Other variables can be explained by variable interpretation. Factor analysis can be done.

3.2 SVR model selection

The kernel function is the key to SVR. Different kernel functions can lead to different SVR. It is very important to choose appropriate kernel function based on the specific data. The following empirical rules can be used for reference: RBF is Select if the characteristic number is far less than the sample. According to the selected sample and the Characteristic factor, RBF is chosen as the kernel function.

The penalty parameter C and RBF kernel parameter g in SVR are the important parameters to improve the generalization ability of the model. The optimization of penalty parameter C and RBF kernel parameter g become the key to improve the model accuracy.

SVR parameter optimization process chart：

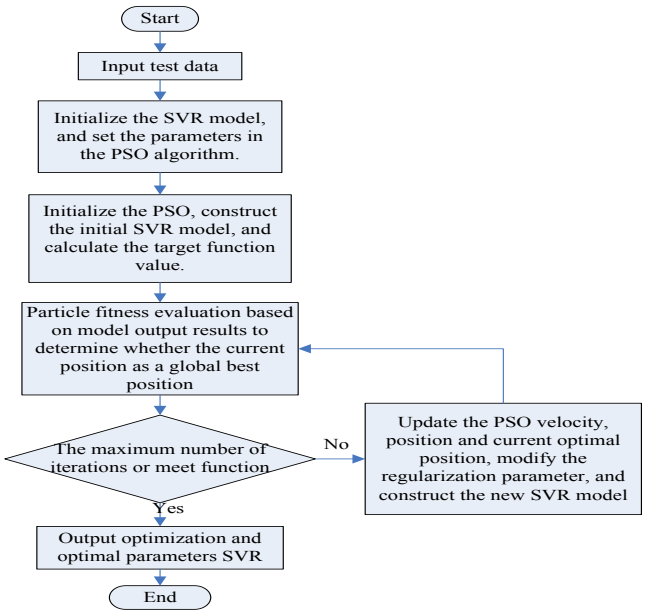

Fig.2 the SVR parameter optimization process table

3.3 SVR model parameter selection

Set the range of C is (-10, 10) ,step size is 1, the range of G is 1, the range of P is (-10,10) step is 1, and the V is 10. Parameters optimization results for Bestc=256.0, Bestg=0.5, Bestp=32.0 [11]

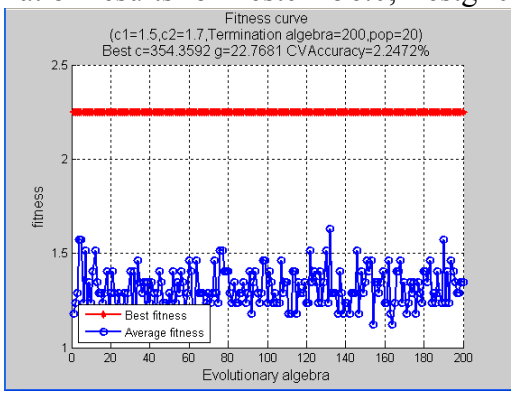

Fig.3 PSO method optimization fitness curve

4. prediction experiments

4.1 Experimental software

LIBSVM is an efficient SVR pattern recognition and regression toolbox which is designed by the (Chih-Jen Lin) professor of National Taiwan University[12] .

4.2 prediction experiment

Put the parameters of penalty parameter C and RBF kernel parameter g which obtained by Grid Traversal Method into the SVR to find the Forecasting model and get the change curve of $PM_{2.5}$. Figure 4 is a comparison between actual data and forecast data:

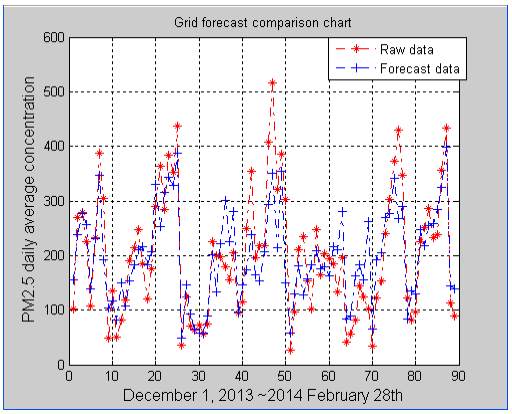

Fig.4 Contrast map of Gird Traversal method optimization and prediction

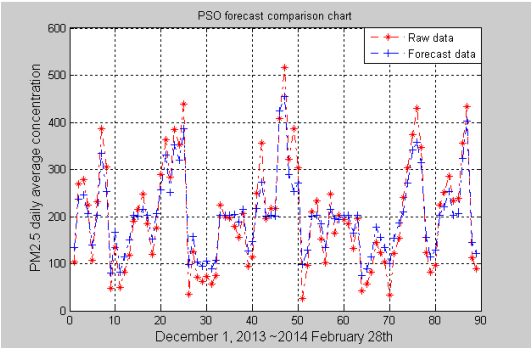

Fig.5 Contrast map of PSO method optimization and prediction

Put the parameters of penalty parameter C and RBF kernel parameter g which obtained by PSO into the SVR to find the Forecasting model and get the change curve of $PM_{2.5}$. Figure 4 is a comparison between actual data and forecast data

4.3 Analysis of experimental results

Figure 4 and 5 is a comparison of the PM2.5's forecast and PM2.5's actual value. The combination of SVR and PSO can capture the nonlinear relationship between PM2.5 and the feature vector well. Experimental results show: The fitting degree between PM2.5 forecast and PM2.5 is higher. In a dramatic change, there is a slightly larger deviation from the forecast curve, but the trend of the forecast curve is consistent with the actual trend. Tracking performance is better. On the whole, SVR combined with PSO is more sensitive to PM2.5 change

5. conclusion

1. In view of the fog haze is increasingly serious and difficult to predict, In this paper, the combination of SVR and genetic algorithm is used to predict the change of PM2.5, By comparing the figure above the predicted value and the actual value of PM2.5 concluded： The model of SVR and PSO parameters optimization model has good effect on PM2.5 prediction.

2. Due to the use of real-time PM2.5 data value and has a larger dimension, the MSE as a result of the forecast evaluation standard is not very accurate. Therefore, graph comparing predicted results show relatively intuitive.

3. Examples confirm that the SVR has a higher accuracy and better training speed for the prediction of PM2.5. Apply some of the less demanding forecast.

4. Because of data limitations, the selection of meteorological factors have certain limitations, ignores the impact of the underlying surface and in the lower circulation etc on haze, Therefore, adding more reasonable input factors is the main direction of improving the accuracy of the model.

References

[1] Zhang R H, Li Q, Zhang R N. 2014. Meteorological conditions for the persistent severe fog and haze event over eastern China in January 2013. Science China: Earth Sciences, 57: 26–35, doi: 10.1007/s11430-013-4774-3.

[2] Zhuo-sen Yang haze pollution induced health effects in humans and the research progress of [J]. Occupation and health, 2014, 30 (17): 2517-2520(in Chinese)

[3] Zhong-jun Xie, haze forecast, world class "Difficult miscellaneous diseases"[N].China weather report, 2013-02-14, (1) (in Chinese)

[4]Chang Tao. Research on Application of support vector machine in air pollution prediction [J].Meteorological Monthly, 2006, 32(12): 61-65 (in Chinese)

[5] Liva Ralaivola, Flovence d' Alche-Bue. Incremental support vector machine Learning: alocal approach. Proceedings of International Conference on Neural Networks, Vienna,Austria, 2001: 322-330.

[6] Nai-yang Deng, Ying-jie Tian, a new method in data mining support vector machine [M].Beijing: Science Press, 2006.224-235 (in Chinese)

[7] Guo-zheng Li, Meng Wang, Hua-jun Zeng. An introduction to support vector machine [M]. Beijing: Publishing House of electronics industry, 2005: 98-105 (in Chinese)

[8] Bai Peng, Zhang Xi-bin, Zhang Bin et al. The support vector machine theory and engineering application [M]. Xi'an: Xi'an electron science and Technology Press, 2008: 41-55 (in Chinese)

[9] clere M and Kennedy J. The Particle Swarm: Explosion, Stability and Convergence in a Mulit-Dimensional Complex Space [J], IEEE Trainsaction on Evolutionary Computation, 2002, 6 (1) : 58-73.

[10] Wei-dong Chen. SLAM algorithm based on improved PSO algorithm for fuzzy adaptive extended Calman filter [J]. Journal of Physics, 2013, 62(17): 1-7.

[11] Bao-guo Xu, Wei-li Xiong. Research on SVR parameter optimization selection method based on PSO [J]. System simulation report, 2006, 18(9): 2442-2445

[12]LIBSVM:a Library for support Vecotr Maehine [OL]. Chih_Chung. http://csie.ntu.edu.tw