Using support vector regression to predict PM$_{10}$ and PM$_{2.5}$

Hou Weizhen$^1$, Li Zhengqiang$^1$, Zhang Yuhuan$^{1,2,*}$, Xu Hua$^{1,2}$, Zhang Ying$^{1,2}$, Li Kaitao$^{1,2}$, Li Donghui$^{1,2}$, Wei Peng$^{1,2,*}$, Ma Yan$^3$

$^1$Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, Beijing, China.

$^2$University of Chinese Academy of Sciences, Beijing, China.

$^3$School of earth sciences and engineering, Hohai University, Nanjing, China.

houwz@irsa.ac.cn

Abstract: Support vector machine (SVM), as a novel and powerful machine learning tool, can be used for the prediction of PM$_{10}$ and PM$_{2.5}$ (particulate matter less or equal than 10 and 2.5 micrometer) in the atmosphere. This paper describes the development of a successive over relaxation support vector regress (SOR-SVR) model for the PM$_{10}$ and PM$_{2.5}$ prediction, based on the daily average aerosol optical depth (AOD) and meteorological parameters (atmospheric pressure, relative humidity, air temperature, wind speed), which were all measured in Beijing during the year of 2010-2012. The Gaussian kernel function, as well as the k-fold crosses validation and grid search method, are used in SVR model to obtain the optimal parameters to get a better generalization capability. The result shows that predicted values by the SOR-SVR model agree well with the actual data and have a good generalization ability to predict PM$_{10}$ and PM$_{2.5}$. In addition, AOD plays an important role in predicting particulate matter with SVR model, which should be included in the prediction model. If only considering the meteorological parameters and eliminating AOD from the SVR model, the prediction results of predict particulate matter will be not satisfying.

Keywords: support vector regression, PM$_{10}$, PM$_{2.5}$, AOD, meteorological parameters

1. Introduction

In recent years, with the rapid development of industrialization and urbanization in China, urban air pollution has been a grow problem, especially for urban communities. Health effects differ upon the size of airborne particulates. In this contribution, PM$_{10}$ and PM$_{2.5}$ (particulate matter less or equal than 10 and 2.5 micrometers respectively) are considered due to its effect on human health. Epidemiologic studies indicate strong links between the concentration of PM$_{10}$ or PM$_{2.5}$ with public morbidity, mortality of respiratory and cardiovascular diseases. Recently, PM has become the primary air pollutant in most major cities in China, which not only threatens people's health, but also causes the decrease of atmospheric visibility and the degradation of the city scenery [1-3].

Support vector machines (SVM) are a new statistical learning technique, based on machine learning and generalization theories, it implies an idea and could be considered as a method to minimize the risk. Besides, a generalization capability makes possible their application to modeling dynamical and

* Corresponding author: Zhang Yuhuan, E-mail: zhangyh@irsa.ac.cn.
non-linear data sets. This study is motivated by a growing popularity of support vector regression (SVR) problems, which leads to better generalization than conventional methods such as artificial neural networks (ANN) [2, 4-8]. This paper presents a study of using the SVR model to investigate of PM$_{10}$ and PM$_{2.5}$, which were measured in Beijing during 2010-2012. The successive over relaxation support vector regressions (SOR-SVR) are trained by performed on the data of PM2.5 (or PM$_{10}$), the aerosol optical depth (AOD) and meteorological parameters (atmospheric pressure, relative humidity, air temperature, wind speed), which were also measured at Beijing at the same period. With the SVR model, based on AOD and meteorological parameters, we can predict the regional particulate matter (PM) in Beijing in China.

2. Support vector regression

2.1. standard support vector regression

In recent years, Support Vector Regression (SVR) problem is a new interest field in support vector machine. In regression problems, we are given a training data set $S = \{(x_1,y_1),\cdots,(x_m,y_m)\} \subseteq \mathbb{R}^n \times \mathbb{R}$, where $x_i \in \mathbb{R}^n$ is the input data and $y_i \in \mathbb{R}$ is the target output. The main goal is to find a function $f(x)$ that can correctly predict the observation values $y$, of the new input points $x$, by learning from the given training data set $S$. A standard regression problem of a liner SVR is

$$\begin{align*}
\min \  & \frac{1}{2} \|\omega\|^2 + C \sum_{i=1}^{m} (\xi_i + \xi_i^*) \\
\text{s.t.} \  & \omega^T x_i + b - y_i \leq \epsilon + \xi_i \\
& y_i - \omega^T x_i - b \leq \epsilon + \xi_i^* \\
& \xi_i \geq 0, \xi_i^* \geq 0, \ i = 1,2,\ldots,m
\end{align*}$$

(1)

where $\omega$ is the normal vector, $b$ is the threshold, $C$ is a regularization constant determining the trade-off between the training error and the generalization performance, $\xi$ and $\xi^*$ are the slack variables, $\epsilon$ is the tolerance (error acceptance). Then the function as

$$f(x) = \omega^T x + b$$

(2)
This problem is called $\varepsilon$-support vector regression ($\varepsilon$-SVR) and a data point $x_i \in \mathbb{R}^n$ is called a support vector if $f(x_i) - y_i \geq \varepsilon$ [4,5].

2.2. SOR-SVR model

For the standard model of SVR, if we append the term $b^2$ to $\omega^T \omega$, that is to say, maximize the margin between the parallel separating planes by optimizing with respect to both $\omega$ and $b$, meanwhile, we change the expression form with matrix and vector, this leads to the following reformulation of the SVR problem as

$$
\min \frac{1}{2} (\omega^T \omega + b^2) + C(\xi^T \xi + \xi'^T \xi')
$$

$$
s.t. \ A\omega + b e - y \leq \varepsilon e + \xi
$$

$$
- A\omega - b e + y \leq \varepsilon e + \xi'
$$

where $A$ is the $m \times n$ matrix with $x_i^T$ as a row, $i=1,2,...,m$, $e$ is an arbitrarily vector with the component as 1, and $y=(y_1, y_2, ..., y_m)$. Its dual problem is

$$
\min \frac{1}{2} (\alpha^T - \alpha'^T) \left( AA^T + ee^T \right) (\alpha^T - \alpha) - y^T(\alpha^* - \alpha) + ee^T(\alpha^* + \alpha) + \frac{1}{2C} (\alpha'^T \alpha^* + \alpha^T \alpha)
$$

$$
s.t. \ \alpha^*, \alpha \geq 0
$$

and

$$
\hat{\alpha}^T = (\alpha'^T \ \alpha^T), H = \begin{bmatrix} A & e \\ -A & -e \end{bmatrix}, I^T = \begin{bmatrix} (y - e\varepsilon)^T, -(y + e\varepsilon)^T \end{bmatrix}, Q = \frac{I}{C} + K(H^T H^T) = L + E + L^T
$$

(5)

Using the successive over relaxation (SOR) method to solve Eq.(6), we get the iterative formula of SOR algorithm as following[4]:

$$
\alpha^{i+1} = (\alpha^i - \omega E^{-1}(Q\alpha^i - e + L(\alpha^{i-1} - \alpha'))),
$$

(7)

where $()$ denotes the nonnegative gradient projection:

$$
((\alpha)_*), = \begin{cases} 0, & \alpha < 0 \\ \alpha, & \alpha > 0 \end{cases}
$$

(8)

Then the regression function can be written as

$$
f(x) = (\alpha^* - \alpha)^T K(A, x) + e^T (\alpha^* - \alpha)
$$

(9)

This SVR problem is called SOR-SVR model and Figure 1(a) further illustrate the diagram of this model.

3. Methodology and results

3.1. Collect and analyze dataset

The data collected to be used in the study refer to air quality in Beijing and contains the particulate matters (PM$_{2.5}$) monitored by U.S. Embassy (United State Embassy in Beijing, China), aerosol optical depth (AOD) measured by the sun photometer Ce318 at RADI (Institute of Remote Sensing and Digital Earth, CAS), and the meteorological data, including air pressure, temperature, relative humidity, and wind speed, which are measured by China Meteorological Administration. The three observation sites are not far away from each other, and those data were measured from 2010 to 2012.
In addition, the PM$_{10}$ data are provided from the Ministry of Environmental Protection of PRC, which were measured in the whole year of 2010.

![Figure 2](image)

**Figure 2.** Time series of daily average data results. (a) pressure and temperature; (b) relative humidity (RH) and wind speed; (c) AOD and PM$_{2.5}$.

| No. | Variable name      | Unit      |
|-----|--------------------|-----------|
| 1   | PM$_{2.5}$ (PM$_{10}$) | μm/m$^3$  |
| 2   | AOD                | --        |
| 3   | Air pressure       | bPa       |
| 4   | temperature        | °C        |
| 5   | Relative humidity (RH) | %         |
| 6   | Wind speed         | m/s       |

Table 1. Variables of dataset

Take the PM$_{2.5}$ prediction as an example, after deleting the outlier from the dataset, we can get the daily average results of PM$_{2.5}$, AOD and meteorological variable respectively, then 300 data are selected for the simulation and discussion. The variable of dataset are presented in table 1, and the time series data are shown in figure 2. For the AOD results, we have already used the Angstrom formula to transform the AOD result at 550nm.

3.2. **Design and test SOR-SVR model**

There are several issues that we need to consider in the SOR-SVR application. First of all, some
parameters must be determined before running the particular algorithm. These parameters are error acceptance ($\varepsilon$), constant ($C$) and kernel specific parameters. In this work, gauss kernel function

$$K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)$$

(10)

was used, where $\sigma$ is the parameter that determines performance in the learning of the kernel function.

As we train and test the SVR model, k-fold cross validation and grid search method are also used, consequently the optimal parameter $\varepsilon$, $C$ and $\sigma$ can be obtained. While the k-fold cross validation method is employed, with the defined value of $\varepsilon$, $C$ and $\sigma$, the dataset is divided into k parts, the k-1 parts among which can be selected as the training dataset and the remaining part as the testing dataset, then the average relative testing error can be gained. In this way, we can train and test the model k times, and then get the average value of the each testing error as final error. With different values of $\varepsilon$, $C$ and $\sigma$, we can search and select the parameter ($\varepsilon$, $C$, $\sigma$) which corresponding to the minimum final error as the optimal parameters in the SOR-SVR model. The diagram of the prediction with SVR model can see figure 1 (b).

3.3 Results

![Figure 3](image1.png)

Figure 3. Using the training dataset as the testing dataset, the comparison results of actual PM2.5 data and predicted result. (a) $\varepsilon=1$, $C=100$, $\sigma=0.1$; (b) $\varepsilon=10$, $C=100$, $\sigma=0.2$

![Figure 4](image2.png)

Figure 4. Prediction of PM2.5. (a) Time series diagram; (b) scatter diagram.

Figure 3 shows the comparison results of actual PM$_{2.5}$ and predicted result while we use the training dataset as the testing dataset for different parameter ($\varepsilon$, $C$, $\sigma$). As we can see in figure 3 (a), though the regression results match with the actual PM$_{2.5}$ in an extremely good condition, the generalization ability of this model is not very good, due to the over-fitting. With the k-fold cross validation, the over-fitting condition can be prevented, meanwhile, combined with grid search method, the optimal parameter ($\varepsilon=0.1$, $C=100$, $\sigma=0.3$) for the good generalization ability can be obtained.

With the optimal parameter, we select 250 days’ data as the training dataset and 50 days’ data as the testing dataset, the predicted result of PM$_{2.5}$ can be found in the figure 4. From the figure, we can see that the predicted result by the SOR-SVM model agree well with the actual data, the correlation coefficient $R^2=0.87$ and the average error is 12.66 $\mu m/m^3$, which can prove that the SOR-SVR model has a good generalization ability to predict PM$_{2.5}$.
Besides, in order to investigate the important role which AOD played in the prediction of PM$_{2.5}$ with SVR model, we also consider using the meteorological parameters (atmospheric pressure, relative humidity, air temperature and wind speed) directly to predict the PM$_{10}$ and PM$_{2.5}$ in the same as above. However, compared with the condition combined with AOD variable, the predicted results are satisfying, the correlation coefficient $R^2$ is about 0.7 and the average error is significantly larger. Due to the reason of space limitations, the detailed results are not shown here.

With the SOR-SVR model, we can further use the AOD retrieval result from the satellite remote sensing, combining with the corresponding meteorological data, to predict the particulate matter in a regional distribution in our next study.

4. Discussion and Conclusions

The potential of applying support vector machine (SVM) in PM$_{10}$ and PM$_{2.5}$ prediction is studied and presented in this paper, based on the daily average aerosol optical depth (AOD) and meteorological parameters such as atmospheric pressure, relative humidity, air temperature and wind speed. The predicting model is developed by using the successive over relaxation support vector regression (SOR-SVR). The Gaussian kernel function, as well as the k-fold crosses validation and grid search method, is used to obtain the optimal parameters ($\varepsilon$, $C$, $\sigma$) to get a better generalization capability. For the final model, 250 days’ daily average data are selected as the training dataset and 50 days’ data as the testing dataset, the predicted results agree well with the actual data, and show that the SOR-SVR model has a good potential and generalization ability to predict PM in the atmosphere. In addition, AOD plays an important role in predicting particulate matter with SVR model, which should be included in the model. If only considering the meteorological parameters and eliminating AOD, the prediction results of predict particulate matter will be not satisfying. With the SOR-SVR model, we can further use the AOD retrieval result from the satellite remote sensing, as well with the corresponding meteorological data, to predict the PM$_{10}$ and PM$_{2.5}$ in a regional distribution in big city such as Beijing in China.

Acknowledgement

This research is supported by National Basic Research Program of China (973 Program) under grant 2010CB950800 (2010CB950801). Besides, authors wish to acknowledge China Meteorological Administration, Ministry of Environmental Protection of PRC and U.S. Embassy in Beijing for providing the meteorological data, PM$_{10}$ and PM$_{2.5}$ data respectively in our research.

References

[1] Wang Z, et al, 2010 Satellite-based estimation of regional particulate matter (PM) in Beijing using vertical-and-RH correcting method Remote Sens. Environ. 114 50-63
[2] Arampongsanuwat S, Meesad P, 2011 Prediction of PM10 using Support Vector Regression IPCSIT 6 120-124
[3] Tai A, Mickley L, Jacob D, 2010 Correlations between fine particulate matter (PM2.5) and meteorological variables in the United States: Implications for the sensitivity of PM2.5 to climate change Atmos. Enviro. 44 3976-3984
[4] Pan M, Hou W, He G, 2006 Improved Support Vector Regression Algorithm based on Successive Over Relaxation JCIS 3 2 1127-1134
[5] Duan H, Hou W, He G, et al, 2007 Predicting Time Series Using Incremental Langrangian Support Vector Regression ISNN 3 812-820
[6] Pan M, He G, Hou W, et al, 2006 Improved Support Vector Machine based on Successive Over Relaxation CIE 4A 15 822-824
[7] Duan H, Shao X, Hou W, et al, 2009 An incremental learning algorithm for Lagrangian support vector machines Pattern Recogn. Lett. 15 30 1384-1391
[8] Artmio S, et al, 2011 Modeling PMx trends contaminants by using support vector machines ACSES Res. in Comput. Sci. 52 114-122