

Prediction of PM₁₀ using Support Vector Regression

Soawalak Arampongsanuwat¹ and Phayung Meesad²⁺

¹ Department of Information Technology, King Mongkut's University of Technology
North Bangkok, Thailand.

² Department of Teacher Training in Electrical Engineering, King Mongkut's University of Technology
North Bangkok, Thailand.

Abstract. This paper describes the development of a support vector regression (SVR) model for the PM₁₀ forecasting in Bangkok. Particulate matter (PM₁₀) with aerodynamic diameter up to 10 μm (PM₁₀) is targeted because these small particles effects to people health and it constitutes the major concern for air quality of Bangkok. The models developed are used to establish the relationships of PM₁₀ with meteorological variables including globe radiation, net radiation, air pressure, rainfall, relative humidity, temperature, wind direction and wind speed as well as the air quality concentrations of Carbon monoxide, Ozone, Nitrogen dioxide and Sulfur dioxide. The data sets examined in the current study were collected by monitoring station operated by Pollution Control Department of Thailand corresponding to PM₁₀ concentrations for the years 2007–2009. In order to provide with an operational air quality forecasting module for PM₁₀, Support Vector Regression method is investigated and applied. The results of this research show the model that set with value $C = 5,000$, $\epsilon = 0.001$ and $\sigma = 0.1$ work out most precise in forecasting over other tested models. Based on the test forecasting data, the mean squared error (MSE) was 1.0588×10^{-10} , which means this model was very satisfactory. The model reports that support vector regression can be used in forecasting PM₁₀ successfully.

Keywords: Prediction, PM₁₀, Support Vector Regression.

1. Introduction

Thailand is facing serious air pollution problems, especially in urban areas, due to rapid industrialization, urbanization and motorization. The government has set the National Ambient Air Quality Standards (NAAQS) and implemented countermeasures for criteria air pollutants such as dust, suspended particulate matters (PM₁₀ and PM_{2.5}), sulfur dioxide (SO₂), carbon monoxide (CO), nitrogen dioxide (NO₂) and ground level ozone (O₃). However, there are reported exceeding the NAAQS at many of the monitoring locations. An emerging air pollution issue in Thailand is air toxic problem resulting from transportation and industrial activities. [1]. Bangkok, the capital of Thailand is considered as one of the most polluted in Thailand. Air pollution arises from the adverse effects on the environment of a variety of contaminants emitted into the atmosphere by natural and man-made processes such as industrial emissions, fixed combustions and vehicular traffic. Air quality result around the Bangkok perimeter was monitored that pollutants exceeding the standards were particulate matters and ozone. It was found that particulate matters remained to be the main problem. Ozone was the secondary problem, whereas other pollutants, such as Sulfur Dioxide (SO₂), Nitrogen Dioxide (NO₂), Carbon Monoxide (CO) were still within the standard [2]. As mentioned, particulate matter is a cause for concern and similar increasing trends have been observed in Bangkok's adjoining provinces and throughout other urban areas in Thailand.

PM₁₀, a category of pollutants including solid and liquid particles having an aerodynamic diameter of up to 10 μm, can be a health hazard for several reasons: it can harm lung tissues and throat, aggravate asthma and increase respiratory illness. Indeed, high PM₁₀ levels have been correlated to increases of hospital admissions for lung and heart disease. These events require drastic measures such as the closing of schools and factories and the restriction of vehicular traffic. The forecasting of such phenomena would allow to take more efficient countermeasures to safeguard citizens health [2-5].

⁺ Corresponding author. Tel.: +66-02-913-2500; fax: +66-02-912-2019.
E-mail address: saowalak.a@hotmail.com, pym@kmutnb.ac.th

Therefore, Air pollution related to high concentration of PM₁₀ is becoming a matter of concern due to its adverse effects on human health. The accurate prediction models of PM₁₀ concentrations are important for proper management, control, and making public warning strategies [3]. The aim of the analysis is forecasting the air pollutants values of meteorological and forecasted data from the main polluted problem areas which are particulate matters (PM₁₀) problems.

In the last decade, artificial neural networks (ANNs) and more recently support vector regression (SVR) have emerged as two attractive tools for nonlinear modeling especially in situations where the development of phenomenological or conventional regression models becomes impractical or cumbersome. In recent years, support vector regression (SVR) which is a statistical learning theory based machine learning formalism is gaining popularity over ANN due to its many attractive features and promising empirical performance. The main difference between conventional ANNs and support vector machines (SVM) lies in the risk minimization principle. Conventional ANNs implement the empirical risk minimization (ERM) principle to minimize the error on the training data, while SVM adheres to the Structural Risk Minimization (SRM) principle seeking to set up an upper bound of the generalization error [6-8].

This study is motivated by a growing popularity of support vector machines (SVM) for regression problems. This process leads to better generalization than conventional methods. This paper presents a study of using the SVM model to investigate of air pollutants, which were measured at Din Dang Monitoring Station during 2007-2009, air pollutant monitoring stations in Thailand territory established by Pollution Control Department (PCD). The SVM was trained by was performed on data of the principal air pollutants Carbon Monoxide, Ozone, Nitrogen Dioxide, Sulfur Dioxide and Particulate Matter (PM₁₀) and meteorological parameters (solar radiation, atmospheric pressure, rain, relative humidity, air temperature, wind velocity and wind direction).

2. Support Vector Machines (SVM)

2.1. SVM methodology

The concept of a maximum margin hyperplane only applies to classification. However, support vector machine algorithms have been developed for numeric prediction that share many of the properties encountered in the classification case: they produce a model that can usually be expressed in terms of a few support vectors and can be applied to non-linear problems using kernel functions.

Similar with linear regression, the basic idea here is to find a function that approximates the training points well by minimizing the prediction error. The crucial difference is that all deviations up to a user-specified parameter x_i are simply discarded. Also, when minimizing the error, the risk of over-fitting is reduced by simultaneously trying to maximize the flatness of the function. Another difference is that what is minimized is normally the predictions' absolute error instead of the squared error used in linear regression. A user-specified parameter x_i defines a tube around the regression function in which errors are ignored.

2.2. Support Vector Regression (SVR) Modeling

Consider a training data set $g = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), (\mathbf{x}_p, y_p)\}$, such that $\mathbf{x}_i \in \mathfrak{R}^n$ is a vector of input variables and $y_i \in \mathfrak{R}$ is the corresponding scalar output (target) value. Here, the modeling objective is to find a regression function, $y = f(\mathbf{x})$, such that it accurately predicts the outputs $\{y\}$ corresponding to a new set of input-output examples, $\{(\mathbf{x}, y)\}$, which are drawn from the same underlying joint probability distribution as the training set. To fulfill the stated goal, SVR considers following linear estimation function [6-8].

$$f(x) = \langle \mathbf{w}, \mathbf{x} \rangle + b \quad (1)$$

where \mathbf{w} denotes the weight vector; b refers to a constant known as “bias”; $f(x)$ denotes a function termed feature, and $\langle \mathbf{w}, \mathbf{x} \rangle$ represents the dot product in the feature space, l , such that $\phi : \mathbf{x} \rightarrow l$, $\mathbf{w} \in l$. The basic concept of support vector regression is to map nonlinearly the original data \mathbf{x} into a higher dimensional feature space and solve a linear regression problem in this feature space.

The regression problem is equivalent to minimize the following regularized risk function:

$$R(f) = \frac{1}{n} \sum_{i=1}^l L(f(x_i) - y_i) + \frac{1}{2} \|\mathbf{w}\|^2 \quad (2)$$

where

$$L(f(x)-y)=\begin{cases} |f(x)-y|-\varepsilon & \text{for } |f(x)-y|\geq\varepsilon \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

Equation (3) is also called ε -insensitive loss function. This function defines a ε -tube. If the predicted value is within the ε -tube, the loss is zero. If the predicted value is outside the tube, the loss is equal to the magnitude of the difference between the predicted value and the radius ε of the tube. ε is a precision parameter representing the radius of the tube located around the regression function (see Fig.1); the region enclosed by the tube is known as “ ε - intensive zone”.

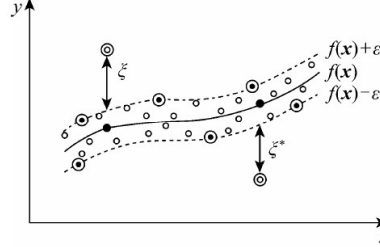


Fig. 1: A schematic diagram of support vector regression using ε -sensitive loss function

The SVR algorithm attempts to position the tube around the data as shown in Fig. 1. By substituting the ε -insensitive loss function into Eq. (2), the optimization object becomes:

$$\text{Minimize} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^l(\xi_i + \xi_i^*) \quad (4)$$

$$\text{subject to} \quad \begin{cases} y_i - \langle \mathbf{w}, \mathbf{x}_i \rangle - b \leq \varepsilon + \xi_i \\ \langle \mathbf{w}, \mathbf{x}_i \rangle + b - y_i \leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases} \quad (5)$$

where the constant $C > 0$ stands for the penalty degree of the sample with error exceeding ε . Two positive slack variables ξ_i, ξ_i^* represent the distance from actual values to the corresponding boundary values of ε -tube. The SVR fits $f(\mathbf{x})$ to the data in a manner such that: the training error is minimized by minimizing ξ_i, ξ_i^* and, \mathbf{w}^2 is minimized to increase the flatness of $f(\mathbf{x})$ or to penalize over complexity of the fitting function. A dual problem can then be derived by using the optimization method to maximize the function:

$$\text{Minimize} \quad -\frac{1}{2}\sum_{i,j=1}^n (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\langle \mathbf{x}_i, \mathbf{x}_j \rangle - \varepsilon \sum_{i=1}^n (\alpha_i + \alpha_i^*) + \sum_{i=1}^n y_i(\alpha_i - \alpha_i^*) \quad (6)$$

$$\text{subject to} \quad \sum_{i=1}^n (\alpha_i - \alpha_i^*) = 0 \quad \text{and} \quad 0 \leq \alpha_i, \alpha_i^* \leq C \quad (7)$$

where α_i, α_i^* are Lagrange multipliers. Owing to the specific character of the above-described quadratic programming problem, only some of the coefficients $(\alpha_i^* - \alpha_i)$ are non-zero and the corresponding input vectors, \mathbf{x}_i , are called support vectors (SVs). The SVs can be thought of as the most informative data points that compress the information content of the training set. The coefficients α and α^* have an intuitive interpretation as forces pushing and pulling the regression estimate $f(\mathbf{x}_i)$ towards the measurements, y_i . The SVM for function fitting obtained by using the above-mentioned maximization function is then given by

$$f(\mathbf{x}) = \sum_{i=1}^n (\alpha_i - \alpha_i^*) \langle \mathbf{x}_i, \mathbf{x} \rangle + b \quad (8)$$

As for the nonlinear cases, the solution can be found by mapping the original problems to the linear ones in a characteristic space of high dimension, in which dot product manipulation can be substituted by a kernel function, i.e. $K(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j)$. In this work, different kernel function is used in the SVR. Substituting $K(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j)$ in Eq. (6) allows us to reformulate the SVM algorithm in a nonlinear paradigm. Finally, we have

$$f(x) = \sum_{i=1}^n (\alpha_i - \alpha_i^*) K \langle \mathbf{x}_i, \mathbf{x} \rangle + b \quad (9)$$

The most known kernel functions used in practice are radial (Gaussian), polynomial, spline or even sigmoidal functions [6, 8]. Till now, it is hard to determine the type of kernel functions for specific data

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|_2^2}{2\sigma^2}\right)$$

patterns. However, any function that satisfies Mercer’s condition by Vapnik [6] can be used as the kernel function. In this work, the Gaussian function is used in SVR.

(10)

3. Methodology

3.1. Collect and analyze data sets

The data was collected to be used in the study referring to air quality and contains carbon monoxide, ozone, nitrogen dioxide, sulfur dioxide, particulate matters (PM₁₀) Meteorological data, including globe radiation, net radiation, air pressure, rainfall, relative humidity, temperature, wind direction and wind speed. This research used data from air quality monitoring stations of the Pollution Control Department from 2007 to 2009.

3.2. Data transformation

The translation of data is the process of adjusting the scope of the data in the appropriate range to be applied in training process. Normalization reduces the value of the data in a smaller scope. By the values in the range [-1, 1] will be used before training.

3.3. Design SVR Model

There are several issues that we need to consider in the SVR application. First of all, some parameters must be determined before running the particular algorithm. These parameters are error acceptance (ϵ), constant (C) and kernel specific parameters. In this work, radial basis function was used where σ is the parameter that determines performance in the learning of the kernel function.

Theoretically, the spread parameter σ greatly affects the prediction performance. Both too large and too small values of σ may lead to poor predictions. Therefore, in practical applications, only the spread parameter σ of Gaussian kernel function needs to be determined during the simulations, while the other two parameters, i.e., C and ϵ , can be set in advance by experiences. In our study, we set $\epsilon = 0.001$ and $C = 5000$. Also, we used different values for the kernel parameter. Specifically, $\sigma = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9$ and 1.0 . Here, the mean square error (MSE) was used as an assessment of deviation between original data and predictions. Generally, the smaller the values of MSE, the better results one can achieve.

4. Experimental Results

In the experiments, we defined two parameters that were $\epsilon = 0.001$ and $C = 5000$. In this way, σ was varied from 0.1 to 1.0 by incrementing 0.1 at a time, as shown in Table I. As the results, σ value of 0.1 produced lowest MSE model which was 1.0588×10^{-10} . Figure 2 shows the comparison of results obtained from the value of PM₁₀ forecasting between support vector regression (predicted data) and actual values.

Table 1: Experimental Results of Change σ (Set C = 5,000, $\epsilon = 0.001$).

σ values	Performance (MSE.)
0.1	1.0588×10^{-10}
0.2	2.4473×10^{-9}
0.3	6.2587×10^{-9}
0.4	9.6685×10^{-9}
0.5	1.2671×10^{-8}
0.6	2.3299×10^{-8}
0.7	3.0815×10^{-8}
0.8	5.6682×10^{-8}
0.9	6.4833×10^{-8}
1.0	8.9096×10^{-8}

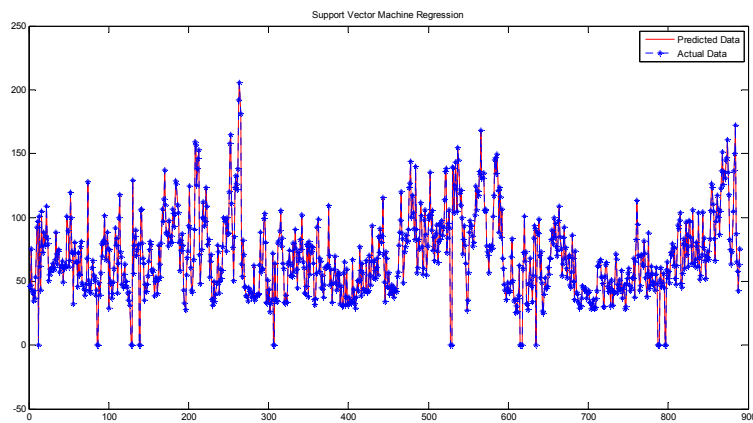


Fig. 2: Comparison of the predicted value from the actual values of SVR.

5. Conclusions

The potential of applying SVM model in ambient air pollutant prediction is studied and presented in this paper. The model was developed by using a technique from support vector regression (SVR). Gaussian Radial Basis Kernel functions was used because it is more suitable for the estimation of function Parameters to define the C (Regularization Parameter) with the ϵ (Error-insensitive) supported by vector machine. Parameters of Gaussian Radial Basis Kernel functions (σ) from the experiment find for the most appropriate using the values in the ascending range 0.1 through 1.0. In the experiment, the model selected has the lowest mean squared error (MSE). The results of this research show the model that set with value $C = 5,000$, $\epsilon = 0.001$ and $\sigma = 0.1$ work out most precise in forecasting over other tested models. Based on the test forecasting data, the mean squared error (MSE) was 1.0588×10^{-10} , which means this model is very satisfactory. The model reports that support vector regression can be used in forecasting PM_{10} successfully.

It can be concluded that SVM model provides a promising alternative and advantage in time series forecast. It offers several advantages. Firstly, it contains fewer free parameters than the other conventional neural network model. In this study, the spread parameter σ is the only factor to be considered in the SVM model once the kernel function is determined. Secondly, due to the adoption of the Structure Risk Minimization Principle, the SVM model provides better prediction than the conventional model. Finally, the SVM model can eliminate the typical drawbacks of conventional neural network models, e.g., over-fitting training and local minima, and proves to be more expandable and robust than the conventional neural network. The application of the SVM method in environmental aspect is a good, interesting attempt; it may be worthy to test its value in more areas.

6. References

- [1] Pollution Control Department (PCD). Development of Environmental and Emission Standards of Volatile Organic Compounds (VOCs) in Thailand. Bangkok: Pollution Control Department, 2007.
- [2] Pollution Control Department. Air Quality and Noise Management Bureau. Thailand State of Environment: The Decade of 2007. Bangkok: Pollution Control Department, 2007.
- [3] Pollution Control Department. Thailand State of Pollution Report 2008. Bangkok: Pollution Control Department, 2008.
- [4] Jiang, D., Zhang, Y., Zeng, Y., Tan, J., Shao D. Progress in developing an ANN model for air pollution index forecast. *Atmospheric Environment* 38: 7055-7064, 2004.
- [5] Ostro, B.D., Eskeland, G.S., Sanchez, J.M., Feyzioglu, T. Air pollution and health effects: a study of medical visits among children in Santiago, Chile. *Environmental Health Perspectives* 107: 69-73, 1999.
- [6] C. Cortes and V. Vapnik. Support-Vector Networks. *Machine Learning* 20, pp.273-297, 1995.
- [7] X.N. Dong, W. Qiang and L. Jin-Chao. Short Term Load Forecasting Model Using Support Vector Machine Based on Artificial Neural Network. *Proceedings of Fourth International Conference on Machine Learning and Cybernetics*: 4260-4265, Guangzhou, 2005.
- [8] A.J Smola and B. Scholkopf. A Tutorial on Support Vector Regression. *NeuroCOLT technical report, NC-TR-98-030*, 1998.