Development of Neural Network Technique for Prediction of PM$_{10}$ Concentration in the Industrial Area, at the East of Thailand

Jatupat Mekparyup and Kidadan Saithanu

Department of Mathematics
Faculty of Science, Burapha University
169 Muang, Chonburi, Thailand
jatupat@buu.ac.th, ksaithan@buu.ac.th

Abstract

The aim of this work is to describe the development of neural network models for prediction of PM$_{10}$ concentration in the industrial area, at the east of Thailand. Both of multilayer perceptron (MLP) and ordinary radial basis function (ORBF) neural network architectures, with air quality and meteorological variable data as the input also with different activation functions, have been applied. To improve the models, the root mean square error (RMSE) of the validation data set of each of model is calculated to compare all these model performances. The result of this study demonstrated that the neural network architecture, 12 input layer nodes, 5 hidden layer nodes and 1 target node, generally provided the good performance as seeing of the MLP 12-5-1 is the best performance with the smallest RMSE value really closed to the ORBF 12-5-1. Additionally, guidelines for appropriate neural network architecture are developed.

Mathematics Subject Classification: 62H20, 68T05, 92B20

Keywords: PM$_{10}$, Neural network architecture, Multilayer perceptron, ordinary radial basis function
1 Introduction

PM$_{10}$ defined as any particulate matter (PM) with median aerodynamic of diameter below 10 $\mu$m is one of the crucial air pollutants which negative effect on human health has been extensively concerned [1]. Many modeling studies have proposed that future changes in regional climate and anthropogenic emissions may influence surface PM$_{10}$ concentration through change in surface temperature, humidity, precipitation and ventilation [2], [3], [4]. The effect of each of these meteorological variables depends on the chemical composition and physical properties of PM$_{10}$ which vary extremely across different geographical regions.

Chonburi and Rayong, industrial areas of the eastern provinces of Thailand, have been many serious increasing air pollution problems including PM$_{10}$ pollution. According to the Air Quality Index (AQI) stipulated by the Thai Environmental Protection Department, the standard of 24 hours PM$_{10}$ level in the air is 120 $\mu$g/m$^3$ [5]. The annual concentration report of Chonburi [6] and the air quality report of Rayong [7] informed the PM$_{10}$ concentration was more frequently exceeded the standard level in many days. Due to a measurement of PM$_{10}$ concentration currently relies on an expensive tool of the air quality monitoring station, this work proposed the neural network model as an alternative equipment for prediction of PM$_{10}$ concentration with development of neural network technique. Neural network model is then generally an alternative method extensively applied because it is not based on any statistical assumptions.

Several papers utilized the multilayer perceptron network [8], [9], [10], [11] as well as the radial basis function network [12] to predict the PM$_{10}$ concentration regarding on the meteorological variables [13], [14] and the air quality variables [11], [12], [15].

2 Data Description

The observational data (2,265 cases) analyzed here is based on the daily PM$_{10}$ concentration and other various independent variables in the climate. The representative of the industrial areas were the two eastern monitoring stations of Thailand, the General Education Centre, Mueang District, Chonburi and the the Map Ta Phut Health Office, Mueang District, Rayong. Nine air quality variables (CO, NO, NO$_2$, NO$_x$, SO$_2$, HC, CH$_4$, NMHC and O$_3$) and seven meteorological variables (pressure, rain, relative humidity (RH), temperature (Temp), sun radiation (SR), wind direction (WD) and wind speed (WS)) were measured for the period 2006-2010. For analysis, two mutually exclusive and distinct data sets were firstly created to train and validate the two predictive neural network architectures, multilayer perceptron and ordinary radial basis function networks. The training data set consisted 1,542 cases (randomly chosen 70% of all data) was used for model training. The validation data set (the remaining data) was utilized to validate the suitability of a data model.
3 Methodology

3.1 Obtaining Influential Variables Associated to PM$_{10}$ Concentration
Correlation analysis was investigated to measure the degree of the linear relationship between PM$_{10}$ concentration and each of independent variables. If the P-value of Pearson Correlation coefficient is less than the chosen significance level, then there is a significant relationship between these two variables.

3.2 Development of Neural Network Technique
The neural network is typically represented using with network architecture as layers (input, hidden and output) of functional nodes or neurons. The input layer contains a node for each input variable. The one or more hidden layers contain hidden nodes. The output layer represents the target so the output node corresponding to the anticipated number of distinct classifications that are to be required by the network. The nodes are intermediate processing units that mathematically transform their inputs. The calculation happens at the hidden and the output layers but the input layer only passes the data to other layers. Several activation functions are used at the computing nodes, for example, arc tangent, hyperbolic tangent, logistic, Gaussian, exponential, softmax function, etc. Hence, there are many crucial factors to be regarded for modeling the neural network model such as the number of hidden layers, the number of hidden layer node and the type of activation function. The neural network model then encompasses various statistical models including a class of flexible nonlinear models.

As of this reason, the neural network model is often applied in many problem areas. Both of the multilayer perceptron (MLP) and the ordinary radial basis function (ORBF) network are regularly utilized in the particular prediction problem. The MLP which typically contained 3 layers is the most common and popular feed-forward type of network architecture because it’s simplicity. The advanced network architecture ORBF which is a feed-forward type with a single hidden layer is difficult to be trained for practitioners [16]. In contrast to the MLP, the hidden nodes in the ORBF are bell-shaped (Gaussian) surfaces centered at positions in the input space. Therefore, when the complicated ORBF network is trained, the activation function at the hidden nodes should be compatible to the activation function at the output nodes.

G. Cybenko ([17]) and K. Hornik, et al. ([18]) recommended the MLP with one hidden layer network is sufficient. However, [19] reported there is no rule of thumb for designing the optimum number of hidden layer node, the significant point is to consider and undertraining or overtraining problems. Thus, this study was designed to have one hidden layer with 3 and 5 nodes for both of the MLP and the ORBF networks.

3.2 Performance Criterion
To determine how well each model would predict PM$_{10}$ concentration in the validation data set, there can be many criterions to apply and measure the
performance of the model. The root mean square error (RMSE) is the most one frequently used. It is defined as Equation 1.

\[
RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}
\]

(1)

where \(y_i\) is the \(i\)th observation of PM\(_{10}\) concentration, \(\hat{y}_i\) is the \(i\)th predicted value of PM\(_{10}\) concentration from the model and \(n\) is the number of observations in the validation data set. The model with the smallest RMSE is the one with the best performance.

4 Results

4.1 Influential Variables Associated to PM\(_{10}\) Concentration
Since the large P-value of Pearson Correlation coefficient test between the PM\(_{10}\) concentration and each of four independent variables (NMHC, pressure, rain and WS) led to conclude that there were no linear association between the PM\(_{10}\) concentration and each of these four variables. Therefore, the PM\(_{10}\) concentration could be predicted with the twelve significant independent variables: CO, NO, NO\(_2\), NO\(_X\), SO\(_2\), HC, CH\(_4\), O\(_3\), RH, Temp, SR and WD.

4.2 Determination of Neural Network Models
Once both of the MLP and ORBF network architectures were designed according to the result of significant independent variables related to the PM\(_{10}\) concentration, the input layer then composed of 12 nodes. In addition, the output layer node would be contained only one node as a representative of the target, the PM\(_{10}\) concentration. Hence, both of the MLP and ORBF network architectures can be illustrated with 12 input nodes, 3 or 5 hidden nodes and only 1 output node (target node). The network architectures to be investigated for this study are the MLP 12–3–1, MLP 12–5–1, ORBF 12–3–1 and ORBF 12–5–1. The MLP models used the hyperbolic tangent and logistic activation functions at the hidden nodes and applied the exponential and identity activation functions at the target node. For the ORBF models, only the Gaussian activation function at the hidden node can be compatible with the exponential and identity activation functions at the target node.

4.3 Comparison of Neural Network Model Performance
To compare the performance of all neural network models, the RMSE of the validation data set of each model was computed and then shown in Table 1.
Development of neural network technique

<table>
<thead>
<tr>
<th>Neural Network</th>
<th>Activation Function at</th>
<th>RMSE of the Validation Data Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Architecture</td>
<td>Hidden Node</td>
<td>Target Node</td>
</tr>
<tr>
<td>MLP 12–3–1</td>
<td>Hyperbolic Tangent</td>
<td>Exponential</td>
</tr>
<tr>
<td>MLP 12–3–1</td>
<td>Logistic</td>
<td>Exponential</td>
</tr>
<tr>
<td>MLP 12–3–1</td>
<td>Hyperbolic</td>
<td>Identity</td>
</tr>
<tr>
<td>MLP 12–3–1</td>
<td>Logistic</td>
<td>Identity</td>
</tr>
<tr>
<td>MLP 12–5–1</td>
<td>Hyperbolic Tangent</td>
<td>Exponential</td>
</tr>
<tr>
<td>MLP 12–5–1</td>
<td>Logistic</td>
<td>Exponential</td>
</tr>
<tr>
<td>MLP 12–5–1</td>
<td>Hyperbolic</td>
<td>Identity</td>
</tr>
<tr>
<td>MLP 12–5–1</td>
<td>Logistic</td>
<td>Identity</td>
</tr>
<tr>
<td>ORBF 12–3–1</td>
<td>Gaussian</td>
<td>Exponential</td>
</tr>
<tr>
<td>ORBF 12–3–1</td>
<td>Gaussian</td>
<td>Identity</td>
</tr>
<tr>
<td>ORBF 12–5–1</td>
<td>Gaussian</td>
<td>Exponential</td>
</tr>
<tr>
<td>ORBF 12–5–1</td>
<td>Gaussian</td>
<td>Identity</td>
</tr>
</tbody>
</table>

Table 1: RMSE of the validation data set of each neural network model

5 Conclusion and Discussion

A conclusion from this study can be summarized as follow: (1) The twelve influential variables affected the PM$_{10}$ concentration in the industrial areas, at the east of Thailand through the P-value of Pearson Correlation coefficient test to be; CO, NO, NO$_2$, NO$_X$, SO$_2$, HC, CH$_4$, O$_3$, RH, Temp, SR and WD. All these variables are in agreement with the results gained by Saithanu and Mekparyup [15]. (2) The best performance model is the MLP 12-5-1 with logistic activation function at the hidden nodes and the exponential activation function at the target node. It presented that this simple neural network model performed better than the advanced neural network model, the ORBF 12-5-1 with Gaussian activation function at the hidden nodes and the identity activation function at the target node. (3) In general, the models with having network architecture of 5 hidden layer nodes provided the small value of the RMSE. (4) Guidelines for determining the best neural network model are depending on the number of hidden layers, the number of hidden layer node and the type of activation function.

Acknowledgements. Authors are appreciative to the Faculty of Science, Burapha University for supporting this research fund. The authors are also grateful to the Air Quality and Noise Management Bureau, Pollution Control Department for kindly providing all data.
REFERENCES


Received: July 1, 2013