Received December 2001
Revised November 2002
Accepted November 2002

Three improved neural network models for air quality forecasting

Wenjian Wang
Faculty of Science, Institute for Information and System Science, Xi'an Jiao Tong University, Xi'an, People's Republic of China
Department of Computer Science, Shanxi University, Taiyuan, People's Republic of China

Zongben Xu
Faculty of Science, Institute for Information and System Science, Xi'an Jiao Tong University, Xi'an, People's Republic of China

Jane Weizhen Lu
Department of Building and Construction, City University of Hong Kong, Hong Kong

Keywords *Modelling, Neural networks, Forecasting*

Abstract *Artificial neural networks (ANN) are appearing as alternatives to traditional statistical modeling techniques in many scientific disciplines. However, the inherent drawbacks of neural networks such as topology specification, undue training expense, local minima and training unpredictability will overlay their merits in engineering applications, especially. In this paper, adaptive radial basis function (ARBF) network and improved support vector machine (SVM) are presented in atmospheric sciences. The principle component analysis (PCA) technique is employed to the ARBF network as well, namely, ARBF/PCA network for the convenience of expression and comparison, so as to fasten the learning process. Comparing with traditional neural network models, the proposed models can automatically determine the size of network and parameters, fasten the learning process and achieve good generalization performances in prediction of pollutant level. The simulation results based on a real-world data set demonstrate the effectiveness and robustness of the proposed methods.*

Nomenclature

a_i	= calculated output	G	= data set
α_i, α_i^*	= Lagrange multipliers	H	= the maximum number of added hidden neurons
b	= constant coefficient		
C	= regularized constant	K	= Kernel function
e_i	= eigenvectors of covariance matrix	L_ε	= ε-insensitive loss function
ε	= insensitive loss coefficient	n	= number of input vectors

Engineering Computations
Vol. 20 No. 2, 2003
pp. 192-210

0264-4401
DOI 10.1108/02644400310465317

The authors wish to thank anonymous reviews for useful comments. This work was supported by the Hi-Tech R&D (863) Program (No. 2001AA113182) and the National Science Foundation (No. 69975016 and 60275019) of China. The provision of original data from the Environmental Protection Department, HKSAR, is also appreciated.

P = covariance matrix
R = regularized risk function
x_i = initial input vectors
x_i' = new input vectors
y_i = computed output
t_i = target output
λ_i = eigenvalues of covariance matrix
μ = center of Gaussian function
σ = spread of Gaussian function
w_i = connection weight
ϕ_i = feature function of inputs
ξ_i, ξ_i^* = slack variables

1. Introduction

With continuous economy development and population increase in metropolitan cities like Hong Kong, Shanghai, etc., a series of severe problems related to environmental pollution such as air pollution, noise pollution, shortage of land resources, waste and sewage disposal have attracted much attention than ever before. Among these, air pollution has direct impact on human health through exposure to pollutants at high concentration level existing in ambient areas. Air pollution control should prevent the situation from worsening in the long run. On the other hand, short-term forecasting of air quality is needed in order to take preventive and evasive action during episodes of airborne pollution. Therefore, as far as air quality forecasting, accurate models are extremely emphasized because such models would allow predicting potential compliance or noncompliance. At present, environmental modeling involves using a variety of approaches. The use of neural networks, in particular, the multilayer perceptron (MLP) neural networks, which can be trained to approximate virtually any smooth, measurable function, have become popular and produced promising results (Boznar *et al.*, 1993; Comrie, 1997; Fan *et al.*, 2000; Gardner and Dorling, 1996; Hadjiiski and Hopke, 2000; Lu *et al.*, 2001, 2002; Reich *et al.*, 1999; Roadknight *et al.*, 1997; Song and Hopke, 1996; Yi and Prybutok, 1996). A detailed review of application of neural networks in the atmospheric science is also available (Gardner and Dorling, 1998).

Above researches have shown that neural networks have salient advantages over traditional statistical methods in air quality forecasting, however, NN-based models still need to improve in order to achieve good prediction performance as effectively and efficiently as expected. In fact, a number of difficulties have been associated with NN use which hampered their effectiveness, efficiency as well as general acceptability in air quality analysis. These difficulties include susceptibility to chaotic behavior, computationally expensive training, local minima, over-fitting problem, training set problem and topology specification problem etc. For example, there are no general rules to determine the size of network and learning parameters, which will greatly affect the prediction performance. This highlights the need to optimize the complexity of the network topology in order to achieve the best "off-sample" performance. Therefore, in particular, for engineering application, e.g. time-series prediction, it is important to achieve an effective model so as to avoid the above-mentioned problems.

In this paper, three improved neural network models are presented. Adaptive radial basis function (ARBF) is a network, which can automatically determine the size of network and only one parameter needs to be estimated. The combination of ARBF and principle component analysis (PCA) (say, ARBF/PCA network) can further simplify the ARBF network, and therefore speed up learning process. Support vector machine (SVM), as a novel type of learning machine based on statistical learning theory, will greatly improve the generalization property of neural networks and automatically determine the size of networks. Here, a real-world data set is used to testify the effectiveness of the proposed methods. The study aims to investigate the prediction of respirable suspended particulate (RSP) concentration level based on six pollutants: RSP, sulphur dioxide (SO_2), nitrogen oxides (NO_X), nitric oxide (NO), nitrogen dioxide (NO_2), and carbon monoxide (CO), and five meteorological variables: wind speed (WS), wind direction (WD), outdoor temperature (OT), solar radiation (SR), and indoor temperature (IT). They are hourly measured during the whole year of 2000 at the Mong Kok Roadside Gaseous Monitory Station, one of the 14 monitory stations established by The Hong Kong Environmental Protection Department (HKEPD) across Hong Kong area.

This paper is organized as follows Section 2 and 3 provide introductions to the ARBF network and ARBF/PCA network, respectively. Section 4 gives an overview to the SVM method used in time series prediction. Section 5 describes the simulation results and comparisons among the three methods. The last section concludes the presented works.

2. The ARBF network

The MLP networks have the well-known ability to learn through training. The back-propagation (BP) algorithm is usually used to train MLP networks in atmospheric and environmental applications. However, for most MLP networks based on BP algorithm, some inherent drawbacks, e.g. slow convergence speed, "over-fitting" during training, local minima, network topology specification and poor generalizing performance, etc., would make it difficult to put these models into practice. Although, some improved approaches such as genetic algorithm (GA), simulated annealing (SA) and other optimization approaches can be employed, a high computational penalty may be needed. This will limit the application of MLP networks.

Radial basis function (RBF) network developed by Broomhead and Lowe (1988), as an effective algorithm for fast learning of MLP networks, possesses good generalization performance and, in the meantime, can avoid over-elaborated, lengthy computing like BP algorithm. Therefore, RBF network is much widely applied than BP-based networks. In RBF network, Gaussian function is often used as the nonlinear transfer function in the unique hidden layer

$$f(x) = e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad (1)$$

where μ and σ denote the center and spread parameters of Gaussian function, respectively.

As mentioned above, there are no general rules to select MLP network architecture – the number of hidden layers and nodes in these layers (in RBF network, only the number of the unique hidden layer needs to be estimated). The number of input and output nodes vary from case to case. Generally, the number of hidden nodes is decided by experiences. As a result, if the number is too small, the network may fail to converge to the minimum during training. Conversely, excessive nodes may result in over-fitting during training and lead to poor generalization performance as well.

To solve the problem of choosing the proper size network, a trial and error method is applied in most cases but with a heavy computation burden and low efficiency. There are mainly two simple adaptive methods in present studies: constructive algorithm and destructive algorithm. The former begins with a small network, e.g. there is no hidden node initially, and neurons and links are gradually added to the model until the network meets the given requirements. On the other hand, the latter starts with a large network, and then deletes neurons and links until one can obtain a suitable network. The synthesis method based on the above two approaches may be more suitable, but high computation cost makes it impossible to accomplish.

In this research, an adaptive constructing method of determining the number of hidden nodes is used. Given a training error, the ARBF network will automatically add nodes and links into the network when needed until the actual error is below the given one. Besides having the features of ordinary RBF networks, the outputs of the ARBF network are driven by the input nodes, whose advantages are visual (Wang, 2000; Wang *et al.*, 1993). The architecture of the ARBF network is shown in Figure 1.

Briefly, the main idea of constructive method proposed in this paper is described in the following.

(1) Adding hidden neurons: the number of added hidden neurons is taken from the range $[1, H]$, where H is the maximum number of added hidden nodes. The hidden neurons are arranged in sequence and the location of each new neuron will not affect subsequent connections. Therefore, if a hidden neuron is added to the network, it can be inserted into a sequence of hidden neurons at any location.

(2) Adding links: when a hidden neuron is added, the links between the added neuron and each input and output neuron will be added. There are no links between the added neuron and other neurons in the hidden layer.

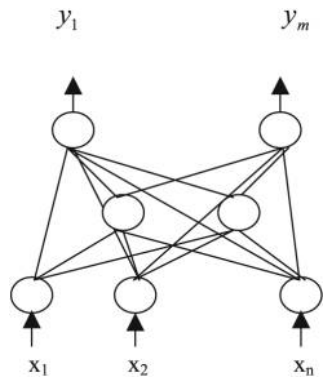

Figure 1.
The structure of three-layer ARBF network

The process of training network will continue until the training error is below the given error. When the training process stops, the size of network may be larger than that of the best network, but there are fewer free parameters which need to be estimated and the algorithm is easy to be implemented.

3. The combination of ARBF network and PCA technique

The ARBF network provides an effective method to automatically determine the network architecture. However, the size, which may be larger than that of the optimal network, will lead to learning slowly and therefore make its advantages be offset, especially in the case of high-dimension input space or in the case of data set having some noisy data. Therefore, in this study, PCA technique is used to pre-process data set so as to make ARBF network more effective and efficient. For convenience of comparison with the ARBF network, we call the method as the second model – ARBF/PCA network.

As we know, PCA is regarded as one of the main feature extraction methods. Here, PCA technique is applied to the data to observe how the pollutants and meteorological data correlate to each other and further find the possibility to reduce the variables for the predictions. Because the data is finite in this real-world database, the traditional matrix algebraic approach can be used to extract principal components through a direct computation of eigenvalues and eigenvectors of the input data covariance matrix.

There are totally 11 parameters supplied from the original database according to all valid data hourly measured during the year of 2000, which include the concentrations of six pollutants and five meteorological parameters. To use the PCA method, the covariance matrix of the studied data for all the 11 variables is calculated firstly. Let x_1, x_2, x_3, x_4, x_5 and x_6 represent pollutant concentrations of SO_2, NO_X, NO, NO_2, CO, and RSP, and x_7, x_8, x_9, x_{10}, x_{11} denote meteorological variables of WS, WD, OT, SR and IT, respectively. After the mean values and standard deviations of valid data for the 11 parameters are calculated, the covariance matrix P of the 11 parameters can be obtained as follows:

$$P=\begin{bmatrix} 1.0000 & 0.4518 & 0.4242 & 0.3510 & 0.2267 & 0.4102 & 0.0702 & -0.1075 & 0.1364 & 0.2998 & -0.0276 \\ 0.4518 & 1.0000 & 0.9722 & 0.6658 & 0.5764 & 0.5168 & -0.0266 & -0.1007 & -0.0967 & 0.0614 & 0.0472 \\ 0.4242 & 0.9722 & 1.0000 & 0.4726 & 0.5297 & 0.3828 & -0.0520 & -0.1019 & -0.0391 & 0.0530 & 0.0670 \\ 0.3510 & 0.6658 & 0.4726 & 1.0000 & 0.4856 & 0.7271 & 0.0624 & -0.0521 & -0.2473 & 0.0600 & -0.0387 \\ 0.2267 & 0.5764 & 0.5297 & 0.4856 & 1.0000 & 0.4375 & -0.1151 & 0.0309 & -0.2242 & -0.0736 & 0.0357 \\ 0.4102 & 0.5168 & 0.3828 & 0.7271 & 0.4375 & 1.0000 & 0.0510 & -0.0809 & -0.1750 & 0.0843 & -0.0910 \\ 0.0702 & -0.0266 & -0.0520 & 0.0624 & -0.1151 & 0.0510 & 1.0000 & -0.2450 & 0.2228 & 0.5183 & -0.0530 \\ -0.1075 & -0.1007 & -0.1019 & -0.0521 & 0.0309 & -0.0809 & -0.2450 & 1.0000 & -0.0676 & -0.0864 & 0.0652 \\ 0.1364 & -0.0967 & -0.0391 & -0.2473 & -0.2242 & -0.1750 & 0.2228 & -0.0676 & 1.0000 & 0.4186 & -0.2805 \\ 0.2998 & 0.0614 & 0.0530 & 0.0600 & -0.0736 & 0.0843 & 0.5183 & -0.0864 & 0.4168 & 1.0000 & -0.1539 \\ -0.0276 & 0.0472 & -0.0387 & -0.0387 & 0.0357 & -0.0910 & -0.0530 & 0.0652 & -0.2805 & -0.1539 & 1.0000 \end{bmatrix} \tag{2}$$

From the covariance matrix P, it can be seen that the elements in the top left squared 6×6 sub-matrix have close values. This indicates that the six pollutants have quite strong correlation with each other. Therefore, although only RSP pollutant is discussed here, similar changing tendency for other five pollutants can be induced.

The eigenvalues λ_i and the eigenvectors e_i, $i = 1, \ldots, 11$, of the covariance matrix P are then computed. The resultant eigenvalues are listed as

$$\begin{aligned} &\lambda_1 = 3.6490, \quad \lambda_2 = 2.0543, \quad \lambda_3 = 1.0677, \quad \lambda_4 = 1.0145, \\ &\lambda_5 = 0.9309, \quad \lambda_6 = 0.7173, \quad \lambda_7 = 0.4958, \quad \lambda_8 = 0.4531, \\ &\lambda_9 = 0.3700, \quad \lambda_{10} = 0.2475, \quad \lambda_{11} = 0. \end{aligned} \tag{3}$$

Here, the first six principal components are selected because the ratio of variance of these six components to the sum of variance is 85.76 percent i.e.

$$\left(\sum_{i=1}^{6} \lambda_i \Big/ \sum_{i=1}^{11} \lambda_i \right)$$

and information in data set may lose a little based on these six components. Then, six new variables, namely, x_i', $i = 1, \ldots, 6$, are defined instead of the original 11 variables from the computed eigenvectors listed as below:

$$\begin{bmatrix} x_1' \\ x_2' \\ x_3' \\ x_4' \\ x_5' \\ x_6' \end{bmatrix} = \begin{bmatrix} 0.0071 & 0.0538 & -0.0953 & -0.0631 & 0.0045 & 0.3913 & 0.3666 & 0.4253 & 0.4364 & 0.4823 & 0.3042 \\ -0.2643 & 0.5660 & 0.4817 & -0.2342 & 0.4833 & 0.0262 & -0.1709 & -0.0238 & -0.0096 & -0.0136 & 0.2421 \\ -0.5440 & -0.0186 & 0.4109 & 0.5182 & -0.4679 & -0.0513 & 0.0808 & -0.0946 & 0.0981 & 0.0576 & 0.1278 \\ 0.5898 & 0.0975 & 0.2568 & 0.0911 & -0.0839 & -0.4398 & -0.0172 & -0.3444 & 0.3924 & 0.2417 & 0.1914 \\ -0.3233 & -0.3458 & 0.1383 & -0.7557 & -0.2348 & -0.1731 & -0.0188 & -0.1727 & 0.2144 & 0.1359 & -0.0775 \\ 0.1751 & -0.0790 & -0.324 & -0.1512 & -0.3683 & 0.2486 & -0.3848 & 0.0152 & -0.2241 & -0.1862 & 0.7131 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \\ x_8 \\ x_9 \\ x_{10} \\ x_{11} \end{bmatrix} \quad (4)$$

In the second model, during training the ARBF network, these new variables x_i' $(i = 1, \ldots, 6)$ are taken as input vectors to predict the RSP concentration level. Such treatment of variables has two well-known advantages. The first advantage is that the number of input variables used in the prediction process can be reduced, so that the structure of predictor and the computational cost can be reduced as well. The second one is that the orthogonalization of input variables for neural networks can be obtained. Such orthogonalization can also make a neural network be easily trained in general.

Hence, theoretically, the ARBF/PCA network should have much simpler architecture, faster training speed than the ARBF network without losing satisfactory predicting performance. The following experiments verify the rule as well.

4. Overview of support vector machine

For any learning problem, generalization performance is one of the most important factors to evaluate the established model. Recently, SVMs developed by Vapnik (1995) have provided another novel approach to improve the generalization property of neural networks. Originally, SVMs were developed for pattern recognition problems. And now, with the introduction of ε-insensitive loss function, SVMs have been extended to solve nonlinear regression estimation and time series prediction problems (Mukherjee *et al.*, 1997; Müller *et al.*, 1997; Vapnik, 1995; Vapnik *et al.*, 1997). Unlike most of the traditional learning machines, which adopt the empirical risk minimization principle (e.g. MLP networks), SVMs implement the structural risk minimization principle, which seeks to minimize an upper bound of the generalization error rather than minimize the training error. This will result in better generalization performance than conventional techniques.

In support vector regression, the basic idea is to map the original data x into a higher dimensional feature space F via a nonlinear mapping ϕ and to do linear regression in this space. Therefore, regression approximation addresses

the problem of estimating a function based on a given set of data $G = \{(x_i, d_i)\}_{i=1}^{l}$ (x_i is input vector, d_i is the desired value), which is produced from the unknown function ϕ. SVMs approximate the function in the following form

$$y = \sum_{i=1}^{l} w_i \phi_i(x) + b, \tag{5}$$

where $\{\phi_i(x)\}_{i=1}^{l}$ are the features of inputs, and $\{w_i\}_{i=1}^{l}$, b are coefficients, respectively. They are estimated by minimizing the regularized risk function

$$R(C) = C\frac{1}{N}\sum_{i=1}^{N} L_\varepsilon(d_i, y_i) + \frac{1}{2}\|w\|^2, \tag{6}$$

where

$$L_\varepsilon(d, y) = \begin{cases} |d - y| - \varepsilon |d - y| \geq \varepsilon \\ 0 \quad \text{otherwise} \end{cases}, \tag{7}$$

and ε is a prescribed parameter.

In equation (6), $L_\varepsilon(d, y)$ is the so-called ε-insensitive loss function, which indicates that it does not penalize errors below ε. The second term, $1/2\|w\|^2$, is used as a measure of function flatness. C is a regularized constant determining the trade-off between the training error and model flatness. Introduction of the slack variables ξ, ξ^* leads equation (6) to the following constrained function.

Minimize:

$$R(w, \xi^*) = \frac{1}{2}\|w\|^2 + C^* \sum_{i=1}^{n} (\xi_i + \xi_i^*) \tag{8}$$

Subjected to:

$$\begin{aligned} & w\phi(x_i) + b - d_i \leq \varepsilon + \xi_i, \\ & d_i - w\phi(x_i) - b_i \leq \varepsilon + \xi_i, \\ & \xi,\ \xi^* \geq 0. \end{aligned} \tag{9}$$

Thus, equation (5) becomes the following explicit form

$$f(x, \alpha_i, \alpha_i^*) = \sum_{i=1}^{l} (\alpha_i - \alpha_i^*) K(x, x_i) + b. \tag{10}$$

In equation (10), Lagrange multipliers α_i, α_i^* satisfy the equality $\alpha_i \times \alpha_i^* = 0$, $\alpha_i \geq 0, \alpha_i^* \geq 0, i = 1, \ldots, l$, and they are obtained by maximizing the dual form of function (8)

$$\phi(\alpha_i, \alpha_i^*) = \sum_{i=1}^{l} d_i(\alpha_i - \alpha_i^*) - \varepsilon \sum_{i=1}^{l} (\alpha_i + \alpha_i^*) - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) K(\alpha_i, \alpha_j) \tag{11}$$

with constrains

$$\sum_{i=1}^{l} (\alpha_i - \alpha_i^*) = 0, \quad 0 \le \alpha_i \le C, \quad 0 \le \alpha_i^* \le C \quad i = 1, \ldots, l. \tag{12}$$

Based on the nature of quadratic programming, only a number of coefficients among α_i and α_i^* will be nonzero, and the data points associated with them could be referred to support vectors.

Because ϕ is usually unknown, a notion, namely, kernel function $K(x_i, x_j)$, is then introduced. The value of $K(x_i, x_j)$ is equal to the inner product of the features $\phi(x_i)$ and $\phi(x_j)$ of two vectors x_i and x_j (say, $K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j)$). Any function satisfying Mercer's theorem (Vapnik, 1995) can be used as the kernel function. A common choice is the Gaussian kernel $K(x_i, x_j) = \exp(-\|x_i - x_j\|^2/(2\sigma^2))$.

In support vector learning, there are three free parameters: σ, C and ε. Since there are no general rules to estimate these parameters, the generalization errors with respect to the spread parameter of Gaussian kernel function σ, C and ε are studied at first.

Figure 2 explains that the prediction error (mean of absolute error) is scarcely influenced by C. When C is enough small, the prediction error is large. As C increases, the error decreases rapidly and when $C \ge 0.5$, it keeps a very small value and hardly varies. In order to make the learning process stable, C should be set up a large value (e.g. $C = 100$).

Figure 3 shows the prediction error versus ε. ε is also found to have little impact on it. With the increment of ε, the prediction error will not change again when ε increases to 0.5. Generally, ε takes a small value, so we can let $\varepsilon = 0.001$.

In theory, σ will greatly affect the level of prediction accuracy. Too small σ (e.g. $\sigma \to 0$) or too large σ (e.g. $\sigma \to \infty$) will lead to very poor generalization results. Our experiments testify the law. From Figure 4, we can see at first the prediction error decreased from large to small values rapidly. After retaining a small value during a certain period, it increased speedily and finally tended to rise slowly when $\sigma \ge 30$. This change law illustrates that there must be a proper σ, which can lead prediction error to the minimum. It could be obtained directly from experiments that when σ is in the range [0.9, 1.5], the mean of absolute error vary a little within the range of [10.2, 10.9]. This means that there

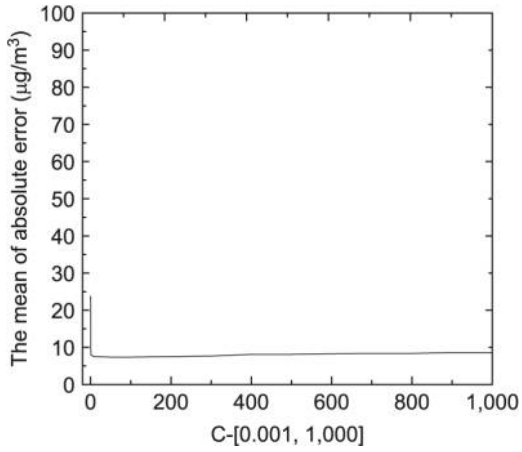

Figure 2.
Prediction error versus C

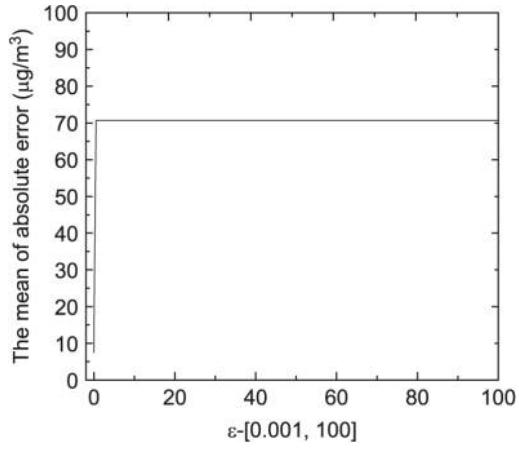

Figure 3.
Prediction error versus ε

exists a certain range of σ, within which the generalization performance is stable. In practical applications, we only need to select an appropriate σ within this range. Due to the changing law of prediction error with σ, gradient descent method can be used to evaluate a proper σ (in the experiment, $\sigma = 1.0$).

It needs to notice that the above conclusions are tenable on assuming that when one parameter is varied, the other two parameters are fixed. If they are all variable, the changing law of each parameter will become too complex to obtain any valuable results. Therefore, this paper aims to provide an effective and practical method to estimate the free parameters of SVM. For the third approach, it is more important that all the free parameters can be estimated on a certain guide and then very good generalization performance can be achieved.

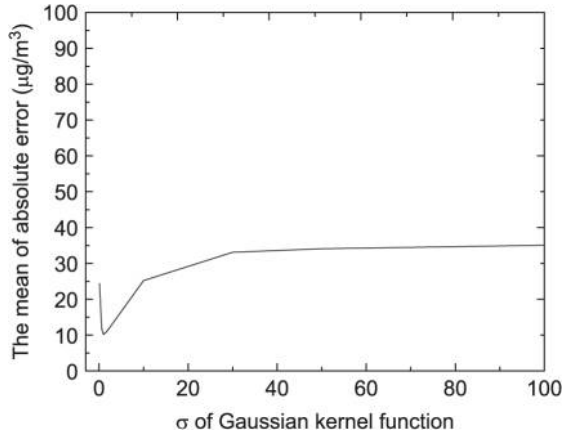

Figure 4.
Prediction error versus σ

Hence, it can avoid the problems that traditional neural networks have to encounter.

5. Simulation results and discussions

5.1 Data set and generalizing criterion

In our experiments, a real-world data set is used. All the air quality data obtained at the Mong Kok monitory station of Hong Kong are based on hourly continuous measurement during the whole year of 2000 (HKEPD, 2000). In this paper, we mainly investigate the short-term prediction of RSP, i.e. 72 h forecasting in advance, due to its practical significance. For the three models, there are 11 input variables, which include six pollutants and five meteorological factors, and RSP concentration is the output vector. After analyzing a total of 8,760 measured data points in the above time series, only 572 data points are invalid for some instrumental or measured reasons. So it is reasonably assumed that deleting these invalid data from the data set does not affect the physical features of the rest ones.

Here, the mean absolute error (MAE), root mean squared error (RMSE) and Willmott's index of agreement (WIA) are used as measurements of derivation between actual and predicted values. They are defined in the following, respectively.

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{abs}(a_i - t_i). \tag{13}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (a_i - t_i)^2}. \tag{14}$$

$$\mathrm{WIA} = \frac{\sum_{i=1}^{n} (a_i - t_i)^2}{\sum_{i=1}^{n} (|a_i^{'}| + |t_i^{'}|)^2}. \tag{15}$$

where,

$$a_i^{'} = a_i - \bar{a}, \quad t_i^{'} = t_i - \bar{a}, \quad \bar{a} = \frac{1}{n}\sum_{i=1}^{n} a_i, \tag{16}$$

and a_i, t_i denote the predicted result and measured one, respectively.

5.2 Simulation results and discussion

All programs are constructed using Matlab 6.0. The algorithms used in ARBF network and ARBF/PCA network are based on Lu *et al.* (2002) and Wang (2000). The improved SVM algorithm for solving time series prediction is implemented in the SVM method in the basis of Wang *et al.* (2002).

5.2.1 Simulation results on different meteorological conditions. In air quality parameter prediction, because the relationship between the pollutants and the meteorological factors is highly complex and extremely nonlinear, it is important to clarify how the prediction results will be affected by these climate variables. At first, the forecasting performances by the three proposed methods are discussed under different meteorological conditions. The data in January and June were selected. The reason to choose the data in January is because it is the month with the highest RSP concentration level due to dry and cold weather. The prevailing WD was NE, and the average relative humidity was about 70 percent during that period, therefore, the other meteorological conditions, e.g. humidity, cloudiness, pressure and level of precipitate etc., can be ignored. In June, all above meteorological factors are active obviously and will greatly affect the pollutant concentration levels. Because in data set, not all needed meteorological variables but only some of them, i.e. WS, WD, OT, SR, IT, have been measured, the experiments will inspect the effectiveness and robustness of the three approaches.

In each selected month, hourly data of the first 10 days (240 data points) are used as training data, and the data of the afterward 3 days (72 data points), representing the hourly data of 11th through 13th day, are used as testing data. Figure 5 gives the 72 h prediction results of RSP concentration in January and June.

From the experiments, it can be seen that, in January, RSP concentration has high level, while, in June, it is obviously reduced. Experimentally, the three approaches produce good performances in the selected 2 months in the case of 72 h prediction in advance. But the forecasting results in January are better than the ones in June because in the latter, other unmeasured meteorology factors such as humidity, cloudiness, pressure and level of precipitate will

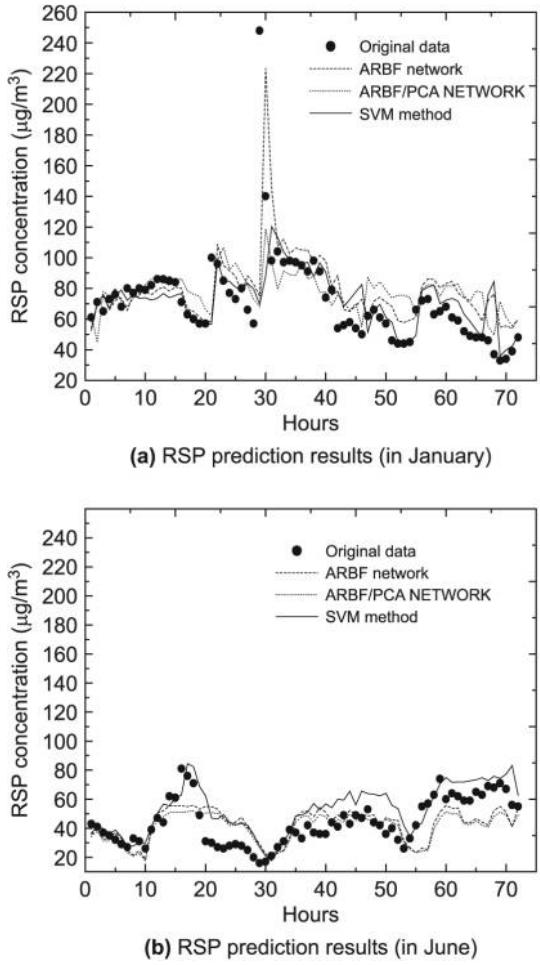

Figure 5.
Comparison of RSP prediction results among the three methods

greatly affect the prediction accuracy of RSP concentration level. Among the three approaches, the SVM method provides the best generalization performances either in January or in June, and the ARBF network gives better prediction results. The ARBF/PCA network is inferior to the other two methods. In overview, all the three approaches express good properties and robustness in different climate conditions.

5.2.2 Simulation results on the training and testing data set. This experiment mainly discusses how the three models fit the training data and how accurate the predictions are on the testing data by the three models. The training data originate from hourly data of the first 10 days in June (240 data points), while the afterward 1,000 data points are used as testing data. Figure 6 shows the prediction results on the training and testing data set.

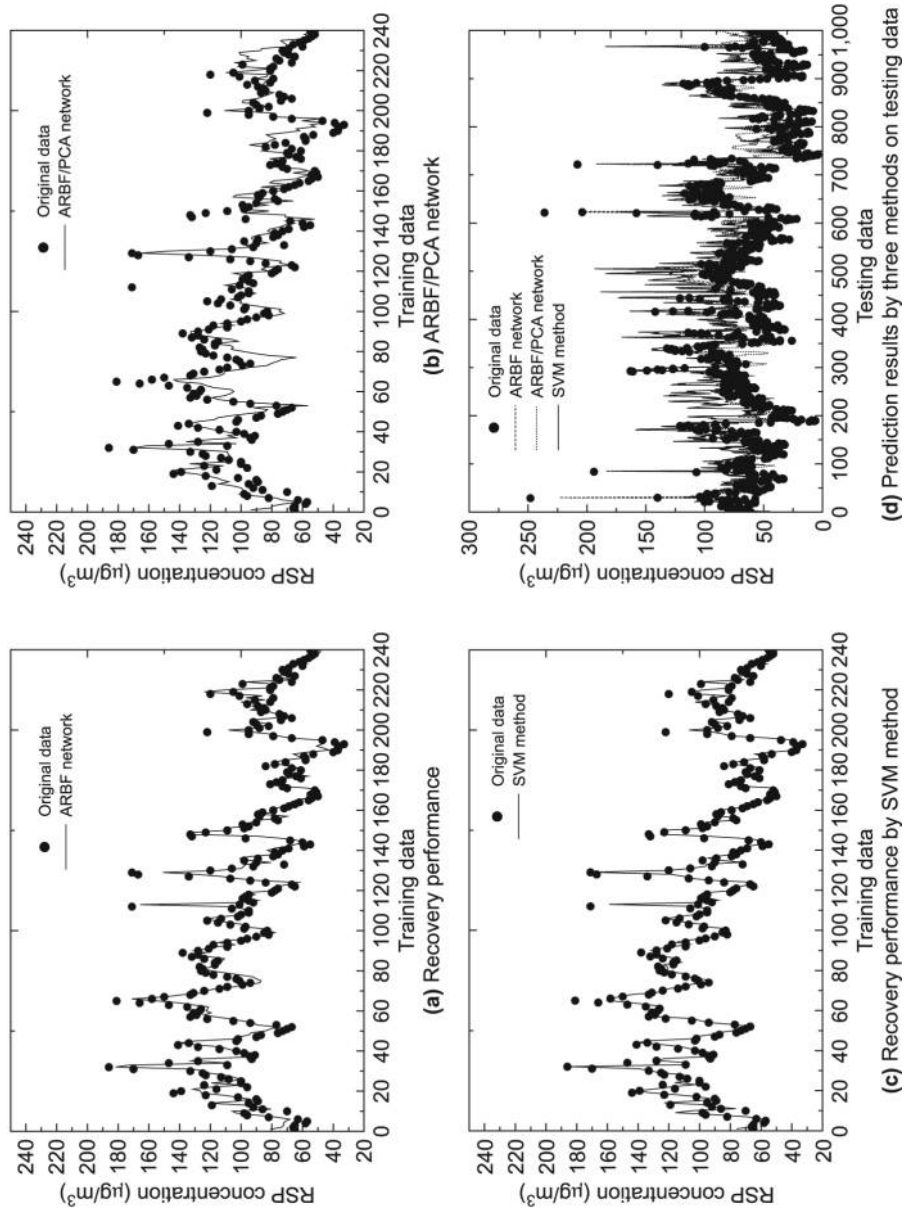

Figure 6.
Comparisons of prediction results among three methods on training and testing data set

For the training data, the differences among the three methods are not evident. It can be directly observed that the recovered values by the SVM method are almost identical to the original data (Figure 6(c)) and the computed results by the ARBF network are also very closer to the actual data (Figure 6(a)). The calculated results by the ARBF/PCA network have some little deviations (Figure 6(b)). Therefore, all the three models can fit the training data well. For the testing data, the three approaches express obvious different performances. Similar to the above experiments, the SVM method still expresses the best prediction property and then the better is the ARBF network. The ARBF/PCA network shows less accuracy than above two networks (Figure 6(d)). Therefore, it can be concluded that SVM method has shown more advantages in generalization performance over traditional neural networks and further more it can achieve good results in some special data points, which have unexpected outputs due to possible influence of meteorological factors (Figures 5(a), (b), 6(d)).

5.2.3 Properties comparison among the three approaches. For evaluating the three approaches, 22 group experiments have been carried out based on the valid data. In each group, there are 312 data points, which are divided into training data set (first 240 data points) and testing data set (rest 72 data points). Table I shows the comparisons among them in generalization performance, the size of network and learning speed.

From Table I, it can be seen that, either for MAE and RMSE or for WIA, SVM method is the best. The reason is that strong theoretical background provides SVM with high generalization capability and avoidance of local minima rather than general neural networks (Vapnik, 1995). While for the simpler network architecture and faster learning speed, the ARBF/PCA network is outstanding, whose average training speed is about 17 and 38 times faster than the ARBF network and SVM method, respectively. The ARBF network is medium between SVM method and the ARBF/PCA network in above aspects.

Figure 7 shows the comparisons of MAE in series experiments among the three approaches. It testified again that SVM method possesses absolute advantages over traditional neural networks in prediction performance. It is important to remember that the usual purpose of training neural networks is to

Table I.
The comparisons among the three approaches

Average	ARBF network	ARBF/PCA network	SVM method
MAE ($\mu g/m^3$)	12.4843	16.4142	6.2840
RMSE ($\mu g/m^3$)	21.4415	26.9766	13.1080
WIA	0.4901	0.4227	0.6843
The number of hidden nodes	46.4	2.7	–
The number of support vectors	–	–	85.5
Learning speed (s)	46.41	2.73	105.37

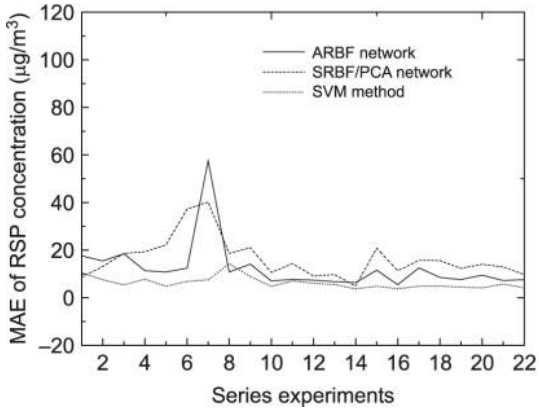

Figure 7.
Comparison of generalization performance among the three methods

achieve good generalization on unseen data (testing data) in prediction applications. It is identical to the principle of SVM and this provides an explanation to SVMs brilliant properties. For traditional neural networks, maximum generalization performance will often achieve before the overall network training error reaches a minimum. Especially, on a noisy set of data, which is a common case in practical applications, the network may be over-trained when global minimum is reached. In our experiments, although over-fitting training does not occur, both the ARBF network and the ARBF/PCA network express inferiority to SVM method in predicting accuracy.

Figure 8 gives the comparisons in learning speed of the three approaches. It is obvious that the training by the ARBF/PCA network is the fastest, whose maximum and minimum CPU-times are 1.5 and 1 s, respectively. The faster training is done by the ARBF network, whose maximum and minimum CPU-times are 43.4 and 1.8 s, respectively. SVM method is the slowest with maximum and minimum CPU-times as 129.2 and 91.3 s, respectively.

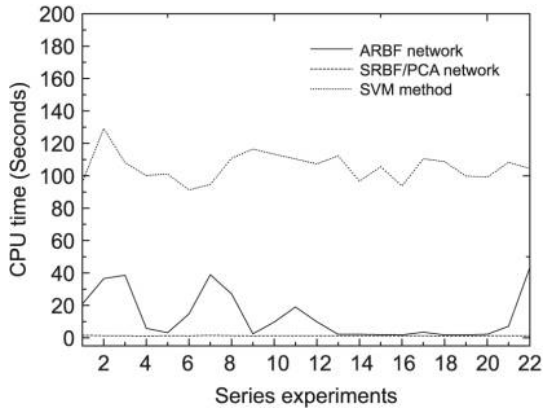

Figure 8.
Comparison of learning speed among the three methods

Figure 9 compares the number of hidden nodes required in the ARBF network and the ARBF/PCA network. The maximum hidden nodes are 129 for the ARBF network and four for the ARBF/PCA network, while the minimum ones are four for the former and one for the latter. The average number of hidden nodes in the latter is almost 32 times less than that in the former model. This demonstrates that before the network is trained, reducing input variables can simplify the network architecture and lessen learning time by losing slight predicting accuracy.

Figure 10 explains that in SVM method, the number of support vectors has little change, which varies in the range [78, 93] for 240 training data points. This means that only about 35.6 percent training data (support vectors) are contributed to the prediction performance. SVM method can automatically find these important data and then determine the network architecture using them. Comparing with standard SVM learning algorithm, the proposed SVM method can adaptively estimate proper free parameters so as to achieve the best prediction results. The experiments have also proven that the presented SVM method is stable.

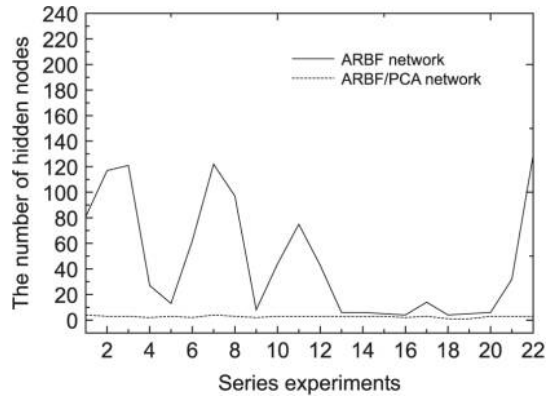

Figure 9.
Comparison of network architecture between ARBF and ARBF/PCA network

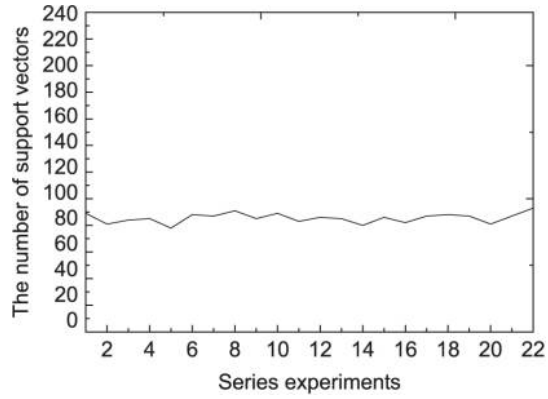

Figure 10.
The change of the number of support vectors versus experiments

6. Conclusion

Because neural networks do not require any prior assumptions regarding the distribution of training data and no decision regarding the relative importance of the various input measurements needs to be made, they can model highly nonlinear functions and can be trained to accurately generalize when presented with new, unseen data. These features of the neural networks make them attractive alternative to developing numerical models and also when choosing between statistical approaches. Therefore, they have many well-known successful applications in the engineering problems.

However, some inherent problems, which neural networks need to encounter, will limit their further applications. These problems include: determination of the network architecture, estimation of suitable learning parameters, selection of training data, local minima, slow learning speed, curse of dimensionality and processing of emergent patterns, etc. In atmospheric sciences, not only successfully modeling the average behavior of a system is needed, but also resolving infrequent, extreme events is important.

In this paper, three improved neural networks are presented. Each one can solve some problems of neural networks. The ARBF network can automatically determine the network architecture and have fast learning speed. The ARBF/PCA network is an improvement of the ARBF network. Beside possesses the advantages of the ARBF network, it can simplify the network architecture and fasten training speed without compromising generalization performance. Despite slow learning speed, the third method, SVM, can provide the best prediction performances, determine the network topology and estimate all the parameters automatically. Moreover, it can process emergent patterns. The simulation results have proven the effectiveness and benefits of the presented approaches. And further studies on providing more effective and practical models will be expected in the near future.

References

Boznar, M., Lesjak, M. and Mlakar, P. (1993), "A neural network-based method for short-term predictions of ambient SO_2 concentrations in highly polluted industrial areas of complex terrain", *Atmospheric Environment (B)*, Vol. 27 No. 2, pp. 221-30.

Broomhead, D. and Lowe, D. (1988), "Multivariable functional interpolation and adaptive networks", *Complex Systems*, Vol. 2, pp. 321-55.

Comrie, A.C. (1997), "Comparing neural networks and regression models for ozone forecasting", *Journal of Air and Waste Management*, Vol. 47, pp. 653-63.

Fan, H.Y., Lu, W.Z. and Xu, Z.B. (2000), "An empirical comparison of three novel genetic algorithms", *Engineering Computations*, Vol. 17 No. 8, pp. 981-1001.

Gardner, M.W. and Dorling, S.R. (1996), "Neural network modelling of the influence of local meteorology on surface ozone concentrations", *Proceedings of the 1st International Conference on GeoComputation*, University of Leeds, pp. 359-70.

Gardner, M.W. and Dorling, S.R. (1998), "Artificial neural networks (the multi-layer feed-forward neural networks) – a review of applications in the atmospheric science", *Atmospheric Environment*, Vol. 30 Nos 14/15, pp. 2627-36.

Hadjiiski, L. and Hopke, P.K. (2000), "Application of artificial neural network to modeling and prediction of ambient ozone concentrations", *Journal of Air and Waste Management Association*, Vol. 50, pp. 894-901.

Hong Kong Environmental Protection Department (2000), Environment Hong Kong.

Lu, W.Z., Fan, H.Y., Lo, S.M. and Wong, J.C.K. (2001), "A particle-swarm-optimization-based neural network approach and its application to environmental modeling", *IAQVEC'2001*, October, Changsha, People's Republic of China.

Lu, W.Z., Wang, W.J. and Xu, Z.B. *et al.* (2002), "Prediction of pollutant levels in Causeway Bay area in Hong Kong using an improved neural network model", *ASCE Journal of Environmental Engineering*, Vol. 128 No. 12, pp. 1146-56.

Mukherjee, S., Osuna, E. and Girosi, F. (1997), "Nonlinear prediction of chaotic time series using a supoort vector machine", *Proc. NNSP*, pp. 511-20.

Müller, K-R., Smola, A.J., Rätsch, G., Schölkopf, B., Kohlmorgen, J. and Vapnik, V. (1997), "Predicting time series with support vector machines", *Proc. ICANN*, pp. 999-1004.

Reich, S.L., Gomez, D.R. and Dawidowski, L.E. (1999), "Artificial neural network for the identification of unknown air pollution sources", *Atmospheric Environment*, Vol. 33, pp. 3045-52.

Roadknight, C.M., Balls, G.R., Mills, G.E. and Palmer_Brown, D. (1997), "Modeling complex environmental data", *IEEE Transactions on Neural Networks*, Vol. 8 No. 4, pp. 852-61.

Song, X.H. and Hopke, P.K. (1996), "Solving the chemical mass balance problem using an artificial neural network", *Environmental Science and Technology*, Vol. 30 No. 2, pp. 531-5.

Vapnik, V. (1995), *The Nature of Statistical Learning Theory*, Wiley, New York, USA.

Vapnik, V., Golowich, S. and Smola, A. (1997), "Support vector machine for function approximation, regression estimation, and signal processing", in Mozer, M., Jordan, M. and Petsche, T. (Eds), *Neural Information Processing Systems*, MIT Press, Cambridge, MA, Vol. 9.

Wang, W.J. (2000), "A high efficiency algorithm of input driven BP networks", *Journal of Systems Science and Systems Engineering*, Vol. 20, pp. 99-101 (in Chinese).

Wang, W.J., Tang, X.C. and Li, W.C. (1993), "A variable structure neural network model and its applications", *IEEE TENCON*, Vol. 20 No. 5, pp. 799-802.

Wang, W.J., Xu, Z.B., Lu, W.Z. and Zhang, X.Y. (2002), "Determination of the spread parameter in the Gaussian kernel for classification and regression", *Neurocomputing* (in press).

Yi, J. and Prybutok, R. (1996), "A neural network model forecasting for prediction of daily maximum ozone concentration in an industrialized urban area", *Environmental Pollution*, Vol. 92 No. 3, pp. 349-57.

This article has been cited by:

1. Carlos Hernán Fajardo-Toro, Josefa Mula, Raul Poler. Adaptive and Hybrid Forecasting Models—A Review 315-322. [Crossref]
2. Elham Ahmadi, Milad Jasemi, Leslie Monplaisir, Mohammad Amin Nabavi, Armin Mahmoodi, Pegah Amini Jam. 2018. New efficient hybrid candlestick technical analysis model for stock market timing on the basis of the Support Vector Machine and Heuristic Algorithms of Imperialist Competition and Genetic. *Expert Systems with Applications* **94**, 21-31. [Crossref]
3. Emad Norouzi, Hesam Moslemzadeh, Soheil Mohammadi. 2018. Maximum entropy based finite element analysis of porous media. *Frontiers of Structural and Civil Engineering* . [Crossref]
4. Huiping Peng, Aranildo R. Lima, Andrew Teakles, Jian Jin, Alex J. Cannon, William W. Hsieh. 2017. Evaluating hourly air quality forecasting in Canada with nonlinear updatable machine learning methods. *Air Quality, Atmosphere & Health* **10**:2, 195-211. [Crossref]
5. Laura E. Boucheron, Amani Al-Ghraibah, R. T. James McAteer. 2015. PREDICTION OF SOLAR FLARE SIZE AND TIME-TO-FLARE USING SUPPORT VECTOR MACHINE REGRESSION. *The Astrophysical Journal* **812**:1, 51. [Crossref]
6. Hyojoo Son, Changmin Kim, Changwan Kim, Youngcheol Kang. 2015. Prediction of government-owned building energy consumption based on an RReliefF and support vector machine model. *Journal of Civil Engineering and Management* **21**:6, 748-760. [Crossref]
7. Dhirendra Mishra, Pramila Goyal. 2015. Development of artificial intelligence based NO 2 forecasting models at Taj Mahal, Agra. *Atmospheric Pollution Research* **6**:1, 99-106. [Crossref]
8. Tzu-Yi Pai, Moo-Been Chang, Shyh-Wei Chen. Application of Computational Intelligence on Analysis of Air Quality Monitoring Big Data 427-441. [Crossref]
9. Qinliang Tan, Yongmei Wei, Minnan Wang, Yuan Liu. 2014. A cluster multivariate statistical method for environmental quality management. *Engineering Applications of Artificial Intelligence* **32**, 1-9. [Crossref]
10. Xiao Fang Du, Stephen C.H. Leung, Jin Long Zhang, K.K. Lai. 2013. Demand forecasting of perishable farm products using support vector machine. *International Journal of Systems Science* **44**:3, 556-567. [Crossref]
11. Ilenia Fronza, Alberto Sillitti, Giancarlo Succi, Mikko Terho, Jelena Vlasenko. 2013. Failure prediction based on log files using Random Indexing and Support Vector Machines. *Journal of Systems and Software* **86**:1, 2-11. [Crossref]
12. Azman Azid, Hafizan Juahir, Mohd Talib Latif, Sharifuddin Mohd Zain, Mohamad Romizan Osman. 2013. Feed-Forward Artificial Neural Network Model for Air Pollutant Index Prediction in the Southern Region of Peninsular Malaysia. *Journal of Environmental Protection* **04**:12, 1-10. [Crossref]
13. A. B. M. Shawkat Ali, Salahuddin Azad. Demand Forecasting in Smart Grid 135-150. [Crossref]
14. Hyojoo Son, Changmin Kim, Changwan Kim. 2012. Hybrid principal component analysis and support vector machine model for predicting the cost performance of commercial building projects using pre-project planning variables. *Automation in Construction* **27**, 60-66. [Crossref]
15. Wei-Chiang Hong. 2012. Application of seasonal SVR with chaotic immune algorithm in traffic flow forecasting. *Neural Computing and Applications* **21**:3, 583-593. [Crossref]
16. Wei-Chiang Hong. 2011. Traffic flow forecasting by seasonal SVR with chaotic simulated annealing algorithm. *Neurocomputing* **74**:12-13, 2096-2107. [Crossref]

17. Wei-Chiang Hong, Yucheng Dong, Chien-Yuan Lai, Li-Yueh Chen, Shih-Yung Wei. 2011. SVR with Hybrid Chaotic Immune Algorithm for Seasonal Load Demand Forecasting. *Energies* **4**:6, 960-977. [Crossref]
18. Wei-Chiang Hong, Yucheng Dong, Feifeng Zheng, Shih Yung Wei. 2011. Hybrid evolutionary algorithms in a SVR traffic flow forecasting model. *Applied Mathematics and Computation* **217**:15, 6733-6747. [Crossref]
19. Wei-Chiang Hong, Yucheng Dong, Feifeng Zheng, Chien-Yuan Lai. 2011. Forecasting urban traffic flow by SVR with continuous ACO. *Applied Mathematical Modelling* **35**:3, 1282-1291. [Crossref]
20. Wei-Chiang Hong, Yucheng Dong, Li-Yueh Chen, Shih-Yung Wei. 2011. SVR with hybrid chaotic genetic algorithms for tourism demand forecasting. *Applied Soft Computing* **11**:2, 1881-1890. [Crossref]
21. Shuang Cang, Nigel Hemmington. 2010. Forecasting U.K. Inbound Expenditure by Different Purposes of Visit. *Journal of Hospitality & Tourism Research* **34**:3, 294-309. [Crossref]
22. Shourong Li, Rongxi Zhou, Xin Ma. The forecast of C0<inf>2</inf> emissions in China based on RBF neural networks 319-322. [Crossref]
23. Wei-Chiang Hong, Yucheng Dong, Li-Yueh Chen, Chien-Yuan Lai. 2010. Taiwanese 3G mobile phone demand forecasting by SVR with hybrid evolutionary algorithms. *Expert Systems with Applications* **37**:6, 4452-4462. [Crossref]
24. JIANSHENG WU, MINGZHE LIU, LONG JIN. 2010. A HYBRID SUPPORT VECTOR REGRESSION APPROACH FOR RAINFALL FORECASTING USING PARTICLE SWARM OPTIMIZATION AND PROJECTION PURSUIT TECHNOLOGY. *International Journal of Computational Intelligence and Applications* **09**:02, 87-104. [Crossref]
25. Xiao-xuan Qi, Jian-wei Ji, Xiao-wei Han, Zhong-hu Yuan. An Approach of Passive Vehicle Type Recognition by Acoustic Signal Based on SVM 545-548. [Crossref]
26. Wei-Mou Hung, Wei-Chiang Hong, Tung-Bo Chen. Application of hybrid genetic algorithm and simulated annealing in a SVR traffic flow forecasting model 728-735. [Crossref]
27. Nicholas Sapankevych, Ravi Sankar. 2009. Time Series Prediction Using Support Vector Machines: A Survey. *IEEE Computational Intelligence Magazine* **4**:2, 24-38. [Crossref]
28. David Shooter, Peter Brimblecombe. 2009. Air quality indexing. *International Journal of Environment and Pollution* **36**:1/2/3, 305. [Crossref]
29. Ligang Zhou, Kin Keung Lai. Adaboosting Neural Networks for Credit Scoring 875-884. [Crossref]
30. Jiansheng Wu, Enhong Chen. A Novel Nonparametric Regression Ensemble for Rainfall Forecasting Using Particle Swarm Optimization Technique Coupled with Artificial Neural Network 49-58. [Crossref]
31. Wei-Chiang Hong. 2008. Rainfall forecasting by technological machine learning models. *Applied Mathematics and Computation* **200**:1, 41-57. [Crossref]
32. Wenjian Wang, Changqian Men, Weizhen Lu. 2008. Online prediction model based on support vector machine. *Neurocomputing* **71**:4-6, 550-558. [Crossref]
33. Dursun Delen, Ramesh Sharda. Artificial Neural Networks in Decision Support Systems 557-580. [Crossref]
34. Wei-Chiang Hong, Young-Jou Lai, Ping-Feng Pai, Shao-Lun Lee, Shun-Lin Yang. Composite of support vector regression and evolutionary algorithms in car-rental revenue forecasting 2872-2878. [Crossref]

35. A. Azadeh, S.F. Ghaderi, V. Ebrahimipour. 2007. An integrated PCA DEA framework for assessment and ranking of manufacturing systems based on equipment performance. *Engineering Computations* **24**:4, 347-372. [Abstract] [Full Text] [PDF]
36. Ping-Feng Pai, Wei-Chiang Hong. 2007. A recurrent support vector regression model in rainfall forecasting. *Hydrological Processes* **21**:6, 819-827. [Crossref]
37. A. Azadeh, S.F. Ghaderi, Y. Partovi Miran, V. Ebrahimipour, K. Suzuki. 2007. An integrated framework for continuous assessment and improvement of manufacturing systems. *Applied Mathematics and Computation* **186**:2, 1216-1233. [Crossref]
38. Wei-Chiang Hong, Ping-Feng Pai. 2007. Potential assessment of the support vector regression technique in rainfall forecasting. *Water Resources Management* **21**:2, 495-513. [Crossref]
39. V. Ebrahimipour, K. Suzuki, A. Azadeh. 2006. An integrated off-on line approach for increasing stability and effectiveness of automated controlled systems based on pump dependability—case study: Offshore industry. *Journal of Loss Prevention in the Process Industries* **19**:6, 542-552. [Crossref]
40. Ping-Feng Pai, Wei-Chiang Hong. 2006. Software reliability forecasting by support vector machines with simulated annealing algorithms. *Journal of Systems and Software* **79**:6, 747-755. [Crossref]
41. Ping-Feng Pai. 2006. System reliability forecasting by support vector machines with genetic algorithms. *Mathematical and Computer Modelling* **43**:3-4, 262-274. [Crossref]
42. Wei-Chiang Hong, Ping-Feng Pai. 2006. Predicting engine reliability by support vector machines. *The International Journal of Advanced Manufacturing Technology* **28**:1-2, 154-161. [Crossref]
43. V. Ebrahimipour, K. Suzuki. 2006. A synergetic approach for assessing and improving equipment performance in offshore industry based on dependability. *Reliability Engineering & System Safety* **91**:1, 10-19. [Crossref]
44. Wei-Chiang Hong, Ping-Feng Pai, Shun-Lin Yang, R. Theng. Highway traffic forecasting by support vector regression model with tabu search algorithms 1617-1621. [Crossref]
45. Ping-Feng Pai, Wei-Chiang Hong. 2005. An Improved Neural Network Model in Forecasting Arrivals. *Annals of Tourism Research* **32**:4, 1138-1141. [Crossref]
46. WENJIAN WANG. 2005. An Incremental Learning Strategy for Support Vector Regression. *Neural Processing Letters* **21**:3, 175-188. [Crossref]
47. Vahid Ebrahimipour, Kazuhiko Suzuki. 2005. An Integrated Model-Based Approach for Evaluating and Improving Safety Instruments in Continuous Process from the Viewpoints of Maintainability, Reliability and Availability—Case Study: Offshore Industry. *JOURNAL OF CHEMICAL ENGINEERING OF JAPAN* **38**:6, 436-445. [Crossref]
48. Wei-Chiang Hong, Ping-Feng Pai, Chen-Tung Chen, Ping-Teng Chang. Recurrent Support Vector Machines in Reliability Prediction 619-629. [Crossref]
49. Ping-Feng Pai, Wei-Chiang Hong, Chih-Sheng Lin. Forecasting Tourism Demand Using a Multifactor Support Vector Machine Model 512-519. [Crossref]