FUZZY DECISION SUPPORT SYSTEM TO THE PREDICTION OF OZONE CONCENTRATIONS

Sungshin Kim, Jaeyong Kim, Chong-Bum Lee[1], and Min-Young Kim[2]

School of Electrical and Computer Engineering, Pusan National University
[1]Dept. of Environmental Science, Kangwon National University
[2]National Institute of Environmental Research

ABSTRACT

Artificial neural networks are used to model the interactions that occur between ozone concentration and environment data. A number of generic methods for analysis and modeling are investigated. Because the mechanism of the ozone concentration is highly complex, nonlinear, and nonstationary, the modeling methods of ozone prediction system have many problems and the results of prediction are not a good performance so far, especially in the high-level ozone concentration.

This paper introduces a modeling method of the ozone prediction system using neuro-fuzzy approaches and fuzzy clustering method. The dynamic polynomial neural network (DPNN) based upon a typical algorithm of GMDH (group method of data handling) is employed for data analysis, identification of nonlinear complex system, and prediction of the ozone concentration. The proposed prediction system is applied to the 19 areas in Seoul, Korea, and the final results are discussed in the senses of root mean-square error (RMSE) and R-square.

1. INTRODUCTION

Recently, one of the major issues of air pollution is the high ozone concentration of the troposphere in summer. Generally, the mechanism of ozone concentration is formed by the photochemical reaction and meteorological variations. In this mechanism, nitric dioxide and hydrocarbon act as pollution materials, and interaction of solar radiation, wind speed, and temperature play the role of meteorological materials. About 10% of ozone exists in the troposphere of the earth. High concentration ozone appears from June to August in summer. Nowadays, developing industries and increasing automobile in our country increase air pollution materials. Therefore, researches of ozone concentration are executed more broadly for air pollution forecasting systems. The variation of high concentration ozone caused by photochemical and specific meteorological characteristics are simulated and predicted daily. There are several conventional methods to predict ozone concentration. For instances, multiple regression model [1] by static methods, a multivariate analyses and artificial neural networks [2] have been developed and applied to predict ozone concentration. Performances of these methods are not sufficient to predict high-level ozone concentration. The causes of these problems are due to the lack of important data. Especially, non-linearity between meteorological parameters and pollution materials of ozone, and complicated dynamics of ozone generation in the troposphere make the difficulty to model the ozone prediction systems. Thus, in this paper, fuzzy clustering and the DPNN are proposed to predict high-level ozone concentration. Fuzzy clustering method as a pre-processing is employed to decide the fuzzy sets of the low and high ozone concentrations. After clustering the two fuzzy sets, the different two models are determined by the DPNN. Finally, the decision making process as a post-processing is applied to forecast the ozone concentration by the compensation of the weight factors to the outputs of the two models.

The proposed forecasting system is adaptively constructed by a successive basic structure of the DPNN. Also important input variables for the final structure of the forecasting system are selected form the possible input variables by a selection criteria. The historical data that consist of pollution material and meteorological information are divided into training data and testing data to identify dynamic system and to prevent overfitting.

The structure of the final model is compact and the computational speed to produce an output is faster than other modeling methods. The proposed method shows that the prediction of the ozone concentration based upon the DPNN gives us a good performance with ability of superior data approximation and self-organization.

2. FUZZY CLUSTERING

Fuzzy clustering is basically known as a method to analyze the specific characteristics of given data [3]. The FCM

0-7803-7090-2/01/$10.00 © 2001 IEEE.

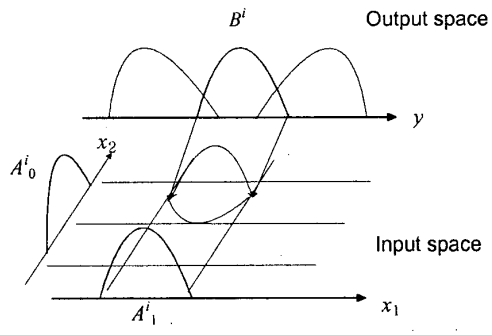


Figure 1. The clustering of the input space into the output spaces.

(Fuzzy C-Means) approach of Bezdek [4] is employed to the fuzzy clustering. In this paper, fuzzy space classification using FCM based on the similar feature of ozone output data is implemented to compute the classified degree of input variables. Figure 1 illustrates that the output space is classified by fuzzy clustering and mapped into the input spaces. The related features between input variables and output data are found by this fuzzy clustering.

3. DYNAMIC POLYNOMIAL NEURAL NETWORK

3.1. The Basic Structure of DPNN

DPNN uses GMDH method [5] based the observed data and prospective input variables to generate a model. This method is widely used for modeling of the dynamic systems, prediction, and artificial intelligent control. As shown in Figure 2, the simple DPNN structure has two inputs and one output at each node.

The polynomial expression between input and output variables of the each node in the DPNN is in Equation (1). The output y_1 and y_2 at each node are expressed as follows:

$$y_1 = w_{01} + w_{11}x_1 + w_{21}x_2 + w_{31}x_1x_2 + w_{41}x_1^2 + w_{51}x_2^2$$
$$y_2 = w_{02} + w_{12}x_3 + w_{22}x_4 + w_{32}x_3x_4 + w_{42}x_3^2 + w_{52}x_4^2 \quad (1)$$

The final output z in Figure 2 is represented by the following polynomial equation:

$$z = w_{03} + w_{13}y_1 + w_{23}y_2 + w_{33}y_1y_2 + w_{43}y_1^2 + w_{53}y_2^2 \quad (2)$$

where, w_{ij} (i=0,1,2,$\cdots$,n. j=0,1,2,$\cdots$,k.) is the coefficient. If input variables of each node are more than three, another

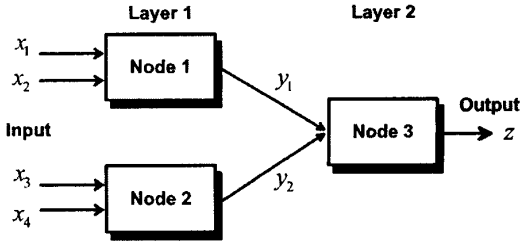


Figure 2. The basic structure of DPNN.

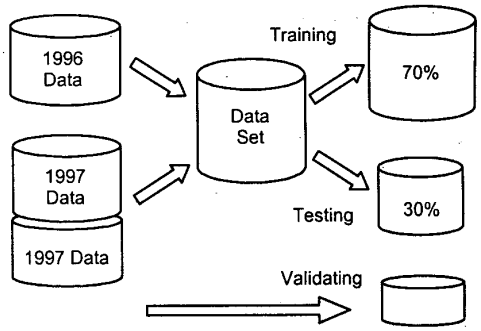


Figure 3. Split of the collected data.

combination terms of input variables are added to the above equation.

The least square method is used to estimate the parameters of each node. And it searches the solution of parameters, which will minimize the objective function formed by error function between actual and node outputs.

Equation (3) shows the object function and the coefficient. At each layer, parameters are solved by the least square method and polynomial functions of current node are structured. This process is repeated until the predefined criterion is satisfied. Thereafter, we could finally get the best function for the best performance.

$$J = \sum_{k=1}^{\#\ of\ data} (z(k) - \hat{z}(k))^2 = \| z - \phi\, w \|^2$$
$$w = (\phi^T \phi)^{-1} \phi^T z \quad (3)$$

3.2. Self-Organization

Another specific characteristic of DPNN is self-organization [6]. The DPNN based on the GMDH method separates data into training data and testing data for

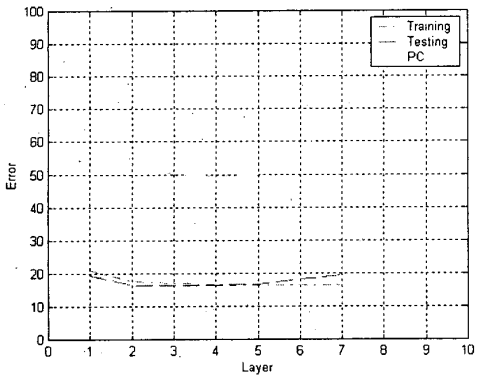


Figure 4. The variation of *PC* corresponding to the increment of layer.

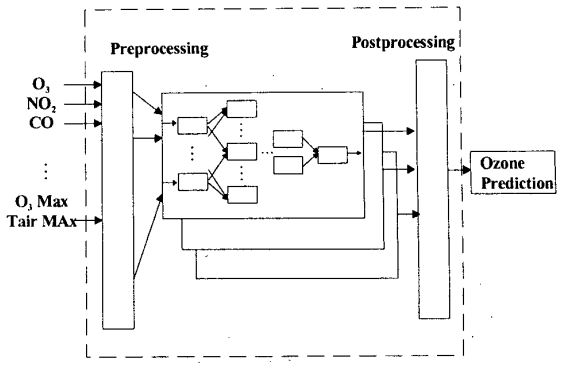


Figure 5. The total structure of the ozone prediction system.

modeling purpose [7] as shown in Figure 3. The purposes of this stage are to identify dynamic system and to prevent overfitting. The DPNN estimates the parameters of each node and composes the network structure of dynamic system using two-separated data sets. Training data set is used to solve the parameters of the function in each node and testing data set is used to evaluate performance of the DPNN. The final network structure is constructed by the relationship of error in training data and testing data. Therefore, The model in each node can select the input of the next node with performance measurements, which is the relationship between training error and testing error at each node. The final network structure is determined based on the performance criterion (*PC*) as shown in Figure 4.

PC could be determined by following Equation (4), where η is existed in the range of $0 \sim 1$. The model performances are also evaluated by Equation (4). This performance criterion can be applied for test data and training data. And also it can test unprepared the new data.

$$e_1^2 = \sum_{i=1}^{n_A} (y_i^A - f_A(x_i^A))^2 / n_A,$$
$$e_2^2 = \sum_{i=1}^{n_B} (y_i^B - f_B(x_i^B))^2 / n_B, \quad (4)$$
$$PC = e_1^2 + e_2^2 + \eta(e_1^2 - e_2^2)^2$$

where, e_1, e_2, n_A, n_B, and y_i indicate training errors, testing errors, the number of training data, the number of testing data and measured outputs, respectively. And $f_A(x_i^A)$ and $f_B(x_i^B)$ are outputs of training data and testing data separately. Total number of data is $n=n_A + n_B$. From the results, the optimized model structure is constructed to minimize the *PC*.

4. SIMULATIONS

Generally, ozone, CO, NO_2, SO_2, TSR, wind speed, wind direction, temperature, solar radiation, humidity, and rain fall are used for the parameters of air pollution materials and meteorological materials in ozone prediction systems. Within the data, the amount of rainfall is normally 0mm at high-level ozone, so it cannot influence the high-level ozone prediction. And TSR and SO_2 are skipped because these values are decreased by the restriction of air pollution material and wind direction is also excluded due to the difficulty of quantification. Thus, ozone, CO, NO_2, wind speed, temperature, humidity, solar radiation, maximum O_3 of previous day, and maximum atmosphere temperature of previous day are chosen as input variables. The total structure of the ozone prediction system is shown in Figure 5. The daily maximum of ozone concentration is generally appeared at $2 \sim 5$ p.m. Therefore, the forecasting of the ozone concentration at $2 \sim 5$ p.m is the goal in this paper. Ozone, CO, and NO_2 of the input variables are collected from the measurement data in the morning and the other data of the input variables are come from the prediction data for $2 \sim 5$ p.m. The forecasting period is from October to September in 1997. The training data and testing data are chosen from May to September in 1996 and from May to July in 1997. When the input variables are applied to predict ozone concentration, those of data are classified by the predicted time or measured time. These data could show as Table 1. The upper column point out the input variables and the higher columns of inner cells indicate the time. The left row is a number of data. Each meaning of input variables is shown in Table 1.

In the data preparation process, the missing data are interpolated by the spline interpolation method. For the decision support system, low-level ozone concentration model and high-level ozone concentration model are

Table 1. Prospective input variables and data structure for forecasting system

		O_3	NO_2	CO	Tair	Rh	Sr	Ws	O_3 Max	Tair Max	O_3
980	245	6	6	6	14	14	14	14			14
		⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	245	7	7	7	15	15	15	15			15
		⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	245	8	8	8	16	16	16	16			16
		⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	245	9	9	9	17	17	17	17			17
		⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

*Tair: Atmosphere temperature, Rh: Relative humidity, Sr: Solar radiation, Ws: Wind speed, O_3 Max: The maximum O_3 of previous day, Tair Max: The maximum atmosphere temperature of previous day

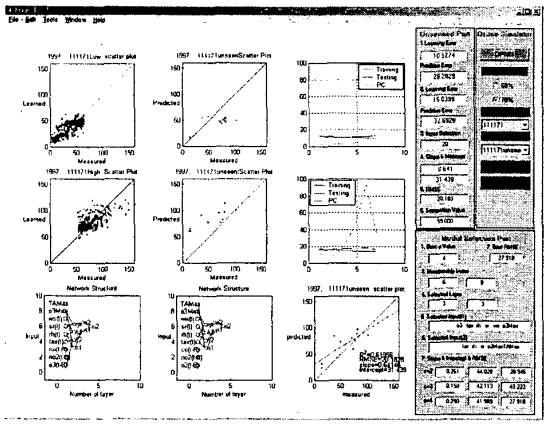


Figure 6. The prediction result using the decision support system (BangHak-Dong).

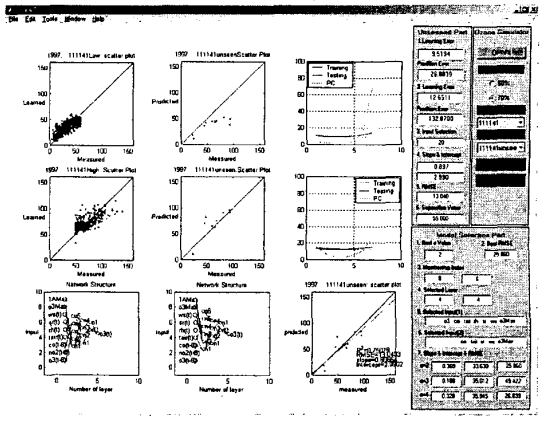


Figure 7. The prediction result of GooEui-Dong.

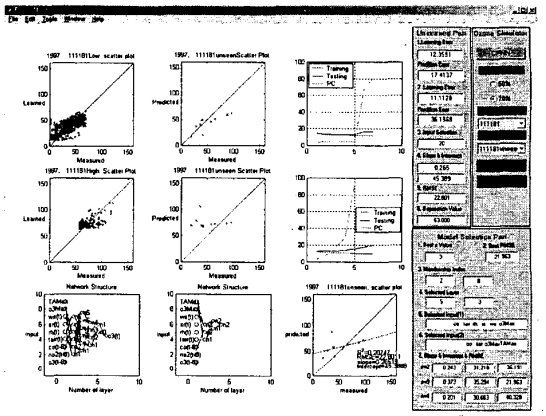


Figure 8. The prediction result of BoolKwang-Dong.

constructed by fuzzy clustering based on the basic training data. In this system, mean values and standard deviation are firstly found with respect to input variables of each model and then required membership functions are selected based on the correlative distance of each membership function.

In the system, prediction data are passed through the preprocessor and then the low and high concentration models fulfill the prediction processing. Thereafter, the final outputs are obtained by the outputs, which are from two models at postprocessor. At the postprocessor, the outputs of models are treated by membership functions formed by the fuzzy clustering.

In the simulations, a number of clusters are applied from 2 to 4. Basically, high-level ozone used the highest value and low-level ozone consists of the other set. Figure 6 shows the ozone prediction from August 1 to 10 in 1997. In this simulation, used model is selected by RMSE that is calculated from the designed the model after clustering the training data. When the number of the cluster is 4, the lowest training RMSE is 27.918, and the prediction RMSE is 20.183.

19 areas in Seoul are predicted based on the proposed prediction system. Figure 7 and Figure 8 are the results of the prediction of GooEui-Dong and BoolKwang-Dong, respectively. As shown in results, Figures 7-9 indicate that the results of the prediction system show a good performance. The other results of the 19 areas in Seoul are summarized in terms of RMSE, slope, and intercept value, as performance measurements in Table 2. RMSE value could be decreased by the sufficient data that include proper information in the high-level ozone concentration.

Slope and intercept values for R-square are displayed in the scatter graphs of ozone observation (x-axis) against the predicted values (y-axis) for each model. Intercept and

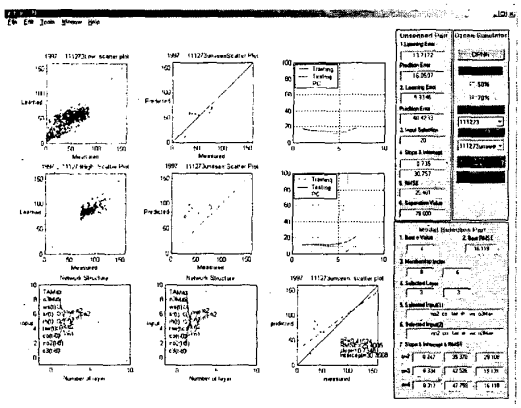


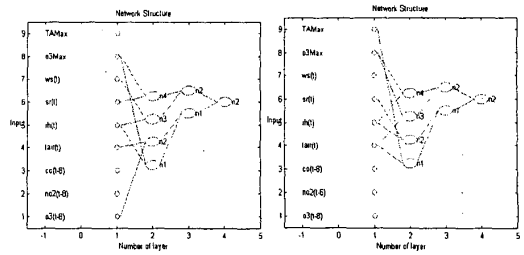
Figure 9. The prediction result of Bangli-Dong.

Table 2. Prediction result of 19 areas in Seoul

Name of Area	RMSE	Slope	Intercept
KwangHwaMoon	21.2221	0.51404	30.5004
HanNam	21.8447	0.28239	35.8997
GooEui	13.0403	0.89664	2.9902
SeongSoo	20.8272	0.42505	30.9287
MyeonMok	18.8231	0.35165	30.3495
Sinseol	15.0591	0.5602	21.8865
GilEum	22.8595	0.30068	38.4053
BangHak	20.1826	0.64146	31.439
BoolKwang	22.8011	0.26519	45.3888
Mapo	17.4159	0.4428	21.9442
HwaGok	27.7074	0.3008	57.1696
GooRo	23.3290	0.13329	35.4893
O-Lyu	16.8801	0.44225	29.7443
MoonRae	21.4642	0.28685	33.8014
ShinRim	20.6881	0.55472	33.8014
DaeChi	21.8937	0.59927	33.339
BanPo	14.3991	0.33134	36.726
JamSil	34.2887	0.42009	45.7232
Bangli	25.4005	0.7348	30.7568

slope values in Table 2 show that both models overpredict low values and underpredict ozone during high pollution events. The two diagonal lines in the plots represent the best-fit regression lines through the data and the line of perfect correspondence between observations and predictions [8].

As a example, the adaptively constructed network models are presented in Figure 10 that is a part of the third row in Figure 6. The network structure in Figure 9(a) is the model of low-level ozone concentration and Figure 9 (b) is the model of high-level ozone concentration. Figure 9 shows that the finally selected inputs in the two models are different. For instance, 6 input variables are selected for the low-level model and 5 input variables are chosen for the high-level model. The polynomial equations at each



(a) The low-level model (b) The high-level model

Figure 10. The Network models of high and low level concentration.

Table 3. The used parameters in the real network model and new named parameters for polynomial equation

Real input variable name of model	Changed variable name for polynomial equation
$O_3(t\text{-}8)$	x_1
$NO_2(t\text{-}8)$	x_2
$CO_2(t\text{-}8)$	x_3
Tair(t)	x_4
Rh(t)	x_5
Sr(t)	x_6
Ws(t)	x_7
O_3 Max	x_8
Tair Max	x_9

node are formulated in Equations (5)-(7). Input variables are 6 as shown in Figure 10(a).

The Equations (5)-(7) is the outputs of each layer. The outputs of the first layer consist of four parts of the second order polynomial equations. The outputs of the second and third layers have two parts of the forth order and one part of the eighth order polynomial equation, respectively. In a word, equation orders are changed with 2 times step.

First layer:

$$\begin{aligned} n_1^1 &= w_{01} + w_{11}x_5 + w_{21}x_8 + w_{31}x_5x_8 + w_{41}x_5^2 + w_{51}x_8^2 \\ n_2^1 &= w_{02} + w_{12}x_1 + w_{22}x_4 + w_{32}x_1x_4 + w_{42}x_1^2 + w_{52}x_4^2 \\ n_3^1 &= w_{03} + w_{13}x_5 + w_{23}x_7 + w_{33}x_5x_7 + w_{43}x_5^2 + w_{53}x_7^2 \\ n_4^1 &= w_{04} + w_{14}x_6 + w_{24}x_8 + w_{34}x_6x_8 + w_{44}x_6^2 + w_{54}x_8^2 \end{aligned} \quad (5)$$

Second layer:

$$\begin{aligned} n_1^2 &= w_{01} + w_{15}n_1^1 + w_{25}n_2^1 + w_{35}n_1^1n_2^1 + w_{45}(n_1^1)^2 + w_{55}(n_2^1)^2 \\ n_2^2 &= w_{02} + w_{16}n_3 + w_{26}n_3^1 + w_{36}n_3^1n_4^1 + w_{46}(n_3^1)^2 + w_{56}(n_4^1)^2 \end{aligned} \quad (6)$$

Output layer:

$$n_3^1 = w_{07} + w_{17}n_1^2 + w_{27}n_2^2 + w_{37}n_1^2n_2^2 + w_{47}(n_1^2)^2 + w_{57}(n_2^2)^2 \quad (7)$$

5. CONCLUSTION

In this paper, the ozone prediction system based on DPNN, fuzzy clustering, and decision support system is developed to improve accuracy in the high-level ozone concentration. The results of the proposed system are getting better than the results of a fixed model because the structure of the proposed system is daily updated with a new incoming data. The selection of the number of the cluster in the fuzzy clustering is very important to split data for low-level and high-level ozone concentration models. Finally, the combined output of the two models that have input selection process and optimized structure will fulfill better performance.

6. REFERENCES

[1] Sungi Lee and Chong-Bum Lee, The development of the AR model to estimate photochemical pollution concentration in Seoul, Meteorological research paper, 1991, pp 71-85

[2] Duc Trung Pham and Liu Xing, Neural Networks for Identification, Prediction and control, Springer-Verlag Inc., 1995.

[3] Sungshin Kim, A Neuro-Fuzzy Approach to Integration and Control of Industrial Processes: Part I , KFIS, 1998, pp 58-69.

[4] James C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum, 1981.

[5] A. G. Ivakhnenko, The Group Method of Data Handling in Prediction Problem, *Soviet Automatic Control*, vol. 9, no. 6, pp. 21-30, 1976.

[6] S. Farlow, ed., Self-Organizing Method in Modeling: GMDH-Type Algorithms, Marcel Deckker, Inc., New York, 1984.

[7] A. G. Ivahnenko, Polynomial theory of complex system, IEEE trans. System. Man and Cybernetic, 1971, pp 364-378.

[8] Greg Spellman, An application of artificial neural networks to the prediction of surface ozone concentrations in the United Kingdom, Applied Geography, 1999, pp123-136.