# Multi-Procedure Ozone Concentration Prediction using Fuzzy Clustering and DPNN

Seong-Pyo Cheon, Seng-Tai Lee, and Sungshin Kim

School of Electrical and Computer Engineering, Pusan National University, Busan, Korea

Pusan National University Changjeon-dong, Keumjeong-ku, Busan 609-735, Korea
E-mail: buzz74@pusan.ac.kr, youandi@pusan.ac.kr, sskim@pusan.ac.kr

## Abstract

Ozone reaction is a complex mechanism because of its intrinsic complexity and nonlinearity. It is not easy to predict its concentration using traditional numerical and statistical methods. Especially, we found that conventional single model could not get corrected ozone concentrations at high range. We propose a multi-procedure model to overcome the complexity and nonlinearity of ozone concentrations. At first, we suggest fuzzy clustering method to divide high- and low- concentration groups for input ozone data in past two years. Then, we determine proper input data group and its data are used to input data for DPNN models. We divide two groups for input ozone data because the results of the proposed model seriously depend upon their inputs. Preprocessing and postprocessing algorithms are applied to the input data as two steps in order to get more accurate and reliable prediction results of ozone concentrations.

Keywords: Fuzzy clustering, DPNN, multi-procedure model

## 1 Introduction

One of the emerging major issues about air pollution is the abnormal ozone concentration of the troposphere in summer. High concentration ozone is strong oxidizing material which is responsible for various adverse effects on both human being and foliage. And it frequently appears metropolis in the daytime from June to August in summer. In general, it revealed that the features on ozone distribution and creation are closely related to the photochemical reaction and meteorological factors. So far, there is not enough information about detailed cause and effect of the ozone. In the ozone reaction mechanism around the troposphere, nitric dioxide and hydrocarbon which are components of exhaust fumes act as precursors and ultraviolet radiation, wind speed, and temperature play the role of meteorological materials. Therefore, ozone is a second pollution material. To reduce harms by high concentration ozone, it needs prediction system which forecasts maximum ozone concentration every morning in summer.

There are some 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. However, it does not show a good prediction performance. We analyses major related reasons as follows: First, ozone is a kind of second pollution material. Due to both first pollution ones and other related factors affected on its behavior and concentration, it is impossible to create complete model which was considered all factors. Second, there are not sufficient data when high concentration ozone appeared. Third, we couldn’t get ultraviolet ray data directly. So, we couldn’t help using solar radiation data that we estimated ultraviolet ray indirectly. Finally, we couldn’t be considered that precursors inflow from long distance. As shown in Figure 1, we suggested multi- procedure ozone concentration forecasting system. In this paper, we introduce a multi-procedure model which consists of preprocessing, dynamic polynomial neural networks (DPNN) models, and postprocessing. Fuzzy clustering method as a preprocessing is employed to decide the fuzzy sets of the low and high ozone concentration. After clustering the two fuzzy sets, the different two models are determined by the DPNN. The decision making process as a postprocessing is applied to forecast the ozone concentration by the compensation of the weight factors to the output 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 from the possible input variables by a selection criterion. The historical data that consist of pollution materials 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.

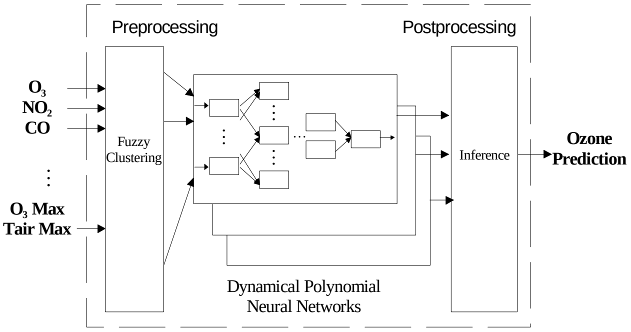

Figure 1. The full structure of the ozone prediction system.

## 2 Fuzzy Clustering

The fuzzy c-mean algorithm (FCM) generalizes the hard

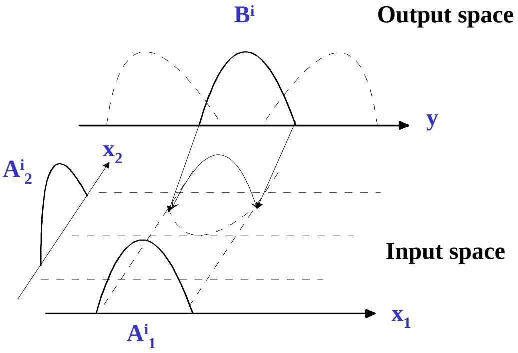

Figure 2. The mapping of input variables into output space

c-mean algorithm to allow a point to partially belong to multiple clusters. Therefore, it produces a soft partition for a given dataset[3]. FCM of Bezdek [4] is normally applied for the fuzzy clustering. In this paper, fuzzy space classification using FCM based on the similar feature of ozone concentration output data is implemented to compute the classified degree of input variables. Figure 2 illustrates that the output space classified by fuzzy clustering is mapped to input spaces, and then the related features between input variables and output data are determined by fuzzy clusters.

## 3 Dynamic Polynomial Neural Network

### 3.1 The Basic Structure of DPNN

DPNN uses GMDH (Group Method Data Handling) method [5] to compose an input/output model based on observed data and variables. This method is widely used for modeling of system, prediction, and artificial intelligent control. As shown in Figure 3, the simple DPNN structure has four inputs and one output at each node. Following polynomial equations between input and output are used at each node in the DPNN. Output $y_1$ and $y_2$ at each node are expressed as follows.

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

The final output $\hat{y}$ is represented by a polynomial equation.

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

where, $w_{ij}$ ($i=0,1,2,\ldots,n, j=0,1,2,\ldots,k$) is the coefficient. If input variables of each node are more than three, other combination terms of input variables are added to the above equation. The least square method is employed to estimate the parameters of each node in the DPNN. And it searches the solution of parameters to minimize the objective function formed by error functions between node outputs and actual target values.

$$J = \sum_{k=1}^{\#\text{ of data}} (y(k) - \hat{y}(k))^2 = \| y - wA \|^2 \qquad (3)$$

$$w = (A^T A)^{-1} A^T y \qquad (4)$$

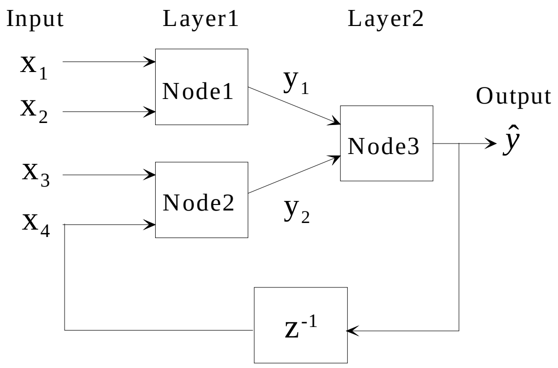

Figure 3. The basic structure of DPNN.

Equation (3) and (4) show the objective function and the coefficient, respectively. Parameters are solved by the least square method and polynomial functions of current node are structured at each of layer. This process is repeated until the criterion is satisfied. Thereafter, we could finally get the best function for the best performance.

### 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 modeling [7]. The purposes of this stage are to identify the behavior of the dynamic system and to prevent overfitting problem. 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 parameter of function of each node and testing data set is used to evaluate the performance of DPNN. The final network structure is constructed by the relationship of error in training data and testing data. Therefore, the DPNN selects the input of the next node under a performance criterion(PC) that is the relationship between training error and testing error at each node. The final network structure is determined as shown in Figure 4. The PC could be determined by following Equation (5), where is existed in the range of *0 ~1*. The model performances are also evaluated by Equation (5). This performance criterion can be applied for the testing data

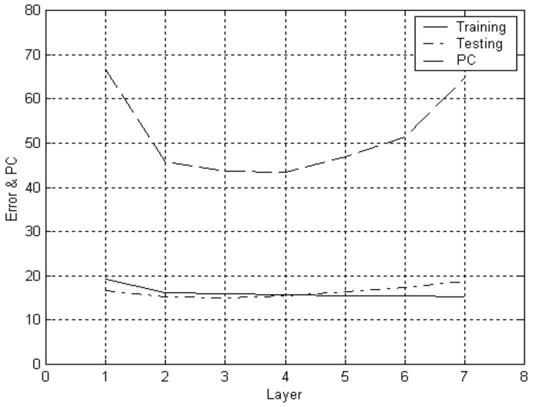

Figure 4. The variation of model performances corresponding to increment of layer.

and the training data. And also it can be used for the unprepared new data.

$$
\begin{aligned}
e_1^2 &= \sum_{i=1}^{n_A} (y_i^A - f_A(x_i^A))^2 / n_A, \\
e_2^2 &= \sum_{i=1}^{n_B} (y_i^B - f_B(x_i^B))^2 / n_B, \qquad (5) \\
PC &= e_1^2 + e_2^2 + \eta(e_1^2 - e_2^2)^2
\end{aligned}
$$

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 at the point of minimized PC.

## 4 Fuzzy Inference in the Postprocessing

High and low levels are classified based on ozone concentration data in the postprocessing and then ozone concentrations are predicted by DPNN. The predicted results are shown in Figure 5. As shown in Figure 5, the predicted results are represented as $Y_L$ and $Y_H$. $Y_i$ and $Y_j$ are calculated by the means and variations of two input variables $X_i$ and $X_j$ ($i \neq j$), respectively. In this case, the two input variables have the longest distance between high and low level concentrations. It means that the $X_i$ and $X_j$ affect strongly to ozone concentrations. The outputs of the $Y_i$ and $Y_j$ influenced by the $X_i$ and $X_j$ are computed by the fuzzy inference in the Equation (6).

$$
\begin{aligned}
Y_i &= \frac{(0.7 \times Y_L) + (0.3 \times Y_H)}{0.7 + 0.3} \qquad (6) \\
Y_j &= \frac{(0.2 \times Y_L) + (0.8 \times Y_H)}{0.2 + 0.8}
\end{aligned}
$$

The final prediction result based upon the decision support system is decided by Equation (7). In this equation, the weights $W_{Xi}$ and $W_{Xj}$ are the relative distances of the input membership functions.

$$\hat{Y} = \frac{(W_{X_i} \times Y_i) + (W_{x_j} \times Y_j)}{W_{x_i} + W_{x_j}} \qquad (7)$$

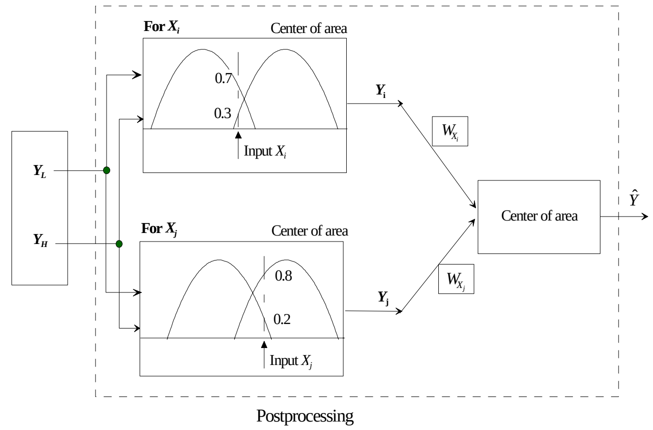

Figure 5. The fuzzy inference in the Postprocessing.

## 5 Performance Assessment

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. Therefore, ozone, CO, $NO_2$, wind speed, temperature, humidity, and solar radiation, maximum $O_3$ of previous day and maximum atmosphere temperature of previous day are chosen as the possible input variables. Because the daily maximum of ozone concentration is appeared at 2～5 p.m., the prediction of the ozone concentration for this time is the goal of this paper. Within the data, ozone, CO, and $NO_2$ are extracted from morning data and the other data are selected at 2～5 p.m. These data structure could show as Table 1. In the first simulation, the number of clusters is applied from 2 to 4. Basically, high-level ozone used the highest value and low-level ozone consists of the other set. Figures 6, 7 and 8 show the ozone prediction results of several areas in Seoul, Korea from August 1 to 10, 1997. Total data are constructed by the data from May to September in 1996 and from May to July in 1997. And the training data and testing data are selected from among the total data. When the input variables are applied to predict ozone concentration, those of data are classified by the predicted time and measured time. 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. In this simulation, a model is chosen based on the *RMSE* that is yielded from the selected model. This model is determined 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. The slope and intercept values for *R-square* are displayed in the scatter graphs of ozone observation (*x*-axis) against the predicted

Table 1. Prospective input variables and data structure

| | | $O_3$ | $NO_2$ | CO | Rh | Sr | Ws | $O_3$ Max | Tair Max | $O_3$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 980 | 245 | 6 | 6 | 6 | 14 | 14 | 14 | PRE | PRE | 14 |
| | | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| | | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| | 245 | 7 | 7 | 7 | 15 | 15 | 15 | PRE | PRE | 15 |
| | | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| | | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| | 245 | 8 | 8 | 8 | 16 | 16 | 16 | PRE | PRE | 16 |
| | | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| | | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| | 245 | 9 | 9 | 9 | 17 | 17 | 17 | PRE | PRE | 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.

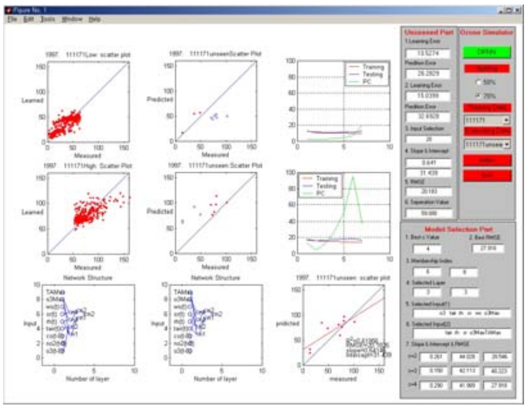

Figure 6. The prediction results for BangHak-Dong.

values (*y*-axis) for each model. The two diagonal lines in the plots represent the best-fit regression and the perfect correspondence between observations and predictions [8]. In the second simulation, the predicted area is Ssang-Mun-Dong in Seoul, which is high-level ozone area in the summer. The predicted period is from May to July in 1999. The training data and testing data are constituted by the data from May to September in 1996, 1997 and 1998. In this ozone prediction, missing data are interpolated by the spline interpolation method. For the decision support system, a low concentration model and a high concentration model are constructed by the fuzzy clustering method based on the basic training data. In this system, mean values and standard deviations are firstly found with respect to input variables of each model and then required membership functions are selected by the correlative distance of each membership function. The number of clusters is applied from 2 to 4. Basically, the high-level ozone uses the highest value and the low-level ozone consists of the other set. Figure 7 shows the result of the ozone prediction the period from May 20 to July 20 in 1999. When the number of the cluster is 4, the lowest training *RMSE* is 15.634, and the prediction *RMSE* is 18.034. In the decision support 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 by the two models at the postprocessing. At the postprocessing, the outputs of the models are handled by membership functions, which are formed by the fuzzy clustering.

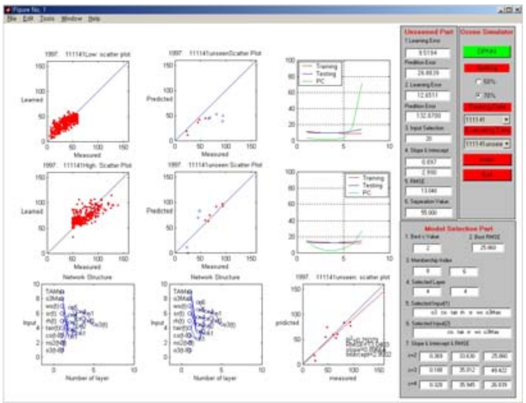

Figure 7. The prediction results for GooEui-Dong.

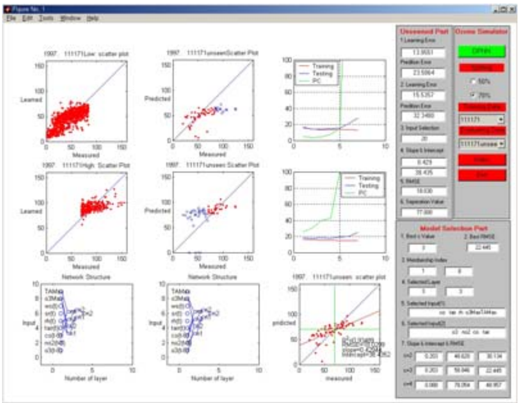

Figure 8. The prediction result for Ssang-Mun-Dong.

## 6 Conclusion

In this paper, we propose the multi-procedure prediction system using various approaching methods. i.e. Fuzzy clustering method, DPNN, and weighted combining postprocessing. The model designed by DPNN, which includes decision support system, is suitable for the high concentration ozone prediction. When the models are daily updated based on new input variables, the prediction performances are getting better than the results by the fixed model. And it is confirmed that the selection of the cluster number in the fuzzy clustering is also very important for the high-level ozone. Finally, the combining over two models, using various input selection and optimizing the structure of models will fulfill better performances.

## 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, 1976, pp 21-30.

[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, pp 123-136.