World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:11, No:11, 2017

Input Data Balancing in a Neural Network PM-10 Forecasting System

Suk-Hyun Yu, Heeyong Kwon

Abstract**—**Recently PM-10 has become a social and global issue. It is one of major air pollutants which affect human health. Therefore, it needs to be forecasted rapidly and precisely. However, PM-10 comes from various emission sources, and its level of concentration is largely dependent on meteorological and geographical factors of local and global region, so the forecasting of PM-10 concentration is very difficult. Neural network model can be used in the case. But, there are few cases of high concentration PM-10. It makes the learning of the neural network model difficult. In this paper, we suggest a simple input balancing method when the data distribution is uneven. It is based on the probability of appearance of the data. Experimental results show that the input balancing makes the neural networks' learning easy and improves the forecasting rates.

Keywords**—**AI, air quality prediction, neural networks, pattern recognition, PM-10.

I. Introduction

AIR pollution comes from a variety of sources like gaseous, liquid, solid waste, or man made products. Such contamination of the atmosphere has a serious effect on human health and the biosphere and also it reduces visibility, and damage materials. Especially, PM-10 (particulate matter) is one of the major pollutants throughout the country. PM-10 is a complex mixture of solid and liquid particles under 10 μm that vary in size and composition. It remains suspended in the air. [1] The size of ambient air particles ranges from approximately 0.005 to 100 μm in diameter. PM-10 is defined as particulate matter with a diameter less than 10 μm. Over the past decades, many health effect studies have shown an association between exposure to PM-10 and increase in daily mortality and symptoms of certain illnesses such as asthma, chronic bronchitis, decreased lung function, and premature death. Sources of PM-10 are numerous; naturally occurring processes and human activities all contribute to PM-10 concentrations. Some sources are natural, such as dust from the earth's surface, sea salt in coastal area, and biologic pollen. Periodic events like forest fires and dust storm can produce large amount of PM-10. In urban areas, PM-10 is mainly produced by combustion from mobile sources such as cars, buses, ships, trucks, and construction equipment, and from stationary sources such as municipal incinerators, thermal power plants, and factory chimneys. Some PM is emitted directly into the atmosphere as particles, while other PM is produced by chemical reactions from the air pollutants in the air. [2] The major developed countries such as USA and UK have developed PM forecasting system using the statistical methods. [3]-[6] Fig. 1 shows whole air quality forecasting system. Conventional forecasting has been conducted with global and/or regional chemical transport model. Recently, efficient neural network PM-10 forecasting models have been introduced. They use the surface and satellite observation data as well as chemical transport modeling results as their inputs. Moreover, neural network models have been improved in performance as they evolved into deep neural networks.

As an AI model, deep neural networks are very good for many application fields. However, it requires good and big data for good performance. Big data could be acquired from the network technologies, like internet and IOT. However, it is very difficult to acquire good data for the PM-10 concentration forecasting. Because there are many cases of low concentration PM-10, but few cases of high. In Seoul metropolitan area, the high concentration of PM-10 during the observation period of 2015 and 2016 (660 days) was only 63 days. The PM-10 data distribution is not even extremely. It is too small for a neural network model to learn and forecast the concentration. So, input balancing is very important for the neural network's learning efficiency. Unevenly distributed data makes the learning result distorted. In this paper, we suggest a simple balancing method when the data distribution is uneven. It is based on the probability of appearance of the data.

The rest of the paper consists of four parts: in section II, we describe the neural network PM-10 forecasting model. Then, we present a detailed algorithm for the input data balancing technique. Finally, experimental results are shown and concluded.

II. Neural Network PM-10 Forecasting Model

Neural networks are non-parametric models which relate unknown inputs with corresponding outputs after learning by examples. In case, relations between inputs and outputs are unknown, like the PM-10 prediction problem, neural networks are effective and powerful tools for solving the problem. There are various models that have different learning rules and architectures as they are applied to different areas. A model capability depends on the connection topology between two layers or among neurons. We need therefore to find an efficient neural model, a network architecture and a learning rule which are adequate for the PM-10 prediction problem [7]-[9].

S. H. Yu is with the Information and Communications Department, Anyang University, Anyang, Korea.

H. Kwon is with the Computer Science and Engineering Department, Anyang University, Anyang, Korea (e-mail: hykwon@anyang.ac.kr).

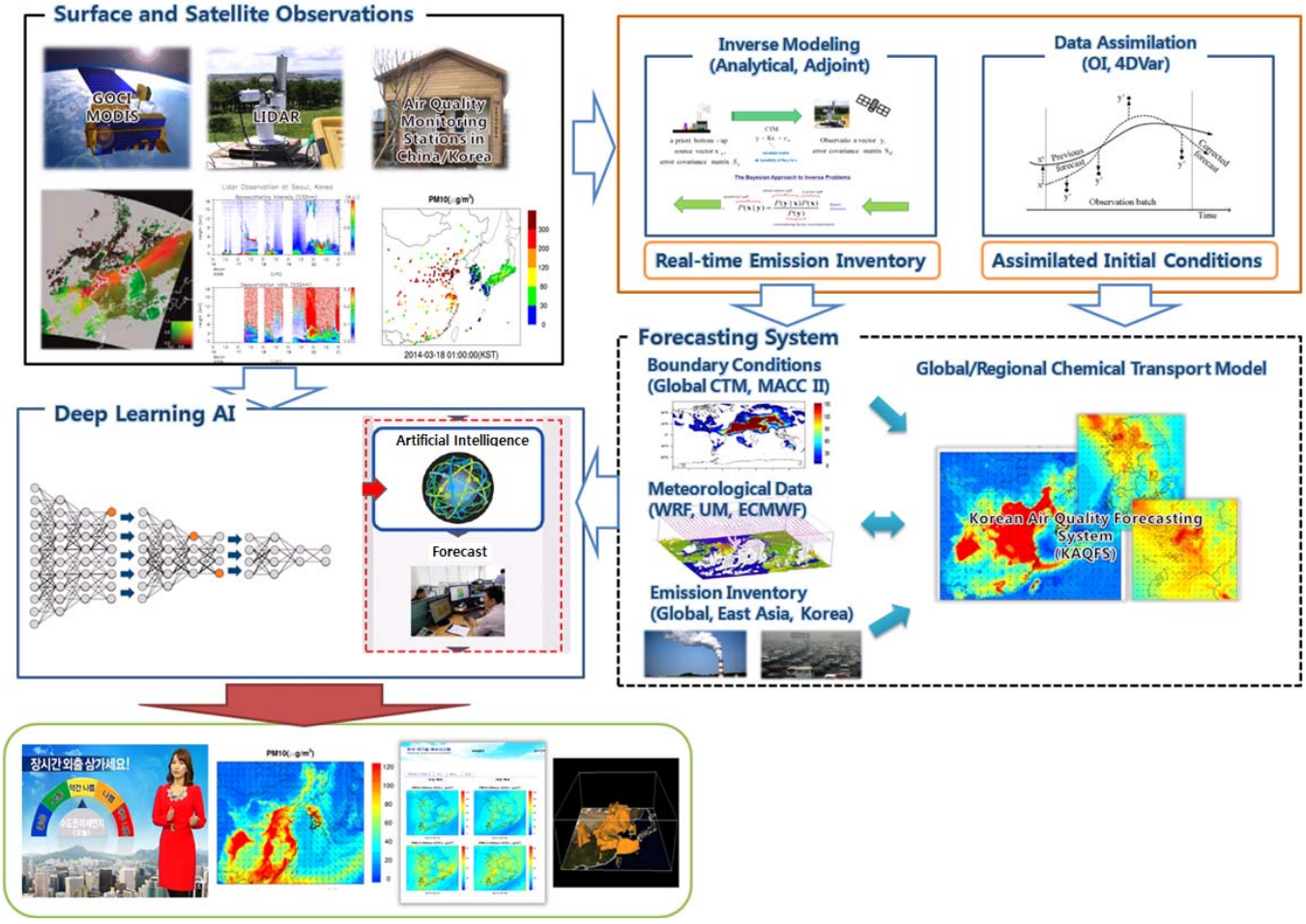


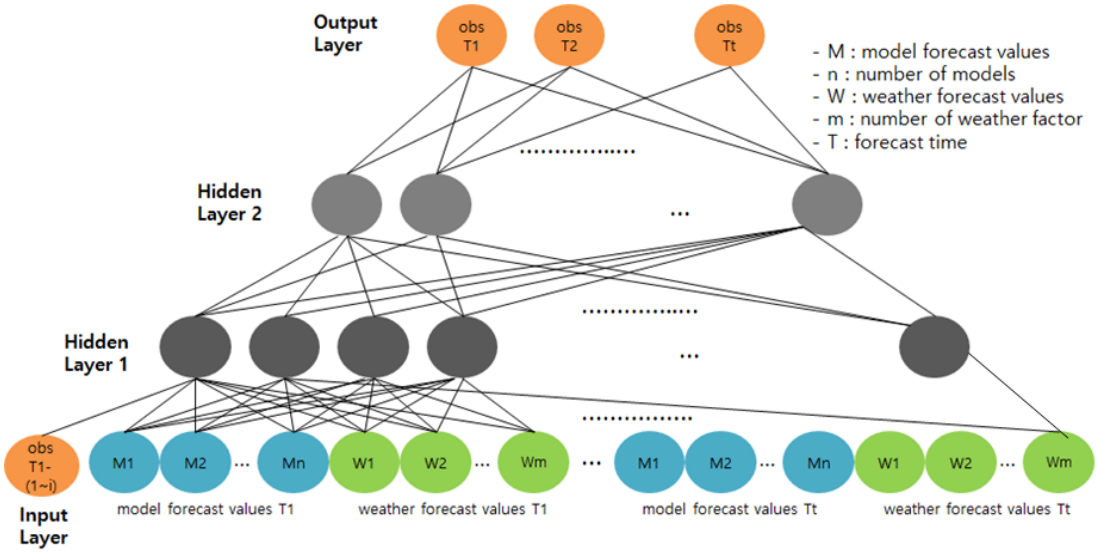
Fig. 1 Air quality forecasting system

In the previous research [1], we used a fully connected MLP neural network with EBP learning rule which is the most popular and has the highest generalization capability. Considering that PM-10 prediction is a typical nonlinear problem, we construct the network with input, hidden-1, hidden-2 and output layers. It was too big to operate in the field and not satisfactory in the accuracy (Fig 2 (a)). So, we propose a new network model. The proposed network is composed with 12 separated networks (Fig 2 (b)). Each network has its own time interval input data. So, the input size could be reduced into 12 times. In addition, the learning speed and the accuracy are enhanced. Fig. 3 shows the proposed neural network's whole structures for PM-10 concentration forecasting.

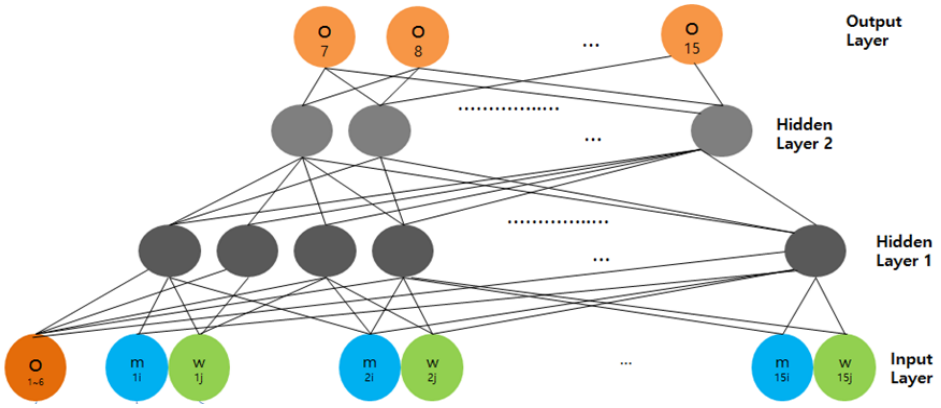
It is important to determine input variables of neural network model for the PM-10 forecasting system, also. In this study, the factors affecting the PM-10 level were selected using the measured data at the monitoring stations. The meteorological factors were wind direction and speed, temperature, humidity, atmospheric pressure, rain fall, mixing height, atmospheric stability and irradiation. The environmental factors were concentrations of SO_2, O_3, NO_2, CO, and PM-10.

III. Input Data Balancing

Neural network model's performance is dependent on lots of factors; network architecture, node size, input parameters, learning algorithm, learning rate, etc. Among them, input parameters are very important to show their ability. However, there are no evident criteria for good data properties. Especially, the data could be distributed unevenly in many application fields. In the case, learning is well conducted for the many cases data, but insufficiently done for the less cases data. Eventually the learning results are distorted largely. For example, high concentration cases are rare, just 10~15% of the annual data, but low cases are very common in the PM-10 forecasting application. It causes a serious imbalance in learning result. The neural network models learn the low cases very well, but cannot learn the high cases. So, we need a new learning method to learn both cases equally.



(a) Integrated model



(b) Separated model for j-th interval

Fig. 2 Neural Network Models

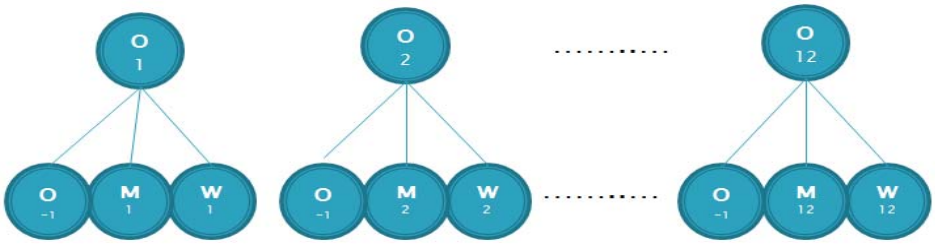


Fig. 3 Neural Network Models (Hidden Layers are omitted). (**M**$_t$: Model predicting parameter values at time t. **W**$_t$: Weather predicting parameter values at time t. **O**$_t$: Observed PM-10 value at time t.)

We introduce a probabilistic input data balancing method in learning process. Balanced learning probability of each i-th interval data, p_i, is defined as bellows.

$$p_i = \frac{q_i^{-1}}{\sum_{i=1}^{N} q_i^{-1}}$$

where, N: PM-10 concentration interval index number, q_i: PM-10 concentration interval i`s data probability. Then, the expectation value of each i-th interval data set is computed as:

$$p_i * n_i = \frac{q_i^{-1}}{\sum_{i=1}^{N} q_i^{-1}} * n_i = \frac{\sum {n_i}/{n_i}}{\sum_{i=1}^{N} q_i^{-1}} * n_i = \frac{\sum n_i}{\sum_{i=1}^{N} q_i^{-1}}$$

It shows that every interval's expectation value is the same. It ensures even learning opportunity of all input data. For example, suppose that, there are four intervals and each of them has four different appearance probabilities, such that N = 4 and

q_i = 1/11, 2/11, 3/11 and 5/11. The sum of all inverse of appearances is $\sum (q_i^{-1}) = 22.4$. So, the balanced input probabilities are p_i = 0.49, 0.25, 0.16, and 0.1. The expectation values of each intervals are 0.49, 0.50, 0.48, and 0.50.

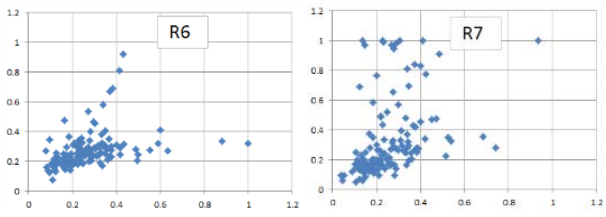
IV. Experiments and Analysis

In order to demonstrate the performance of the proposed method, two data sets are tested in a MLP neural network models, 4 layers, 20 input nodes, 10 and 5 hidden nodes, 1 output node, 0.09 learning rate and 60,000 epochs. The first dataset has a typical MLP network inputs. It uses past five interval's PM-10 measurements and nine meteorological prediction data (wind direction, speed, temperature, humidity, rain fall, atmospheric pressure, mixing height, atmospheric stability and irradiation) and the numerical models' air quality prediction data. Learning data are from 2 years (2015~2016) measurements in Seoul metropolitan area. Evaluations are conducted with 4 months data of 2017. They are unevenly distributed. The cases of high PM-10 concentration are very rare. We present the data distribution in Table 1. Here, 0.4 (interval 4) or more is a high PM-10 concentration. It means high concentration data is under 7% of whole data. But, in the second one, we use probabilities of appearance of the data. The amount of low concentration data (under 0.4) are 13 times of those of high data. Therefore, when the data are low concentration, it is ruled out with probability of $(1 - p_i)$.

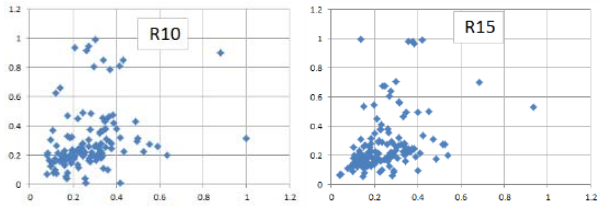
TABLE I
Data Distribution (2015~2016, Seoul)

Interval	0	1	2	3	4
Concentration	0~	25~	50~	75~	100~
Freq.	1,504	3,990	2,631	1,114	361
Prob.(q)	0.15	0.40	0.27	0.11	0.04
Prob.(p)	0.002	0.001	0.001	0.003	0.009
Interval	5	6	7	8	9
Concentration	125~	150~	175~	200~	225~
Freq.	136	78	39	4	43
Prob.(q)	0.01	0.007	0.004	0.001	0.004
Prob.(p)	0.023	0.039	0.079	0.771	0.071

Figs. 3 and 4 show two results of two networks. The results of the networks with balanced input data, in Fig. 5, are more precise than those of the imbalanced networks as Fig. 4. In this paper, we study on the data distribution effects of MLP neural network learning model to affect the PM-10 prediction performance. The proposed learning model is more efficient and accurate when the data are distributed unevenly

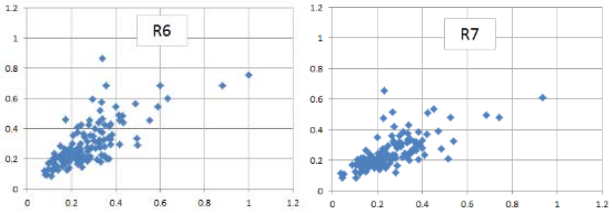


(a) Interval 6(0h~6h) (b) Interval 7(7h~12h)

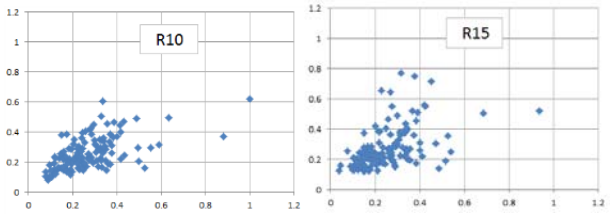


(c) Interval 10(24h~30h) (d) Interval 15(54h~60h)

Fig. 4 No input balancing results



(a) Interval 6(0h~6h) (b) Interval 7(7h~12h)



(c) Interval 10(24h~30h) (d) Interval 15(54h~60h)

Fig. 5 Input balancing results

V. Conclusion

In this paper, we present an input data balancing method in a neural network PM-10 concentration prediction system. We study on the data distribution effects of MLP neural network learning model to affect the PM-10 prediction performance. Neural network prediction model is very good for some application area. But, it requires good learning data. It means that the data for a neural network should be even for all cases. However, there are few bad days, high concentration of PM-10, It tend to be heavily biased. So, the neural network system could not learn the situations well. We introduce data appearance probability to solve the infrequent case. It provides even opportunities (expectation value) for all data. Experimental results show that the performance of the network with balanced input data is more precise than those of the non-balanced input data.

References

[1] Hee-Yong Kwon, S.H. Yu, Y.S. Koo, and E.Y. Ha, 'PM-10 Forecasting using Neural Networks Model', CIMCA'2008 Proc., pp.60~60, 2008
[2] https://nepis.epa.gov/Exe/ZyPURL.cgi?Dockey=2000F0ZT.TXT, Sep, 2017.
[3] Ian G. McKendry: 'Evaluation of Artificial Neural Networks for Fine Particulate Pollution (PM10 and PM2.5) Forecasting', Journal. of Air & Waste Management Association, Sep, 2002.
[4] T. S. Dye, D. S. Miller, C. B. Anderson, C. P. MacDonald, C. A. and Knoderer, B. S. Thompson: 'PM2.5 Forecasting Method Development and Operations for Salt Lake City, Utah', 2003 National Air Quality Conference, U.S. EPA, pp 1-18, 2003.

[5] M. Benjamin and J. Rousseau: 'Winter INFO-SMOG Program Forecast for the Greater Montreal Area', 2003 National Air Quality Conference, U.S. EPA, pp 19-23, 2003.
[6] Use of Time-Series Analysis to Examine the Link Between Photochemistry and PM Concentrations in Chicago, http://capita.wustl.edu/NEARDAT/WebLinks /pmupdate.htm.
[7] Air Pollution Forecasting in the UK, http: //www.airquality.co.uk/ archive/reports/ list.php.
[8] Ana Russo, Pedro G. Lin, Frank Raischel, Ricardo Trigo, and Manuel Mendes, 'Neural network forecast of daily pollution concentration using optimal meteorological data at synoptic and local scales', Atmospheric Pollution Research, pp.540~549, 6, 2015.
[9] Madhavi Anushka Elangasinghe, Naresh Singhal, Kim N. Dirks, Jennifer A. Salmond, 'Development of an ANN–based air pollution forecasting system with explicit knowledge through sensitivity analysis', Atmospheric Pollution Research, pp.696~708, 5, 2014.