Estimation of Air Pollutant Index (API) of Klang Valley using Artificial Neural Network (ANN)

Roshaslinie binti Amdam @ Ramli
Faculty of Electrical Engineering,
Universiti Teknologi MARA,
40450 Shah Alam, Selangor.
Email: linie_ee220@yahoo.com.sg

Abstract— The air pollution problems has received more attention during the last decades whereby there has been a significant increase in public awareness of the potential dangers caused by chemical pollutants and their effects both human beings and the environment. To overcome these problems, the need for accurate estimates of air pollutant index (API) becomes important. To achieve such estimation tasks, the use of artificial neural network (ANN) is regarded as an effective technique. The purpose of this paper, ANN trained with feed-forward back-propagation algorithm is used to estimate the air pollutant index (API). The API system normally includes the major air pollutants which are ozone (O_3), carbon monoxide (CO), nitrogen dioxide (NO_2), sulphur dioxide (SO_2) and suspended particulate matter of less than 10 microns in size (PM10). This method uses the past raw data values to estimate the API. The data collected comprises of data for the previous three month, beginning from October 2006 for Klang Valley areas which are Shah Alam, Klang, Petaling Jaya and Kuala Lumpur. The results indicate that the ANN model estimated API with good accuracy to more than 90%.

Keywords—Air Pollutant Index, Artificial Neural Network, estimation

I. INTRODUCTION

Air pollution is the presence of undesirable material in air, in quantities large enough to produce harmful effects. The World Health Organisation (WHO) estimates that 500,000 people die prematurely each year because of exposure to ambient concentration of airborne particulate matter [1]. Worldwide air pollution is responsible for a large number of deaths and cases of respiratory disease. There are many natural sources of air pollution such as eruption of volcanoes, biological decay and lightning-caused forest fire [2]. Naturally, the Earth already has its own air pollution loading. However, industrialization or just everyday routines has become added burden to the existing air pollution loading. Sources of air pollution are issue from industrial and development activities, motor vehicles, power generation, everyday routine and open burning [3], [4].

Here, a training technique of Back Propagation Neural Network (BPNN) was applied. The performance of this estimation is examined with the real world data that provided by the Department of Environmental (DOE). The selection of the input variables, ANN architecture, collection of training sample and processing of data, are discussed in this paper.

Among the various method of ANN, the feed-forward back-propagation neural network is employed in this study.

A. Air Pollution Index

The air quality in Malaysia is described in terms of Air Pollutant Index (API). The API is an indicator of air quality and was developing based on scientific assessment to indicate in an easily understood manner, the presence of pollutants and its impact on health. The API system of Malaysia closely follows the Pollutant Standard Index (PSI) develop by the United States Environmental Protection Agency (US-EPA) [5].

The air pollutant index scale and terms used in describing the air quality levels are as follows:

TABLE I
THE API SCALE AND TERMS USED IN DESCRIBING THE AIR QUALITY LEVELS.

API scale	Air quality
0 – 50	Good
51 – 100	Moderate
101 – 200	Unhealthy
201 – 300	Very unhealthy
301 - 500	Hazardous
Above 500	Emergency

The Continuous Air Quality Monitoring (CAQM) stations measure the concentration of 5 major pollutants in the ambient air which are O_3, CO, NO_2, SO_2 and PM10. These concentrations are measured continuously on hourly basis. Then, all the concentrations have to be converted to a sub pollutant index, before the API can be determined. The sub pollutant index values are obtained from a set of equations given by the DOE [5].

Sub index for O_3 = Concentration *1000 (1)

Sub index for CO = Concentration * 11.11111 (2)

Sub index for NO_2 = Concentration * 588.23529 (3)

Sub index for SO_2 = Concentration * 2500 (4)

Sub index for PM10 = 50 + [(Concentration -50)*0.5] (5)

After the sub index was calculated, the highest index values among the 5 major pollutants is taken as the API value for the day.

B. Artificial Neural Network (ANN)

An artificial neural network (ANN) is a mathematical model or computational model based on biological neural networks. It consists of an interconnected group of artificial neurons and processes information using a connectionist approach to computation [6]. In most cases an ANN is an adaptive system that changes its structure based on external or internal information that flows through the network during the learning phase [7].

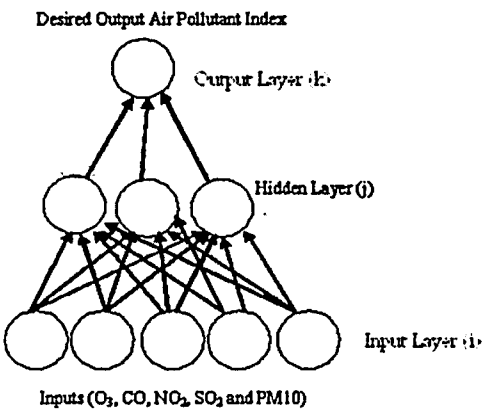

Figure.1. The structure of a multi-layer neural network.

The neural architecture used in this work is the feed-forward back-propagation consisting of only two layers of multiple neurons (the input and hidden layer) and of a single neuron in the output layer. The input layer is the only one made up of linear neurons. This architecture, where every non-linear neuron (of the hidden and output layer) is connected with every neuron of the previous layer by weighted links and is activated by the *sigmoid* transfer function in equation (6)

$$F(x) = 1/(1+e^{-x}) \quad (6)$$

where is citied as the most suitable for air pollution estimation [2]. And sigmoid is monotonic S-shaped function that maps numbers in the interval to a finite interval such as (-1,+1) or (0,1) [8].

The feed-forward refers to the propagation of the information through the net (from the input layer to the output neuron). The back-propagation refers to the learning algorithm, which with the aim of decreasing the error, proceeds by iterations from the output neuron back to the input layer [2], [9].

ANN is particularly useful for investigations on large data sets and for the problems with the input/output relationships only partially known. In fact, ANN can overcome these difficulties because they are model-free working under the only hypothesis that the inputs variables (experimental space) form an almost complete phase space [10], [11].

II. METHODOLOGY

The data used for forecast model using ANN is divided into two categories. The first data is known as training set and the second data is known as testing set. By using the Matlab R2008a, the training and the testing of ANN was carried out. A two-layer feedforward network is created. The first layer has two *tansig* neurons, and the second layer has one *purelin* neuron. The *Levenberg-Marquardt (trainlm)* network training function and the *Supervised Learning* network testing function are to be used.

A. Preprocessing of Data

The raw data for three months, beginning from October 2006 to December 2006 were taken for training and testing purpose. The data were divided into which is 76% & of the data is for training and 24 % percent for testing. These make up total of 1680 data sets for training purpose and 528 data sets for testing purposes. The data must be normalized in order to made the model easily convergence and fall in the range of zero to one.

B. Training of Data

The forecast model for API consists of five input nodes and one output nodes. For this project, five major air pollutant sub index were selected to be the input nodes. The input variables are sub index of O_3, CO, NO_2, SO_2 and PM10. The output variable is the Air Pollutant Index (API). The data was train until achieved the desired output. This training data pattern was fed to the network. The fully trained network was saved in order to use for testing purpose.

C. Testing of Data

Testing process is performed in order to check the performance of the trained network. In the testing process, it been carried out by feeding the saved trained network with series untrained data. During the testing process, 528 untrained data were fed into networks as a new input and the network provides the result.

It is important for the learning rate, the learning momentum, and the number of iteration to be tuned to the optimum in order to achieve rapid learning during the teaching process. Several combinations of learning rate, learning momentum and the number of iteration exist, which can be applied to attain the convergence of error and also the respective optimal values.

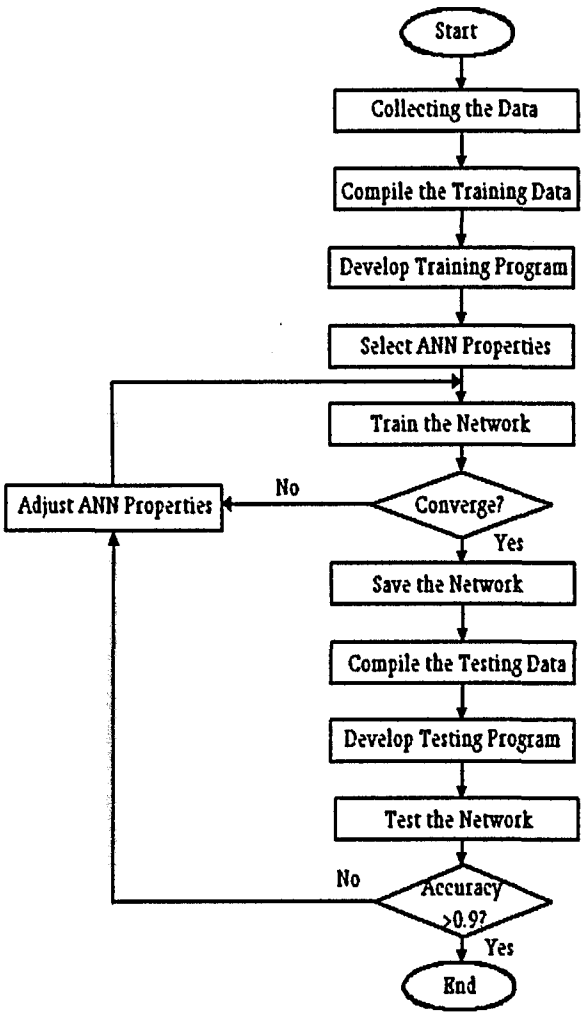

Figure.2. Flowchart for develop the ANN

The accuracy can be determine using equation (7)

$$\text{Accuracy} = \frac{\text{Actual Value} - \text{Estimated Value}}{\text{Actual Value}} \times 100\% \qquad (7)$$

III. RESULTS AND DISCUSSION

The estimation of Air Pollution Index of Klang Valley for October to December 2006 was obtained after training is evaluated by using the test data set. The test was carried out using data from Department of Environment. Four separate tests were carried out using data obtained from four different locations in Klang Valley. The training data set contained of 420 data for each four locations making a total of 1680 of training data set. For testing purposes, a total of 528 data is used in four separate tests.

Figure 3 shows the comparison between the actual value and estimated value of Shah Alam for test data set of October to December 2006. The optimum setting value from the training data set is tuned with the number of hidden nodes is 5, learning rate at 0.52, momentum at 0.77, and iteration at 5000. After the testing process, the performance met goal at 12 iterations and the accuracy of this estimation is found to be 90.33% to 93.56%.

Figure 4 shows the comparison between the actual value and estimated value of Klang for test data set of October to December 2006. The optimum setting value from the training data set is tuned with the number of hidden nodes is 8, learning rate at 0.32, momentum at 0.57, and iteration at 4000. After the testing process, the performance met goal at 21 iterations and the accuracy of this estimation is found to be 91.47% to 95.11%.

Figure 5 shows the comparison between the actual value and estimated value of Petaling Jaya for test data set of October to December 2006. The optimum setting value from the training data set is tuned with the number of hidden nodes is 7, learning rate at 0.42, momentum at 0.63, and iteration at 7000. After the testing process, the performance met goal at 15 iterations and the accuracy of this estimation is found to be 92.21% to 96.89%.

Figure 6 shows the comparison between the actual value and estimated value of Kuala Lumpur for test data set of October to December 2006. The optimum setting value from the training data set is tuned with the number of hidden nodes is 10, learning rate at 0.48, momentum at 0.33, and iteration at 6000. After the testing process, the performance met goal at 24 iterations and the accuracy of this estimation is found to be 90.51% to 96.21%.

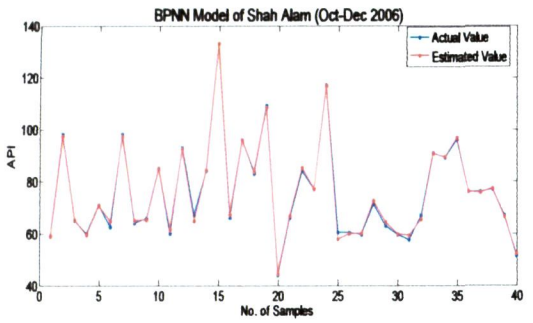

Figure.3. Actual and Estimated Value of API in Shah Alam for testing data set of October to December 2006

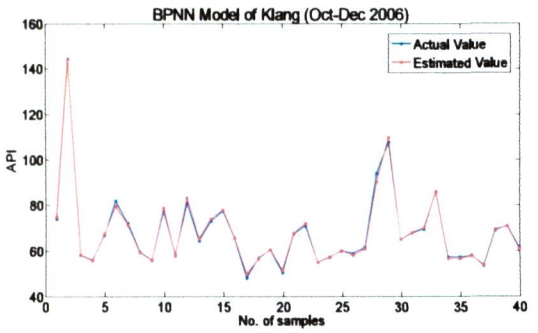

Figure.4. Actual and Estimated Value of API in Klang for testing data set of October to December 2006

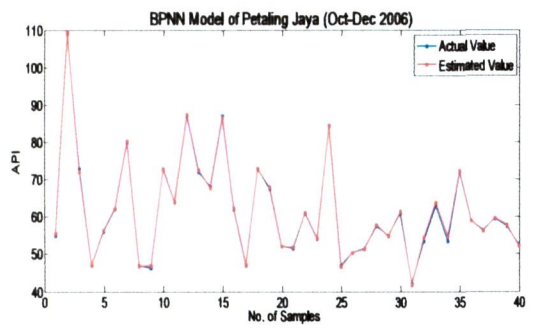

Figure.5. Actual and Estimated Value of API in Petaling Jaya for testing data set of October to December 2006

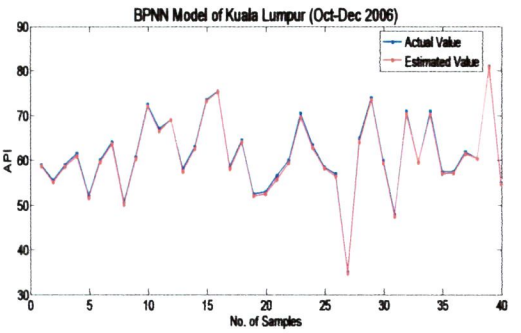

Figure.6. Actual and Estimated Value of API in Kuala Lumpur for testing data set of October to December 2006

It can be concluded that during the training process, the range of number hidden nodes is between 5-10, learning rate between 0.32-0.52, the learning momentum between 0.33-0.77 and the number of iteration are tuned within the range of 4000-7000 were adjusted to determine the best tuned set for optimal performance.

Backpropagation training with too small a learning rate will make agonizingly slow process. Too large a learning rate will proceed much faster, but may simply produce oscillations between relatively poor solutions. Typical values for the learning rate parameter are between 0 and 1.

Momentum can be helpful in speeding the convergence and the values are between 0 and 0.9. The number of iteration important in order to train the data until the performance goal met.

The performance met goal for testing process is between 12-24 iterations. The average API scale for Shah Alam, Klang, Petaling Jaya and Kuala Lumpur are between 51-100 and shows that the air quality is moderate.

From the result obtained using ANN algorithm, the BPNN Model for Petaling Jaya shows a good performance results lays between 92.21% and 96.89%.

IV. CONCLUSION

In this paper, the Back-propagation Neural Network (BPNN) model of Artificial Neural Network is used to estimate the Air Pollutant Index (API). This model was applied for four different locations in Klang Valley and good performance of estimation was obtained. The air pollution is closely related to health of human. Hence once air pollution index (API) is hazardous, it will affect the health of many people. By using this BPNN, we can able to give warning immediately if the air quality indicates that very unhealthy, hazardous or emergency which is more than 200 in API. Then, the BPNN are suitable for investigations

working on large data sets and for problems in which the inputs and corresponding values are known but the relationships between the inputs and the outputs are difficult to understand with usual analysis techniques. Besides that, if many parameters and air data will be taken, this BPNN will be forecast the future API value. It would also be possible of making the estimation of the API at CAQM stations itself. Therefore, this BPNN model can be used to support Department of Environmental (DOE) giving faster and efficient analysis.

V. FUTURE DEVELOPMENT

For the future development, it is recommended that more air data samples from monitoring station will be taken and tested. It will help the Artificial Neural Network (ANN) to estimate the Air Pollution Index (API) more accurate and improved. Besides the Back-propagation Neural Network (BPNN) method, other methods such as Radial Basis Function Network (RBFN), Regression Neural Network (RNN) and Modular Neural Network (MNN) can be used in order to compare the effectiveness in estimating the Air Pollution Index (API). Furthermore, the ANN also can be used to estimate the water quality index.

ACKNOWLEDGMENT

The author would like to thank Mrs. Norhayati Hamzah for her guidance and support in conducting this project. These thanks also goes to everyone who is directly or indirectly involve in contributing the ideas especially to Mrs. Mahanijah Md Kamal and Mrs. Kama Azura. And special thanks to Department of Environment (DOE) for their co-operations in providing the required air data samples.

REFERENCES

[1] F.Benvenuto and A.Marani, "Neural Networks for Environmental Problems: Data Quality Control and Air Pollution Nowcasting" Global Nest: the Int.J, vol.2, pp. 281-292, 2000.

[2] Mahanijah Md Kamal, Rozita Jailani and Ruhizan Liza Ahmad Shauri, "Prediction of Ambient Air Quality Based on Neural Network Technique," *4th Student Conference on Research and Development,* June 2006.

[3] Hamdy K.Elminir and Hala Abdel-Galil, "Estimation of Air Pollutant Concentrations from Meteorological Parameters using Artificial Neural Network," *Journal of Electrical Engineering,* vol. 57, no. 2, p.p 105-110, 2006.

[4] Bruno Ando, Salvatore Baglio, Salvatore Graziani and Nicola Pitrone, "Models for Air Quality Manangement and Assestment", *IEEE Transactions on Systems, Man and Cybernetics,*" vol. 30, no. 3, August 2000.

[5] "A Guide to Air Pollution Index in Malaysia (API)," Department of Environment (DOE), Ministry of Science, Technology and the Environment, 1997.

[6] S.Agatonovic-Kustrin and R. Beresford, "Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research," *Journal of Pharmaceutical and Biomedical Analysis,* vol. 22, pp. 717-727, Issue 5, June 2000.

[7] Amir F. Atiya, Suzan M. El-Shoura, Samir I. Shaheen, Mohamed S. El-Sherif and R. Beresford, "A Comparison Between Neural-Network Forecasting Techniques—Case Study: River Flow Forecasting," IEEE Transactions on Neural Networks, vol. 10, no. 2, March 1999

[8] Waylon Collins and Philippe Tissot, "Use of an Artificial Neural Network to Forecast Thunderstorm Location: Performance Enhancement Attempts", *J3.4,* American Meteorological Society Conference, 2001.

[9] Okyay Kaynak and Serdar Iplikci, *"An Algorithm for Fast Convergence in Training Neural Networks,"* Bogazici University, Electrical and Electronics Engineering Department, Bebek, 80815, Istanbul, Turkey.

[10] Andrew C. Comrie, University of Arizono, Tucson, Arizona, *"Comparing Neural Networks and Regression Models for Ozone Forecasting,"* Journal of the Air & Wate Management Association, vol.47, June 1997, 653-663.

[11] Charalambous, C., "*Conjugate gradient algorithm for efficient training of artificial neural networks,*" IEEE Proceedings, Vol. 139, No. 3, pp. 301-310, June 1992.