Indian Journal of Science and Technology, Vol 8(35), DOI: 10.17485/ijst/2015/v8i35/56235, December 2015
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Prediction of Air Pollution in Tehran based on Evolutionary Models

Masoume Asghari Esfandani and Hossein Nematzadeh*

Department of Computer Engineering, Sari Branch, Islamic Azad University, Sari, Iran;
Asghari.masoume@yahoo.com, nematzadeh@iausari.ac.ir

Abstract

With respect to the increasing problems of air pollution due to urban development, pollution control is necessary. The purpose of this study is to predict the density of particulate matter less than 10 microns (PM_{10}), to plan and reduce its effects and to avoid reaching a crisis situation. For this purpose, the data of air pollutants and meteorological parameters recorded at Aghdasiyeh Weather Quality Control Station and Mehrabad Weather Station in Tehran were used as input parameters. Next, Artificial Neural Network with Back Propagation (BP), its hybrid with GA (BP-GA) and PSO (BP-PSO) were used and ultimately the performance of these three models was compared with each other. It was concluded that BP-PSO has the highest accuracy and performance. In addition it was also found that the results are more accurate for shorter time periods and this is because the large fluctuation of data in long-term returns negative effect on network performance. Also unregistered data have negative effect on predictions.

Keywords: Air Pollution, Algorithm, Artificial Neural Networks, Genetic Algorithm, Particle Swarm Optimization PM10, Tehran

1. Introduction

Air pollution is one of the world's problems with the development of industrialization and increasing the number of cities, the amount and intensity is increasing day by day[1]. Tehran's main air pollutants include: CO, SO_2, HC, O_3, NO_X and PM that 80% of car fuel and the remainder are created by factories and homes heating equipment. One of the most effective actions to control and reduce air pollution is to estimate the pollutants density and to describe the state of air quality in comparison with the standard conditions[2]. This paper tries to estimate and predict the air pollution of Tehran with three approaches. First basic ANN was used with randomly generated weights. Second, GA was applied to generate the initial weights of ANN. Third PSO was used to produce the initial weights. The results finally showed that the hybrid method of PSO and ANN have better performance.

Neural Networks or more specifically artificial Neural Networks rooted in many fields of Science. Neurology, Mathematics, Statistics, Physics, Computer Science and Engineering are examples of mentioned sciences[1–3]. Most recently Multi Layer Perceptron (MLP) has been widely used to predict pollutants so that in most large cities around the world for MLP has been used to predict air pollutant. The results of several studies that have been done in this context also show that the performance of Neural Networks is better in comparison with traditional statistical methods such as multivariate regression and auto regression models4. In[4] developed a method to predict Uberlandia Brazil air pollutions using Neural Networks. The research direction in the field tends develop tools for modeling the distribution of air pollution in near future. In[5] tried to predict PM_{10} hourly concentration using neural networks in four major stations in Athens. In6 have done the same research in Milan Italy using Artificial Neural Network. In[7] proposed a two day ahead prediction with concentration on five particles in Palermo Italy. In Belgium country Data between 1997and 2000 have used to predict the average concentration of particulate matter

*Author for correspondence

for next day and there were some efforts to predict the air pollution index in Shanghai and Santiago using Neural Network as well. In[8] presented a model based on neural network which was able to predict daily average concentration of PM_{10} in a densely populated area of Tehran. The method had warning system in order to reduce their unnecessary trips in polluted areas in Tehran. In[9] proposed an Artificial Neural Network Model to predict the annual PM_{10} greenhouse gas emissions. In that research Artificial Neural Networks was trained by using following variables: Gross domestic product, Gross domestic energy consumption, Burning wood, the motorized, manufacture of paper and paper board, production of sawn timber, production of copper, production of aluminum, production of pig iron and crude steel production. The results show a very good performance of the ANN model in contrast to the Multivariate regression model.

2. Methods

In this section the techniques of Artificial Neural Network, genetic algorithm and particle swarm optimization which have been used in research, briefly presented and introduced. All the implementations and simulations have been done through MATLAB Toolbox.

2.1 Back-Propagation Neural Network

The neural network model is built to estimate air pollution from forward multilayer network with back-propagation learning algorithm, which is a supervised learning method. The network structure consisting of an inputl ayer, a middle layer and an output layer. Figure 1 shows the proposed back-propagation neural network model. Table 1 shows the characteristics of neural networks.

After training (training ends after 25 epochs), ANN would be tested with unused data in the training phase and consequently the results and network performance would be assessed.

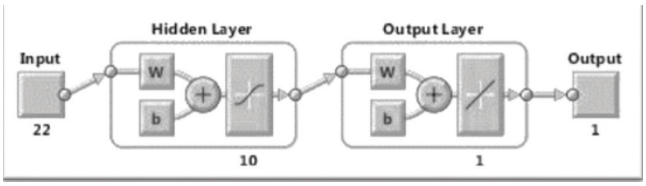

Figure 1. Model of Back-Propagation neural network (BP).

Table 1. Specification of Back-Propagation neural network (BP)

Trial and error (2-10)	The number ofneurons inmiddle layer
3	Trying
Sigmoid	Activationfunction ofthe hiddenlayer
Linear	Activationfunctionof outputlayer
levenberg marquardt function	Training the network
max_fail=6	Stop condition

2.2 Genetic Algorithm

Since the back propagation error algorithm is very slow for real problems, genetic algorithm is used to select the initial weight. In other words, by using neural network and combining it with genetic algorithm the performance (speed of achieving better solutions) and precision of results would be increased. In this research, in both training and testing phase of genetic algorithm was used to optimize the basic ANN behavior. The objective function is Z = fit nn (w), in which the input are the initial weights that should be calculated and the output is the summation of errors that should be minimized. The specifications of the GA used in the hybrid approach of ANN – GA is presented in Table 2. Figure 2 shows the development of genetic algorithm during 300 generations, the black dots are the best of the 20 chromosomes the blue dots are the average of 20 chromosomes in each generation.

Genetic algorithm calculates the initial weights for using in Artificial Neural Network. After training (Training test ends after 20 epochs), Network with data that is not used in the training would be assessed and its performance would be checked using statistical index.

2.3 PSO Algorithm

As what has been done for GA, PSO algorithm has also been used to obtain the initial weights of the neural network. All the particles have a fitness value that are achieved through objective. The objective function is Z = fit_nn (w), in which the input are the initial weights that should be calculated and the output is the summation of errors that should be minimized. The specifications of the PSO used in the hybrid approach of ANN – PSO is presented in Table 3.

PSO calculates the initial weights for using in Artificial Neural Network. After training (Training test ends after 17 epochs), network with data that is not used in the training would be assessed and its performance would be checked using statistical index. The general structure and methodology of the research was shown in Figure 3. In this study, a neural network trained with random weights and a time Using Genetic Algorithm Paid to the training of the neural network and again using PSO algorithm the neural network was trained. The general structure and the methodology of the research was presented in Figure 3.

Table 2. Specification of (ANN +GA)

Array ofreal numbers	View (encoded) chromosome
20	The initial population
300	Number of generations
0.8	Probability of crossover
0.03	Probability of mutation
Z=fit_nn(w)	The objective function
Roulette wheel	Selection function
number of generations=300	Stop condition

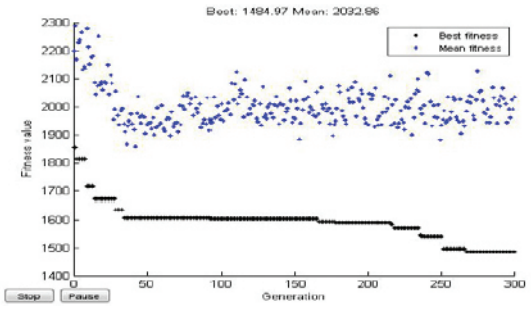

Figure 2. The development of genetic algorithms.

Table 3. Specification of (ANN +PSO)

20	The initial population
Z=fit_nn(w)	The objective function
2	c1 , c2
1	W
0.99	wdamp
300	The number of iteration

3. Simulation and Evaluation

The information in Aghdasiyeh Weather Quality Control Station and Mehrabad Weather Station in Tehran from 2007 to 2013 was collected as a real case study in this paper. Aghdasiyeh station was selected because it had more complete course of records in its database. Table 4 shows the location of the stations under study.

The information of 2400 days (from 2007 to 2013) was used. Eleven parameters have been selected as input parameters to our models. These parameters were year, month, day, minimum temperature, mean temperature, maximum temperature, humidity, velocity, day of week, holidays from Mehrabad station and PM_{10} from Aghdasiyeh station. To clean existing data and review the situation and quality control the following preprocessing issues were considered:

- Controlling suspicious data and their comparison with the same data in previous and following days.
- On some days, air pollution data were not registered which leads to have a gap. This can happen due to a mistake in the data recording device. These data were excluded from the study. Thus the information of 2400 days decreased to 1362 days means that 1038 days air pollution data were not recorded.

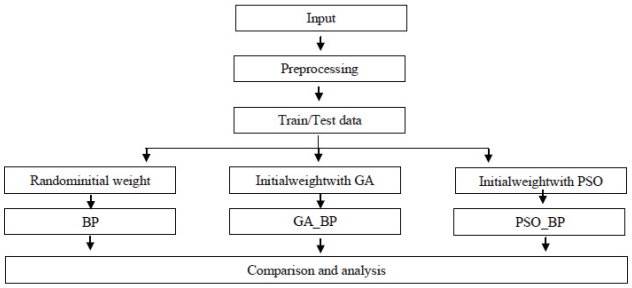

Figure 3. The general structure of the algorithm.

Table 4. Aghdasiyeh and Mehrabad stations

Station	Stationlocation	Latitude	Longitude
Aghdasiyeh	Nobonyad Plaza, Shahid Langari Road	43.75° 40′ 35′′	15.12° 20′ 51′′
Mehrabad	In the vicinity of northern Tehran, Shahid langari Roadside	35° 47′ 57′′	5° 29′ 7′′

- Normalizing data through conversion to range of [0, 1]. Normalization of data prevents to have larger weights. To do so, the Equation 1 is used:

$$X = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

Where x_{min} is minimum and x_{max} the maximumin putvectorx and Xisits normalization. The input data after preprocessing were divided to train and test data. 80% percent of the input data were selected as training set (almost 1090 individuals) and 20% have been selected as testing set (almost 272 individuals). The next step is assessment and evaluation of the accuracy of the models. The evaluation is done based on four famous criteria: Mean Square Error, Root Mean Square Error, Mean Absolute Error and assessment coefficient (R2) which are shown in Equations 2, 3 and 4.

$$RESE = \sqrt{\sum_{i=1}^{n} \frac{(c_i - m_i)^2}{n}}, MSE = (RMSE)^2 \quad (2)$$

In which C_i is optimal value that has been estimated by the model, m_i the amount which has been calculated and n the number of data pairs which have been observed. RMSE value is usually positive and the ideal value equals to zero. The algebraic sign of the MAE indicates either the error value is positive or negative. In Equation 3, assuming MAE is positive (negative) shows that the estimated value is higher (lower) than the measured value. The ideal value equals to zero. In Equation 4, R^2 shows the dependence between two data groups. The ideal value for *R2* equals to one. The closer *R2* to one, more dependent the data groups are:

$$MAE = \sum_{i=1}^{n} \frac{(c_i - m_i)}{n} \quad (3)$$

$$R^{2=}\left[1 - \frac{\sum_{i=1}^{n} |()c_i - m_i|^2}{\sum_{i=1}^{n} (m_i)^2}\right] * 100 \quad (4)$$

For simulation and implementation purpose Microsoft Excel 2013was used for pre-processing and data preparation (eliminate suspicious cases, data normalization, etc.) as well as Matlab 2013 for implementing ANN with Back Propagation (BP), ANN with Back Propagation (BP) and its hybrid with GA (BP-GA), ANN with Back Propagation (BP) and its hybrid with PSO (BP-PSO). The evaluation of three methods was shown in Table 5. According to Table 5 BP-PSO is the best model among the three models since it has the smallest amount of MSE, RMSE and MAE in testing set. It also has the greatest R^2.

Table 5. Evaluation of BP, BP-GA and BP-PSO models

	TRAIN				TEST			
	R^2	MSE	RMSE	MAE	R^2	MSE	RMSE	MAE
BP_PSO	0.69275	846.2977	29.0912	19.3619	0.5573	1648.7025	40.6042	23.7135
BP_GA	0.74823	714.7516	26.7348	17.9929	0.54889	1756.7358	41.9134	25.7154
BP	0.69793	832.0611	28.8455	18.7362	0.53932	1778.8447	42.1764	25.5921

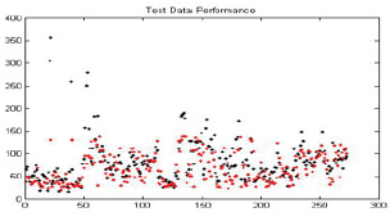

Figure 4. Distribution of test data in the BP-PSO.

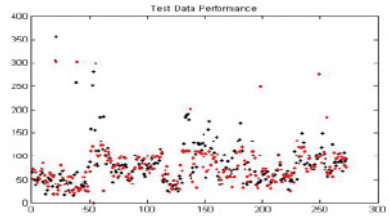

Figure 5. Distribution of test data in the BP-GA.

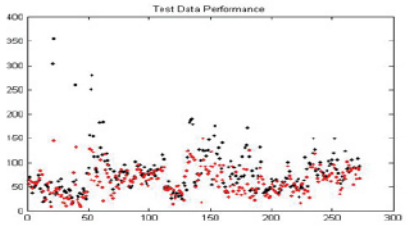

Figure 6. Distribution of test data in the BP.

4. Conclusions and Suggestions

In this paper three models have been proposed to predict Tehran air pollution based on information from Aghdasiyeh Weather Quality Control Station and Mehrabad Weather Station. The accuracy and performance of the three models in decreasing order can

be placed as: BP-PSO, BP-GA and BP. In other words, the error rate increases. The lack of input data does not affect the predictive ability of the models considerably. It should be mentioned that having more input data and solving the problem of data fluctuation could led to have better predictions. One of the main limitations of the research is that the prediction models have more accurate results for shorter period of time rather than longer period of time. Two future works are identified for this research. First, more input data can be fed to the network in order to have better accuracy. This research mostly focused on PM_{10}. Second, other heuristic algorithms like swarm intelligence algorithms can be used to increase the performance and accuracy.

5. References

1. Dimitriou K, Paschalidou A, Kassomenos P. Assessing air quality with regards to its effect on human health in the European Union through air quality indices. Ecological Indicators. 2013 Apr; 27:108–15.
2. Nayak PC, Sudheer KP, Rangan DM, Ramasastri S. Short-term flood forecasting with a neuro fuzzy model. Water Reso Res. 2005 Apr; 41(4).
3. Haykin S. Neural network, a comprehensive foundation, Prentice Hall International Inc, Second Edition.1999.
4. Lira TS, Barrozo MAS, Assis AJ. Air quality prediction in Uberlandia, Brazil using linear models and neural networks. 17th European Symposium on Computer Aided Process-Eng. 2007.
5. Grivas G, Chaloulakou A. Artificial Neural Network model for prediction of PM_{10} hourly concentration in Great Area of Athens, Greece. Atomospheric Environ. 2006 Mar; 40(7):1216–.29.
6. Cecchetti M, Corani G, Guariso G. Artificial Neural Network Prediction of PM_{10} in the Milan area, Inte IEMSS International Congress Osnabrack. 2004.
7. Bruelli U, Piazza V, Pignato L, Sorbello F, Vitabile S. Two days ahead prediction of daily maximum concentration of SO_2, O_3, PM_{10}, NO_2, CO in the urban area of Palermo, Italy, Atom Env. 2007 May; 41(14):2967–95.
8. Nejadkoorki F, Baroutian S. Forecasting extreme PM_{10} concentrations using Artificial Neural Networks. 2012; 6(1):277–84.
9. Antanasijević DZ, Pocajt VV, Povrenović DS, Ristić MD, Perić-Grujić AA. PM10 emission forecasting using Artificial Neural Networks and genetic algorithm input variable optimization. Science of the Total Environment. 2013 Jan; 443:511–9.