Journal of AI and Data Mining
Vol 4, No 1, 2016, 49-54 ***10.5829/idosi.JAIDM.2016.04.01.06***

Predicting air pollution in Tehran: Genetic algorithm and back propagation neural network

M. Asghari Esfandani and H. Nematzadeh*

Department of Computer Engineering, Sari Branch, Islamic Azad University, Sari, Iran.

Received 03 January 2015; Accepted 18 November 2015
**Corresponding author: nematzadeh@iausari.ac.ir (H. Nematzadeh).*

Abstract
Suspended particles have deleterious effects on human health and one of the reasons why Tehran is effected is its geographically location of air pollution. One of the most important ways to reduce air pollution is to predict the concentration of pollutants. This paper proposed a hybrid method to predict the air pollution in Tehran based on particulate matter less than 10 microns (PM_{10}), and the information and data of Aghdasiyeh Weather Quality Control Station and Mehrabad Weather Station from 2007 to 2013. Generally, 11 inputs have been inserted to the model, to predict the daily concentration of PM_{10}. For this purpose, Artificial Neural Network with Back Propagation (BP) with a middle layer and sigmoid activation function and its hybrid with Genetic Algorithm (BP-GA) were used and ultimately the performance of the proposed method was compared with basic Artificial Neural Networks along with (BP) Based on the criteria of - R^2-, RMSE and MAE. The finding shows that BP-GA $R^2 = 0.54889$ has higher 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. Microsoft Excel and Matlab 2013 conducted the simulations.

Keywords: *Artificial Neural Networks, Genetic Algorithm, Air Pollution, PM_{10}.*

1. Introduction

Air pollution IS one of the biggest environmental problems in Tehran. Several factors are involved in Tehran pollution and their geographical factors are more important. Daily large amounts of toxic gases, Types of pollutants, Suspended hazardous materials vehicles, Factories, industrial sites, power plants and refineries, and numerous residential units are added to the air of the city [12]. Air pollution is a serious risk to the environment and causes serious respiratory and skin diseases, especially for the elderly and children. The environmental and health problems caused by air pollution in large cities have become a major challenge.[10] Air pollution is one of the world's problems with the development of industrialization and with increasing the number of cities, the amount and intensity day by day. [6] Tehran's main air pollutants include: CO, SO2, HC, O3, NOX and PM that 80% of car fuel and the remainder are created by factories and homes heating equipment. Particulate matter affects on human health, such as the impact of the lungs, respiratory system, asthma and deaths. It was reported that the detrimental effects of particulate matter on human health, is mostly due to being exposed to concentrations of particles. 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 [7]. This paper tries to estimate and predict the air pollution of Tehran with two approaches. First, basic ANN was used with randomly generated weights. Second, GA was applied to generate the initial weights of ANN. The results finally showed that the hybrid method of GA and ANN have better performance. The structure of the paper is as follows: Section two introduces algorithms and techniques in the literature related to the study. Section 3 describes the main methodology of the research. Section 4 discusses the research model and its estimation method and research databases and also the results

are presented. Finally, section 5 provides conclusions and future works.

2. Literature review

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 [5,6,7]. 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 these methods that have been applied for different pollutants are good. 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 models [1]. Taisa and Barrozo (2007) developed a method to predict Uberlandia Brazil air pollutions using neural networks [1]. The research direction in the field tends develop tools for modeling the distribution of air pollution in near future. Gryvas and Chaloulakou (2006) tried to predict PM10 hourly concentration using neural networks in four major stations in Athens [3]. Cecchetti et al (2004) have done the same research in Milan Italy using Artificial Neural Network [4] Bruelli et al (2007) proposed a two-day ahead prediction with concentration on five particles in Palermo Italy [2]. In Belgium country Data between 1997 and 2000 have used to predict the average concentration of particulate matter for the next day and there were some efforts to predict the air pollution index in Shanghai and Santiago using neural network as well. Nejadkoorki and Baroutian (2012) 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 [8]. The method had a warning system in order to reduce their unnecessary trips in polluted areas in Tehran. Davar et al (2013), proposed an Artificial Neural Network Model to predict the annual PM_{10} greenhouse gas emissions. In that research artificial neural networks, were trained by using following variables: Gross domestic product, Gross domestic energy consumption, Burning wood, the motorized, manufacture of paper and paperboard, 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. [9] Information about the three stations Fatemi and Aghdasie and bazar is intended to predict PM_{10}. During the years 1779-1781, the neural network is used MLP. The answers are compared with the values obtained from multivariate regression model and the results represent MLP method is superior [11].

3. Proposed methods

In this section, the techniques of artificial neural network, and genetic algorithm used in research are briefly presented and introduced.

3.1. 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 consists of an input layer, with 22 neurons (11*2=22), in which we have 11 variables and 2 is the number of days of study (Our goal is to use the data from yesterday and today to predict PM_{10}). The output layer represents the concentration of PM_{10}. The number of neurons of the intermediate layer is calculated by trial and error, The number of neurons in hidden layer will vary from 2 to 10 (Trying=3) and Each test is done 3 times. Finally, we compare these with the best, and each one was better, the number of hidden layer neurons is. When the sixth consecutive epoch had this error are increased, train stops. Also an output layer represents the concentration of PM_{10}. Figure 1 shows the proposed back-propagation neural network model. And 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.

3.2. Genetic algorithm

Since the back propagation error algorithm is very slow for real problems, genetic algorithm is used to select the initial weight. Genetic algorithm is a heuristic optimization method, which acts on the basis of evaluation in nature and searches for the final solution among a population of potential solutions [13].

In other words, using neural network and combining it with genetic algorithm the performance (speed of achieving better solutions) and precise results would be increased. Indeed, as we have our own neural network, this time Genetic Algorithm calculates its initial weight. 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 BP-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. The general structure and the methodology of the research are presented in figure 3.

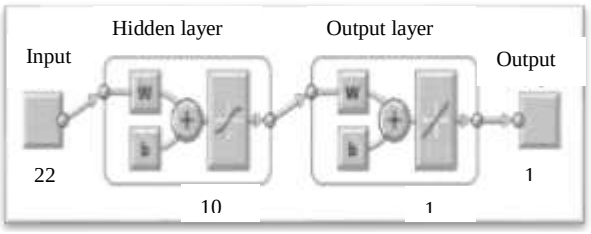

Figure 1. Model of back-propagation neural network (BP).

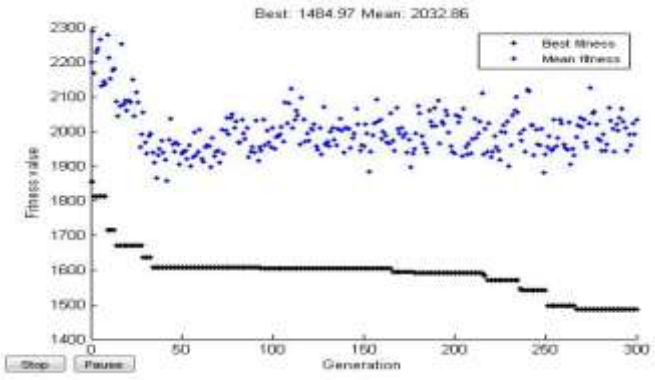

Figure 2. The development of genetic algorithms.

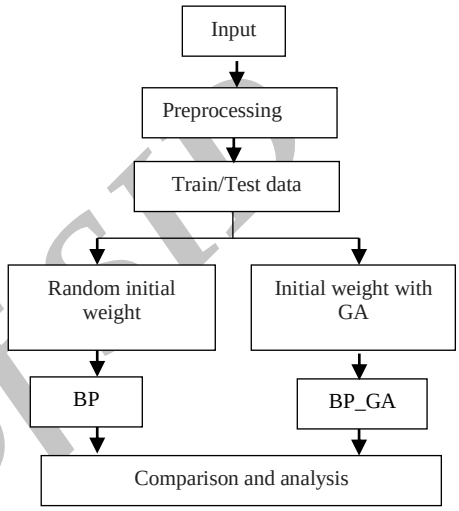

Figure 3. Research methodology.

Table 1. Specification of (BP).

Value	Concept
Trial and error (2-10)	The number of neurons in middle layer
3	Trying
Sigmoid	Activation function of the hidden layer
Linear	Activation function of output layer
levenberg marquardt function	Training the network
max_fail=6	Stop condition

Table 2. Specification of (BP - GA).

Value	Concept
Array of real 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

4. Simulation and evaluation

Not all air quality monitoring stations have recorded a continuous concentration of pollutants; therefore, in this study, those stations have been ignored. Aghdasiyeh station period only was chosen because it has a more complete course of data in hours (2000-2014). The meteorological station and the airport having a period of more perfect location closer to the air pollution monitoring stations, were studied (2007-2013) The data were recorded on a daily basis. When two air pollution stations and air quality data with our accession will join and make our final data from 2007 to 2013. Unique factors rainfall,

relative humidity, wind speed, temperature play a decisive role on the spread of pollutants, especially particulate matter. For example, humidity has a negative impact on air particulates. Input variables to network included variable time (Day , month , year , weekdays and holidays) and meteorological variables (Minimum temperature , average temperature , maximum temperature , humidity , wind speed and PM_{10}).The difference between these two sets of data is that, time variables do not need prediction when working with models and they can basically inserted to the model as inputs. However, meteorological data should be inserted to the model as predicted data. 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 3 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, week day, holidays from Mehrabad station and PM10 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 led to a gap were not registered. 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.
- Normalizing data through conversion was to a range of [0, 1]. Normalization of data prevents to have larger weights. To do so, (1) is used:

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

where, x_{min} is minimum and x_{max} the maximum input vector x, and X is its 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 a testing set (almost 272 individuals). The next step is to assess and evaluate 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) shown in (2),(3), and (4).

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

In which Ci is optimal value that has been estimated by the model, mi 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 the error value is positive or negative. In (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 (4), R^2 shows the dependence between two data groups. The ideal value for R^2 equals to one. The closer R^2 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 2013 was 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). The evaluation of two methods was shown in tables 4 and 5. According to tables 4 and 5 BP-GA is the better model among in comparison with BP since it has the smallest amount of MSE, RMSE, and MAE in the testing set. It also has the greatest R^2. Figures 4 and 5 show the distribution of test data in three models in which black dots are real answers and red dots are predictions.

5. Conclusions and suggestions

As discussed in previous studies [11] to forecast air pollution in Tehran on ANN and linear regression were used, and artificial neural networks and genetic algorithms to predict the composition of PM_{10} in Tehran have not been used. The results of this paper can be compared with the results of previous studies. This is because both types of input data as well as the methods are similar but, the comparison

algorithms vary. In this regard, the paper on Tehran air quality prediction using a combination of neural networks and genetic algorithms better than the results obtained in [11].

They say the results of the 2006 $R^2 = 0.57$ for 2005 and 2006 are separated $R^2 = 0.54$ and for 2004 to 2006 apart seized $R^2 = 0.5$ and concluded that the results for one year is better than two years and three years, as two years are better than three years.

Table 3. Aghdasiyeh and Mehrabad stations.

Station	Station location	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	35º 47′ 57′′	5º 29′ 7′′

Table 4. Evaluation of BP, BP-GA (Train phase).

Method	TRAIN			
	R^2	MSE	RMSE	MAE
BP_GA	0.74823	714.7516	26.7348	17.9929
BP	0.69793	832.0611	28.8455	18.7362

Table 5. Evaluation of BP, BP-GA (Test phase).

Method	TEST			
	R^2	MSE	RMSE	MAE
BP_GA	0.54889	1756.7358	41.9134	25.7154
BP	0.53932	1778.8447	42.1764	25.5921

However, we examined data for seven years (2007-2013) and found that R^2 =0.54889. It is clear that with neural network $R^2 = 0.53$ and the combination of neural networks and genetic algorithms $R = 0.54889$ and because we got better results data for seven years rather than. One, two, and three years. The results are shown in table 6.

In this paper, two 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 two models are decreased respectively: 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 lead 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 more accurate result. This research mostly focused on PM_{10}. Second, other heuristic algorithms like swarm intelligence algorithms can be used to increase the performance and accuracy.

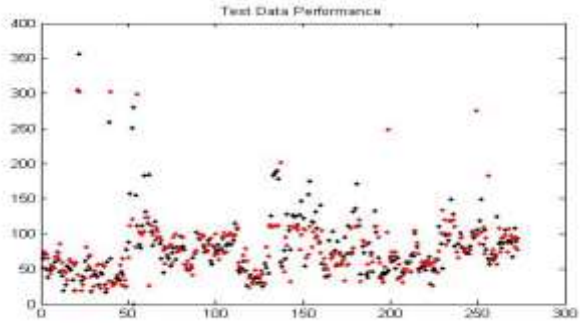

Figure 4. Distribution of test data in the BP-GA.

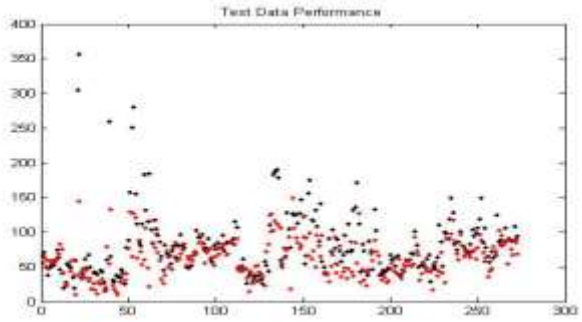

Figure 5. Distribution of test data in the BP.

Table 6. Comparison of previous studies and proposed methods.

Location	RMSE	R^2	*Method*	Year
Proposed method	41.9134	0.54889	BP_GA	2007-2013
	42.1764	0.53932	BP	2007-2013
[11]	46.2596	0.575	BP	2006
	54.3565	0.546	BP	2006 – 2005
	54.7014	0.5	BP	2007 – 2005

References

[1] Lira, T. S., Barrozo, A. S. & Assis A. J. (2007). Air quality prediction in Uberlandia, Brazil, using linear models and neural networks, 17th European Symposium on Computer Aided Process-Eng, vol. 24, pp. 51-56.

[2] Bruelli, U., Piazza, V., Pignato, L., Sorbello, F. & Vitabile, S. (2007). Two days ahead prediction of daily maximum concentration of SO2, O3, PM10, NO2, CO in the urban area of Palermo, Italy, Atom. Env, Vol. 41, no. 14, pp. 2967-2995.

[3] Grivas, A. & Chaloulakou. A. (2006). Artificial neural network model for prediction of PM10 hourly concentration, In Great Area of Athens, Greece, Atomospheric Environ, vol. 40, pp. 1216-1229.

[4] Cecchetti, M., Corani,G. & Guariso, G. (2004). Artificial neural network prediction of PM10 in the

Milan area, Inte. IEMSS International Congress Osnabrack.

[5] Haykin, S. (1999). Neural network, a comprehensive foundation, Prentice Hall International Inc, Second Edition.

[6] Dimitriou, K., Paschalidou, A. & Kassomenos, P. (2013). Assessing air quality with regards to its effect on human health in the European Union through air quality indices, EcologicalIndicators, vol. 27, pp. 108-115.

[7] Nayak, P., Sudheer, K. P., Rangan, D. M. & Ramasastri, K. S. (2005). Short-term flood forecasting with a neuro fuzzy model, Water Reso Res, pp. 2517-253.

[8] Nejadkoorki, F. & Baroutian, S. (2012). Forecasting Extreme PM10 concentrations Using Artificial Neural Networks", Pages: 277-284.

[9] Davor, Z.A., Viktor, V.P., Dragan, S.P., Mirjana, Đ. R. & Aleksandra, A. P. (2013). PM10 emission forecasting using artificial neural networks and genetic algorithm input variable optimization", Science of the Total Environment, Pages: 511-519.

[10] Tavakoli, M, & Esmaili,A. (2013). Artificial neural networks to predict the concentration of particulate matter air of Tehran, The second national conference on environmental protection and coarser, Hamedan.

[11] Soltaniye, M, Moslehi,P. & Yari, M. (2012). The concentration of suspended particles in the air in Tehran predicted by neural network models and compare multiple regression model, Sanati Sharif University, Tehran.

[12] Rahimi, N, (2012). The effect of geographic factors on air pollution in Tehran, Air Pollution Management Conference, Sharif University.

[13] Majidnezhad, V. (2014). A novel hybrid method for vocal fold pathology diagnosis based on Russian language, Journal of AI and Data Mining, Shahroud university , vol. 2, no. 2, pp. 141-147.

نشریه هوش مصنوعی و داده کاوی

دانشگاه شاهرود

پیش‌بینی آلودگی هوای تهران: الگوریتم ژنتیک و شبکه عصبی پس انتشار

معصومه اصغری اسفندانی و حسین نعمت زاده *

گروه مهندسی کامپیوتر، شعبه ساری، دانشگاه آزاد اسلامی، ساری، ایران.

ارسال ۲۰۱۵/۰۱/۰۳؛ پذیرش ۲۰۱۵/۱۱/۱۸

چکیده:

ذرات معلق اثرات زیانباری بر روی سلامتی دارد و یکی از دلایلی که تهران آسیب‌پذیر است موقعیت جغرافیایی آن است. یکی از راه‌های مهم کاهش آلودگی هوا پیش‌بینی غلظت آلاینده هاست. این مقاله یک روش ترکیبی برای پیش‌بینی آلودگی هوای تهران بر اساس ذرات کمتر از ۱۰ میکرون ارائه داده است. اطلاعات و داده‌ها از ایستگاه کنترل کیفیت هوای اقدسیه و ایستگاه هواشناسی مهرآباد از سال ۲۰۰۷ تا ۲۰۱۳ جمع آوری شده است. بطور کلی برای پیش‌بینی غلظت روزانه ذرات معلق مدل پیشنهادی دارای ۱۱ ورودی است. برای این منظور شبکه عصبی مصنوعی پس انتشار با یک لایه پنهان و تابع فعال سازی سیگموید و ترکیب آن با الگوریتم ژنتیک استفاده شد و نهایتا کارایی روش ترکیبی پیشنهادی با شبکه عصبی بر اساس معیارهای R^2,RMSE, MAE مقایسه شد. یافته‌ها نشان می‌دهد که روش ترکیبی دارای $R^2 = ۰/۵۴۸۸۹$ می‌باشد و دارای دقت و کارایی بالاتری است. همچنین نشان داده شد که نتایج برای بازه‌های زمانی کوتاه تردقیق‌تر است که این به دلیل نوسان بالای داده‌ها در مدت زمان زیاد است که تاثیر منفی بر روی کارایی شبکه می‌گذارد. همچنین داده‌های ثبت نشده نیز تاثیر منفی بر روی پیش‌بینی دارد. مایکروسافت اکسل و متلب ۲۰۱۳ برای شبیه‌سازی استفاده شده‌اند.

کلمات کلیدی: شبکه عصبی مصنوعی، الگوریتم ژنتیک، آلودگی هوا، ذرات کمتر از ۱۰ میکرون.