International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 4 Issue: 2 153 - 157

Data Mining In the Prediction of Impacts of Ambient Air Quality Data Analysis in Urban and Industrial Area

S. Christy
Research Scholar, Dept. of C.S.E.
BIHER University
Chennai, Tamil Nadu, India
christymelwyn @ gmail.com

Dr. V. Khanaa
Dean Information,
BIHER University
Chennai, Tamil Nadu, India
Drvkannan62@yahoo.com

Abstract—Air pollution caused due to the introduction of particulate matters, biological molecules and other harmful materials into the Earth's atmosphere. Pollution brings vital diseases, death to humans, damages other living organisms such as vegetations, animals, natural environment and built environment. Data mining concerned with finding hidden patterns inside largely available data, so that the information retrieved can be transformed into usable knowledge. The Air Quality Index is an indicator of air quality standards around Chennai. It is based on air pollutants that have bad effects on human health and the environment. Growing use of vehicles in the city and growing industrial activities on the outskirt of city cause more air pollution. The problem of air pollution is becoming a major concern for the health of the population. The ambient air quality data collected from Central Pollution Control Board and Tamil Nadu Pollution Control Board ambient air quality data available in the websites. Air quality is monitored by air quality monitoring stations in Chennai through the use of wireless sensors deployed in huge numbers around the city and industrial areas. The four years of data from the year 2012 to 2015 are collected from various monitoring stations and processed. Data mining tool is used for the prediction, forecasting and support in making effective decision. Artificial Neural Network model in Data mining techniques analyzed the data using Feed Forward Neural Networks and Multilayer Perceptron neural network models. The pattern obtained from these models could serve as an important reference for the Government policy makers in devising future air pollution standard policies.

Keywords- *Data mining; Data analysis; Monitoring stations; Decision Support*

*****

I. INTRODUCTION

Data mining, known as knowledge discovery in databases (KDD) is the process of discovering useful knowledge from large amount of data stored in databases, data warehouses, or other information repositories[2]. Data understanding starts with data collection and proceeds with activities to identify data quality problems, and to discover missing values into the data. Data preparation construct the data to be modelled from the collected data. The modelling phase applies various modelling techniques, and determines the optimal values for parameters in models. The evaluation phase evaluates the model for the problem requirements[4]. Data mining technology is used to identify the national air quality distribution of Chennai, whose hourly air quality data are continuously collected and archived through a network of several stations. Major composition of air pollution are Suspended particulate matter(PM_{10},$PM_{2.5}$), sulphur dioxide(SO_2), oxides of nitrogen(NOX), carbon monoxide(CO), volatile organic compounds, sulphur trioxide(SO_3) and lead(PB). Four years data collected from CPCB and TNPCB are processed and analyzed with data mining techniques and provide decision support to policy makers.

II. AIR POLLUTION MONITORING SYSTEM

Air pollution monitoring system is considered as a very complex task but it is very important. Traditionally data collectors were used to collect data periodically and this was very time consuming and quite expensive. The use of Wireless Sensor Networks can make air pollution monitoring less complex and more instantaneous readings can be obtained[7]. Currently, the Air Monitoring Unit in Chennai lacks resources and makes use of bulky instruments. This reduces the flexibility of the system and makes it difficult to ensure proper control and monitoring. Air Quality Modelling is an attempt to predict or simulate the ambient concentrations of contaminants in the atmosphere. These models are used as quantitative tools to correlate cause and effect of concentration levels found in an area. They are also used to support laws and regulations designed to protect air quality. The models have the subjects of extensive evaluation to determine their performance under a variety of meteorological conditions, the wireless sensor network air pollution monitoring system comprises of an array of sensor nodes and a communications system which allows the data to reach a server. The sensor nodes gather data autonomously and the data network is used to pass data to one or more base stations and forward it to a sensor network server. The system sends commands to the

nodes to get the data, and also send out data whenever required.

III. MINING EPA DATA

The EPA (Environmental Protection Administration) of Chennai runs Chennai Air Quality Monitoring Network (CAQMN) which is composed of several air quality monitoring stations to automatically collect and monitor air quality every week. More stations are set up in the industrial area, thus possibly have higher air pollution. Five types of the priority pollutants are recorded: PM_{10} (suspended particulate), $S0_2$ (sulphur dioxides), $N0_2$ (nitrogen dioxide), CO (carbon monoxide), and 0_3 (ozone). The EPA maintains a Web site for publishing archived and real-time pollutant information and forecasting. For instance, the homogeneous regions could be varied when the scale of temporal data is changed from small scale (e.g., hourly, daily, etc.) to large scale (e.g., monthly, seasonally, or annually). The selection of an appropriate scale is dependent on the application purpose. The data are collected from online CPCB and TNPCB websites.

IV. ARTIFICIAL NEURAL NETWORK DATA MINING

Artificial neural network have found various applications in the field of environmental engineering. Models have also been developed for air pollution data optimizing the process for prediction of vehicular emissions. The most popular ANN is feed-forward back-propagation, multi-layer perceptron (MLP) neural network. The development of ANN model consists of six steps. They are variable selection, Formation of Training, Testing, Validation data sets, Network modeling and Neural network training[5].

V. ARFF FILE FORMAT

The data obtained from online CPCB and TNPCB are stored in Microsoft Excel sheet with FILENAME.CSV format. The data value will be more than 16000 instances. The pollutants are taken as the field name. The file can be opened in WEKA tool for further processing and analysing. The data has to be pre processed and the data stored in Weka Explorer with FILENAME.ARFF file format. This data file can be accessed for weka tool for further analysis. The data is available from year 2012 to 2015. The huge volume of data can be accessed and processed using the WEKA tool.

VI. FEED FORWARD NEURAL NETWORKS (FFNN)

The simplest feed forward neural networks (FFNN), consists of three layers: input layer, hidden layer and output layer. In each layer there were one or more processing elements. A processing element receives inputs from outside world or the previous layer. There are connections between the Processing elements in each layer that have a weight (parameter) associated with them. This parameter is adjusted during training. Information travels in the forward direction through the network, there are no feedback loops[6]. The feed-forward back-propagation MLP for development of ANN model to predict daily maximum pollutants concentration in Chennai.

VII. BACK PROPAGATION ALGORITHM:

Back propagation Algorithm is a common method of teaching artificial neural networks how to perform a given task. The back propagation algorithm, artificial neurons are organized in layers, and send their signals forwardly, and then the errors are propagated backwardly. The back propagation algorithm uses supervised learning, compute the result and then the error is calculated.

The output for the MLP model was the daily maximum 1-hr pollutant level. All input dataset were normalized to provide values between 0.05 and 0.95 using the following formula:

$$P_i' = \frac{0.9(p_i - p_{\min})}{p_{\max} - p_{\min}} + 0.05$$

where P' transformed values, P_i actual observation values, P_{min} and P_{max} are the minimum and maximum values of observation values. Normalization of input data was performed for two reasons: to provide commensurate data range so that the models were not dominated by any variable that happened to be expressed in large numbers: and, to avoid the asymptotes of the sigmoid function. Once the best network is found, all the transformed data are transformed back into their original value by the formula:

$$P_i = \frac{(P_{\max} - P_{\min})(P_i' - 0.05)}{0.9} + P_{\min}$$

Before an MLP model can be utilized for predicting, the number of hidden layer and hidden nodes, and connection weights between neurons of the MLP network were determined by an iterative process in training (learning) stage with the training dataset of 361 patterns until the training error, measured by performance statistical indicators, is below the given error. The initial values of the weights are randomly selected and they can be both negative and positive values. In addition, activation function used in the hidden and output layers was determined by the required degree of accuracy of the problem under study. The activation function selected for the layers were logistic sigmoid for hidden layer and linear for the output layer. The number of hidden layers and hidden nodes were tried and increased systematically, checking each time if the prepared neural network obtained the stable performance error in the

performance plot. The best MLP network was the optimum found by the iterative process. The trained MLP network model was used to test the model's performance with testing dataset of 120 patterns. The resulting predictions were then compared with observed data, and performance statistical indicators were calculated.

VIII. MULTIVARIATE REGRESSION MODEL:

Multivariate regression, also known as ordinary least squares, is the most popular technique to obtain a linear input-output model for a given data set. The preliminary regression model has the general form:

$$Y = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_k X_k + \varepsilon$$

where Y stand for the predictand variable Y (e.g., daily maximum pollution level), β_i, i = 0, 1, 2, ….k, are called the regression coefficients (parameters), X_i is a set of k predictor variables X with matching β coefficients, and ε is a residual error.

To further assess the accuracy of the developed MLP network, its predictions were compared to linear regression model. An LR model between the eight input variables and the output (domain peak pollutants) was performed using a stepwise regression analysis on the first dataset to determine the coefficients of the above equation. A least-squares analysis was carried out, with the objective of finding the best linear equation that fit the dataset. The developed regression model was also tested performance with the testing dataset.

IX. LINEAR REGRESSION MODEL:

The stepwise regression procedure on the first dataset showed that PM_{10}, $PM_{2.5}$, SO_2, NO_2, CO, O_3 were important to predict daily maximum pollutants levels. The best single variable among the six independent variables was the nitrogen dioxide. The second-best single variable was maximum SO_2. Each step of forward stepwise regression procedure is shown in the Table 1. There are two factors that attribute the strength of correlation between PM_{10} and $PM_{2.5}$. High air temperature is an excellent indication of environmental conditions conductive to pollutants formation and accumulation. In addition, the photochemical reaction rates are highly temperature dependent.

Table 1: Forward Stepwise regression results

Steps	Set of variables	Coefficient of correlation, R_r^2
1	NO_2	0.2
2	NO_2, SO_2	0.273
3	NO_2, SO_2, PM_{10}	0.315
4	NO_2, SO_2, PM_{10}, $PM_{2.5}$	0.351
5	NO_2, SO_2, PM_{10}, $PM_{2.5}$, CO	0.371
6	NO_2, SO_2, PM_{10}, $PM_{2.5}$, CO, O_3	0.396

The following linear regression model (LR) was found to give the best fit, with the mean absolute error (MAE) was 12.67 ppb, the root mean square error (RMSE) was 15.02 ppb, the coefficient of determination (R^2) was 0.29, and the index of agreement (*d*) was 0.74 . A scatter plot for this model with the training and testing sets, showing the predicted versus the actual pollutant concentrations are given in Figure 1 and Figure 2. Based on the results of iterative process in training stage, it was found that the architecture of the best MLP network contains 6 input layer neurons, 10 hidden neurons for the first hidden layer, 14 hidden neurons for the second hidden layer and 1 output layer neuron. The scatter plots of predicted and observed pollutant concentrations for the training and testing sets. The mean absolute error (MAE) and the root mean square error (RMSE) for the training dataset were 15.32 and 0.012 ppbv, respectively. The corresponding errors for the testing dataset were 17.54 and 0.014 ppbv, respectively. To further check the accuracy of the developed MLP model, a plot of predicted versus observed pollutant concentrations was shown in Figure 3 and 4. The predicted values are in good agreement with the recorded Pollutant concentrations, indicating that the maximum Pollutants levels are captured fairly well by the MLP model.

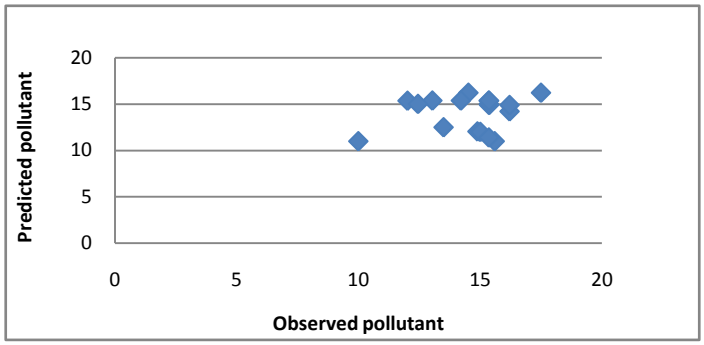

Figure 1: Training dataset Scatter plots of observed versus predicted pollutant levels of regression model.

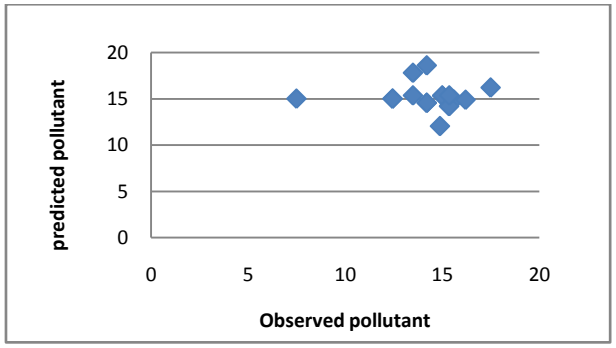

Figure 2: Testing dataset Scatter plots of observed versus predicted pollutant levels of regression model.

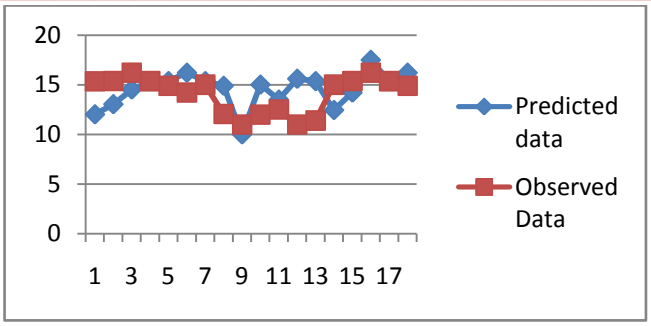

Figure 3: Comparison of observed and predicted pollutants for the training dataset of the MLP model.

X. COMPARATIVE ANALYSIS OF THE DEVELOPED MODELS

The relative effectiveness of the models are examined in predicting pollutant levels using the testing data set. The performance of the developed models was evaluated using statistical indicators and graphical comparisons.

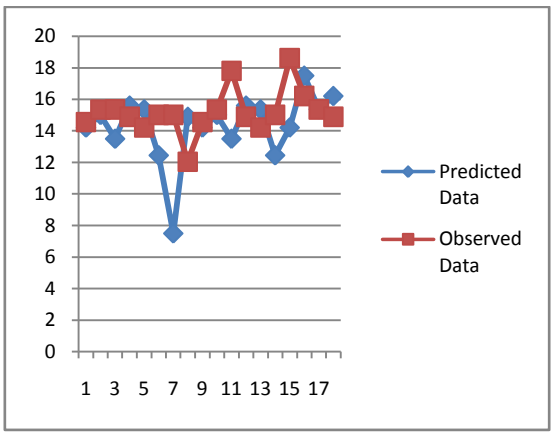

Figure 4: Comparison of observed and predicted pollutants for the testing dataset of the MLP model.

Table 2: Performance statistical indicators for the developed models

	MLP		LR	
Indicators	Training	Testing	Training	Testing
MAE (ppb)	5.32	7.54	12.67	12.56
RMSE(ppb)	0.012	0.014	15.02	14.35
R^2	0.134	0.121	0.29	0.31
D	0.92	0.89	0.74	0.68

It can be seen that the MLP model clearly gave the better results according to all statistical indicators. In terms of the MAE and the RMSE values, the MLP model performs better than the regression model for both datasets. Figure 4 shows the linear regression model performed significantly less well at predicting high pollutant level concentrations. The reason for the underestimation is that the problem of fitting of regression coefficients is solved using a "least-squares" criterion. A direct consequence is that the LR model, by nature, does not make any distinction between low and high levels of the values. The regression analysis process aims at modeling the "average" behavior for the predict and (output) variable, whereas with regards to air quality standards, the prediction of extreme pollutant levels is much more important from the health perspective. Despite the strong nonlinear character of the phenomena, the MLP gives rather good predictions. The data are processed using data mining tool and give results which help the policy maker in taking effective decisions in order to control air pollution created in various parts of Chennai.

XI. CONCLUSION

Air pollution play hazardous role in the health of the humans and plants. The effects of air pollution on health are very complex as there are many different sources and their individual effects vary from one to the other. The ambient air quality is assessed from various parts of Chennai and industrial area. The online data has been collected from Central Pollution Control Board (CPCB), Tamil Nadu Pollution Control Board(TNPCB) ambient air quality data for the past four years from 2012 to 2015. The data are pre processed and Data can be further processed by data mining tool and proper decision support can be given to the policy makers. The government has since adopted an array of measures to combat this problem. The prediction of Air pollution in urban and industrial area of Chennai using data mining could serve as an important reference for the policy maker in formulating future policies. The NAAQ(National Ambient Air Quality) standards of 2009, which superseded the earlier standard has more stringent values. The trend analysis shows that the norms are adhered and maintained so as to meet the new standards. This work paves way for the formation of new standards in the future so as to enhance the sustainable development. In future this research can be extended to predict the air pollution outside of Chennai and in other states.

ACKNOWLEDGMENT

The authors would like to thank Central Pollution Control Board, Tamil Nadu Pollution Control Board for online Data.

REFERENCES

[1] Sarah N. Kohail, Alaa M. El-Halees, Implementation of Data Mining Techniques for Meteorological Data Analysis, International Journal of Information and Communication Technology Research, Volume 1 No. 3, July 2011.

[2] Fayyad, U. M., Piatetsky-Shapiro, G., & Smyth P. (1996). The KDD process for extracting useful knowledge from volumes of data. Communications of the ACM, 39(11), 27–34.

[3] Li S., and Shue L., "Data mining to aid policy making in air pollution management," Expert Systems with Applications, vol. 27, pp. 331-340, 2004.
[4] Pyle, D. (1999). Data preparation for data mining. Los Altos, CA: Morgan Kaufmann.
[5] Amrender Kumar, "Artificial Neural Networks for Data mining", New Delhi.
[6] Fayyad, Usama, Ramakrishna " Evolving Data mining into solutions for Insights", communications of the ACM 45, no. 8
[7] Kavi K. Khedo, Rajiv Perseedoss and Avinash Mungur, A Wireless Sensor Network Air Pollution Monitoring System, International journal of Wireless and mobile network, Vol 2, issue 2, 2010.
[8] Haykin, S., Neural Networks, Prentice Hall International Inc., 1999
[9] Khajanchi, Amit, Artificial Neural Networks: The next intelligence
[10] Agrawal, R., Imielinski, T., Swami, A., "Database Mining: A Performance Perspective", IEEE Transactions on Knowledge and Data Engineering, pp. 914-925, December 1993
[11] Berry, J. A., Lindoff, G., Data Mining Techniques, Wiley Computer Publishing, 1997 (ISBN 0-471-17980-9).
[12] Berson, "Data Warehousing, Data-Mining & OLAP", TMH
[13] Bhavani,Thura-is-ingham, "Data-mining Technologies,Techniques tools & Trends", CRC Press
[14] http://www.cpcb.gov.in/CAAQM/frmCurrent Data
[15] http://www.tnpcb.gov.in/ambient_airquality
[16] Dr. Yashpal Singh, Alok Singh Chauhan, 'Neural Networks In Data Mining', Bundelkhand Institute of Engineering & Technology, Jhansi, Institute of Management, Allahabad, India, 2009.