ISSN: 0974-2115
www.jchps.com Journal of Chemical and Pharmaceutical Science

Impacts of ambient air quality data analysis in urban and industrial area helps in policy making

Christy S, Khanaa V
Bharath University, Chennai, Tamil Nadu, India,
***Corresponding author: E-Mail: christymelwyn @ gmail.com, Drvkannan62@yahoo.com**

ABSTRACT

Air pollution affect our health and environment in many different ways. In past few years, the heavy environmental loading has led to the deterioration of air quality in urban and industrial areas in Chennai. The task of controlling and improving the air quality is essential for a developing country. Data mining in ambient air quality data is concerned with finding the information inside the largely available data. The information retrieved can be transformed into usable knowledge. 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 websites. Air quality is monitored by air quality monitoring stations deployed in huge numbers using wireless sensors around the city and industrial areas in Chennai. The four years of data from January 2012 to December 2015 are collected from various monitoring stations and processed. Data mining technique is used in prediction, forecasting and support in making effective decision for future plan. The data can be analyzed by the data mining techniques using 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

1. INTRODUCTION

Data mining is the knowledge discovery in databases (KDD). It is the process of discovering useful knowledge from large amount of data stored in databases, data warehouses, or other repositories of information Fayyad, 1996. The understanding of Data starts with data collection and proceeds with activities to identify data quality problems, and to discover missing values into the data. Data preparation constructs the data to be modelled from the collected data. The modelling phase determines the optimal values for parameters in models by applying various modelling techniques. The evaluation phase evaluates the model for the problem requirements Pyle, 1999. Hourly air quality data are continuously collected through a network of several stations in Chennai and processed using data mining techniques. 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 websites are analyzed and processed with data mining techniques and provide effective decision support to policy makers of government agencies.

2. MATERIALS AND METHODS

Wireless sensor nodes: Air pollution monitoring system is considered as a very complex task. It is very important for controlling and maintaining the quality of air. Traditionally data collectors used to go to the spot and collect data periodically. This was time consuming and also quite expensive. The use of Wireless Sensor Networks (WSN) make air pollution monitoring less complex. More instantaneous readings can be obtained Kavi, 2010 by using WSN. The Air Monitoring Unit in Chennai lacks resources and makes use of bulky instruments, which reduces the flexibility of the system and makes it difficult to ensure proper control and monitoring. Air Quality Modelling is used to predict or simulate the ambient concentrations of pollutants in the atmosphere. They are also used as quantitative tools to find the cause and effect of concentration levels and to support laws and regulations designed to protect air quality. The models have the extensive evaluation to determine their performance under a variety of meteorological conditions. The air pollution monitoring system consists of number of wireless sensor nodes and communications system, allows the data to reach a server. The system sends commands to the nodes to get the data periodically, and also send out data for processing whenever required.

Air quality monitoring network: The Environmental Protection Administration (EPA) of Chennai runs Chennai Air Quality Monitoring Network (CAQMN) which is composed of several air quality monitoring stations. These stations automatically collect and monitor air quality every week. More stations are set up in urban and industrial area, which have higher air pollution. Five types of the priority pollutants are recorded: Suspended particulate (PM_{10}), Sulphur dioxides ($S0_2$), Nitrogen dioxide ($N0_2$), Carbon monoxide (CO) and Ozone (O_3). The Environmental Protection Administration maintains a Web site for publishing archived and real-time pollutant information and forecasting. The homogeneous regions can be varied when the scale of data is changed from small scale that is hourly, daily, etc., to large scale monthly, seasonally, or annually. The selection of an appropriate scale is dependent on the requirement of data. The data are collected from online CPCB and TNPCB websites.

Data mining tool: Weka tool is used to analyse the ambient air pollution data of urban and industrial areas. This tool is open source and freely available, provides different algorithms for data mining and machine learning. It is platform-independent. It provides flexible facilities for scripting experiments. Artificial neural network have large

number of applications in the field of environmental engineering. Air pollution data optimizing models have been developed in the tool for prediction of air pollution in urban and industrial areas. Feed-forward back-propagation, multi-layer perceptron (MLP) neural network are ANN models used. 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.

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 15000 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.

Feed Forward Neural Networks (FFNN): The simplest feed forward neural networks (FFNN) model, consists of three layers. They are input layer, hidden layer and output layer. One or more processing elements present in each layer. A processing element receives inputs from previous layer or other sources. The connections between the processing elements in each layer have a parameter associated with each other. This parameter is adjusted during training. Information get travels through the network in the forward direction, there are no feedback loops (Fayyad, 1996). The feed-forward back-propagation MLP for development of ANN model used to predict daily maximum pollutants concentration in Chennai.

Back propagation algorithm: Back propagation algorithm is a 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 always 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 Pi' transformed values, P_i actual observation values, P_{min} and P_{max} are the minimum observation values and maximum observation values. Normalization of input data was performed for two reasons: to provide appropriate 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 suitable network is found, all the transformed data are transformed back into their original value by the formula:

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

The number of hidden layers and hidden nodes, and connection weights between neurons of the MLP network were determined before an MLP model can be utilized for predicting. It is obtained by an iterative process in training stage with the training dataset of various patterns. The training errors can be measured by performance statistical indicators and should be below the given error. The initial values of the weights are randomly selected and they can be both negative and positive values. The activation function used in the hidden and output layers was determined. By the iterative process the optimum best MLP network was found. The trained MLP network model was used to test the model's performance with testing dataset of 160 patterns. The resulting predictions were found, performance statistical indicators were calculated and then compared with observed data.

Multivariate regression model: Multivariate regression, 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:

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

Where Y stand for the predict 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 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.

Linear regression model: The 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 nitrogen dioxide was the best single variable among

the five independent variables. The SO_2 was the next-best single variable. 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 environmental condition for pollutants formation and accumulation. The photochemical reaction rates are highly dependent on temperature.

Table 1: Forward Stepwise regression results

Steps	Set of variables	Coefficient of correlation, R_r^2
1	NO_2	0.200
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

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. Based on the results of iterative process in training stage, it was found that the architecture of the best MLP network contains 7 input layer neurons, 10 hidden neurons for the first hidden layer. There are 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 check the accuracy of the developed MLP model, predicted versus observed pollutant concentrations was shown in Figure 1. The predicted values are in good agreement with the recorded pollutant concentrations, indicating that the maximum pollutants levels are captured by the MLP model.

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

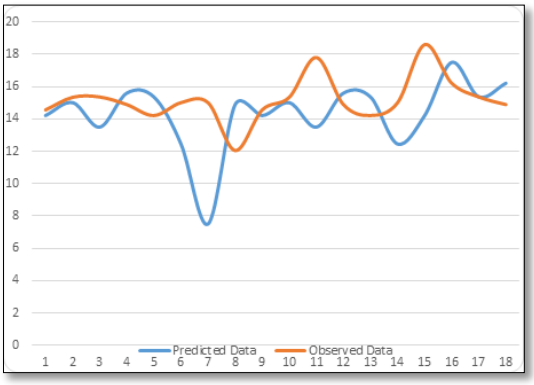

Figure.1.Comparison of observed and predicted pollutants for the testing dataset of the MLP model

Table.2.Performance statistical indicators for the developed models

Indicators	MLP		LR	
	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 Multi-Layer Perceptron model performs better than the regression model for both datasets. The linear regression model performed significantly less well at predicting high pollutant level concentrations is shown in Figure 1. 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 does not make any difference between low and high levels of the values. The regression analysis process moderate the behavior for the predicted variable and output variable, whereas with regards to air quality standards, the prediction of extreme pollutant levels is 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.

2. CONCLUSION

Air pollution play dangerous role in the health of the humans and plants. The effects of air pollution on health are very complex. There are many different sources and their individual effects of pollutants 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 further processed by data mining tool and proper decision support can be given to the policy makers. The government has since adopted many 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 for protecting our environment. 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.

3. ACKNOWLEDGMENT

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

REFERENCES

Agrawal R, Imielinski T, Swami A, Database Mining, A Performance Perspective, IEEE Transactions on Knowledge and Data Engineering, 1993, 914-925.

Christy S, Khanaa V, Analysis Of Air Pollution Data Using Weka Tool, Journal of Research in Computer Science, Engineering and Technology, 1(4), 2016, 146 - 148.

Christy S, Khanaa V, Data Mining In The Prediction Of Impacts Of Ambient Air Quality Data Analysis In Urban And Industrial Area, International Journal on Recent and Innovation Trends in Computing and Communication(IJRITCC), 4(2), 2016, 153 – 157.

Christy S, Khanaa V, Rajkumar S, Wireless Sensor Networks helps in the Prediction of Air Pollution, International Journal of Applied Engineering and Research, 2015, 327 – 330.

Christy S, Khanaa V, The Effects of Air Pollution on Human Health, International Journal of Mathematics and Computer Applications Research, 6(1), 2016, 51-58.

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

http://www.cpcb.gov.in/CAAQM/frmCurrent Data

http://www.tnpcb.gov.in/ambient_airquality

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.

Li S and Shue L, Data mining to aid policy making in air pollution management, Expert Systems with Applications, vol. 27, 2004, 331-340.

Pyle D, Data preparation for data mining, Los Altos, CA, Morgan Kaufmann 1999.

Sarah N, Kohail, Alaa, M El-Halees, Implementation of Data Mining Techniques for Meteorological Data Analysis, International Journal of Information and Communication Technology Research, 1(3), 2011.