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

Ambient Air prominence information Analysis Collision/Population

K. Uma*
Department of SITE, VIT University, Vellore, India.
***Corresponding author: E-Mail: dr.k.uma.phd@gmail.com**

ABSTRACT

Air pollution is affected our healthiness and atmosphere in numerous ways. In the earlier period little years, the heavy environmental loading has led to the deterioration of atmosphere excellence in industrial area near Chennai. The task of controlling and improving atmosphere eminence is essential for a developing country. Ambient space eminence data mining is a shape of data-mining concerned with finding the information inside the largely available statistics, hence to the data retrieved can be transformed into usable knowledge. The difficulty of air contamination is becoming a major concern for the fitness of the population. The ambient atmosphere eminence data collected from Central Pollution Control Board and Tamil Nadu Pollution Control Board ambient air quality statistics obtainable in the websites. Air excellence is scrutinized by air eminence observed stations deployed in huge numbers using wireless sensors around the city and industrial areas in Chennai. The past four years of statistics from the year 2012 towards 2015 are collected from various monitoring stations and processed. Data mining device is used for the calculation, forecasting and sustain in making efficient result. Artificial Neural- Network model (ANNM) using data mining methods verified the information by neural system 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.

KEY WORDS: Data mining, Data analysis, monitoring stations, Decision Support.

1. INTRODUCTION

Data mining, known as knowledge-discovery with databases (KDD) is the procedure of discovering helpful information from large amount of data stored in databases, data/information warehouses, or other data's repositories. Data understanding starts with data collection and proceeds with activities to identify data eminence trouble, and to discover missing values into the data. Data preparation constructs the statistics 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. Data-mining knowledge is used to spot the state air quality allocation of Chennai, whose hourly air excellence information is always collected through a network of several stations. Major composition of air contamination are undecided particulate material (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 information composed from CPCB and TNPCB are developed and verified with data-mining schemes and provide decision support to policy makers.

Wireless Sensor Nodes: Air effluence monitoring scheme is measured as a very multifaceted task other than it is very significant. Conventionally statistics collectors are worn to gather statistics. They used to go to the spot and collect data periodically and this was extremely time overriding and also quite expensive. The use of Wireless Sensor Networks can construct air effluence monitoring fewer complex and extra instantaneous values can be obtained. At present, the Air Monitoring component in Chennai lack of possessions and create using bulky instruments. This decreases the suppleness of the scheme and makes it hard to ensure proper control and monitoring. Air Quality replica is used to forecast or simulate the ambient concentrations of contaminant in the ambience. They are also used as quantitative tools to find the cause and effect of concentration levels and to support laws and regulations designed to defend air eminence. The models have the theme of widespread estimate to decide their performance below a selection of meteorological conditions. The air contamination monitoring scheme encompass of a collection of wireless sensor nodes and infrastructure system which permits the data to reach a server. The system sends commands to the nodes to get the data, and also send out data whenever required.

Air quality monitoring network: The Environmental-Protection organization (EPO) of Chennai runs Chennai Air eminence Monitoring Network (CAEMN) which is composed of several air excellence checking stations. These stations automatically collect and monitor air quality every week. More stations are set up in city and business region, 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 (0_3). The Environmental Protection government maintains a Web site for publishing archived and real-time pollutant information and forecasting. The homogeneous regions could be varied when the scale of chronological statistics 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 information is composed from online CPCB and TNPCB websites.

Data- mining tool: Weka tool is worn to analyse the ambient atmosphere effluence facts of urban and business region. It provides numerous different schemes for data-mining and machine-learning. It is free open source and generously available. It is platform dependent. It provides lithe amenities for scripting experiments. ANN has huge number of applications in the field of environmental engineering. Air pollution data optimizing reproduction have

been developed in the process for calculation of air contamination in municipal and business 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, configuration of guidance, 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 towards 2015. The huge volume of information must accessed and processed using the WEKA tool.

Feed forward neural networks (FFNN): The simplest feed-forward neural networks (FFNN), consists of three layers: input layer, secreted layer and output layer. In each layer there were one or more processing elements. A processing element receives inputs from previous layer or other sources. The relations between the dispensation fundamentals in each layer have a parameter associated with each other. This parameter is familiar during education. In order travels in the forward path through the network, there are no criticism loops. The feed-forward back-propagation MLP for development of ANN model used to forecast daily maximum pollutants concentration in Chennai.

Back propagation algorithm (BPA): BPA is a widespread method of training artificial neural networks how to execute a known mission. The back broadcast algorithm, reproduction neurons are ordered in layers, and transmit their signals directly, and then the mistakes are propagated backwardly. The BPA uses supervised learning, calculate the effect and then the fault 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:

$$P_i' = \frac{0.9(p_i - p_{\min})}{p_{\max} - p_{\min}} + 0.05 \quad (1)$$

Where Pi' transformed values, P_i actual observation values, P_{min} and P_{max} are the minimum and maximum values of observation values. Normalization of key information was performed for two reasons: to provide appropriate data range, therefore the models were not dominated by any variable that happened to be expressed in large numbers and, to evade the asymptotic of the sigmoid function. Once the top system is originate, all the transformed data are transformed reverse into their unique value by the formula:

$$P_i = \frac{(P_{\max} - P_{\min})(P_i' - 0.05)}{0.9} + P_{\min} \quad (2)$$

The number of concealed layer with hidden nodes, and connection weights between neurons of the MLP network were determined before an MLP form can be utilized for predicting. It is obtained by an iterative procedure in preparation phase with the training dataset of various patterns. The training error can be measured by performance statistical indicators and should be below the given error. The initial values of the weights are randomly chosen and it can be both negative and positive values. The opening function used in the concealed and result 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.

2. PROPOSED METHODS

Multivariate regression model: Multivariate regression, also acknowledged 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:

$$Y = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_k X_k + \varepsilon \quad (3)$$

Where Y stand for the forecast 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 additional assess the accurateness of the developed MLP network; its predictions were compared to linear regression model. An LR replica among the eight key variables and the result (domain peak pollutants) was performed using a stepwise regression analysis on the first dataset to conclude the coefficients of the above equation. A least-squares analysis was carried out, with the objective of result the finest linear equation that fit the dataset. The developed deterioration models are also tested performance with the sample information set.

Linear regression model: The step-wise regression process 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 greatest solitary variable between the six self-determining variables is the nitrogen dioxide [13+0]. The next better unique variable was maximum SO_2. Each

stage of forward step-wise deterioration process 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. 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	0.200
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 deterioration model (LDM) 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 outcome of iterative procedure in preparation stage, it was establish that the structural design of the most excellent MLP network contains 7 input layer neurons, 10 hidden neurons for the first hidden layer. There are 14 hidden neurons for the following hidden height and 1 output layer neuron. The disperse plots of forecasted and experimental pollutant attentiveness for the training and testing sets. The MAE and the RMSE for the preparation testing-set were 15.32 and 0.012 ppbv, respectively. The corresponding errors for the testing dataset were 17.54 and 0.014 ppbv, respectively. To supplementary check the accuracy of the urban MLP model, a plot of expected against observed pollutant concentrations was shown in Figure 1. The predicted values are in good agreement with the recorded Pollutant absorptions, representative that the highest Pollutants stages are measured fairly well by the MLP model.

3. EXPERIMENTAL RESULTS

Overall Evaluation 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.

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

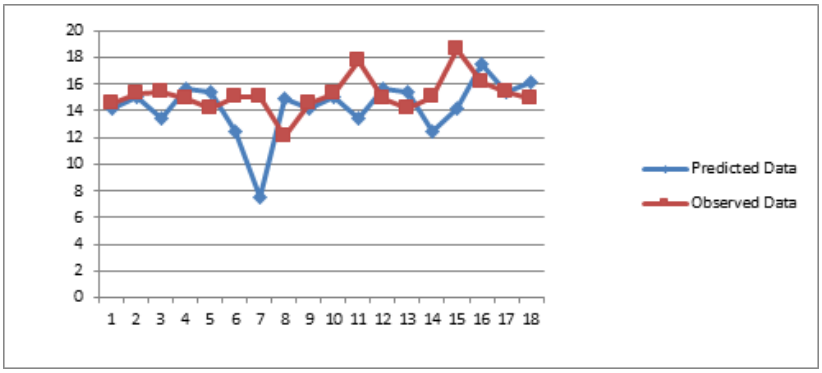

Figure.1. Comparison of observed and predicted pollutants for the testing dataset of the MLP model

It can be seen that the MLP model clearly gave the better results according to all statistical indicators. The MAE and the RMSE values, the MLP replica achieves enhanced than the regression model for both datasets. Figure 1 shows the linear deterioration form performed significantly less well at predicting high pollutant level concentrations. The reason for the underestimation is that the trouble of fitting of regression coefficients is solved using a "least-squares" criterion. A straight result is that the LR model, by nature, does not make any distinction between low and high levels of the values. The regression investigation procedure intend at moderate behavior for the predict and output changeable with regards to atmosphere quality principles, the calculation of extreme pollutant levels is a large amount of vital from the health perspective. Despite the strong nonlinear character of the phenomena,

the MLP gives rather good predictions. The data are processed using data extraction tool and give results which help the policy maker in taking effective results in order to control air pollution created in different divisions of Chennai.

4. CONCLUSION

Air pollution play dangerous role in the fitness of the humans and plants. The property of air contamination on physical condition is incredibly complex. There are numerous dissimilar sources and their entity effects of pollutants vary from one to the other. The atmosphere cleanliness is assessed from different partitions of Chennai and industrial area. The online data has been composed from Central-Pollution Control Board (CPCB), Tamil Nadu-Pollution Control Board (TNPCB) ambient atmosphere eminence statistics for the precedent four years as of 2012 towards 2015. The information is pre-processed and statistics can be further processed by data extraction 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 forecast of Air effluence in metropolitan and developed vicinity 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 given the norms are adhered and maintained to consequently 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 extensive to calculate the air pollution outside of Chennai and in other states.

5. 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.

Berson, Alex, and Stephen J Smith, Data warehousing, data mining, and OLAP. McGraw-Hill, Inc., 1997.

Fayyad UM, 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.

Fayyad, Usama, and Ramasamy Uthurusamy, Evolving data into mining solutions for insights, Communications of the ACM, 45(8), 2002, 28-31.

Kavi K Khedo, Rajiv Perseedoss, and Avinash Mungur, A Wireless Sensor Network Air Pollution Monitoring System, International journal of Wireless and mobile network, 2(2), 2010.

Kumar, Amrender, Artificial Neural Networks for Data Mining, IASRI, Library Avenue, Pusa, New Delhi-110012.

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

Peters, Annette, Increased particulate air pollution and the triggering of myocardial infarction, Circulation, 103(23), 2001, 2810-2815.

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.

Singh, Yashpal, and Alok Singh Chauhan, Neural networks in data mining, Journal of Theoretical and Applied Information Technology, 5(6), 2009, 36-42.