International Journal of Pure and Applied Mathematics
Volume 116 No. 9 2017, 249-254
ISSN: 1311-8080 (printed version); ISSN: 1314-3395 (on-line version)
url: http://www.ijpam.eu
Special Issue

ijpam.eu

IMPACTS OF AMBIENT AIR QUALITY DATA ANALYSIS IN URBAN AND INDUSTRIAL AREA HELPS IN POLICY MAKING

[1]R.Velvizhi, [2]M.S.Keerthikha
ASST.PROF/CSE, BHARATH UNIVERSITY
[1]velvizhi.cse@bharath.unive.ac.in, [2]keerthika.cse@bharath.unive.ac.in

Abstract: Air pollution can affect our health and environment in many ways. In the past few years, the heavy environmental loading has led to the deterioration of air quality in industrial area near Chennai. The task of controlling and improving air quality is essential for a developing country. Ambient air quality data mining is a form of data mining concerned with finding the information inside the largely available data, so that 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 ambient air quality data available in the 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 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 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

Consuming and also quite expensive. The use of Data mining, known as knowledge Wireless Sensor Networks can make air pollution discovery in databases (KDD) is the process of monitoring less complex and more instantaneous discovering useful knowledge from large amount of readings can be obtained [7]. Currently, the Air data stored in databases, data warehouses, or other Monitoring Unit in Chennai lacks resources and information repositories [2]. Data understanding makes use of bulky instruments. This reduces the starts with data collection and proceeds with flexibility of the system the subjects problem requirements [4]. Data mining technology of extensive evaluation to determine their is used to identify the national air quality performance under a variety of meteorological distribution of Chennai, whose hourly air quality conditions. The air pollution monitoring system data are continuously collected through a network comprises of an array of wireless sensor nodes and of several stations. Major composition of air communications system which allows the data to pollution are suspended particulate matter reach a server. The system sends commands to the (PM , PM), sulphur dioxide(SO), oxides of nodes to get the data, and also send out data nitrogen(NOX), carbon monoxide(CO), volatile whenever required[1-3].organic compounds, sulphur trioxide(SO3) and lead(PB). Four years data collected from CPCB and

1.1 Air quality monitoring network:

TNPCB are processed and analyzed with data mining techniques and provide decision support to The Environmental Protection Administration policy makers. (EPA) of Chennai runs Chennai Air Quality Monitoring Network.

1.2 Wireless sensor nodes

composed of several air quality monitoring Air pollution monitoring system stations. These stations automatically collect considered as a very complex task but it is verymonitor air quality every week. More stations are important. Traditionally data collectors were usedset up in urban and industrial area, which have to collect data. They used to go to the spot andhigher air pollution. Five types of the priority collect data periodically and this was very time pollutants are recorded: Suspended particulate (M10), Sulphur dioxides (S02), Nitrogen dioxide (N02), Carbon monoxide(CO) and Ozone(03). The Environmental Protection Administration maintains a Web site for publishing archived and real-time pollutant information and

forecasting. The homogeneous regions could be varied when the scale of temporal data is changed from small scale that is hourly, daily, etc., to large scale monthly, seasonally, or annually. The selection of an appropriate scaleis dependenton the requirement of data. The data are collected from online CPCB and TNPCB websites[4,5].

1.3 Data mining tool

Weka tool is used to analyse the ambient air pollution data of urban and industrial area. It provides many different algorithms for data mining and machine learning. It is open source and freely available. 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 [6,7]. Air pollution data optimizing Models have been developed in the process 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[5].

1.4 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 analyzing[8,9]. 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.

1.5 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 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 travels in the forward direction through the network, there are no feedbackloops[10,11]]. Thefeed-forwardback-ropagation MLP for development of ANN model used to predict daily maximum pollutants concentration in Chennai.

1.6 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[12,13].

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:

$$Pi' = 0.9(pi - pmin) + 0.05$$

where Pi' transformed values, Pi actual observation values, Pmin and Pmax are the minimum and maximum values of 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 best network is found, all the transformed data are transformed back into their original value by the formula:

$$(P ax - P in)(P' - 0.05)$$

The number of hidden layer 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 error 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[14,15,16]. The resulting predictions were calculated and then compared with observed data.

1.7 Multivariate 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:

$Y = \beta o + \beta 1X1 + 2X2 + \beta 3X3 + ... + \beta kXk + \varepsilon$

where Y stand for the predict variable Y (e.g., daily maximum pollution level), βi, i = 0, 1, 2, ….k, are called the regression coefficients (parameters), Xi 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[17,18]. The developed regression model was also tested performance with the testing dataset.

1.8 Linear regression model

The stepwise regression procedure on the first dataset showed that PM10, PM2.5, SO2, NO2, CO, O3 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 SO2. Each step of forward stepwise regression procedure is shown in the Table 1. There are two factors that attribute the strength of correlation between PM10 and PM2.5. High air temperature is an environmental condition for pollutants formation and accumulation. In addition, the photochemical reaction rates are highly temperature dependent.

The following linear regression model (LR) was found to give the best fit, with the mean

Steps	Set of variables	Coefficient of correlation, Rr2
1	NO2	0.200
2	NO2, SO2	0.273
3	NO2, SO2, PM10	0.315
4	NO2, SO2, PM10, PM2.5	0.351
5	NO2, SO2, PM10, PM2.5, CO	0.371

absolute error (MAE) was 12.67 ppb, the root mean square error (RMSE) was 15.02 ppb, the coefficient of determination (R2) 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 further check the accuracy of the developed MLP model, a plot of 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 fairly well by the MLP model.

1.9 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.

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
R2	0.134	0.121	0.29	0.31
D	0.92	0.89	0.74	0.68

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 moderate behavior for the predict and output variable[19,20], 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.

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 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 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.

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

References

[1]Udayakumar R., Kaliyamurthie K.P., Khanaa, Thooyamani K.P., Data mining a boon: Predictive system for university topper women in academia, World Applied Sciences Journal, v-29, i-14, pp-86-90, 2014.

[2]Kaliyamurthie K.P., Parameswari D., Udayakumar R., QOS aware privacy preserving location monitoring in wireless sensor network, Indian Journal of Science and Technology, v-6, i-SUPPL5, pp-4648-4652, 2013.

[3]Brintha Rajakumari S., Nalini C., An efficient cost model for data storage with horizontal layout in the cloud, Indian Journal of Science and Technology, v-7, i-, pp-45-46, 2014.

[4]Brintha Rajakumari S., Nalini C., An efficient data mining dataset preparation using aggregation in relational database, Indian Journal of Science and Technology, v-7, i-, pp-44-46, 2014.

[5]Khanna V., Mohanta K., Saravanan T., Recovery of link quality degradation in wireless mesh networks, Indian Journal of Science and Technology, v-6, i-SUPPL.6, pp-4837-4843, 2013.

[6]Khanaa V., Thooyamani K.P., Udayakumar R., A secure and efficient authentication system for distributed wireless sensor network, World Applied Sciences Journal, v-29, i-14, pp-304-308, 2014.

[7]Udayakumar R., Khanaa V., Saravanan T., Saritha G., Retinal image analysis using curvelet transform and multistructure elements morphology by reconstruction, Middle - East Journal of Scientific Research, v-16, i-12, pp-1781-1785, 2013.

[8]Khanaa V., Mohanta K., Saravanan. T., Performance analysis of FTTH using GEPON in direct and external modulation, Indian Journal of Science and Technology, v-6, i-SUPPL.6, pp-4848-4852, 2013.

[9]Kaliyamurthie K.P., Udayakumar R., Parameswari D., Mugunthan S.N., Highly secured online voting system over network, Indian Journal of Science and Technology, v-6, i-SUPPL.6, pp-4831-4836, 2013.

[10]Thooyamani K.P., Khanaa V., Udayakumar R., Efficiently measuring denial of service attacks using appropriate metrics, Middle - East Journal of Scientific Research, v-20, i-12, pp-2464-2470, 2014.

[11]R.Kalaiprasath, R.Elankavi, Dr.R.Udayakumar, Cloud Information Accountability (Cia) Framework Ensuring Accountability Of Data In Cloud And Security In End To End Process In Cloud Terminology, International Journal Of Civil Engineering And Technology (Ijciet)Volume 8, Issue 4, Pp. 376–385, April 2017.

[12]R.Elankavi, R.Kalaiprasath, Dr.R.Udayakumar, A fast clustering algorithm for high-dimensional data, International Journal Of Civil Engineering And Technology (Ijciet), Volume 8, Issue 5, Pp. 1220–1227, May 2017.

[13]R. Kalaiprasath, R. Elankavi and Dr. R. Udayakumar. Cloud. Security and Compliance - A Semantic Approach in End to End Security, International Journal Of Mechanical Engineering And Technology (Ijmet), Volume 8, Issue 5, pp-987-994, May 2017.

[14]Thooyamani K.P., Khanaa V., Udayakumar R., Virtual instrumentation based process of agriculture by automation, Middle - East Journal of Scientific Research, v-20, i-12, pp-2604-2612, 2014.
[15]Udayakumar R., Thooyamani K.P., Khanaa, Random projection based data perturbation using geometric transformation, World Applied Sciences Journal, v-29, i-14, pp-19-24, 2014.
[16]Udayakumar R., Thooyamani K.P., Khanaa, Deploying site-to-site VPN connectivity: MPLS Vs IPSec, World Applied Sciences Journal, v-29, i-14, pp-6-10, 2014.
[17] 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.
[18] 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.